MonetDB Server Reference Manual

Version 5.0

The MonetDB Development Team

Table of Contents

1 General Introduction

The MonetDB reference manual serves as the primary entry point to locate information on its functionality, system architecture, services and best practices on using its components.

The manual is produced from a Texinfo framework file, which collects and organizes bits-and-pieces of information scattered around the many source components comprising the MonetDB software family. The Texinfo file is turned into a HTML browse-able version using *makeinfo* program. The PDF version can be produced using *pdflatex*. Alternative formats, e.g., XML and DocBook format, can be readily obtained from the Texinfo file.

The copyright(2008) on the MonetDB software, documentation and logo is owned by CWI. Other trademarks and copyrights referred to in this manual are the property of their respective owners.

1.1 Intended Audience

The MonetDB reference manual is aimed at application developers and researchers with an intermediate level exposure to database technology, its embedding in host environments, such as C, Perl, Python, PHP, or middleware solutions based on JDBC and ODBC.

The bulk of the MonetDB reference manual deals with the techniques deployed in the back-end for the expert user and researcher. Judicious use of the programming interfaces and database kernel modules for domain specific tasks lead to high-performance solutions. The grand challenge for the MonetDB development team is to assemble a sufficient and orthogonal set of partial solutions to accommodate a wide variety of front-ends.

Feedback on the functionality provided is highly appreciated, especially when you embark on a complex programming project. If the envisioned missing functionality is generally applicable it makes sense to contribute it to the community. Share your comments and thoughts through the MonetDB mailing list held at SourceForge.

1.2 How to read this manual

The reference manual covers a lot of ground, which at first reading may be slightly confusing. The material is presented in a top-down fashion. Starting at installing the system components, SQL & XQuery and the application interface layer, it discusses the MAL software stack at length. Forward references are included frequently to point into the right direction for additional information.

If you are interested in technical details of the MonetDB system, you should start reading Section 1.70 [Design Overview], page 31. Two reading tracks are possible. The Chapter 3 [MAL Reference], page 48 language and subsequent sections describe the abstract machine and MAL optimizers to improve execution speed. It is relevant for a better understanding of the query processing behavior and provides an entry point to built new languages on top of the database kernel. The tutorial on SQL to MAL compilation provides a basis for developing your own language front-end.

The second track, The Inner Core describes the datastructures and operations exploited in the abstract machine layer. This part is essential for developers to aid in bug fixing and to extend the kernel with new functionality. For most readers, however, it can be skipped without causing problems to develop efficient applications.

1.3 Features and Limitations

In this section we give a short overview of the key features to (not) consider the MonetDB product family. In a nutshell, its origin in the area of data-mining and data-warehousing makes it an ideal choice for high volume, complex query dominant applications. MonetDB was not designed for high-volume secure OLTP settings initially.

It is important to recognize that the MonetDB language interfaces are primarily aimed at experienced system programmers and administrators. End-users are advised to use any of the open-source graphical SQL workbenches to interact with the system.

1.3.1 When to consider MonetDB ?

A high-performance database management system. MonetDB is an easy accessible open-source DBMS for SQL-[XQuery-]based applications and database research projects. Its origin goes back over a decade, when we decided that the database hotset - the part used by the applications - can be largely held in main-memory or where a few columns of a broad relational table are sufficient to handle a request. Further exploitation of cache-conscious algorithms proved the validity of these design decisions.

A multi-model system. MonetDB supports multiple query language front-ends. Aside from its proprietary language, called the MonetDB Assembly Language (MAL), it aims at ANSI SQL-2003 and W3C XQuery with update facilities. Their underlying logical data model and computational scheme differs widely. The system is designed to provide a common ground for both languages and it is prepared to support languages based on yet another data model or processing paradigm.

A column-store based database kernel. MonetDB is built on the canonical representation of database containers, called Binary Association Tables (BATs). MonetDB is known as one of the principal COLUMN-STORES, as typically, a separate BAT is used for each table column. The datastructures are geared towards efficient representation when they mimic an n-ary relational scheme.

This led to an architecture where the traditional page-pool is replaced by one with a much larger granularity based on BATs. They are sizeable entities -up to hundreds of megabytes- swapped into memory upon need. The benefit of this approach has been shown in numerous papers in the scientific literature.

A broad spectrum database system. MonetDB is continuously developed to support a broad application field. Although originally developed for Analytical CRM products, it is now being used at the low-end scale as an embedded relational kernel and projects are underway to tackle the huge database problems encountered in scientific databases, e.g. astronomy.

An extendable database system. MonetDB has been strongly influenced by the scientific experiments to understand the interplay between algorithms and hardware features. It has turned MonetDB into an extensible database system. It proves valuable in those cases where an application specific and critical component makes all the difference between slow and fast implementation.

An open-source software system. MonetDB has been developed over many years of research at CWI, whose charter ensures that results are easily accessible to others. Either through publications in the scientific domain or publication of the software components involved. The MonetDB mailing list is the access point to a larger audience for advice.

A subscription to the mailing list helps the developer team to justify their hours put into MonetDB's development and maintenance.

1.3.2 When not to consider MonetDB ?

There are several areas where MonetDB has not yet built a reputation. They are the prime candidates for experimentation, but also areas where application construction may become risky. Mature products or commercial support may then provide a short-term solution, while MonetDB programmers team works on filling the functional gaps. The following areas should be considered with care:

Persistent object caches. The tendency to develop applications in Java and C based on a persistent object model, is a no-go area for MonetDB. Much like other database engines, the overhead of individual record access does not do justice to the data structures and algorithms in the kernel. They are chosen to optimize bulk processing, which always comes at a price for individual object access.

Nevertheless, MonetDB has been used from its early days in a commercial application, where the programmers took care in maintaining the Java object-cache. It is a route with great benefits, but also one where sufficient manpower should be devoted to perform a good job.

High-performance financial OLTP. MonetDB was originally not designed for highly concurrent transaction workloads. It was decided to make ACID hooks explicit in the query plans generated by the front-end compilers. Given the abundance of main memory nowadays and the slack CPU cycles to process database requests, it may be profitable to consider serial execution of all OLTP transactions.

The SQL implementation provides full transaction control and recovery.

Security. MonetDB has not been designed with a strong focus on security. The major precautions have been taken, but are incomplete when access to the hosting machine is granted or when direct access is granted to the Monet Assembly Language features. The system is preferably deployed in a sand-boxed environment where remote access is encapsulated in a dedicated application framework.

Scaling over multiple machines. MonetDB does not provide a centralized controlled, distributed database infrastructure yet. Instead, we move towards an architecture where multiple autonomous MonetDB instances are joining together to process a large and distributed workload.

In the multimedia applications we have exploited successfully the inherent data parallelism to speedup processing and reduce the synchronization cost. The underlying platforms were Linux-based cluster computers with sizeable main memories.

1.3.3 What are key features of MonetDB

The list below provides a glimpse on the technical characteristics and features of the MonetDB software packages.

The software characteristics for the MonetDB packages are:

- The kernel source code is written in ANSI-C and POSIX compliant.
- The application interface libraries source code complies with the latest language versions.

- The source code is written in a literate programming style, to stimulate proximity of code and its documentation.
- The source code is compiled and tested on many platforms with different compiler options to ensure portability.
- The source code is based on the GNU toolkit, e.g. Automake, Autoconf, and Libtool for portability.
- The source code is heavily tested on a daily basis, and scrutinized using the Valgrind toolkit.

The heart is the MonetDB server, which comes with the following innovative features.

- A fully decomposed storage scheme using memory mapped files.
- It supports scalable databases, 32- and 64-bit platforms.
- Connectivity is provided through TCP/IP sockets and SSH on many platforms.
- Index selection, creation and maintenance is automatic.
- The relational operators materialize their results and are self-optimizing.
- The operations are cache- and memory-aware with supreme performance.
- The database back-end is multi-threaded and guards a single physical database instance.

1.3.4 Size Limitations for MonetDB

The maximal database size supported by MonetBD depends on the underlying processing platform, e.g., a 32- or 64-bit processor, and storage device, e.g., the file system and disk raids.

The number of columns per tables is practically unlimited. The storage space limitation depends only on the maximal file size. For each column is mapped onto a file, whose limit is dictated by the operating system and hardware platform.

The number of concurrent user threads is a configuration parameter. Middleware solutions are adviced to serialize access to the database when large number of users are expected to access the database.

1.4 A Brief History of MonetDB

The Dark Ages [1979-1992] The development of the MonetDB software family goes back as far as the early eighties when the first relational kernel, called Troll, was delivered to a larger audience. It was spread over ca 1000 sites world-wide and became part of a software case-tool until the beginning of the nineties. None of the code of this system has survived, but several ideas and experiences on how to obtain a fast kernel by simplification and explicit materialization found their origin during this period.

The second part of the eighties was spent on building the first distributed main-memory database system in the context of the national project PRISMA. A fully functional system of 100 processors and a wealthy 1GB of main memory showed the road to develop database technology from a different perspective.

The Early Days [1993-1995] Immediately after the PRISMA project was termed dead, a new database kernel based on Binary Association Tables (BATs) was laid out. The original target was to aim for better support of scientific databases with their then archaic file structures.

The Data Distilleries Era [1996-2003] The datamining projects running as of 1993 called for better database support. It culminated in the spin-off Data Distilleries, which based their analytical customer relationship suite on the power provided by the early MonetDB implementations. In the years following, many technical innovations were paired with strong industrial maturing of the software base. Data Distilleries became a subsidiary of SPSS in 2003 and its development activity was shifted to Chicago in 2007.

The Open-Source Challenge [2003-2007] Moving MonetDB Version 4 into the open-source field required a large number of extensions to the code base. It became utmost important to support a mature implementation of the SQL-99 standard, and the bulk of application programming interfaces (PHP,JDBC,Perl,ODBC). The result of this activity was the first official release in 2004 and the release of the XQuery front-end in 2005. The XQuery code generator used grew out of a student summer project (`"milprint_summer"`) and proved that scalable, high-performance XQuery processing on a relation DBMS is possible.

The Road Ahead [2008- This manual describes the MonetDB Version 5 release, the result of a multi-year activity to clean up the software stack and to better support both simple and complex database requests.

The Future New versions in the MonetDB software family are under development. Extensions and renovation of the kernel are studied in the X100 project. Its Volcano-style interpreter aims to provide performance in I/O-dominant and streaming settings using vectorized processing and Just-In-Time (de)compression.

The scene of distributed database is (again) addressed in the Armada project, but not using the traditional centralized administration focus. Instead the Armada project seeks the frontiers of autonomous database systems, which still provide a coherent functional view to its users. In its approach it challenges many dogmas in distributed database technology, such as the perspective on global consistency, the role of the client in managing the distributed world, and the way resources are spread.

The MonetDB software framework provides a rich setting to pursue these alleys of database research. We hope that many may benefit from our investments, both research and business wise.

1.5 Manual Generation

The MonetDB code base is a large collection of files, scattered over the system modules. Each source file is written in a literal programming style, which physically binds documentation with the relevant code sections. The utility program Mx processes the files marked *.mx to extract the code sections for system compilation or to prepare for a pretty printed listing.

The reference manual is based on *Texinfo* formatted documentation to simplify generation for different rendering platforms. The components for the reference manual are extracted by

```
Mx -i -B -H1 <filename>.mx
```

which generates the file <filename>.bdy.texi. These pieces are collected and glued together in a manual framework, running *makeinfo* to produce the desired output format. The

Texinfo information is currently limited to the documentation, it could also be extended to process the code.

A printable version of an *.mx file can be produced using the commands:

```
Mx  <filename>.mx
pdflatex <filename>.tex
```

1.5.1 Conventions and Notation

The typographical conventions used in this manual are straightforward. Monospaced text is used to designate names in the code base and examples. *Italics* is used in explanations to indicate where a user supplied value should be substituted.

Snippets of code are illustrated in small caps font. The interaction with textual client interfaces uses the default prompt-setting of the underlying operating system.

Keywords in the MonetDB interface languages are case sensitive; SQL keywords are not case sensitive. No distinction is made in this manual.

1.5.2 Additional Resources

Although this reference manual aims to be complete for developing applications with MonetDB, it also depends on additional resources for a full understanding.

This reference manual relies on external documentation for the basics of its query languages SQL, XQuery, its application interfaces, PHP, Perl, Pyhton, and its middleware support, JDBC and ODBC. Examples are used to illustrate their behaviour in the context of MonetDB only. The resource locations identified below may at times proof valuable.

Perl DBI	http://www.perl.org/
PHP5	http://www.php.net/
Python	http://www.python.org/
XQuery	http://www.w3c.org/TR/xquery/

The primary source for additional information is the MonetDB website, http://monetdb.cwi.nl/, and the code base itself. Information on the background of its architecture can be found in the library of scientific publications.

1.6 Downloads and Installation

For most people a binary distribution package is sufficient. The prime decision is to select either the SQL or XQuery product line. They currently rely on different and incompatible back-end servers.

The binary distribution contains all components for MonetDB application development, i.e. a back-end server, an SQL or XQuery compiler, and the client libraries. These components are packaged conveniently for several platforms in the download section at SourceForge. It can be installed in a private directory or in the Linux/Windows compliant default folder location.

The Developers distribution is meant for source experimentation and functional enhancements. A "stable" version is prepared regularly. It means that special care has been taken to assure that errors reported during the nightly builds have been solved on the platforms of interest. Major bug fixes are also applied to the latest stable version, while functional enhancements are kept for the next release or the daily builds.

The Experts distribution is meant for MonetDB kernel software developers only. They should have a clear understanding of Linux development tools, e.g. automake, config, CVS, and team-based software development and the interdependencies of the MonetDB components.

If you encounter errors during the installation, please have a look at the MonetDB mailing list for common errors and some advice on how to proceed.

1.6.1 Developers Distribution

Developers interested in source code to be linked with the MonetDB libraries or running on non-supported platforms may use the nightly builds tarballs to assemble a working system. Alternatively, they can check out the latest STABLE of CURRENT version from CVS.

The easiest way is use the All-In-One scripts provided. A quick installation based on the nightly tarballs and super source tarball is supported by the monetdb-install.sh script. Check it out and run it in an empty directory should be sufficient on most Linux platforms to get going. It takes about 10-20 minutes to install and compile from scratch on a modern PC. The nightly build source distribution comes with the complete test-bench to assure that changes do not affect (in as far as they get tested) its stability. A single stable release is maintained for external users while we concurrently work on the next release. Older versions are not actively maintained by the development team.

1.6.2 Experts

The experts may want more control than provided by the developer distribution support. Set up of a fully functional system requires downloading and installation of the latest packages from SourceForge. The compatibility table below illustrates the packages in the CVS repository.

	MonetDB/SQL	MonetDB/XQuery	Clients	JDBC & XR-PCwrapper
java				x
buildtools	x	x	x	
clients	x	x	x	
MonetDB	x	x		
MonetDB4		x		
MonetDB5	x			
SQL	x			
Pathfinder		x		

Thanks to the GNU AUTOCONF and AUTOMAKE tools, the MonetDB software runs on a wide variety of hardware/software platforms. The MonetDB development team uses many of the platforms available to perform automated nightly regression testing. For more details see The Test Web.

The MonetDB code base -with daily builds available for users preferring living at the edge- evolves quickly. Application developers, however, may tune into the MonetDB mailing list to be warned when a major release has become available, or when detected errors require a patch.

1.7 How To Start with MonetDB

This document helps you compile and install the MonetDB suite from scratch on Unix-like systems (this includes of course Linux, but also MacOS X and Cygwin). This document is meant to be used when you want to compile and install from CVS source. When you use the prepared tar balls, some of the steps described here should be skipped.

In case you prefer installing a pre-compiled binary distribution, please check out the binary distribution.

This document assumes that you are planning on compiling and installing MonetDB on a Unix-like system (e.g., Linux, IRIX, Solaris, AIX, Mac OS X/Darwin, or CYGWIN). For compilation and installation on a native Windows system (NT, 2000, XP) see the instructions in the file ../buildtools/doc/windowsbuild.rst.

1.8 The Suite

The MonetDB software suite consists of the following parts which need to be built in the correct order:

`buildtools`
: Tools used only for building the other parts of the suite. These tools are only needed when building from CVS. When building from the source distribution (i.e. the tar balls), you do not need this.

`MonetDB`
: Fundamental libraries used in the other parts of the suite.

`clients`
: Libraries and programs to communicate with the server(s) that are part of the suite.

`MonetDB4`
: The MIL-based server. This is required if you want to use XML/XQuery (pathfinder), and can be used with SQL.

`MonetDB5`
: The MAL-based server. This can be used with and is recommended for SQL.

`pathfinder`
: The XML/XQuery engine built on top of MonetDB4.

`sql`
: The SQL server built on top of (targeted on) either MonetDB4 or MonetDB5.

MonetDB4 and MonetDB5 are the basic database engines. One or the other is required, but you can have both. Pathfinder currently needs MonetDB4, sql can run on both MonetDB4 and MonetDB5 (the latter is recommended).

The order of compilation and installation is important. It is best to use the above order (where pathfinder and sql can be interchanged) and to configure-make-make install each package before proceeding with the next.

1.9 Prerequisites

`CVS`
: You only need this if you are building from CVS. If you start with the source distribution from SourceForge you don't need CVS.

  You need to have a working CVS. For instructions, see the SourceForge documentation and look under the heading CVS Instructions.

`Python` MonetDB uses Python (version 2.0.0 or better) during configuration of the software. See http://www.python.org/ for more information. (It must be admitted, version 2.0.0 is ancient and has not recently been tested, we currently use 2.4 and newer.)

`autoconf/automake/libtool`

MonetDB uses GNU autoconf (>= 2.57) and automake (>= 1.5) during the Bootstrap phase, and libtool (>= 1.4) during the Make phase. autoconf and automake are not needed when you start with the source distribution.

`standard software development tools`

To compile MonetDB, you also need to have the following standard software development tools installed and ready for use on you system:

- a C compiler (e.g. GNU's `gcc`);
- GNU `make` (`gmake`) (native `make` on, e.g., IRIX and Solaris usually don't work).

The following are not needed when you start with the source distribution:

- a C++ compiler (e.g. GNU's `g++`);
- a lexical analyzer generator (e.g., `lex` or `flex`);
- a parser generator (e.g., `yacc` or `bison`).

The following are optional. They are checked for during configuration and if they are missing, the feature is just missing:

- swig
- perl
- php

`buildtools (Mx, mel, autogen, and burg)`

These tools are not needed when you start with the source distribution.

Before building any of the other packages from the CVS sources, you first need to build and install the buildtools. Check out buildtools with

```
cvs -d:pserver:anonymous@monetdb.cvs.sourceforge.net:/cvsroot/monetdb checkout bu
```

and follow the instructions in the README file, then proceed with MonetDB. For this step only you need the C++ compiler.

`libxml2` The XML parsing library libxml2 is only used by XML/XQuery (pathfinder). The library is used for:

1. the XML Schema import feature of the Pathfinder compiler, and
2. the XML document loader (runtime/shredder.mx).

If libxml2 is not available on your system, the Pathfinder compiler will be compiled without XML Schema support. The XML document loader will not be compiled at all in that case. Current Linux distributions all come with libxml2.

1.10 Space Requirements

The packages take about this much space:

buildtools	1.5 MB	8 MB	2.5 MB
MonetDB	2 MB	21 MB	4 MB
clients	9 MB	25 MB	10 MB
MonetDB4	35.5 MB	50 MB	14 MB
MonetDB5	26 MB	46 MB	12 MB
sql	100 MB	22.5 MB	8 MB
pathfinder	130 MB	43 MB	12 MB

Some of the source packages are so large because they include lots of data for testing purposes.

1.11 Getting the Software

There are two ways to get the source code:

1. checking it out from the CVS repository on SourceForge;
2. downloading the pre-packaged source distribution from SourceForge.

The following instructions first describe how to check out the source code from the CVS repository on SourceForge; in case you downloaded the pre-packaged source distribution, you can skip this section and proceed to Configure and Make.

1.12 CVS checkout

This command should be done once. It records a password on the local machine to be used for all subsequent CVS accesses with this server.

```
cvs -d:pserver:anonymous@monetdb.cvs.sourceforge.net:/cvsroot/monetdb login
```

Just type RETURN when asked for the password.

Then get the software by using the command:

```
cvs -d:pserver:anonymous@monetdb.cvs.sourceforge.net:/cvsroot/monetdb checkout \
buildtools MonetDB clients MonetDB4 MonetDB5 pathfinder sql
```

This will create the named directories in your current working directory. Then first follow the instructions in `buildtools/README` before continuing with the others. Naturally, you don't need to check out packages you're not going to use.

Also see the SourceForge documentation for more information about using CVS.

1.13 Bootstrap, Configure and Make

Before executing the following steps, make sure that your shell environment (especially the variables `PATH`. `LD_LIBRARY_PATH`, and `PYTHONPATH`) is set up so that the tools listed above can be found. Also, set up PATH to include the *prefix*/bin directory where *prefix* is the prefix is where you want everything to be installed, and set up PYTHONPATH to include the *prefix*/lib/*python2.X* directory where *python2.X* is the version of Python being used. It is recommended to use the same *prefix* for all packages. Only the *prefix*/lib/*python2.X* directory for buildtools is needed in PYTHONPATH.

In case you checked out the CVS version, you have to run `bootstrap` first; in case you downloaded the pre-packaged source distribution, you should skip `bootstrap` and start with `configure` (see Configure).

For each of the packages do all the following steps (bootstrap, configure, make, make install) *before* proceeding to the next package.

1.14 Bootstrap

This step is only needed when building from CVS.

In the top-level directory of the package type the command (note that this uses `autogen.py` which is part of the `buildtools` package — make sure it can be found in your `$PATH`):

```
./bootstrap
```

1.15 Configure

Then in any directory (preferably a *new, empty* directory and *not* in the `MonetDB` top-level directory) give the command:

```
.../configure [<options>]
```

where `...` is replaced with the (absolute or relative) path to the `MonetDB` top-level directory.

The directory where you execute `configure` is the place where all intermediate source and object files are generated during compilation via `make`.

By default, MonetDB is installed in `/usr/local`. To choose another target directory, you need to call

```
.../configure --prefix=<prefixdir> [<options>]
```

Some other useful `configure` options are:

`--enable-debug`
: enable full debugging default=[see Configure defaults and recommendations below]

`--enable-optimize`
: enable extra optimization default=[see Configure defaults and recommendations below]

`--enable-assert`
: enable assertions in the code default=[see Configure defaults and recommendations below]

`--enable-strict`
: enable strict compiler flags default=[see Configure defaults and recommendations below]

`--enable-warning`
: enable extended compiler warnings default=off

`--enable-profile`
: enable profiling default=off

`--enable-instrument`
: enable instrument default=off

`--with-mx=<Mx>`
: which Mx binary to use (default: whichever Mx is found in your PATH)

`--with-mel=<mel>`
: which mel binary to use (default: whichever mel is found in your PATH)

`--enable-bits=<#bits>`
: specify number of bits (32 or 64) default is compiler default

`--enable-oid32`
: use 32-bit OIDs on 64-bit systems default=off

You can also add options such as `CC=<compiler>` to specify the compiler and compiler flags to use.

Use `configure --help` to find out more about `configure` options.

The `--with-mx` and `--with-mel` options are only used when configuring the sources as retrieved through CVS.

1.16 Configure defaults and recommendations

For convenience of both developers and users as well as to comply even more with open source standards, we now set/use the following defaults for the configure options

```
--enable-strict, --enable-assert, --enable-debug, --enable-optimize
```

When compiling from CVS sources (as mainly done by developers):

```
strict=yes  assert=yes  debug=yes  optimize=no (*)
```

When compiling from packaged/distributed sources (i.e., tarballs) (as mainly done by users):

```
strict=no   assert=no   debug=no   optimize=no (*)
```

For building binary distributions (RPMs):

```
strict=no   assert=no   debug=no   optimize=yes
```

`(*)` IMPORTANT NOTE:

Since `--enable-optimize=yes` is no longer the default for any case except binary packages, it is *strongly recommended* to (re)compile everything from scratch, *explicitly configured* with

```
--enable-debug=no --enable-assert=no --enable-optimize=yes
```

in case you want/need to run any performance experiments with MonetDB!

Please note: `--enable-X=yes` is equivalent to `--enable-X`, and `--enable-X=no` is equivalent to `--disable-X`.

1.17 Make

In the same directory (where you called `configure`) give the command

```
make
```

to compile the source code. Please note that parallel make runs (e.g. `make -j2`) are currently known to be unsuccessful.

1.18 Testing the Build

This step is optional and only relevant for the packages clients, MonetDB4, MonetDB5, pathfinder, and sql.

If `make` went successfully, you can try

```
make check
```

This will perform a large number of tests, some are unfortunately still expected to fail, but most should go successfully. At the end of the output there is a reference to an HTML file which is created by the test process that shows the test results.

1.19 Install

Give the command

```
make install
```

By default (if no `--prefix` option was given to `configure` above), this will install in `/usr/local`. Make sure you have appropriate privileges.

1.20 Testing the Installation

This step is optional and only relevant for the packages clients, MonetDB4, MonetDB5, pathfinder, and sql.

Make sure that *prefix*/bin is in your `PATH`. Then in the package top-level directory issue the command

```
Mtest.py -r [--package=<package>]
```

where *package* is one of `clients`, `MonetDB4`, `MonetDB5`, `sql`, or `pathfinder` (the `--package=<package>` option can be omitted when using a CVS checkout; see

```
Mtest.py --help
```

for more options).

This should produce much the same output as `make check` above, but uses the installed version of MonetDB.

You need write permissions in part of the installation directory for this command: it will create subdirectories `var/dbfarm` and `Tests`.

1.21 Usage

The MonetDB4 and MonetDB5 engines can be used interactively or as a server. The XQuery and SQL back-ends can only be used as servers.

To run MonetDB4 interactively, just run:

```
Mserver
```

To run MonetDB5 interactively, just run:

```
mserver5
```

The disadvantage of running the systems interactively is that you don't get readline support (if available on your system). A more pleasant environment can be had by using the system as a server and using `mclient` to interact with the system. For MonetDB4 use:

```
Mserver --dbinit 'module(mapi); mil_start();'
```

When MonetDB5 is started as above, it automatically starts the server in addition to the interactive "console".

In order to use the XQuery back-end, which is only available with MonetDB4, start the server as follows:

```
Mserver --dbinit 'module(pathfinder);'
```

If you want to have a MIL server in addition to the XQuery server, use:

```
Mserver --dbinit 'module(pathfinder); mil_start();'
```

In order to use the SQL back-end with MonetDB4, use:

```
Mserver --dbinit 'module(sql_server);'
```

If you want to have a MIL server in addition to the SQL server, use:

```
Mserver --dbinit 'module(sql_server); mil_start();'
```

In order to use the SQL back-end with MonetDB5, use:

```
mserver5 --dbinit 'include sql;'
```

Once the server is running, you can use `mclient` to interact with the server. `mclient` needs to be told which language you want to use, but it does not need to be told whether you're using MonetDB4 or MonetDB5. In another shell window start:

```
mclient -l<language>
```

where *language* is one of `mil`, `mal`, `sql`, or `xquery`. If no `-l` option is given, `mil` is the default.

With `mclient`, you get a text-based interface that supports command-line editing and a command-line history. The latter can even be stored persistently to be re-used after stopping and restarting `mclient`; see

```
mclient --help
```

for global details and

```
mclient -l<language> --help
```

for language-specific details.

At the `mclient` prompt some extra commands are available. Type a single question mark to get a list of options. Note that one of the options is to read input from a file using `<`. This interferes with XQuery syntax. This is a known bug.

1.22 Troubleshooting

`bootstrap` fails if any of the requisite programs cannot be found or is an incompatible version.

`bootstrap` adds files to the source directory, so it must have write permissions.

During `bootstrap`, warnings like

```
Remember to add `AC_PROG_LIBTOOL' to `configure.in'.
You should add the contents of `/usr/share/aclocal/libtool.m4' to `aclocal.m4'.
configure.in:37: warning: do not use m4_patsubst: use patsubst or m4_bpatsubst
configure.in:104: warning: AC_PROG_LEX invoked multiple times
configure.in:334: warning: do not use m4_regexp: use regexp or m4_bregexp
```

```
automake/aclocal 1.6.3 is older than 1.7.
Patching aclocal.m4 for Intel compiler on Linux (icc/ecc).
patching file aclocal.m4
Hunk #1 FAILED at 2542.
1 out of 1 hunk FAILED -- saving rejects to file aclocal.m4.rej
patching file aclocal.m4
Hunk #1 FAILED at 1184.
Hunk #2 FAILED at 2444.
Hunk #3 FAILED at 2464.
3 out of 3 hunks FAILED -- saving rejects to file aclocal.m4.rej
```

might occur. For some technical reasons, it's hard to completely avoid them. However, it is usually safe to ignore them and simply proceed with the usual compilation procedure. Only in case the subsequent `configure` or `make` fails, these warning might have to be taken more seriously. In any case, you should include the `bootstrap` output whenever you report (see Reporting Problems) compilation problems.

`configure` will fail if certain essential programs cannot be found or certain essential tasks (such as compiling a C program) cannot be executed. The problem will usually be clear from the error message.

E.g., if `configure` cannot find package XYZ, it is either not installed on your machine, or it is not installed in places that `configure` searches (i.e., `/usr`, `/usr/local`). In the first case, you need to install package XYZ before you can `configure`, `make`, and install MonetDB. In the latter case, you need to tell `configure` via `--with-XYZ=<DIR>` where to find package XYZ on your machine. `configure` then looks for the header files in <DIR>/include, and for the libraries in <DIR>/lib.

In case one of `bootstrap`, `configure`, or `make` fails — especially after a `cvs update`, or after you changed some code yourself — try the following steps (in this order; if you are using the pre-packaged source distribution, you can skip steps 2 and 3):

(In case you experience problems after a `cvs update`, first make sure that you used `cvs update -dP` (or have a line `update -dP` in your `~/.cvsrc`); `-d` ensures that cvs checks out directories that have been added since your last `cvs update`; `-P` removes directories that have become empty, because all their file have been removed from the cvs repository. In case you did not use `cvs update -dP`, re-run `cvs update -dP`, and remember to always use `cvs update -dP` from now on (or simply add a line `update -dP` to your `~/.cvsrc`)!)

1. In case only `make` fails, you can try running:

   ```
   make clean
   ```

   in your build directory and proceed with step 5; however, if `make` then still fails, you have to re-start with step 1.
2. Clean up your whole build directory (i.e., the one where you ran `configure` and `make`) by going there and running:

   ```
   make maintainer-clean
   ```

   In case your build directory is different from your source directory, you are advised to remove the whole build directory.
3. Go to the top-level source directory and run:

   ```
   ./de-bootstrap
   ```

and type `y` when asked whether to remove the listed files. This will remove all the files that were created during `bootstrap`. Only do this with sources obtained through CVS.

4. In the top-level source directory, re-run:

   ```
   ./bootstrap
   ```

   Only do this with sources obtained through CVS.

5. In the build-directory, re-run:

   ```
   configure
   ```

   as described above.

6. In the build-directory, re-run:

   ```
   make
   make install
   ```

   as described above.

If this still does not help, please contact us.

1.23 Reporting Problems

Bugs and other problems with compiling or running MonetDB should be reported using the bug tracking system at SourceForge (preferred) or emailed to monet@cwi.nl; see also http://monetdb.cwi.nl/Development/Bugtracker/index.html. Please make sure that you give a *detailed* description of your problem!

1.24 Building MonetDB On Windows

In this document we describe how to build the MonetDB suite of programs on Windows using the sources from our source repository at SourceForge. This document is mainly targeted at building on Windows XP on a 32-bit architecture, but there are notes throughout about building on Windows XP x64 which is indicated with Windows64.

1.25 Introduction

The MonetDB suite of programs consists of a number of components which we will describe briefly here. The general rule is that the components should be compiled and installed in the order given here, although some components can be compiled and installed in a different order. Unless you know the inter-component dependencies, it is better to stick to this order. Also note that before the next component is built, the previous ones need to be installed. The section names are the names of the CVS modules on SourceForge.

1.26 buildtools

The buildtools component is required in order to build the sources from the CVS repository. If you get the pre-packaged sources (i.e. the one in tar balls), you don't need the buildtools component (although this has not been tested on Windows).

1.27 MonetDB

Also known as the MonetDB Common component contains the database kernel, i.e. the heart of MonetDB, and some generally useful libraries. This component is required.

1.28 clients

Also known as the MonetDB Client component contains a library which forms the basis for communicating with the MonetDB server components, and some interface programs that use this library to communicate with the server. This component is required.

1.29 MonetDB4

The deprecated (but still used) database server MonetDB4 Server. This component is still required for the MonetDB XQuery (pathfinder) component. This is the old server which uses MIL (the MonetDB Interface Language) as programming interface. This component is only required if you need MIL or if you need the MonetDB XQuery component.

1.30 MonetDB5

The MonetDB5 Server component is the new database server. It uses MAL (the MonetDB Algebra Language) as programming interface. This component is required if you need MAL or if you need the MonetDB SQL component.

1.31 sql

Also known as MonetDB SQL, this component provides an SQL frontend to MonetDB5. This component is required if you need SQL support.

1.32 pathfinder

Also known as MonetDB XQuery, this component provides an XQuery query engine on top of a relational database. You can store XML documents in the database and query these documents using XQuery. This component is required if you need XML/XQuery support.

1.33 java

Also known as MonetDB Java, this component provides both the MonetDB JDBC driver and the XRPC wrapper. This component is optional.

1.34 geom

The geom component provides a module for the MonetDB SQL frontend. This component is optional.

1.35 testing

The testing component contains some files and programs we use for testing the MonetDB suite. This component is optional.

1.36 Prerequisites

In order to compile the MonetDB suite of programs, several other programs and libraries need to be installed. Some further programs and libraries can be optionally installed to enable optional features. The required programs and libraries are listed in this section, the following section lists the optional programs and libraries.

1.37 CVS (Concurrent Version System)

All sources of the MonetDB suite of programs are stored using CVS at SourceForge. You will need CVS to get the sources. We use CVS under Cygwin, but any other version will do as well.

1.38 Compiler

The suite can be compiled using one of the following compilers:

- Microsoft Visual Studio .NET 2003 (also known as Microsoft Visual Studio 7);
- Microsoft Visual Studio 2005 (also known as Microsoft Visual Studio 8);
- Intel(R) C++ Compiler 9.1 (which actually needs one of the above);
- Intel(R) C++ Compiler 10.1 (which also needs one of the Microsoft compilers).

Note that the pathfinder component can currently not be compiled with any of the Microsoft compilers. It can be compiled with the Intel compiler.

Not supported anymore (but probably still possible) are the GNU C Compiler gcc under Cygwin. Using that, it (probably still) is possible to build a version that runs using the Cygwin DLLs, but also a version that uses the MinGW (Minimalist GNU for Windows) package. This is not supported and not further described here.

1.39 Python

Python is needed for creating the configuration files that the compiler uses to determine which files to compile. Python can be downloaded from http://www.python.org/. Just download and install the Windows binary distribution.

On Windows64 you can use either the 32-bit or 64-bit version of Python.

1.40 Bison

Bison is a reimplementation of YACC (Yet Another Compiler Compiler), a program to convert a grammar into working code.

A version of Bison for Windows can be gotten from the GnuWin32 project at http://gnuwin32.sourceforge.net/. Click on the Packages link on the left and then on Bison, and get the Setup file and install it.

1.41 Flex

Flex is a fast lexical analyzer generator.

A version of Flex for Windows can be gotten from the GnuWin32 project at http://gnuwin32.sourceforge.net/. Click on the Packages link on the left and then on Flex, and get the Setup file and install it.

1.42 Pthreads

Get a Windows port of pthreads from ftp://sources.redhat.com/pub/pthreads-win32/. You can download the latest pthreads-*-release.exe which is a self-extracting archive. Extract it, and move or copy the contents of the Pre-built.2 folder to `C:\Pthreads` (so that you end up with folders `C:\Pthreads\lib` and `C:\Pthreads\include`).

On Windows64, in a command interpreter, run `nmake clean VC` in the extracted `pthreads.2` folder with the Visual Studio environment set to the appropriate values, e.g. by executing the command `Open Visual Studio 2005 x64 Win64 Command Prompt`. Then copy the files `pthreadVC2.dll` and `pthreadVC2.lib` to `C:\Pthreads\lib`.

1.43 Diff

Diff is a program to compare two versions of a file and list the differences. This program is not used during the build process, but only during testing. As such it is not a strict prerequisite.

A version of Diff for Windows can be gotten from the GnuWin32 project at http://gnuwin32.sourceforge.net/. Click on the Packages link on the left and then on DiffUtils (note the name), and get the Setup file and install it.

1.44 Patch

Patch is a program to apply the output of diff to the original. This program is not used during the build process, but only for testing, and then only to approve results that were different from what was expected. As such it is not a strict prerequisite.

A version of Patch for Windows can be gotten from the GnuWin32 project at http://gnuwin32.sourceforge.net/. Click on the Packages link on the left and then on Patch, and get the Setup file and install it.

1.45 PsKill

PsKill is a program to kill (terminate) processes. This program is only used during testing to terminate tests that take too long.

PsKill is part of the Windows Sysinternals. Go to the Process Utilities, and get the PsKill package. PsKill is also part of the PsTools package and the Sysinternals Suite, so you can get those instead. Extract the archive, and make sure that the folder is in your `Path` variable when you run the tests.

1.46 PCRE (Perl Compatible Regular Expressions)

The PCRE library is used to extend the string matching capabilities of MonetDB. The PCRE library is required for the MonetDB5 component.

Download the source from http://www.pcre.org/. In order to build the library, you will need a program called `cmake` which you can download from http://www.cmake.org/. Follow the Download link and get the Win32 Installer, install it, and run it. It will come up with a window where you have to fill in the location of the source code and where to build the binaries. Fill in where you extracted the PCRE sources, and some other directory (I used a `build` directory which I created within the PCRE source tree). You need to configure some

PCRE build options. I chose to do build shared libs, to match newlines with the `ANYCRLF` option, and to do have UTF-8 support and support for Unicode properties. When you're satisfied with the options, click on Configure, and then on Generate. Then in the build directory you've chosen, open the PCRE.sln file with Visual Studio, and build and install. Make sure you set the Solution Configuration to Release if you want to build a releasable version of the MonetDB suite. The library will be installed in `C:\Program Files\PCRE`.

For Windows64, select the correct compiler (`Visual Studio 9 2008 Win64`) and proceed normally. When building the 32 bit version on Windows64, choose `C:/Program Files (x86)/PCRE` for the `CMAKE_INSTALL_PREFIX` value, otherwise choose `C:/Program Files/PCRE`.

1.47 OpenSSL

The OpenSSL library is used during authentication of a MonetDB client program with the MonetDB server. The OpenSSL library is required for the MonetDB5 component, and hence implicitly required for the clients component when it needs to talk to a MonetDB5 server.

Download the source from http://www.openssl.org/. We used the latest stable version (0.9.8k). Follow the instructions in the file `INSTALL.W32` or `INSTALL.W64`.

Fix the `OPENSSL` definitions in `MonetDB\NT\winrules.msc` so that they refer to the location where you installed the library and call `nmake` with the extra parameter `HAVE_OPENSSL=1`.

1.48 libxml2

Libxml2 is the XML C parser and toolkit of Gnome.

This library is only a prerequisite for the pathfinder component, although the MonetDB5 component can also make use of it.

The home of the library is http://xmlsoft.org/. But Windows binaries can be gotten from http://www.zlatkovic.com/libxml.en.html. Click on Win32 Binaries on the right, and download libxml2, iconv, and zlib. Install these in e.g. `C:\`.

Note that we hit a bug in version 2.6.31 of libxml2. See the bugreport. Use version 2.6.30 or 2.6.32.

On Windows64 you will have to compile libxml2 yourself (with its optional prerequisites iconv and zlib, for which see below).

Edit the file `win32\Makefile.msvc` and change the one occurrence of `zdll.lib` to `zlib1.lib`, and then run the following commands in the `win32` subdirectory, substituting the correct locations for the iconv and zlib libraries:

```
cscript configure.js compiler=msvc prefix=C:\libxml2-2.6.30.win64 ^
 include=C:\iconv-1.11.win64\include;C:\zlib-1.2.3.win64\include ^
 lib=C:\iconv-1.11.win64\lib;C:\zlib-1.2.3.win64\lib iconv=yes zlib=yes
nmake /f Makefile.msvc
nmake /f Makefile.msvc install
```

After this, you may want to move the file `libxml2.dll` from the `lib` directory to the `bin` directory.

1.49 geos (Geometry Engine Open Souce)

Geos is a library that provides geometric functions. This library is only a prerequisite for the geom component.

There are no Windows binaries available (not that I looked very hard), so to get the software, you will have to get the source and build it yourself.

Get the source tar ball from http://trac.osgeo.org/geos/#Download and extract somewhere. All the versions I have tried (up to 3.1.1) miss one essential file to build on Windows, so in addition get the file `nmake.opt` from http://svn.osgeo.org/geos/branches/3.1/ and copy it to the top of the extracted source directory. Then build using:

```
nmake /f makefile.vc
```

Then install the library somewhere, e.g. in `C:\geos-3.1.win32`:

```
mkdir C:\geos-3.1.win32
mkdir C:\geos-3.1.win32\lib
mkdir C:\geos-3.1.win32\bin
mkdir C:\geos-3.1.win32\include
mkdir C:\geos-3.1.win32\include\geos
copy source\geos_c_i.lib C:\geos-3.1.win32\lib
copy source\geos_c.dll C:\geos-3.1.win32\bin
copy source\headers C:\geos-3.1.win32\include
copy source\headers\geos C:\geos-3.1.win32\include\geos
copy capi\geos_c.h C:\geos-3.1.win32\include
```

1.50 Optional Packages

1.51 iconv

Iconv is a program and library to convert between different character encodings. We only use the library.

The home of the program and library is http://www.gnu.org/software/libiconv/, but Windows binaries can be gotten from the same site as the libxml2 library: http://www.zlatkovic.com/libxml.en.html. Click on Win32 Binaries on the right, and download iconv. Install in e.g. `C:\`. Note that these binaries are quite old (libiconv-1.9.2, last I looked).

On Windows64 you will have to compile iconv yourself. Get the source from the iconv website and extract somewhere. Note that with the 1.12 release, the libiconv developers removed support for building with Visual Studio but require MinGW instead, which means that there is no support for Windows64. In other words, get the latest 1.11 release.

Build using the commands:

```
nmake -f Makefile.msvc NO_NLS=1 DLL=1 MFLAGS=-MD PREFIX=C:\iconv-1.11.win64
nmake -f Makefile.msvc NO_NLS=1 DLL=1 MFLAGS=-MD PREFIX=C:\iconv-1.11.win64 install
```

Fix the `ICONV` definitions in `MonetDB\NT\winrules.msc` so that they refer to the location where you installed the library and call `nmake` with the extra parameter `HAVE_ICONV=1`.

1.52 zlib

Zlib is a compression library which is optionally used by both MonetDB and the iconv library. The home of zlib is http://www.zlib.net/, but Windows binaries can be gotten from the same site as the libxml2 library: http://www.zlatkovic.com/libxml.en.html. Click on Win32 Binaries on the right, and download zlib. Install in e.g. `C:\`.

On Windows64 you will have to compile zlib yourself. Get the source from the zlib website and extract somewhere. Open the Visual Studio 6 project file `projects\visualc6\zlib.dsw` and click on `Yes To All` to convert to the version of Visual Studio which you are using. Then add a x64 Solution Platform by selecting `Build -> Confguration Manager...`, in the new window, in the pull down menu under `Active solution platform:` select `<New...>`. In the pop up window select `x64` for the new platform, copying the settings from `Win32` and click on `OK`. Set the `Active solution configuration` to `DLL Release` and click on `Close`. Then build by selecting `Build -> Build Solution`. Create the directory where you want to install the binaries, e.g. `C:\zlib-1.2.3.win64`, and the subdirectories `bin`, `include`, and `lib`. Copy the files `zconf.h` and `zlib.h` to the newly created `include` directory. Copy the file `projects\visualc6\win32_dll_release\zlib1.lib` to the new `lib` directory, and copy the file `projects\visualc6\win32_dll_release\zlib1.dll` to the new `bin` directory.

1.53 Perl

Perl is only needed to create an interface that can be used from a Perl program to communicate with a MonetDB server.

We have used ActiveState's ActivePerl distribution (release 5.10.0.1003). Just install the 32 or 64 bit version and compile the clients component with the additional `nmake` flags `HAVE_PERL=1 HAVE_PERL_DEVEL=1 HAVE_PERL_SWIG=1` (the latter flag only if SWIG is also installed).

1.54 PHP

PHP is only needed to create an interface that can be used from a PHP program to communicate with a MonetDB server.

Download the Windows installer and source package of PHP 5 from http://www.php.net/. Install the binary package and extract the sources somewhere (e.g. as a subdirectory of the binary installation).

In order to get MonetDB to compile with these sources a few changes had to be made to the sources:

- In the file `Zend\zend.h`, move the line

  ```
  #include <stdio.h>
  ```

  down until just *after* the block where `zend_config.h` is included.
- In the file `main\php_network.h`, delete the line

  ```
  #include "arpa/inet.h"
  ```

We have no support yet for Windows64.

1.55 SWIG (Simplified Wrapper and Interface Generator)

We use SWIG to build interface files for Perl and Python. You can download SWIG from http://www.swig.org/download.html. Get the latest swigwin ZIP file and extract it somewhere. It contains the `swig.exe` binary.

1.56 Java

If you want to build the java component of the MonetDB suite, you need Java. Get Java from http://java.sun.com/, but make sure you do *not* get the latest version. Get the Java Development Kit 1.5. Our current JDBC driver is not compatible with Java 1.6 yet, and the XRPC wrapper is not compatible with Java 1.4 or older.

In addition to the Java Development Kit, you will also need Apache Ant which is responsible for the actual building of the driver.

1.57 Apache Ant

Apache Ant is a program to build other programs. This program is only used by the java component of the MonetDB suite.

Get the Binary Distribution from http://ant.apache.org/, and extract the file somewhere.

1.58 Build Environment

1.59 Placement of Sources

For convenience place the various MonetDB packages in sibling subfolders. You will need at least:

- buildtools
- MonetDB
- clients
- one or both of MonetDB4, MonetDB5

Optionally:

- sql (requires MonetDB5)
- pathfinder (requires MonetDB4)

Apart from buildtools, all packages contain a subfolder `NT` which contains a few Windows-specific source files, and which is the directory in which the Windows version is built. (On Unix/Linux we recommend to build in a new directory which is not part of the source tree, but on Windows we haven't made this separation.)

1.60 Build Process

We use a command window `cmd.exe` (also known as `%ComSpec%`) to execute the programs to build the MonetDB suite. We do not use the point-and-click interface that Visual Studio offers. In fact, we do not have project files that would support building using the Visual Studio point-and-click interface.

We use a number of environment variables to tell the build process where other parts of the suite can be found, and to tell the build process where to install the finished bits.

In addition, you may need to edit some of the `NT\rules.msc` files (each component has one), or the file `NT\winrules.msc` in the MonetDB component which is included by all `NT\rules.msc` files.

1.61 Environment Variables

1.62 Compiler

Make sure that the environment variables that your chosen compiler needs are set. A convenient way of doing that is to use the batch files that are provided by the compilers:

- Microsoft Visual Studio .NET 2003 (also known as Microsoft Visual Studio 7):
  ```
  call "%ProgramFiles%\Microsoft Visual Studio .NET 2003\Common7\Tools\vsvars32.bat"
  ```
- Microsoft Visual Studio 2005 (also known as Microsoft Visual Studio 8):
  ```
  call "%ProgramFiles%\Microsoft Visual Studio 8\Common7\Tools\vsvars32.bat"
  ```
- Intel(R) C++ Compiler 10.1.013:
  ```
  call "C:%ProgramFiles%\Intel\Compiler\C++\10.1.013\IA32\Bin\iclvars.bat"
  ```

When using the Intel compiler, you also need to set the `CC` and `CXX` variables:

```
set CC=icl -Qstd=c99 -GR- -Qsafeseh-
set CXX=icl -Qstd=c99 -GR- -Qsafeseh-
```

(These are the values for the 10.1 version, for 9.1 replace `-Qstd=c99` with `-Qc99`.)

1.63 Internal Variables

- `MONETDB_SOURCE` - source folder of the MonetDB component
- `CLIENTS_SOURCE` - source folder of the clients component
- `MONETDB4_SOURCE` - source folder of the MonetDB4 component
- `MONETDB5_SOURCE` - source folder of the MonetDB5 component
- `SQL_SOURCE` - source folder of the sql component
- `PATHFINDER_SOURCE` - source folder of the pathfinder component
- `MONETDB_BUILD` - build folder of the MonetDB component (i.e. `%MONETDB_SOURCE%\NT`)
- `CLIENTS_BUILD` - build folder of the clients component (i.e. `%CLIENTS_SOURCE%\NT`)
- `MONETDB4_BUILD` - build folder of the MonetDB4 component (i.e. `%MONETDB4_SOURCE%\NT`)
- `MONETDB5_BUILD` - build folder of the MonetDB5 component (i.e. `%MONETDB5_SOURCE%\NT`)
- `SQL_BUILD` - build folder of the sql component (i.e. `%SQL_SOURCE%\NT`)
- `PATHFINDER_BUILD` - build folder of the pathfinder component (i.e. `%PATHFINDER_SOURCE%\NT`)
- `MONETDB_PREFIX` - installation folder of the MonetDB component
- `CLIENTS_PREFIX` - installation folder of the clients component

- `MONETDB4_PREFIX` - installation folder of the MonetDB4 component
- `MONETDB5_PREFIX` - installation folder of the MonetDB5 component
- `SQL_PREFIX` - installation folder of the sql component
- `PATHFINDER_PREFIX` - installation folder of the pathfinder component

We recommend that the various `PREFIX` environment variables all point to the same location (all contain the same value) which is different from the source and build folders.

1.64 PATH and PYTHONPATH

Extend your `Path` variable to contain the various folders where you have installed the prerequisite and optional programs. The `Path` variable is a semicolon-separated list of folders which are searched in succession for commands that you are trying to execute (note, this is an example: version numbers may differ):

```
rem Python is required
set Path=C:\Python25;C:\Python25\Scripts;%Path%
rem Bison and Flex (and Diff)
set Path=%ProgramFiles%\GnuWin32\bin;%Path%
rem for testing: pskill
set Path=%ProgramFiles%\PsTools;%Path%
rem Java is optional, set JAVA_HOME for convenience
set JAVA_HOME=%ProgramFiles%\Java\jdk1.5.0_13
set Path=%JAVA_HOME%\bin;%ProgramFiles%\Java\jre1.5.0_13\bin;%Path%
rem Apache Ant is optional, but required for Java compilation
set Path=%ProgramFiles%\apache-ant-1.7.0\bin;%Path%
rem SWIG is optional
set Path=%ProgramFiles%\swigwin-1.3.31;%Path%
```

In addition, during the build process we need to execute some programs that were built and installed earlier in the process, so we need to add those to the `Path` as well. In addition, we use Python to execute some Python programs which use Python modules that were also installed earlier in the process, so we need to add those to the `PYTHONPATH` variable:

```
set Path=%BUILDTOOLS_PREFIX%\bin;%Path%
set Path=%BUILDTOOLS_PREFIX%\Scripts;%Path%
set PYTHONPATH=%BUILDTOOLS_PREFIX%\Lib\site-packages;%PYTHONPATH%
```

Here the variable `BUILDTOOLS_PREFIX` represents the location where the buildtools component is installed. This variable is not used internally, but only used here as a shorthand.

For testing purposes it may be handy to add some more folders to the `Path`. To begin with, all DLLs that are used also need to be found in the `Path`, various programs are used during testing, such as diff (from GnuWin32) and php, and Python modules that were installed need to be found by the Python interpreter:

```
rem Pthreads DLL
set Path=C:\Pthreads\lib;%Path%
rem PCRE DLL
set Path=C:\Program Files\PCRE\bin;%Path%
rem PHP binary
set Path=C:\Program Files\PHP;%Path%
```

```
if not "%MONETDB_PREFIX%" == "%SQL_PREFIX%" set Path=%SQL_PREFIX%\bin;%SQL_PREFIX%\lib;%SQL_
set Path=%MONETDB4_PREFIX%\lib\MonetDB4;%Path%
if not "%MONETDB_PREFIX%" == "%MONETDB4_PREFIX%" set Path=%MONETDB4_PREFIX%\bin;%MONETDB4_P
if not "%MONETDB_PREFIX%" == "%CLIENTS_PREFIX%" set Path=%CLIENTS_PREFIX%\bin;%CLIENTS_PREF
set Path=%MONETDB_PREFIX%\bin;%MONETDB_PREFIX%\lib;%Path%

set PYTHONPATH=%CLIENTS_PREFIX%\share\MonetDB\python;%PYTHONPATH%
set PYTHONPATH=%MONETDB_PREFIX%\share\MonetDB\python;%PYTHONPATH%
set PYTHONPATH=%SQL_PREFIX%\share\MonetDB\python;%PYTHONPATH%
```

1.65 Compilation

1.66 Building and Installing Buildtools

The buildtools component needs to be built and installed first:

```
cd ...\buildtools
nmake /nologo /f Makefile.msc "prefix=%BUILDTOOLS_PREFIX%" install
```

where, again, the `BUILDTOOLS_PREFIX` variable represents the location where the buildtools component is to be installed.

1.67 Building and Installing the Other Components

The other components of the MonetDB suite are all built and installed in the same way. Do note the order in which the components need to be built and installed: MonetDB, clients, MonetDB4/MonetDB5, sql/pathfinder. There is no dependency between MonetDB4 and MonetDB5. MonetDB4 is a prerequisite for pathfinder, and pathfinder can use MonetDB5 (there is some very preliminary support). Sql requires MonetDB5.

For each of the components, do the following:

```
cd ...\<component>\NT
nmake /nologo NEED_MX=1 ... "prefix=%..._PREFIX%"
nmake /nologo NEED_MX=1 ... "prefix=%..._PREFIX%" install
```

Here the first `...` needs to be replaced by a list of parameters that tells the system which of the optional programs and libraries are available. The following parameters are possible:

- `DEBUG=1` - compile with extra debugging information
- `NDEBUG=1` - compile without extra debugging information (this is used for creating a binary release);
- `HAVE_ICONV=1` - the iconv library is available;
- `HAVE_JAVA=1` - Java and Apache Ant are both available;
- `HAVE_LIBXML2=1` - the libxml2 library is available;
- `HAVE_MONETDB4=1` - for sql and pathfinder: MonetDB4 was compiled and installed;
- `HAVE_MONETDB5=1` - for sql and pathfinder: MonetDB5 was compiled and installed;
- `HAVE_MONETDB5_XML=1` - for sql and pathfinder: MonetDB5 was compiled with the xml2 library available (HAVE_LIBXML2=1), and hence provides XML support (i.e., module xml);

- `HAVE_MONETDB5_RDF=1` - for sql and pathfinder: MonetDB5 was compiled with the raptor library available (HAVE_RAPTOR=1), and hence provides RDF support (i.e., module rdf);
- `HAVE_RAPTOR=1` - the raptor library is available;
- `HAVE_NETCDF=1` - the netcdf library is available;
- `HAVE_OPENSSL=1` - the OpenSSL library is available;
- `HAVE_PERL=1` - Perl is available;
- `HAVE_PERL_DEVEL=1` - Perl development is possible (include files and libraries are available–also need `HAVE_PERL=1`);
- `HAVE_PERL_SWIG=1` - Perl development is possible and SWIG is available (also need `HAVE_PERL=1`);
- `HAVE_PHP=1` - PHP is available;
- `HAVE_PROBXML=1` - compile in support for probabilistic XML (an experimental extension to the pathfinder component);
- `HAVE_PYTHON=1` - Python is available.

In addition, you can add a parameter which points to a file with extra definitions for `nmake`. This is very convenient to define where all packages were installed that the build process depends on since you then don't have to edit any of the `rules.msc` files in the source tree:

- `"MAKE_INCLUDEFILE=..."` - file with extra `nmake` definitions.

It is recommended to at least put the `MAKE_INCLUDEFILE` parameter with argument in double quotes to protect any spaces that may appear in the file name.

The contents of the file referred to with the `MAKE_INCLUDEFILE` parameter may contain something like:

```
bits=32
PTHREAD_INCS=-IC:\Pthreads\include
PTHREAD_LIBS=C:\Pthreads\lib\pthreadVC2.lib
PHP_SRCDIR=C:\Program Files\PHP\php-5.2.6
PHP_INSTDIR=C:\Program Files\PHP
LIBPERL=C:\Perl
LIBPCRE=C:\Program Files\PCRE
LIBICONV=C:\iconv-1.11.win32
LIBZLIB=C:\zlib-1.2.3.win32
LIBXML2=C:\libxml2-2.6.32+.win32
```

1.68 Building Installers

Installers can be built either using the full-blown Visual Studio user interface or on the command line. To use the user interface, open one or more of the files `MonetDB4-XQuery-Installer.sln`, `MonetDB5-SQL-Installer.sln`, `MonetDB-ODBC-Driver.sln`, and `MonetDB5-Geom-Module.sln` in the installation folder and select `Build -> Build Solution`. To use the command line, execute one or more of the commands in the installation folder:

```
devenv MonetDB4-XQuery-Installer.sln /build
devenv MonetDB5-SQL-Installer.sln /build
devenv MonetDB-ODBC-Driver.sln /build
devenv MonetDB5-Geom-Module.sln /build
```

In both cases, use the solutions (`.sln` files) that are appropriate.

There is an annoying bug in Visual Studio on Windows64 that affects the MonetDB5-Geom-Module installer. The installer contains code to check the registry to find out where MonetDB5/SQL is installed. The bug is that the 64 bit installer will check the 32-bit section of the registry. The code can be fixed by editing the generated installer (`.msi` file) using e.g. the program `orca` from Microsoft. Open the installer in `orca` and locate the table `RegLocator`. In the Type column, change the value from `2` to `18` and save the file. Alternatively, use the following Python script to fix the `.msi` file:

```
# Fix a .msi (Windows Installer) file for a 64-bit registry search.
# Microsoft refuses to fix a bug in Visual Studio so that for a 64-bit
# build, the registry search will look in the 32-bit part of the
# registry instead of the 64-bit part of the registry.  This script
# fixes the .msi to look in the correct part.

import msilib
import sys
import glob

def fixmsi(f):
    db = msilib.OpenDatabase(f, msilib.MSIDBOPEN_DIRECT)
    v = db.OpenView('UPDATE RegLocator SET Type = 18 WHERE Type = 2')
    v.Execute(None)
    v.Close()
    db.Commit()

if __name__ == '__main__':
    for f in sys.argv[1:]:
        for g in glob.glob(f):
            fixmsi(g)
```

1.68.1 Daily Builds

Next to functionality and performance, stability and portability are first class goals of the MonetDB project. Pursuing these goals requires to constantly monitor the evolving MonetDB code base. For this purpose, we developed a test environment that automatically compiles and tests MonetDB (and its most prominent add-on packages) every night on a variety of system configurations.

Software patches and functional enhancements are checked into the repositories on a daily basis. A limited set of distribution packages is prepared to disseminate the latest to developers and application programmers as quickly as possible. Such builds may, however, contain bugs or sometimes even break old functionality.

The TestWeb provides access to the test web-site that summarizes the results of the Automated Testing activities on various platforms. It is a good starting point before picking up a daily build version.

Up to three versions of MonetDB are tested daily on all available platforms:

- the *cutting edge* development version ("Current"), i.e. the head of the main CVS branch;
- the canditate for the next (feature) release ("Candidate"), i.e. the head of the latest release candidate branch; and
- the latest release version ("Stable"), i.e. the head of the most recent release branch.

The test reports consist of three overview pages ("cross-check-lists") revealing the results of

1. all compilation steps (bootstrap, configure, make, make install),
2. testing via "make check" (using debugmask 10, i.e., exhaustive monitoring and correction of physical BAT properties is enabled in the server), and
3. testing via "Mtest.py -d0 -r" (using debugmask 0, i.e., all debugging is switched off in the server).

1.68.1.1 Stability

With a (code-wise) complex system like MonetDB, modifying the source code — be it for fixing bugs or for adding new features — always bears the risk of breaking or at least altering some existing functionality. To facilitate the task of detecting such changes, small test scripts together with their respective correct/expected ("stable") output are collected within the CVS repository of MonetDB. Given the complexity of MonetDB, there is no way to do anything close to "exhaustive" testing, hence, the idea is to continuously extend the test collection. E.g., each developer should add some tests as soon as she/he adds new functionality. Likewise, a test script should be added for each bug report to monitor whether/when the bug is fixed, and to prevent (or at least detect) future occurrences of the same bug. The collection consists for hundreds of test scripts, each covering many micro-functionality tests.

To run all the tests and compare their current output to their stable output, a tool called Mtest is included in the MonetDB code base. Mtest recursively walks through the source tree, runs tests, and checks for difference between the stable and the current output. As a result, Mtest creates the web interface that allows convenient access to the differences encountered during testing. Each developer is supposed to run "Mtest" (respectively "make check") on his/her favorite development platform and check the results before checking in her/his changes. During the automatic daily tests, "make check" and "Mtest" are run on all testing platforms and the TestWeb is generated to provide convenient access to the results.

1.68.1.2 Portability

Though Fedora Linux on AMD Athlon PC's is our main development platform at CWI, we do not limit our attention to this single platform. Supporting a broad range of hardware and software platforms is an important concern.

Using standard configuration tools like automake, autoconf, and libtool, we have the same code base compiling not only on various flavors of Unix (e.g., Linux, Cygwin, AIX, IRIX, Solaris, MacOS X) but also on native Windows. Furthermore, the very code base compiles with a wide spectrum of (C-) compilers, ranging from GNU's gcc over several native Unix compilers (IBM, SGI, Sun, Intel, Portland Group) to Microsoft's Visual Studio and Visual Studio .NET on Windows.

On the hardware side, we have (had) MonetDB running on "almost anything" from a Intel StrongARM-based Linux PDA with 64 MB of flash memory to an SGI Origin2000 with 32 MIPS R12k CPU's and a total of 64 GB of (shared) main memory.

1.69 Development Roadmap

In this section we summarize the MonetDB development roadmap. The information is organized around the major system components. A precise timeline can not be given. It depends too much on the available resources and urgency (= pressure) by our research needs and clients.

1.69.1 Server Roadmap

The MonetDB server code base is continously being improved. A few areas under development in the kernel area are:

- *Parallelism* Exploitation of multi-core systems calls for renewed attention to parallel processing of the MonetDB kernel. Stress testing of concurrent processing may reveal race conditions hereto undetected.
- *Streaming Data* A separete area is support for streaming database functionality. It requires additions to the way we support io-channels and schedule query plans.
- *Functional Enhancements* Support for geographical application is underway. It consists of a concise library for managing geometric types.

1.69.2 SQL Roadmap

The long term objective for the SQL front-end is to provide all features available in SQL:2003. The priority for individual features is determined in an ad hoq way. The SQL features scheduled for implementation and those that won't be supported in the foreseeable future are shown below.

Our current assessment of the features planned for upcoming releases, in order of priority, are:

- *Window functions* Datawarehousing and data mining applications require support for windowing functions, e.g. (x() OVER (partition by order by)
- *Full text retrieval* A full text retrieval support function consists of a special constructed index over text appearing in multiple columns of a relational table. This index is built using well-known Information Retrieval techniques, such as stemming, keyword recognition, and stop-word reduction. Several IR projects are underway, which enhance MonetDB with IR capabilities.
- *Support for multi-media objects* MonetDB has been used in several multi-media projects, but mostly to store and manipulate derived features. Multimedia objects can be stored as unprotected URLs, i.e. there is no guarantee the object referred to exists upon answering a query. The functionality should be extended with image, audio, and video types.

- *Replication Service* A single-write multiple-read distributed replication service is prepared for release mid 2007. It will provide both the concept of merge tables and selective replication of tuples to different servers.
- *GIS support* Support for geographical application is underway. It consists of a concise library for managing geometric types.
- *General column and table constraint enforcement*
- *Internationalization of the character sets*
- *Full outer-join queries*

The database back-end architecture prohibits easy implementation of several SQL-99 features. Those on the list below are not expected to be supported.

- Cursor based processing, because the execution engine is not based on the iterator model deployed in other engines. A simulation of the cursor based scheme would be utterly expensive from a performance point of view.
- Multi-level transaction isolation levels. Coarse grain isolation is provided using table level locks.

1.69.3 Embedded MonetDB Roadmap

The embedded MonetDB software family provides support for both SQL and XQuery. The software has been tuned to run on small scale hardware platforms.

A broader deployment of the embedded technology requires both extensions in the distributed MonetDB versions and its replication services. Continual attention is given to the memory footprint and cpu/io resource consumptions on embedded devices.

A separate project, called the Datacell, is underway and geared at providing a streaming environment for embedded applications.

1.70 MonetDB Version 5

The MonetDB product family consists of a large number of components developed within our group over the last decade. Some components have already been shipped to happy customers, some are still in the making, and others have found a resting place in the attic.

The MonetDB architecture is designed to accommodate a wide-spectrum of standardized query language front-ends (SQL, XQuery), a variety of query transformation schemes, and different execution platforms (interpreted materialized or pipelined, dynamic compilation).

MonetDB Version 5 is a major release of our software infrastructure. The most notable differences are its greatly improved software stack and a new interface language, which turns the database server back-end into an abstract database machine with its associated assembly language (MAL). It supports backward compatibility of interfaces, tools, and source sharing where feasible within the limited scope of resources available.

In the remainder of this section we shortly introduce the MonetDB Version 5 design considerations and a quick overview of its architecture.

1.71 Design Considerations

Redesign of the MonetDB software stack was driven by the need to reduce the effort to extend the system into novel directions and to reduce the Total Execution Cost (TEC).

The TEC is what an end-user or application program will notice. The TEC is composed on several cost factors:

- A) API message handling
- P) Parsing and semantic analysis
- O) Optimization and plan generation
- D) Data access to the persistent store
- E) Execution of the query terms
- R) Result delivery to the application

Choosing an architecture for processing database operations pre-supposes an intuition on how the cost will be distributed. In an OLTP setting you expect most of the cost to be in (P,O), while in OLAP it will be (D,E,R). In a distributed setting the components (O,D,E) are dominant. Web-applications would focus on (A,E,R).

Such a simple characterization ignores the wide-spread differences that can be experienced at each level. To illustrate, in D) and R) it makes a big difference whether the data is already in the cache or still on disk. With E) it makes a big difference whether you are comparing two integers, evaluation of a mathematical function, e.g., Gaussian, or a regular expression evaluation on a string. As a result, intense optimization in one area may become completely invisible due to being overshadowed by other cost factors.

The Version 5 infrastructure is designed to ease addressing each of these cost factors in a well-defined way, while retaining the flexibility to combine the components needed for a particular situation. It results in an architecture where you assemble the components for a particular application domain and hardware platform.

The primary interface to the database kernel is still based on the exchange of text in the form of queries and simply formatted results. This interface is designed for ease of interpretation, versatility and is flexible to accommodate system debugging and application tool development. Although a textual interface potentially leads to a performance degradation, our experience with earlier system versions showed that the overhead can be kept within acceptable bounds. Moreover, a textual interface reduces the programming effort otherwise needed to develop test and application programs. The XML trend as the language for tool interaction supports our decision.

1.72 Architecture Overview

The architecture is built around a few independent components: the MonetDB server, the merovigian, and the client application. The MonetDB server is the heart of the system, it manages a single physical database on one machine for all (concurrent) applications. The merovigian program works along side a single server, keeping an eye on its behavior. If the server accidently crashes, it is this program that will attempt an automatic restart.

The top layer consists of applications written in your favorite language. They provide both specific functionality for a particular product, e.g., Proximity, and generic functionality, e.g., the Aquabrowser or Dbvisualizer. The applications communicate with the server using de-facto standard interface packaged, i.e., JDBC, ODBC, Perl, PHP, etc.

The middle layer consists of query language processors such as SQL and XQuery. The former supports the core functionality of SQL'99 and extends into SQL'03. The latter is based on the W3C standard and includes the XUpdate functionality. The query language processors each manage their own private catalog structure. Software bridges, e.g., import/export routines, are used to share data between language paradigms.

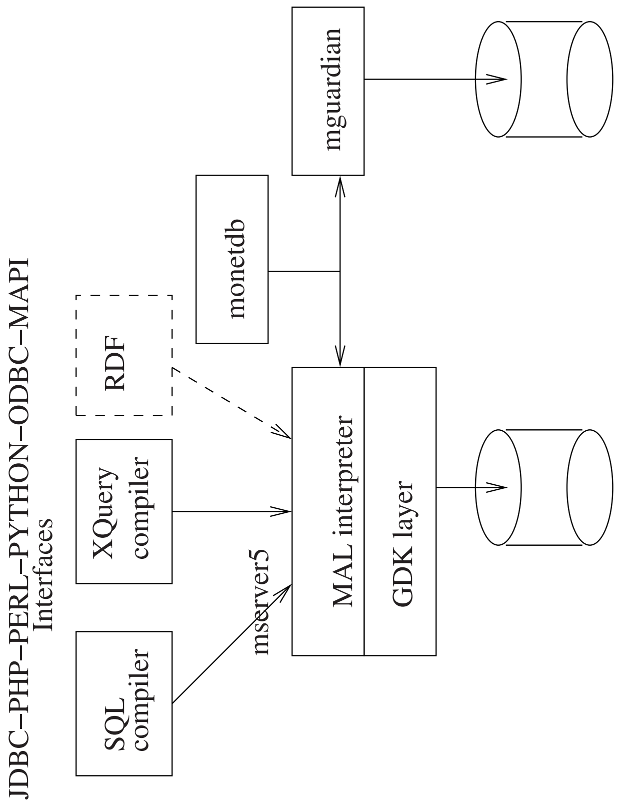


Figure 2.1

1.73 MonetDB Assembly Language (MAL)

The target language for a query compiler is the MonetDB Assembly Language (MAL). It was designed to ease code generation and fast interpretation by the server. The compiler produces algebraic query plans, which are turned into physical execution plans by the MAL optimizers.

The output of a compiler is either an ASCII representation of the MAL program or the compiler is tightly coupled with the server to save parsing and communication overhead.

A snippet of the MAL code produced by the SQL compiler for the query SELECT COUNT(*) FROM TABLES is shown below. It illustrates a sequences of relational operations against a table column and producing a partial result.

```
...
    _22:bat[:oid,:oid]  := sql.bind_dbat("tmp","_tables",0);
    _23 := bat.reverse(_22);
    _24 := algebra.kdifference(_20,_23);
    _25 := algebra.markT(_24,0:oid);
    _26 := bat.reverse(_25);
```

```
        _27 := algebra.join(_26,_20);
        _28 := bat.setWriteMode(_19);
        bat.append(_28,_27,true);
    ...
```

MAL supports the full breath of computational paradigms deployed in a database setting. It is language framework where the execution semantics is determined by the code transformations and the final engine choosen.

The design and implementation of MAL takes the functionality offered previously a significant step further. To name a few:

- All instructions are strongly typed before being executed.
- It supports polymorphic functions. They act as templates that produce strongly typed instantiations when needed.
- Function style expressions where each assignment instruction can receive multiple target results; it forms a point in the dataflow graph.
- It supports co-routines (Factories) to build streaming applications.
- Properties are associated with the program code for ease of optimization and scheduling.
- It can be readily extended with user defined types and function modules.

1.74 Execution Engine

The execution engine comes in several flavors. The default is a simple, sequential MAL interpreter. For each MAL function call it creates a stack frame, which is initialized with all constants found in the function body. During interpretation the garbage collector ensures freeing of space consumptive tables (BATs) and strings. Furthermore, all temporary structures are garbage collected before the funtion returns the result.

This simple approach leads to an accumulation of temporary variables. They can be freed earlier in the process using an explicit garbage collection command, but the general intend is to leave such decisions to an optimizer or scheduler.

The execution engine is only called when all MAL instructions can be resolved against the available libraries. Most modules are loaded when the server starts using a bootstrap script MAL_INIT.MX Failure to find the startup-file terminates the session. It most likely points to an error in the MonetDB configuration file.

During the boot phase, the global symbol table is initialized with MAL function and factory definitions, and loading the pre-compiled commands and patterns. The libraries are dynamically loaded by default. Expect tens of modules and hundreds of operations to become readily available.

Modules can not be dropped without restarting the server. The rational behind this design decision is that a dynamic load/drop feature is often hardly used and severely complicates the code base. In particular, upon each access to the global symbol table we have to be prepared that concurrent threads may be actively changing its structure. Especially, dropping modules may cause severe problems by not being able to detect all references kept around. This danger required all accesses to global information to be packaged in a critical section, which is known to be a severe performance hindrance.

1.75 Session Scenarios

In MonetDB multiple languages, optimizers, and execution engines can be combined at run time to satisfy a wide user-community. Such an assemblage of components is called a *scenario* and consists of a *reader*, *parser*, *optimizer*, *tactic scheduler* and *engine*. These hooks allow for both linked-in and external components.

The languages supported are SQL, XQuery, and the Monet Assembly Language (MAL). The default scenario handles MAL instructions, which is used to illustrate the behavior of the scenario steps.

The MAL reader component handles interaction with a front-end to obtain a string for subsequent compilation and execution. The reader uses the common stream package to read data in large chunks, if possible. In interactive mode the lines are processed one at a time.

The MAL parser component turns the string into an internal representation of the MAL program. During this phase semantic checks are performed, such that we end up with a type correct program.

The code block is subsequently sent to an MAL optimizer. In the default case the program is left untouched. For other languages, the optimizer deploys language specific code transformations, e.g., foreign-key optimizations in joins and remote query execution. All optimization information is statically derived from the code blocks and possible catalogues maintained for the query language at hand. Optimizers leave advice and their findings in properties in the symbol table, see Section 3.12 [Property Management], page 62.

Once the program has thus been refined, the MAL scheduler prepares for execution using tactical optimizations. For example, it may parallelize the code, generate an ad-hoc user-defined function, or prepare for efficient replication management. In the default case, the program is handed over to the MAL interpreter without any further modification.

The final stage is to choose an execution paradigm, i.e. interpretative (default), compilation of an ad-hoc user defined function, dataflow driven interpretation, or vectorized pipe-line execution by a dedicated engine.

A failure encountered in any of the steps terminates the scenario cycle. It returns to the user for a new command.

1.76 Scenario management

Scenarios are captured in modules; they can be dynamically loaded and remain active until the system is brought to a halt. The first time a scenario XYZ is used, the system looks for a scenario initialization routine XYZINITSYSTEM() and executes it. It is typically used to prepare the server for language specific interactions. Thereafter its components are set to those required by the scenario and the client initialization takes place.

When the last user interested in a particular scenario leaves the scene, we activate its finalization routine calling XYZEXITSYSTEM(). It typically perform cleanup, backup and monitoring functions.

A scenario is interpreted in a strictly linear fashion, i.e. performing a symbolic optimization before scheduling decisions are taken. The routines associated with each state in the scenario may patch the code so as to assure that subsequent execution can use a different scenario, e.g., to handle dynamic code fragments.

The building blocks of scenarios are routines obeying a strict name signature. They require exclusive access to the client record. Any specific information should be accessible from there, e.g., access to a scenario specific state descriptor. The client scenario initialization and finalization brackets are XYZINITCLIENT() and XYZEXITCLIENT().

The XYZPARSER(CLIENT C) contains the parser for language XYZ and should fill the MAL program block associated with the client record. The latter may have been initialized with variables. Each language parser may require a catalog with information on the translation of language specific datastructures into their BAT equivalent.

The XYZOPTIMIZER(CLIENT C) contains language specific optimizations using the MAL intermediate code as a starting point.

The XYZTACTICS(CLIENT C) synchronizes the program execution with the state of the machine, e.g., claiming resources, the history of the client or alignment of the request with concurrent actions (e.g., transaction coordination).

The XYZENGINE(CLIENT C) contains the applicable back-end engine. The default is the MAL interpreter, which provides good balance between speed and ability to analysis its behavior.

1.77 Server Management

This section presents the basics to manage a collection of MonetDB database servers. The system is designed to run out of the box for most end-users.

Additional finetuning by database administrators may be required in those cases where the MonetDB software is centrally made available or when mission critial databases are kept on highly-reliable production platforms.

1.77.1 Start and Stop the Server

On Windows starting a MonetDB server is done by simply clicking: 'Start' -> 'Programs' -> 'MonetDB5 ' -> 'MonetDB SQL Server'. This will start the MonetDB Server with SQL support in a separate window. Although this window comes with an interactive prompt, you should (unless you know what you are doing) keep this window minimized.

```
# MonetDB server v5.4, based on kernel v1.20.0
# Serving database 'demo'
# Compiled for i686-pc-win32/32bit with 32bit OIDs dynamically linked
# Copyright (c) 1993-2008 CWI, all rights reserved
# Visit http://monetdb.cwi.nl/ for further information
#warning: please don't forget to set your vault key!
#(see C:\Program Files\CWI\MonetDB5\etc\monetdb5.conf)
# Listening for connection requests on mapi:monetdb://127.0.0.1:50000/
# MonetDB/SQL module v2.20.2 loaded
>
```

The database is created in a default location with the name `demo`.

The textual interface shipped with the server can be started by clicking: 'Start' -> 'Programs' -> 'MonetDB5 ' -> 'MonetDB SQL Client'.

If you plan to make the server accessible from remote locations then the configuration file should be editted. See Section 1.77.4 [Database Configuration], page 38 for more details.

On UNIX-like systems, MonetDB/SQL comes with the following programs: MEROVINGIAN, MSERVER5, MONETDB and MCLIENT. MEROVINGIAN is a daemon process that controls a collection of database servers, i.e. MSERVER5 processes, each looking after a single physical database. Start this program to gain access to your MonetDB database farm. MEROVINGIAN is designed to be used in a system initialisation script in production environments.

With MEROVINGIAN running in the background, managing the databases and their connections is greatly simplified. After a fresh install the next step would typically be to create your first database, e.g. demo.

The program MONETDB is your aid here. It can create/destroy databases and it provides options to inspect the stability/liveliness of all database servers. Database servers can be temporarily closed for external access for maintenance, allowing for checkpointing. Let's create the demo database:

```
shell> monetdb create demo
successfully created database 'demo'
```

The status of all database servers can be inspected using:

```
shell> monetdb status
      name       state     uptime       health       last crash
demo            stopped
```

This report is helpful to determine possible instabilities and heavy loaded servers. In this case, it indicates that our database exists, but that no server is running yet.

```
shell> monetdb start demo
starting database 'demo'... done
shell> monetdb status demo
      name       state     uptime       health       last crash
demo            running       1m 18s  100%,  0s  -
```

Users can now establish a connection using any of the user interfaces. The most common one is MCLIENT, which provides a light-weight textual interface. For example, the statements below illustrate a short session. The session is closed using the mclient console command \q.

```
shell> mclient -lsql --database=demo
sql>CREATE USER "voc" WITH PASSWORD 'voc' NAME 'VOC Explorer' SCHEMA "sys";
sql>CREATE SCHEMA "voc" AUTHORIZATION "voc";
sql>ALTER USER "voc" SET SCHEMA "voc";
sql>\q
```

See for a more complete session VOC demo.

A database can be closed for maintenance. Doing so blocks any new non-administrator clients to connect to the server.

```
shell> monetdb lock demo
```

The effect is that only the system administrator can gain access to the server. All other users are warned using the message 'Database temporarily unavailable for maintenance' upon an attempt to connect. A database under maintenance will also not be automatically started by MEROVINGIAN if a client requests access to it, while not running.

After maintenance has been completed, the database server can be opened for connections using `monetdb release demo`.

For more details on merovingian and monetdb inspect their manual pages.

1.77.2 Database Dumps

An ASCII-based database dump is a safe scheme to transport a database to another platform or to migrate to an (incompatible) new version of MonetDB. This feature is standard available in MCLIENT.

1.77.3 Server Architecture

Maintenance of the MonetDB servers is based on a clear separation of tasks between multiple processes, directories, and their dependencies.

The anchor point for MonetDB is a directory (or folder) called the *dbfarm*. It contains sub-directories, one for each database. Similar, the database logs and checkpoint anchor points are *dblogs* and *dbarchive*. They should preferably be mounted on different storage devices.

Access restrictions are inherited from the operating system authorization scheme. It may proof useful to introduce a separate account for controlling access to MonetDB resources.

1.77.4 Database Configuration

The database environment is described in a configuration file, which is used by server-side applications, e.g. MEROVINGIAN and MSERVER5. A default version is created upon system installation in PREFIX/ETC/MONETDB5.CONF. Below we illustrate its most important components, for the remaining details look at the configuration file itself.

- prefix=/ufs/myhome/monet5/Linux
- exec_prefix=${prefix}
- dbfarm=${prefix}/var/MonetDB5/dbfarm
- monet_mod_path=${exec_prefix}/lib(64)/MonetDB5
- mal_init=${prefix}/lib(64)/MonetDB5/mal_init.mal
- sql_init=${exec_prefix }/lib/MonetDB5/sql_init.sql
- merovingian_log=${prefix}/var/merovingian.log
- mapi_open=false

The header consist of system wide information. The PREFIX and EXEC_PREFIX describe the location where MonetDB has been installed. MONET_MOD_PATH tells where to find the libraries. The bootstrap file for the kernel is given by MAL_INIT. These arguments are critical for a proper working server.

The option SQL_INIT is a comma separated list of SQL files to be executed upon system restart. It is primarily used to make SQL library functions known to all users.

The logs are typically stored on a different storage medium to protect the database against accidental hardware loss.

Client connections are limited to those originating from the same machine. To make the database accessible from remote sites the option MAPI_OPEN should be set to TRUE.

1.77.5 Checkpoint and Recovery

Safeguarding the content of your database against disasters, both hardware and malicious use, requires carefully planned steps. The first line of defense is to keep the database logs physically separated from the database store itself, e.g. on different disks. The second line of defense is to regularly create a database dump or full checkpoint. This is a consolidated snapshot and should be stored away at a failure independent location, e.g. a vault. Since a dump is a rather expensive operation, the third line of defense is to keep differential lists from the last dump based on the update logs. It forms a basis to rollback to a known correct state.

We are working on this topic

At the moment the best way to make a checkpoint is to make a database dump while the database is under maintenance. Use the monetdb utility to lock the database before dumping its contents.

1.77.6 Embedded Server

The Embedded Server version is optimized for running on small board computers as a database back-end for a single client. It is of particular interest if you need database functionality within a limited application setting, e.g a self-contained database distributed as part of the application. Within this context, much of the code to facilitate and protect concurrent use of the kernel can be disabled. For example, locking of critical resources in the kernel is not needed anymore, which results in significant performance improvements.

The approach taken is to wrap a server such that the interaction between client code and server can still follow the Mapi protocol. It leads to a C-program with calls to the Mapi library routines, which provides some protection against havoc behaviour. From a programming view, it differs from a client-server application in the startup and (implicit) termination.

You normally only have to change the call MAPI_CONNECT() into EMBEDDED_SQL() (or EMBEDDED_MAL()). It requires an optional argument list to refine the environment variables used by the server. In combination with the header file EMBEDDEDCLIENT.H it provides the basis to compile and link the program.

The behavior of an embedded SQL program can be simulated with a server started as follows:

```
mserver5 --set embedded=yes --dbinit="include sql;" &
```

As a result, the server starts in 'daemon' mode, loads the SQL support library, and waits for a connection. Only one connection is permitted.

1.77.6.1 Mbedded Example

A minimalistic embedded application is shown below. It creates a temporary table in the database, fills it, and retrieves the records for some statistics gathering.

The key operation is EMBEDDED_SQL() which takes an optional environment argument list. Upon success of this call, there will be a separate server thread running in the same user space to handle the database requests. A short-circuit interaction is established between the application and the kernel using in memory buffers.

The body of the program consists of the Mapi calls. It terminates with a call to MAPI_DISCONNECT() which lets the MonetDB thread gracefully die.

The tight coupling of application and kernel code also carries some dangers. Many of the MonetDB data structures can be directly accessed, or calls to the kernel routines are possible. It is highly advised to stick to the Mapi interaction protocol. It gives a little more protection against malicious behavior or unintended side-effects.

```
#include <embeddedclient.h>
#include <stdlib.h>

#define die(dbh,hdl) (hdl?mapi_explain_result(hdl,stderr): \
              dbh?mapi_explain(dbh,stderr): \
              fprintf(stderr,"command failed\n"), \
              exit(-1))

int
main()
{
    Mapi dbh;
    MapiHdl hdl = NULL;
    int i;

    dbh = embedded_sql(NULL, 0);
    if (dbh == NULL || mapi_error(dbh))
        die(dbh, hdl);

    /* switch off autocommit */
    if (mapi_setAutocommit(dbh, 0) != MOK || mapi_error(dbh))
        die(dbh, NULL);

    if ((hdl = mapi_query(dbh, "create table emp (name varchar(20), age int)")) == NUL
        mapi_error(dbh))
        die(dbh, hdl);
    if (mapi_close_handle(hdl) != MOK)
        die(dbh, hdl);

    for (i = 0; i < 1000; i++) {
        char query[100];

        snprintf(query, 100, "insert into emp values('user%d', %d)", i, i % 82);
        if ((hdl = mapi_query(dbh, query)) == NULL || mapi_error(dbh))
            die(dbh, hdl);
        if (mapi_close_handle(hdl) != MOK)
            die(dbh, hdl);
    }

    if ((hdl = mapi_query(dbh, "select * from emp")) == NULL || mapi_error(dbh))
        die(dbh, hdl);
    i = 0;
    while (mapi_fetch_row(hdl)) {
        char *age = mapi_fetch_field(hdl, 1);

        i = i + atoi(age);
    }
    if (mapi_error(dbh))
        die(dbh, hdl);
    if (mapi_close_handle(hdl) != MOK)
        die(dbh, hdl);
    printf("The footprint is %d Mb \n", i);

    mapi_disconnect(dbh);
    return 0;
}
```

The embedded MonetDB engine is available as the library LIBEMBEDDED_SQL.A (and LIBEMBEDDED_MAL.A) to be linked with a C-program. Provided the programming environment have been initialized properly, it suffices to prepare the embedded application using

```
    gcc -g myprog.c -o myprog \
        `monetdb-sql-config --cflags --libs` \
        `monetdb-clients-config --cflags --libs` \
        `monetdb-config --cflags --libs` \
        `monetdb5-config --cflags --libs` \
        -lMapi -lembeddedsql5
```

The configuration parameters for the server are read from its default location in the file system. In an embedded setting this location may not be accessible. It requires calls to MO_ADD_OPTION() before you asks for the instantiation of the server code itself. The code snippet below illustrate how our example is given hardwired knowledge on the desired settings:

```
    main(){
        opt *set = NULL;
        int setlen = 0;
    ...
     if (!(setlen = mo_builtin_settings(&set)))
            usage(prog);
    ...
     /* needed to prevent the MonetDB config file from being used */
        setlen = mo_add_option(&set, setlen, opt_config, "dbfarm", ".");
        setlen = mo_add_option(&set, setlen, opt_config, "dbname", "demo");
    ...
        setlen = mo_system_config(&set, setlen);
        mid = embedded_mal(set, setlen);
```

For a complete picture see the sample program in the distribution.

1.77.6.2 Limitations for Embedded MonetDB

In embedded applications the memory footprint is a factor of concern. The raw footprint as delivered by the Unix SIZE command is often used. It is, however, also easily misleading, because the footprint depends on both the code segments and buffered database partitions in use. Therefore it makes sense to experiment with a minimal, but functionally complete application to decide if the resources limitations are obeyed.

The minimal static footprint of MonetDB is about 16 Mb (+ ca 4Mb for SQL). After module loading the space quickly grows to about 60Mb. *This footprint should be reduced.*

The embedded application world calls for many, highly specialized enhancements. It is often well worth the effort to carve out the functionality needed from the MonetDB software packages. The easiest solution to limit the functionality and reduce resource consumption is to reduce the modules loaded. This requires patches to the startup scripts.

The benefit of an embedded database application also comes with limitations. The one and foremost limitation of embedded MonetDB is that the first application accessing the database effectively locks out any other concurrent use. Even in those situations where concurrent applications merely read the database, or create privately held tables.

2 Client Interfaces

Clients gain access to the Monet server through a internet connection or through its server console. Access through the internet requires a client program at the source, which addresses the default port of a running server. The functionality of the server console is limited. It is a textual interface for expert use.

At the server side, each client is represented by a session record with the current status, such as name, file descriptors, namespace, and local stack. Each client session has a dedicated thread of control, which limits the number of concurrent users to the thread management facilities of the underlying operating system. A large client base should be supported using a single server-side client thread, geared at providing a particular service.

The number of clients permitted concurrent access is a compile time option. The console is the first and is always present. It reads from standard input and writes to standard output.

Client sessions remain in existence until the corresponding communication channels break or its retention timer expires The administrator and owner of a sesssion can manipulate the timeout with a system call.

2.1 The Mapi Client Utility

The `mclient` program is the universal command-line tool that implements the MAPI protocol for client-server interaction with MonetDB.

On a Windows platform it can be started using start->MonetDB->MonetDB SQL Client. Alternatively, you can use the command window to start `mclient.exe`. Be aware that your environment variables are properly set to find the libraries of interest.

On a Linux platform it provides readline functionality, which greatly improves user interaction. A history can be maintained to ease interaction over multiple sessions.

A `mclient` requires minimally a language and host or port argument. The default setting is geared at establishing a guest connection to a SQL or XQuery database at a default server running on the localhost. The `-h hostname` specifies on which machine the MonetDB server is running. If you communicate with a MonetDB server on the same machine, it can be omitted.

The timer switch reports on the round-about time for queries sent to the server. It provides a first impression on the execution cost.

```
Usage: mclient --language=(sql|xquery|mal|mil) [ options ]

Options are:
 -d database | --database=database  database to connect to
 -e          | --echo              echo the query
 -f kind     | --format=kind       specify output format {dm,xml} for XQuery, or {csv,t
 -H          | --history           load/save cmdline history (default off)█
 -h hostname | --host=hostname     host to connect to
 -i          | --interactive       read stdin after command line args
 -l language | --language=lang     {sql,xquery,mal,mil}
 -L logfile  | --log=logfile       save client/server interaction
 -P passwd   | --passwd=passwd     password
 -p portnr   | --port=portnr      port to connect to
 -s stmt     | --statement=stmt   run single statement
 -t          | --time             time commands
 -X          | --Xdebug           trace mapi network interaction
 -u user     | --user=user        user id
 -?          | --help             show this usage message
 -| cmd      | --pager=cmd        for pagination

SQL specific opions
 -r nr       | --rows=nr          for pagination
 -w nr       | --width=nr         for pagination
 -D          | --dump             create an SQL dump

XQuery specific options
 -C colname  | --collection=name  collection name
 -I docname  | --input=docname    document name, XML document on standard input
```

The default `mapi_port` TCP port used is 50000. If this port happens to be in use on the server machine (which generally is only the case if you run two MonetDB servers on it), you will have to use the `-p port` do define the port in which the `mserver` is listening. Otherwise, it may also be omitted. If there are more than one mserver running you must also specify the database name `-d database`. In this case, if your port is set to the wrong database, the connection will be always redirect to the correct one. Note that the default port (and other default options) can be set in the server configuration file.

Within the context of each query language there are more options. They can be shown usin the command \? or using the commandline.

For SQL there are several knobs to tune for a better rendering of result tables (\w).

```
shell> mclient -lsql --help
\?       - show this message
\<file   - read input from file
\>file   - save response in file, or stdout if no file is given
\|cmd    - pipe result to process, or stop when no command is given
\h       - show the readline history
\t       - toggle timer
\e       - echo the query in sql formatting mode
\D table- dumps the table, or the complete database if none given.
\d table- describe the table, or the complete database if none given.
\A       - enable auto commit
\a       - disable auto commit
\f       - format using a built-in renderer {csv,tab,raw,sql,xml}
\w#      - set maximal page width (-1=raw,0=no limit, >0 max char)
\r#      - set maximum rows per page (-1=raw)
\L file - save client/server interaction
\X       - trace mclient code
\q       - terminate session
```

2.1.1 Online help

The textual interface MCLIENT supports a limited form of online help commands. The argument is a (partial) operator call, which is looked up in the symbol table. If the pattern includes a '(' it also displays the signature for each match. The ARGUMENT TYPES and ADDRESS attributes are also shown if the call contains the closing bracket ')'.

```
>?bat.is
bat.isSynced
bat.isCached
bat.isPersistent
bat.isTransient
bat.isSortedReverse
bat.isSorted
bat.isaSet
bat.isaKey
>?bat.isSorted(
command bat.isSorted(b:bat[:any_1,:any_2]):bit
>?bat.isSorted()
command bat.isSorted(b:bat[:any_1,:any_2]):bit address BKCisSorted;
Returns whether a BAT is ordered on head or not.
```

The module and function names can be replaced by the wildcard character '*'. General regulat pattern matching is not supported.

```
>?*.print()
command color.print(c:color):void
pattern array.print(a:bat[:any_1,:any_2],b:bat[:any_1,:int]...):void
pattern io.print(b1:bat[:any_1,:any]...):int
pattern io.print(order:int,b:bat[:any_1,:any],b2:bat[:any_1,:any]...):int
pattern io.print(val:any_1):int
pattern io.print(val:any_1,lst:any...):int
pattern io.print(val:bat[:any_1,:any_2]):int
```

The result of the help command can also be obtained in a BAT, using the commands MANUAL.HELP. Keyword based lookup is supported by the operation MANUAL.SEARCH; Additional routines are available in the INSPECT module to built reflexive code.

2.2 Jdbc Client

The textual client using the JDBC protocol comes with several options to fine-tune the interaction with the database server. A synopsis of the calling arguments is given below

```
java -jar ${prefix}/share/MonetDB/lib/jdbcclient.jar  \
        [-h host[:port]] [-p port] \
[-f file] [-u user] [-l language] [-b [database]] \
[-d [table]] [-e] [-X<opt>]
```

or using long option equivalents –host –port –file –user –language –dump –echo –database. Arguments may be written directly after the option like -p50000.

If no host and port are given, localhost and 50000 are assumed. An .MONETDB file may exist in the user's home directory. This file can contain preferences to use each time the program is started. Options given on the command line override the preferences file. The .MONETDB file syntax is <option>=<value> where option is one of the options host, port, file, mode debug, or password. Note that the last one is perilous and therefore not available as command line option. If no input file is given using the -f flag, an interactive session is started on the terminal.

NOTE The JDBC protocol does not support the SQL DEBUG <query>, PROFILE <query>, and TRACE <query> options. Use the mclient tool instead. OPTIONS

`-h --host`
: The hostname of the host that runs the MonetDB database. A port number can be supplied by use of a colon, i.e. -h somehost:12345.

`-p --port`
: The port number to connect to.

`-f --file`
: A file name to use either for reading or writing. The file will be used for writing when dump mode is used (-d –dump). In read mode, the file can also be an URL pointing to a plain text file that is optionally gzip compressed.

`-u --user`
: The username to use when connecting to the database.

`-d --database`
: Try to connect to the given database (only makes sense if connecting to a DatabasePool, M5 or equivalent process).

`-l --language`
: Use the given language, for example 'xquery'.

`--help`
: This screen.

`--version`
: Display driver version and exit.

`-e --echo`
: Also outputs the contents of the input file, if any.

`-q --quiet`
: Suppress printing the welcome header.

`-D --dump`
: Dumps the given table(s), or the complete database if none given.

EXTRA OPTIONS

`-Xdebug`
: Writes a transmission log to disk for debugging purposes. If a file name is given, it is used, otherwise a file called monet<timestamp>.log is created. A given file will never be overwritten; instead a unique variation of the file is used.

`-Xembedded`
: Uses an "embedded" server instance. The argument to this option should be in the form of path/to/mserver:dbname[:dbfarm[:dbinit]].

`-Xhash`
: Use the given hash algorithm during challenge response. Supported algorithm names: SHA1, MD5, plain.

`-Xoutput`
: The output mode when dumping. Default is sql, xml may be used for an experimental XML output.

`-Xbatching`
: Indicates that a batch should be used instead of direct communication with the server for each statement. If a number is given, it is used as batch size.

I.e. 8000 would execute the contents on the batch after each 8000 read rows. Batching can greatly speedup the process of restoring a database dump.

3 MonetDB Assembly Language (MAL)

The primary textual interface to the Monetdb kernel is a simple, assembly-like language, called MAL. The language reflects the virtual machine architecture around the kernel libraries and has been designed for speed of parsing, ease of analysis, and ease of target compilation by query compilers. The language is not meant as a primary programming language, or scripting language. Such use is even discouraged.

Furthermore, a MAL program is considered a specification of intended computation and data flow behavior. It should be understood that its actual evaluation depends on the execution paradigm choosen in a scenario. The program blocks can both be interpreted as ordered sequences of assembler instructions, or as a representation of a data-flow graph that should be resolved in a dataflow driven manner. The language syntax uses a functional style definition of actions and mark those that affect the flow explicitly. Flow of control keywords identify a point to chance the interpretation paradigm and denote a synchronization point.

MAL is the target language for query compilers, such as the SQL and XQuery front-ends. Even simple SQL queries generate a long sequence of MAL instructions. They represent both the administrative actions to ensure binding and transaction control, the flow dependencies to produce the query result, and the steps needed to prepare the result set for delivery to the front-end.

Only when the algebraic structure is too limited (e.g. updates), or the database back-end lacks feasible builtin bulk operators, one has to rely on more detailed flow of control primitives. But even in that case, the basic blocks to be processed by a MAL back-end are considered large, e.g. tens of simple bulk assignment instructions.

The remainder of this chapter provide a concise overview of the language features and illustrative examples.

3.1 MAL Literals

Literals in MAL follow the lexical conventions of the programming language C. A default type is attached, e.g. the literal 1 is typed as an INT value. Likewise, the literal 3.14 is typed FLT rather than DBL.

A literal can be coerced to another type by tagging it with a type classifier, provided a coercion operation is defined. For example, 1:LNG marks the literal as of type LNG. and "1999-12-10":DATE creates a DATE literal.

MonetDB comes with the hardwired types BIT, BTE, CHR, WRD, SHT, INT, LNG, OID, FLT, DBL, STR and BAT, the bat identifier. The kernel code has been optimized to deal with these types efficiently, i.e. without unnecessary function call overheads. In addition, the system supports temporal types DATE, DAYTIME, TIME, TIMESTAMP, TIMEZONE, extensions to deal with IPv4 addresses and URLs using INET, URL, and several types to interact more closely with the kernel LOCK, SEMPHORE. This list can be extended with user defined types.

3.2 MAL Variables

Variables are denoted by identifers and implicitly defined upon first use. They take on a type through a type classifier or inherit it from the context in which they are first used, see Section 3.8 [MAL Type System], page 57.

Variables are organized into two classes, starting with and without an underscore. The latter are reserved as MAL parser tempoaries, whose name aligns with an entry in the symbol table. In general they can not be used in MAL programs, but they may become visible in MAL program listings or during debugging.

3.3 Instructions

A MAL instruction has purposely a simple format. It is syntactically represented by an assignment, where an expression (function call) delivers results to multiple target variables. The assignment patterns recognized are illustrated below.

```
(t1,..,t32) := module.fcn(a1,..,a32);
t1 := module.fcn(a1,..,a32);
t1 := v1 operator v2;
t1 := literal;
(t1,..,tn) := (a1,..,an);
```

Operators are grouped into user defined modules. Ommission of the module name is interpreter as the USER module.

Simple binary arithmetic operations are merely provided as a short-hand, e.g. the expression T:=2+2 is converted directly into T:= CALC.+(2,2).

Target variables are optional. The compiler introduces temporary variables to hold the result of the expression upon need. They won't show up when you list the MAL program unless it is used elsewhere.

For parsing simplicity, each instruction fits on a single line. Comments start with a sharp '#' and continues to the end of the line. They are retained in the internal code representation to ease debugging of compiler generated MAL programs.

The data structure to represent a MAL block is kept simple. It contains a sequence of MAL statements and a symbol table. The MAL instruction record is a code byte string overlaid with the instruction pattern, which contains references into the symbol tables and administrative data for the interpreter.

This method leads to a large allocated block, which can be easily freed. Variable- and statement- block together describe the static part of a MAL procedure. It carries enough information to produce a listing and to aid symbolic debugging.

3.4 MAL Flow-of-control

The flow of control within a MAL program block can be changed by tagging a statement with either RETURN, YIELD, BARRIER, CATCH, LEAVE, REDO, or EXIT.

The flow modifiers RETURN and YIELD mark the end of a call and return one or more results to the calling environment. The RETURN and YIELD are followed by a target list or an assignment, which is executed first.

The BARRIER (CATCH) and EXIT pair mark a guarded statement block. They may be nested to form a proper hierarchy identified by their primary target variable, also called the control variable.

The LEAVE and REDO are conditional flow modifiers. The control variable is used after the assignment statement has been evaluated to decide on the flow-of-control action to be taken. Built-in controls exists for booleans and numeric values. The barrier block is opened when the control variable holds true, when its numeric value >= 0, or when it is a non-empty string. The NIL value blocks entry in all cases.

Once inside the barrier you have an option to prematurely LEAVE it at the exit statement or to REDO interpretation just after the corresponding barrier statement. Much like 'break' and 'continue' statements in the programming language C. The action is taken when the condition is met.

The EXIT marks the exit for a block. Its optional assignment can be used to re-initialize the barrier control variables or wrap-up any related administration.

The barrier blocks can be properly nested to form a hierarchy of basic blocks. The control flow within and between blocks is simple enough to deal with during an optimizer stage. The REDO and LEAVE statements mark the partial end of a block. Statements within these blocks can be re-arranged according to the data-flow dependencies. The order of partial blocks can not be changed that easily. It depends on the mutual exclusion of the data flows within each partial block.

Common guarded blocks in imperative languages are the for-loop and if-then-else constructs. They can be simulated as follows.

Consider the statement FOR(I=1;I<10;I++) PRINT(I). The (optimized) MAL block to implement this becomes:

```
     i:= 1;
barrier B:= i<10;
     io.print(i);
     i:= i+1;
redo B:= i<10;
exit B;
```

Translation of the statement IF(I<1) PRINT("OK"); ELSE PRINT("WRONG"); becomes:

```
    i:=1;
barrier ifpart:= i<1;
    io.print("ok");
exit ifpart;
barrier elsepart:= i>=1;
    io.print("wrong");
exit elsepart;
```

Note that both guarded blocks can be interchanged without affecting the outcome. Moreover, neither block would have been entered if the variable happens to be assigned NIL.

The primitives are sufficient to model a wide variety of iterators, whose pattern look like:

```
barrier i:= M.newIterator(T);
    elm:= M.getElement(T,i);
    ...
    leave i:= M.noMoreElements(T);
    ...
    redo i:= M.hasMoreElements(T);
exit i:= M.exitIterator(T);
```

The semantics obeyed by the iterator implementations is as follows. The redo expression updates the target variable *i* and control proceeds at the first statement after the barrier when the barrier is opened by *i*. If the barrier could not be re-opened, execution proceeds with the first statement after the redo. Likewise, the leave control statement skips to the exit when the control variable *i* shows a closed barrier block. Otherwise, it continues with the next instruction. Note, in both failed cases the control variable is possibly changed.

A recurring situation is to iterate over the elements in a BAT. This is supported by an iterator implementation for BATs as follows:

```
barrier (idx,hd,tl):= bat.newIterator(B);
    ...
    redo (idx,hd,tl):= bat.hasMoreElements(B);
exit (ids,hd,tl);
```

Where idx is an integer to denote the row in the BAT, hd and tl denote values of the current element.

3.5 Exception handling

MAL comes with an exception handling mechanism, similar in style as found in modern programming languages. Exceptions are considered rare situations that alter the flow of control to a place where they can be handled. After the exceptional case has been handled the following options exist a) continue where it went wrong, b) retry the failed instruction, c) leave the block where the exception was handled, or d) pass the exception to an enclosing call. The current implementation of the MAL interpreter only supports c) and d).

3.5.1 Exception control

The exception handling keywords are: CATCH and RAISE The CATCH marks a point in the dataflow where an exception raised can be dealt with. Any statement between the point where it is raised and the catch block is ignored. Moreover, the CATCH ... EXIT block is ignored when no errors have occurred in the preceeding dataflow structure. Within the catch block, the exception variable can be manipulated without constraints.

An exception message is linked with a exception variable of type string. If this variable is defined in the receiving block, the exception message can be delivered. Otherwise, it implicitly raises the exception in the surrounding scope. The variable ANYexception can be used to catch them irrespective of their class.

After an exception has been dealt with the catch block can be left at the normal EXIT with the option LEAVE or continue after the failed instruction using a REDO. The latter case assumes the caught code block has been able to provide an alternative for the failed instruction. [todo, the redo from failed instruction is not implemented yet]

Both LEAVE and REDO are conditional flow of control modifiers, which trigger on a non-empty string variable. An exception raised within a catch-block terminates the function and returns control to the enclosing environment.

The argument to the catch statement is a target list, which holds the exception variables you are interested in.

The snippet below illustrates how an exception raised in the function IO.READ is caught using the exception variable IOerror. After dealing with it locally, it raises a new exception FATALERROR for the enclosing call.

```
io.write("Welcome");
...
catch IOerror:str;
print("input error on reading password");
raise FATALerror:= "Can’t handle it";
exit IOerror;
```

Since CATCH is a flow control modifier it can be attached to any assignment statement. This statement is executed whenever there is no exception outstanding, but will be ignored when control is moved to the block otherwise.

3.5.2 Builtin exceptions

The policy implemented in the MAL modules, and recognized by the interpreter, is to return a string value by default. A NULL return value indicates succesful execution; otherwise the string encodes information to analyse the error occurred.

This string pattern is strictly formatted and easy to analyse. It starts with the name of the exception variable to be set, followed by an indication where the exception was raise, i.e. the function name and the program counter, and concludes with specific information needed to interpret and handle the exception.

For example, the exception string 'MALException:Admin.main[2]:address of function missing' denotes an exception raised while typechecking a MAL program.

The exceptions captured within the kernel are marked as 'GDKerror'. At that level there is no knowledge about the MAL context, which makes interpretation difficult for the average programmer. Exceptions in the MAL language layer are denoted by 'MALerror', and query language exceptiosn fall in their own class, e.g. 'SQLerror'. Exceptions can be cascaded to form a trail of exceptions recognized during the exection.

3.6 Functions

MAL comes with a standard functional abstraction scheme. Functions are represented by MAL instruction lists, enclosed by a FUNCTION signature and END statement. The FUNCTION signature lists the arguments and their types. The END statement marks the end of this sequence. Its argument is the function name.

An illustrative example is:

```
function user.helloWorld(msg:str):str;
    io.print(msg);
    msg:= "done";
    return msg;
end user.helloWorld;
```

The module name 'user' designates the collection to which this function belongs. A missing module name is considered a reference to the current module, i.e. the last module or atom context openend. All user defined functions are assembled in the module USER by default.

The functional abstraction scheme comes with several variations: COMMANDS, PATTERNS, and FACTORIES. They are discussed shortly.

3.6.1 Polymorphic Functions

Polymorphic functions are characterised by type variables denoted by :ANY and an optional index. Each time a polymorphic MAL function is called, the symbol table is first inspected for the matching strongly typed version. If it does not exists, a copy of the MAL program is generated, whereafter the type variables are replaced with their concrete types. The new MAL program is immediately type checked and, if no errors occured, added to the symbol table.

The generic type variable :ANY designates an unknown type, which may be filled at type resolution time. Unlike indexed polymorphic type arguments, :ANY type arguments match possibly with different concrete types.

An example of a parameterised function is shown below:

```
function user.helloWorld(msg:any_1):any_1;
    io.print(msg);
    return user.helloWorld;
end helloWorld;
```

The type variables ensure that the return type equals the argument type. Type variables can be used at any place where a type name is permitted. Beware that polymorphic typed variables are propagated throughout the function body. This may invalidate type resolutions decisions taken earlier (See Section 3.8 [MAL Type System], page 57).

This version of HELLOWORLD can also be used for other arguments types, i.e. BIT,SHT,LNG,FLT,DBL,.... For example, calling HELLOWORLD(3.14:FLT) echoes a float value.

3.6.2 C functions

The MAL function body can also be implemented with a C-function. They are introduced to the MAL type checker by providing their signature and an ADDRESS qualifier for linkage.

We distinguish both COMMAND and PATTERN C-function blocks. They differ in the information accessible at run time. The COMMAND variant calls the underlying C-function, passing pointers to the arguments on the MAL runtime stack. The PATTERN command is passed pointers to the MAL definition block, the runtime stack, and the instruction itself. It can be used to analyse the types of the arguments directly.

For example, the definitions below link the kernel routine BKCINSERT_BUN with the function BAT.INSERT(). It does not fully specify the result type. The IO.PRINT() pattern applies to any BAT argument list, provided they match on the head column type. Such a polymorphic type list may only be used in the context of a pattern.

```
command bat.insert(b:bat[:any_1,:any_2], ht:any_1, tt:any_2)
:bat[:any_1,:any_2]
address BKCinsert_bun;

pattern io.print(b1:bat[:any_1,:any]...):int
address IOtable;
```

3.7 Factories

A convenient programming construct is the co-routine, which is specified as an ordinary function, but maintains its own state between calls, and permits re-entry other than by the first statement.

The random generator example is used to illustrate its definition and use.

```
factory random(seed:int,limit:int):int;
    rnd:=seed;
    lim:= limit;
barrier lim;
    leave lim:= lim-1;
    rnd:= rnd*125;
    yield rnd:= rnd % 32676;
    redo lim;
exit lim;
end random;
```

The first time this factory is called, a *plant* is created in the local system to handle the requests. The plant contains the stack frame and synchronizes access.

In this case it initializes the generator. The random number is generated and YIELD as a result of the call. The factory plant is then put to sleep. The second call received by the factory wakes it up at the point where it went to sleep. In this case it will find a REDO statement and produces the next random number. Note that also in this case a seed and limit value are expected, but they are ignored in the body. This factory can be called upon to generate at most 'limit' random numbers using the 'seed' to initialize the generator. Thereafter it is being removed, i.e. reset to the original state.

A cooperative group of factories can be readily constructed. For example, assume we would like the random factories to respond to both RANDOM(SEED,LIMIT) and RANDOM(). This can be defined as follows:

```
factory random(seed:int,limit:int):int;
    rnd:=seed;
    lim:= limit;
barrier lim;
    leave lim:= lim-1;
    rnd:= rnd*125;
    yield rnd:= rnd % 32676;
    redo lim;
exit lim;
end random;

factory random():int;
barrier forever:=true;
    yield random(0,0);
    redo forever;
exit forever;
end random;
```

3.7.1 Factory Ownership

For simple cases, e.g. implementation of a random function, it suffices to ensure that the state is secured between calls. But, in a database context there are multiple clients active. This means we have to be more precise on the relationship between a co-routine and the client for which it works.

The co-routine concept researched in Monet 5 is the notion of a 'factory', which consists of 'factory plants' at possibly different locations and with different policies to handle requests. Factory management is limited to its owner, which is derived from the module in which it is placed. By default Admin is the owner of all modules.

The factory produces elements for multiple clients. Sharing the factory state or even remote processing is up to the factory owner. They are set through properties for the factory plant.

The default policy is to instantiate one shared plant for each factory. If necessary, the factory can keep track of a client list to differentiate the states. A possible implementation would be:

```
factory random(seed:int,clientid:int):int;
    clt:= bat.new(:int,:int);
    bat.insert(clt,clientid,seed);
barrier always:=true;
    rnd:= algebra.find(clt,clientid);
catch   rnd; #failed to find client
    bat.insert(clt,clientid,seed);
    rnd:= algebra.find(clt,clientid);
exit    rnd;
    rnd:= rnd * 125;
    rnd:= rnd % 32676;
    algebra.replace(clt,clientid,rnd);
    yield rnd;
    redo always;
exit always;
end random;
```

The operators to built client aware factories are, FACTORIES.GETCALLER(), which returns a client index, FACTORIES.GETMODULE() and FACTORIES.GETFUNCTION(), which returns the identity of scope enclosed.

To illustrate, the client specific random generator can be shielded using the factory:

```
factory random(seed:int):int;
barrier always:=true;
    clientid:= factories.getCaller();
    yield user.random(seed, clientid);
    redo always;
exit always;
end random;
```

3.7.2 Complex Factories

The factory scheme can be used to model a volcano-style query processor. Each node in the query tree is an iterator that calls upon the operands to produce a chunk, which are combined into a new chunk for consumption of the parent. The prototypical join(R,S) query illustrates it. The plan does not test for all boundary conditions, it merely implements a nested loop. The end of a sequence is identified by a NIL chunk.

```
factory query();
    Left:= sql.bind("relationA");
    Right:= sql.bind("relationB");
    rc:= sql.joinStep(Left,Right);
barrier rc!= nil;
    io.print(rc);
    rc:= sql.joinStep(Left,Right);
    redo rc!= nil;
exit rc;
end query;

#nested loop join
factory sql.joinStep(Left:bat[:any,:any],Right:bat[:any,:any]):bat[:any,:any];
    lc:= bat.chunkStep(Left);
barrier outer:= lc != nil;
    rc:= bat.chunkStep(Right);
    barrier inner:= rc != nil;
        chunk:= algebra.join(lc,rc);
        yield chunk;
        rc:= bat.chunkStep(Right);
        redo inner:= rc != nil;
    exit inner;
    lc:= bat.chunkStep(Left);
    redo outer:= lc != nil;
exit outer;
    # we have seen everything
    return nil;
end joinStep;

#factory for left branch
factory chunkStepL(L:bat[:any,:any]):bat[:any,:any];
    i:= 0;
    j:= 20;
    cnt:= algebra.count(L);
barrier outer:= j<cnt;
    chunk:= algebra.slice(L,i,j);
    i:= j;
    j:= i+ 20;
    yield chunk;
    redo loop:= j<cnt;
```

```
exit outer;
    # send last portion
    chunk:= algebra.slice(L,i,cnt);
    yielD chunk;
    return nil;
end chunkStep;

#factory for right leg
factory chunkStepR(L:bat[:any,:any]):bat[:any,:any];
```

So far we haven't re-used the pattern that both legs are identical. This could be modeled by a generic chunk factory. Choosing a new factory for each query steps reduces the administrative overhead.

3.7.3 Materialized Views

An area where factories might be useful are support for materialized views, i.e. the result of a query is retained for ease of access. A simple strategy is to prepare the result once and return it on each successive call. Provided the arguments have not been changed. For example:

```
factory view1(l:int, h:int):bat[:oid,:str];
    a:bat[:oid,:int]:= bbp.bind("emp","age");
    b:bat[:oid,:str]:= bbp.bind("emp","name");
barrier always := true;
    lOld := l;
    hOld := h;
    c := algebra.select(a,l,h);
    d := algebra.semijoin(b,c);
barrier available := true;
    yield d;
    leave available := calc.!=(lOld,l);
    leave available := calc.!=(hOld,h);
    redo available := true;
exit available;
    redo always;
exit always;
end view1;
```

The code should be extended to also check validity of the BATs. It requires a check against the last transaction identifier known.

The Factory concept is still rather experimental and many questions should be considered, e.g. What is the lifetime of a factory? Does it persists after all clients has disappeared? What additional control do you need? Can you throw an exception to a Factory?

3.8 Type implementation

3.9 MAL Type System

The MAL type module overloads the atom structure managed in the GDK library. For the time being, we assume GDK to support at most 127 different atomic types. Type composition is limited to at most two builtin types to form a BAT. Furthermore, the polymorphic type ANY can be qualified with a type variable index ANY$I, where I is a digit (1-9). Beware, the TYPE_any is a speudo type known within MAL only.

3.10 Type Resolution

Given the interpretative nature of many of the MAL instructions, when and where type resolution takes place is a critical design issue. Performing it too late, i.e. at each instruction call, leads to performance problems if we derive the same information over and over again. However, many built-in operators have polymorphic typed signatures, so we cannot escape it altogether.

Consider the small illustrative MAL program:

```
function sample(nme:str, val:any_1):bit;
   c := 2 * 3;
   b := bbp.bind(nme);  #find a BAT
   h := algebra.select(b,val,val);
   t := aggr.count(h);
   x := io.print(t);
   y := io.print(val);
end sample;
```

The function definition is polymorphic typed on the 2nd argument, it becomes a concrete type upon invocation. The system could attempt a type check, but quickly runs into assumptions that generally do not hold. The first assignment can be type checked during parsing and a symbolic optimizer could even evaluate the expression once. Looking up a BAT in the buffer pool leads to an element :BAT[*ht,tt*] where *ht* and *tt* are runtime dependent types, which means that the selection operation can not be type-checked immediately. It is an example of an embedded polypmorphic statement, which requires intervention of the user/optimizer to make the type explicit before the type resolver becomes active. The operation COUNT can be checked, if it is given a BAT argument. This assumes that we can infer that 'h' is indeed a BAT, which requires assurance that ALGEBRA.SELECT produces one. However, there are no rules to avoid addition of new operators, or to differentiate among different implementations based on the argument types. Since PRINT(T) contains an undetermined typed argument we should postpone typechecking as well. The last print statement can be checked upon function invocation.

Life becomes really complex if the body contains a loop with variable types. For then we also have to keep track of the original state of the function. Or alternatively, type checking should consider the runtime stack rather than the function definition itself.

These examples give little room to achieve our prime objective, i.e. a fast and early type resolution scheme. Any non-polymorphic function can be type checked and marked type-safe upon completion. Type checking polymorphic functions are post-poned until a concrete type instance is known. It leads to a clone, which can be type checked and is entered into the symbol table.

3.10.1 User Defined Types

MonetDB supports an extensible type system to accomodate a wide spectrum of database kernels and application needs. The type administration keeps track of their properties and provides access to the underlying implementations.

MAL recognizes the definition of a new type by replacing the MODULE keyword with ATOM. Atoms definitions require special care, because their definition and properties should be communicated with the kernel library. The commands defined in an ATOM block are screened as of interest to the library.

MonetDB comes with the hardwired types BIT, CHR, SHT, INT, LNG, OID, FLT, DBL, STR and BAT, the representation of a bat identifier. The kernel code has been optimized to deal with these types efficiently, i.e. without unnecessary function call overheads.

A small collection of user-defined ATOM types is shipped with the sysem. They implement types considered essential for end-user applications, such as COLOR, DATE, DAYTIME, TIME, TIMESTAMP, TIMEZONE, BLOB, and INET, URL. They are implemented using the type extension mechanism described below. As such, they provide examples for future extensions. A concrete example is the 'blob' datatype in the MonetDB atom module library(see ../modules/atoms/blob.mx)

3.10.2 Defining your own types

For the courageous at heart, you may enter the difficult world of extending the kernel library. The easiest way is to derive the atom modules from one shipped in the source distributed. More involved atomary types require a study of the documentation associated with the atom structures (gdk_atoms), because you have to develop a handful routines complying with the signatures required in the kernel library. They are registered upon loading the ATOM module.

3.11 Boxed Variables

Clients sessions often come with a global scope of variable settings. Access to these global variables should be easy, but they should also provide protection against concurrent update when the client wishes to perform parallel processing. Likewise, databases, query languages, etc. may define constants and variables accessible, e.g., relational schemas, to a selected user group.

The approach taken is to rely on persistent object spaces as pioniered in Lynda and -later- JavaSpaces. They are called boxes in MonetDB and act as managed containers for persistent variables.

Before a client program can interact with a box, it should open it, passing qualifying authorization information and parameters to instruct the box-manager of the intended use. A built-in box is implicitly opened when you request for its service.

At the end of a session, the box should be closed. Some box-managers may implement a lease-scheme to automatically close interaction with a client when the lease runs out. Likewise, the box can be notified when the last reference to a leased object ceases to exist.

A box can be extended with a new object using the function deposit(name) with name a local variable. The default implementation silently accepts any new definition of the box. If the variable was known already in the box, its value is overwritten.

A local copy of an object can be obtained using the pattern 'take(name,[param])', where name denotes the variable of interest. The type of the receiving variable should match the one known for the object. Whether an actual copy is produced or a reference to a shared object is returned is defined by the box manager.

The object is given back to the box manager calling 'release(name)'. It may update the content of the repository accordingly, release locks, and move the value to persistent store. Whatever the semantics of the box requires. [The default implementation is a no-op]

Finally, the object manager can be requested to 'discard(name)' a variable completely. The default implementation is to reclaim the space in the box.

Concurrency control, replication services, as well as access to remote stores may be delegated to a box manager. Depending on the intended semantics, the box manager may keep track of the clients holding links to this members, provide a traditional 2-phase locking scheme, optimistic control, or check-out/check-in scheme. In all cases, these management issues are transparant to the main thread (=client) of control, which operates on a temporary snapshot. For the time being we realize the managers as critical code sections, i.e. one client is permitted access to the box space at a time.

Fo example, consider the client function:

```
function myfcn():void;
    b:bat[:oid,:int] := bbp.take("mytable");
    c:bat[:int,:str] := sql.take("person","age");
    d:= intersect(b,c);
    io.print(d);
    u:str:= client.take(user);
    io.print(u);
    client.release(user);
end function;
```

The function binds to a copy from the local persistent BAT space, much like bat-names are resolved in earlier MonetDB versions. The second statement uses an implementation of take that searches a variable of interest using two string properties. It illustrates that a box manager is free to extend/overload the predefined scheme, which is geared towards storing MAL variables.

The result bat C is temporary and disappears upon garbage collection. The variable U is looked up as the string object user.

Note that BATs B and C need be released at some point. In general this point in time does not coincide with a computational boundary like a function return. During a session, several bats may be taken out of the box, being processed, and only at the end of a session being released. In this example, it means that the reference to b and c is lost at the end of the function (due to garbarge collection) and that subsequent use requires another take() call. The box manager bbp is notified of the implicit release and can take garbage collection actions.

The box may be inspected at several times during a scenario run. The first time is when the MAL program is type-checked for the box operations. Typechecking a take() function is tricky. If the argument is a string literal, the box can be queried directly for the objects' type. If found, its type is matched against the lhs variable. This strategy fails in the situation when at runtime the object is subsequently replaced by another typed-instance in the box. We assume this not to happen and the exceptions it raises a valuable advice to reconsider the programming style.

The type indicator for the destination variable should be provided to proceed with proper type checking. It can resolve overloaded function selection.

Inspection of the Box can be encoded using an iterator at the MAL layer and relying on the functionality of the box. However, to improve introspection, we assume that all box implementations provide a few rudimentary functions, called objects(arglist) and dir(arglist). The function objects() produces a BAT with the object names, possibly limited to those identified by the arglist.

The world of boxes has not been explored deeply yet. It is envisioned that it could play a role to import/export different objects, e.g., introduce xml.take() which converts an XML document to a BAT, jpeg.take() similer for an image.

Nesting boxes is possible. It provides a simple containment scheme between boxes, but in general will interfere with the semantics of each box.

Each box has (should) have an access control list, which names the users having permission to read/write its content. The first one to create the box becomes the owner. He may grant/revoke access to the box to users on a selective basis.

3.11.1 Session Box

Aside from box associated with the modules, a session box is created dynamically on behalf of each client. Such boxes are considered private and require access by the user name (and password). At the end of a session they are closed, which means that they are saved in persistent store until the next session starts. For example:

```
function m():void;
    box.open("client_name");
    box.deposit("client_name","pi",3.417:flt);
    f:flt := box.take("client_name","pi");
    io.print(t);
    box.close("client_name");
end function;
```

In the namespace it is placed subordinate to any space introduced by the system administrator. It will contain global client data, e.g., user, language, database, port, and any other session parameter. The boxes are all collected in the context of the database directory, i.e. the directory <dbfarm>/box

3.11.2 Garbage Collection

The key objects managed by MonetDB are the persistent BATs, which call for an efficient scheme to make them accessible for manipulation in the MAL procedures taking into account a possibly hostile parallel access.

Most kernel routines produce BATs as a result, which will be referenced from the runtime stack. They should be garbage collected as soon as deemed possible to free-up space. By default, temporary results are garbage collected before returning from a MAL function.

3.11.3 Globale Environment

The top level interaction keeps a 'box' with global variables, i.e. each MAL statement is interpreted in an already initialized stack frame. This causes the following problems: 1) how to get rid of global variables and 2) how to deal with variables that can take 'any' type. It is illustrated as follows:

```
f:= const.take("dbname");
io.print(f);
```

When executed in the context of a function, the answer will be simple [nil]. The reason is that the expecteed type is not known at compilation time. The correct definition would have been

```
f:str:= const.take("dbname");
io.print(f);
```

3.12 Property Management

Properties come in several classes, those linked with the symbol table and those linked with the runtime environment. The former are determined once upon parsing or catalog lookup. The runtime properties have two major subclasses, i.e. reflective and prescriptive. The reflective properties merely provide a fast cache to information aggregated from the target. Prescriptive properties communicate desirable states, leaving it to other system components to reach this state at the cheapest cost possible. This multifacetted world makes it difficult to come up with a concise model for dealing with properties. The approach taken here is an experimental step into this direction.

This MAL_PROPERTIES module provides a generic scheme to administer property sets and a concise API to manage them. Its design is geared towards support of MAL optimizers, which typically make multiple passes over a program to derive an alternative, better version. Such code-transformations are aided by keeping track of derived information, e.g. the expected size of a temporary result or the alignment property between BATs.

Properties capture part of the state of the system in the form of an simple term expression (NAME, OPERATOR, CONSTANT). The property model assumes a namespace built around Identifiers. The operator satisfy the syntax rules for MAL operators. Conditional operators are quite common, e.g. the triple (count, <, 1000) can be used to denote a small table.

The property bearing objects in the MAL setting are variables (symbol table entries). The direct relationship between instructions and a target variable, make it possible to keep the instruction properties in the corresponding target variable.

Variables properties The variables can be extended at any time with a property set. Properties have a scope identical to the scope of the corresponding variable. Ommision of the operator and value turns it into a boolean valued property, whose default value is TRUE.

```
b{count=1000,sorted}:= mymodule.action("table");
name{aligngroup=312} := bbp.take("person_name");
age{aligngroup=312} := bbp.take("person_age");
```

The example illustrates a mechanism to maintain alignment information. Such a property is helpful for optimizers to pick an efficient algorithm.

MAL function signatures. A function signature contains a description of the objects it is willing to accept and an indication of the expected result. The arguments can be tagged with properties that 'should be obeyed, or implied' by the actual arguments. It extends the typing scheme used during compilation/optimization. Likewise, the return values can be tagged with properties that 'at least' exist upon function return.

```
function test(b:bat[:oid,:int]{count<1000}):bat[:oid,:int]{sorted}
   #code block
end test
```

These properties are informative to optimizers. They can be enforced at runtime using the operation OPTIMIZER.ENFORCERULES() which injects calls into the program to check them. An assertion error is raised if the property does not hold. The code snippet

```
z:= user.test(b);
```

is translated into the following code block;

```
mal.assert(b,"count","<",1000);
z:= user.test(b);
mal.assert(z,"sorted");
```

How to propagate properties? Property inspection and manipulation is strongly linked with the operators of interest. Optimizers continuously inspect and update the properties, while kernel operators should not be bothered with their existence. Property propagation is strongly linked with the actual operator implementation. We examine a few recurring cases.

V:=W; Both V and W should be type compatible, otherwise the compiler will already complain.(Actually, it requires V.type()==W.type() and ~V.isaConstant()) But what happens with all others? What is the property propagation rule for the assignment? Several cases can be distinguished:

I) W has a property P, unknown to V. II) V has a propery P, unknown to W. III) V has property P, and W has property Q, P and Q are incompatible. IV) V and W have a property P, but its value disaggrees.

case I). If the variable V was not initialized, we can simply copy or share the properties. Copying might be too expensive, while shareing leads to managing the dependencies. case II) It means that V is re-assigned a value, and depending on its type and properties we may have to 'garbage collect/finalize' it first. Alternatively, it could be interpreted as a property that will hold after assignment which is not part of the right-hand side expression. case III) if P and Q are type compatible, it means an update of the P value. Otherwise, it should generates an exception. case IV) this calls for an update of V.P using the value of W.P. How this should be done is property specific.

Overall, the policy would be to 'disgard' all knowledge from V first and then copy the properties from W.

[Try 1] V:= fcn(A,B,C) and signature fcn(A:int,B:int,C:int):int The signature provides several handles to attach properties. Each formal parameter could come with a list of 'desirable/necessary' properties. Likewise, the return values have a property set. This leads to the extended signature function fcn(A:T,....,B:T): (C:T...D:T) where each Pi denotes a property set. Properties P1..Pn can be used to select the proper function variant. At its worst, several signatures of fcn() should be inspected at runtime to find one with matching properties. To enable analysis and optimization, however, it should be clear that once the function is finished, the properties Pk..Pm exist.

[Try 2] V:= fcn(A,B,C) and signature fcn(A:int,B:int,C:int):int The function is applicable when a (simple conjuntive) predicate over the properties of the actual arguments holds. A side-effect of execution of the function leads to an update of the property set associated with the actual arguments. An example:

```
function fcn (A:int,B:bat[int,int],C:int):int
        ?
```

[Try 3] Organize property management by the processor involved, e.g. a cost-based optimizer or a access control enforcer. For each optimizer we should then specify the 'symbolic' effect of execution of instructions of interest. This means 'duplication' of the instruction set.

Can you drop properties? It seems possible, but since property operations occur before actual execution there is no guarantee that they actually take place.

[case: how to handle sort(b:bat):bat as a means to propagate] [actually we need an expression language to indicate the propety set, e.g. sort(b:bat):bat which first obtains the properties of b and extends it with sorted. A nested structure emerge

Is it necessary to construct the property list intersection? Or do we need a user defined function to consolidate property lists?]

Aside, it may be valuable to collect information on e.g. the execution time of functions as a basis for future optimizations. Rather then cluttering the property section, it makes sense to explicitly update this information in a catalog.

3.13 Properties at the MAL level

Aside from routines targeted as changing the MAL blocks, it should be possible to reason about the properties within the language itself. This calls for gaining access and update. For example, the following snippet shows how properties are used in a code block.

```
B := bbp.new(int,int);
I := properties.has(B,);
J := properties.get(B,);
print(J);

properties.set(B,,2315);
barrier properties.has(B,);
exit;
```

These example illustrate that the property manipulations are executed throug patterns, which also accept a stack frame.

Sample problem with dropping properties:

```
    B := bbp.new(int,int);
barrier tst:= randomChoice()
    I := properties.drop(B,);
exit tst;
```

3.14 The cost model problem

An important issue for property management is to be able to pre-calculate a cost for a MAL block. This calls for an cost model implementation that recognizes instructions of interest, understands and can deal with the dataflow semantics, and

For example, selectivity estimations can be based on a histogram associated with a BAT. The code for this could look like

```
B:= new(int,int);
properties.add(B,);
Z:= select(B,1,100);
```

Addition of a property may trigger its evaluation, provided enough information is available (e.g. catalog). The instruction triggers the calls properties.set(B,), properties.set(B,), and properties.set(B,) once a property evaluation engine is ran against the code block. After assignment to Z, we have to propagate properties properties.update(B,).

3.15 SQL case

To study the use of properties in the complete pipeline SQL-execution we contrive a small SQL problem. The person table is sorted by name, the car table is unsorted.

```
create table person(name varchar not null,
address varchar);
create table car(name varchar,
model varchar,
milage int not null);
select distinct name, model, milage
from person, car
where car.name= person.name
  and milage>60000;
```

3.16 Implementation rules

Properties can be associated with variables, MAL blocks, and MAL instructions. The property list is initialized upon explicit request only, e.g. by the frontend parser, a box manager, or as a triggered action.

Every property should come with a function that accepts a reference to the variable and updates the property record. This function is activated either once or automatically upon each selection.

3.17 Property ADT implementation

addProperty(O,P) adds property P to the list associated with O. If O represents a compound structure, e.g. a BAT, we should indicate the component as well. For example, addProperty(O,P,Ia,...Ib) introduces a property shared by the components Ia..Ib (indicated with an integer index.

hasProperty(O,P) is a boolean function that merely checks existence hasnotProperty(O,P) is the dual operation.

setProperty(O,P,V) changes the propety value to V. It may raise a PropertyUpdateViolation exception when this can not be realized. Note, the property value itself is changed, not the object referenced.

getProperty(O,P) retrieves the current value of a property. This may involve calling a function or running a database query.

setPropertyAttribute(O,P,A) changes the behavior of the property. For example, the attribute 'freeze' will result in a call to the underlying function only once and to cache the result for the remainder of the objects life time.

3.18 Predefined properties

The MAL language uses a few defaults, recognizable as properties

unsafe	function has side effects.Default, unsafe=off
read	data can be read but not updated
append	data can be appended

4 The MAL Interpreter

The MAL interpreter always works in the context of a single user session, which provides for storage access to global variables and modules.

4.1 MAL API

The linkage between MAL interpreter and compiled C-routines is kept as simple as possible. Basically we distinguish four kinds of calling conventions: CMDcall, FCNcall, FACcall, and PATcall. The FCNcall indicates calling a MAL procedure, which leads to a recursive call to the interpreter.

CMDcall initiates calling a linked function, passing pointers to the parameters and result variable, i.e. f(ptr a0,..., ptr aN) The function returns a MAL-SUCCEED upon success and a pointer to an exception string upon failure. Failure leads to raise-ing an exception in the interpreter loop, by either looking up the relevant exception message in the module administration or construction of a standard string.

The PATcall initiates a call which contains the MAL context, i.e. f(MalBlkPtr mb, MalStkPtr stk, InstrPtr pci) The mb provides access to the code definitions. It is primarilly used by routines intended to manipulate the code base itself, such as the optimizers. The Mal stack frame pointer provides access to the values maintained. The arguments passed are offsets into the stack frame rather than pointers to the actual value.

4.2 Exception handling

Calling a built-in or user-defined routine may lead to an error or a cached status message to be dealt with in MAL. To improve error handling in MAL, an exception handling scheme based on CATCH-EXIT blocks. The CATCH statement identifies a (string-valued) variable, which carries the exception message from the originally failed routine or RAISE exception assignment. During normal processing CATCH-EXIT blocks are simply skipped. Upon receiving an exception status from a function call, we set the exception variable and skip to the first associated CATCH-EXIT block. MAL interpretation then continues until it reaches the end of the block. If no exception variable was defined, we should abandon the function alltogether searching for a catch block at a higher layer.

Exceptions raised within a linked-in function requires some care. First, the called procedure does not know anything about the MAL interpreter context. Thus, we need to return all relevant information upon leaving the linked library routine.

Second, exceptional cases can be handled deeply in the recursion, where they may also be handled, i.e. by issueing an GDKerror message. The upper layers merely receive a negative integer value to indicate occurrence of an error somewhere in the calling sequence. We then have to also look into GDKerrbuf to see if there was an error raised deeply inside the system.

The policy is to require all C-functions to return a string-pointer. Upon a successfull call, it is a NULL string. Otherwise it contains an encoding of the exceptional state encountered. This message starts with the exception identifer, followed by contextual details.

```
    mal_export str catchKernelException(Client cntxt, str ret);
```

```
str catchKernelException(Client cntxt, str ret){
 str z;
 if(cntxt->errbuf && cntxt->errbuf[0] ) {
  if (ret != MAL_SUCCEED){
   z= (char*) GDKmalloc( strlen(ret)+strlen(cntxt->errbuf)+2);
   strcpy(z,ret);
   if (z[strlen(z)-1] != ' n') strcat(z," n");
   strcat(z,cntxt->errbuf);
  } else {
   /* trap hidden (GDK) exception */
   z= (char*) GDKmalloc( strlen("GDKerror:")+strlen(cntxt->errbuf)+2);
   sprintf(z,"GDKerror:%s n",cntxt->errbuf);
  }
  cntxt->errbuf[0] = ' 0';
 } else z = ret;
 return z;
}
```

4.3 Garbage collection

Garbage collection is relatively straightforward, because most values are retained on the stackframe of an interpreter call. However, two storage types and possibly user-defined type garbage collector definitions require attention: BATs and strings.

A key issue is to deal with temporary BATs in an efficient way. References to bats in the buffer pool may cause dangling references at the language level. This appears as soons as your share a reference and delete the BAT from one angle. If not carefull, the dangling pointer may subsequently be associated with another BAT

All string values are private to the VALrecord, which means they have to be freed explicitly before a MAL function returns. The first step is to always safe the destination variable before a function call is made.

All operations are responsible to properly set the reference count of the BATs being produced or destroyed. The libraries should not leave the physical reference count being set. This is only allowed during the execution of a GDK operation. All references should be logical.

4.4 MAL runtime stack

The runtime context of a MAL procedure is allocated on the runtime stack of the corresponding interpreter. Access to the elements in the stack are through index offsets, determined during MAL procedure parsing.

The scope administration for MAL procedures is decoupled from their actual runtime behavior. This means we are more relaxed on space allocation, because the size is determined by the number of MAL procedure definitions instead of the runtime calling behavior. (See mal_interpreter for details on value stack management)

The variable names and types are kept in the stack to ease debugging. The underlying string value need not be garbage collected. Runtime storage for variables are allocated on the stack of the interpreter thread. The physical stack is often limited in size, which calls for safeguarding their value and garbage collection before returning. A malicious procedure or implementation will lead to memory leakage.

A system command (linked C-routine) may be interested in extending the stack. This is precluded, because it could interfere with the recursive calling sequence of procedures. To accommodate the (rare) case, the routine should issue an exception to be handled by the interpreter before retrying. All other errors are turned into an exception, followed by continuing at the exception handling block of the MAL procedure.

5 The MAL Optimizer

One of the prime reasons to design the MAL intermediate language is to have a high-level description for database queries, which is easy to generate by a front-end compiler and easy to decode, optimize and interpret.

An optimizer needs several mechanisms to be effective. It should be able to perform a symbolic evaluation of a code fragment and collect the result in properties for further decision making. The prototypical case is where an optimizer estimates the result size of a selection.

Another major issue is to be able to generate and explore a space of alternative evaluation plans. This exploration may take place up front, but can also be ran at runtime for query fragments.

5.1 The Optimizer Landscape

A query optimizer is often a large and complex piece of code, which enumerates alternative evaluation plans from which 'the best' plan is selected for evaluation. Limited progress has been made sofar to decompose the optimizer into (orthogonal) components, because it is a common believe in research that a holistic view on the problem is a prerequisite to find the best plan. Conversely, commercial optimizers use a cost-model driven approach, which explores part of the space using a limited (up to 300) rewriting rules.

Our hypothesis is that query optimization should be realized with a collection of query optimizer transformers (QOT), each dedicated to a specific task. Furthermore, they are assembled in scenarios to support specific application domains or achieve a desired behavior. Such scenarios are selected on a session basis, a query basis, or dynamically at runtime; they are part of the query plan.

The query transformer list below is under consideration for development. For each we consider its goal, approach, and expected impact. Moreover, the minimal prerequisites identify the essential optimizers that should have done their work already. For example, it doesn't make sense to perform a static evaluation unless you have already propagated the constants using Alias Removal.

Constant expressions Goal: to remove scalar expressions which need be evaluated once during the query lifetime. Rationale: static expressions appear when variables used denote literal constants (e.g. 1+1), when catalog information can be merged with the plan (e.g. max(B.salary)), when session variables are used which are initialized once (e.g. user()). Early evaluation aids subsequent optimization. Approach: inspect all instructions to locate static expressions. Whether they should be removed depends on the expected re-use, which in most cases call for an explicit request upon query registration to do so. The result of a static evaluation provides a ground for alias removal. Impact: relevant for stored queries (MAL functions) Prereq: alias removal, common terms

Relational Expression Optimizer Goal: to evaluate a relational plan using properties of BATs, such as being empty or forming an aligned group. These optimizations assume that the code generator can detect properties while compiling e.g. an SQL query. Impact: high Prereq:

Alias Removal Goal: to reduce the number of variables referenceing the same value, thereby reducing the analysis complexity. Rationale: query transformations often result in replacing the right-hand side expression with a result variable. This pollutes the code block with simple assignments e.g. V:=T. Within the descendant flow the occurrence of V could be replaced by T, provided V is never assigned a new value. Approach: literal constants within a MAL block are already recognized and replaced by a single variable. Impact: medium.

Common Term Optimizer Goal: to reduce the amount of work by avoiding calculation of the same operation twice. Rationale: to simplify code generation for front-ends, they do not have to remember the subexpressions already evaluated. It is much easier to detect at the MAL level. Approach: simply walk through the instruction sequence and locate identical patterns. (Enhance is with semantic equivalent instructions) Impact: High Prereq: Alias Removal

Dead Code Removal Goal: to remove all instructions whose result is not used Rationale: due to sloppy coding or alternative execution paths dead code may appear. Als XML Pathfinder is expected to produce a large number of simple assignments. Approach: Every instruction should produce a value used somewhere else. Impact: low

Heuristic Rule Rewrites Goal: to reduce the volume as quick as possible. Rationale: most queries are focussed on a small part of the database. To avoid carrying too many intermediates, the selection should be performed as early as possible in the process. This assumes that selectivity factors are known upfront, which in turn depends on histogram of the value distribution. Approach: locate selections and push them back/forth through the flow graph. Impact: high

Join Path Optimizer Goal: to reduce the volume produced by a join sequence Rationale: join paths are potentially expensive operations. Ideally the join path is evaluated starting at the smallest component, so as to reduce the size of the intermediate results. Approach: to successfully reduce the volume we need to estimate their processing cost. This calls for statistics over the value distribution, in particular, correlation histograms. If statistics are not available upfront, we have to restore to an incremental algorithm, which decides on the steps using the size of the relations. Impact: high

Operator Sort Goal: to sort the dataflow graph in such a way as to reduce the cost, or to assure locality of access for operands. Rationale: A simple optimizer is to order the instructions for execution by permutation of the query components Approach: Impact:

Singleton Set Goal: to replace sets that are known to produce precisely one tuple. Rationale: Singleton sets can be represented by value pairs in the MAL program, which reduces to a scalar expression. Approach: Identify a set variable for replacement. Impact:

Range Propagation Goal: look for constant ranges in select statements and propagate them through the code. Rationale: partitioned tables and views may give rise to expressions that contain multiple selections over the same BAT. If their arguments are constant, the result of such selects can sometimes be predicted, or the multiple selections can be cascaded into a single operation. Impact: high, should be followed by alias removal and dead code removal

Result Cacher Goal: to reduce the processing cost by keeping track of expensive to compute intermediate results Rationale: Approach: result caching becomes active after an instruction has been evaluated. The result can be cached as long as its underlying operands remain unchanged. Result caching can be made transparent to the user, but affects the other quer optimizers. Impact: high

Vector Execution Goal: to rewrite a query to use a cache-optimal vector implementation Rationale: processing in the cache is by far the best you can get. However, the operands may far exceed the cache size and should be broken into pieces followed by a staged execution of the fragments involved. Approach: replace the query plan with fragment streamers Impact:

Staged Execution Goal: to split a query plan into a number of steps, such that the first response set is delivered as quickly as possible. The remainder is only produced upon request. Rationale: interactive queries call for quick response and an indication of the processing time involved to run it too completion. Approach: staged execution can be realized using a fragmentation scheme over the database, e.g. each table is replaced by a union of fragments. This fragmentation could be determined upfront by the user or is derived from the query and database statistics. impact: high

Code Parallizer Goal: to exploit parallel IO and cpu processing in both SMP and MPP settings. Rationale: throwing more resources to solve a complex query helps, provided it is easy to determine that parallel processing recovers the administrative overhead Approach: every flow path segment can be handled by an independent process thread. Impact: high

Query Evaluation Maps Goal: to avoid touching any tuple that is not relevant for answering a query. Rationale: the majority of work in solving a query is to disgard tuples of no interest and to find correlated tuples through join conditions. Ideally, the database learns these properties over time and re-organizes (or builts a map) to replace disgarding by map lookup. Approach: piggyback selection and joins as database fragmentation instructions Impact: high

MAL Compiler (tactics) Goal: to avoid interpretation of functional expressions Rationale: interpretation of arithmetic expressions with an interpreter is always expensive. Replacing a complex arithmetic expressin with a simple dynamically compiled C-functions often pays off. Especially for cached (MAL) queries Approach: Impact: high

Dynamic Query Scheduler (tactics) Goal: to organize the work in a way so as to optimize resource usage Rationale: straight interpretation of a query plan may not lead to the best use of the underlying resources. For example, the content of the runtime cache may provide an opportunity to safe time by accessing a cached source Approach: query scheduling is the last step before a relation algebra interpreter takes over control. The scheduling step involves a re-ordering of the instructions within the boundaries imposed by the flow graph. impact: medium

Aggregate Groups Goal: to reduce the cost of computing aggregate expressions over times Rationale: many of our applications call for calculation of aggregates over dynamically defined groupings. They call for lengtly scans and it pays to piggyback all aggregate calculates, leaving their result in the cache for later consumption (eg the optimizers) Approach: Impact: High

Data Cube optimizer Goal: to recognize data cube operations Rationale: Approach: Impact:

Demand Driven Interpreter (tactics) Goal: to use the best interpreter and libraries geared at the task at hand Rationale: Interpretation of a query plan can be based on different computational models. A demand driven interpretation starts at the intended output and 'walks' backward through the flow graph to collect the pieces, possibly in a pipelined fashion. (Vulcano model) Approach: merely calls for a different implementation of the core operators Impact: high

Iterator Strength Reduction Goal: to reduce the cost of iterator execution by moving instructions out of the loop. Rationale: although iteration at the MAL level should be avoided due to the inherent low performance compared to built-in operators, it is not forbidden. In that case we should confine the iterator block to the minimal work needed. Approach: inspect the flowgraph for each iterator and move instructions around. Impact: low

Accumulator Evaluation Goal: to replace operators with cheaper ones. Rationale: based on the actual state of the computation and the richness of the supporting libraries there may exists alternative routes to solve a query. Approach: Operator rewriting depends on properties. No general technique. The first implementation looks at calculator expressions such as they appear frequently in the RAM compiler. Impact: high Prerequisite: should be called after common term optimizer to avoid clashes. Status: Used in the SQL optimizer.

Code Inliner Goal: to reduce the calling depth of the interpreter and to obtain a better starting point for code squeezing Rationale: substitution of code blocks (or macro expansion) leads to longer linear code sequences. This provides opportunities for squeezing. Moreover, at runtime building and managing a stackframe is rather expensive. This should be avoided for functions called repeatedly. Impact: medium Status: Used in the SQL optimizer to handle SQL functions.

Code Outliner Goal: to reduce the program size by replacing a group with a single instruction Rationale: inverse macro expansion leads to shorter linear code sequences. This provides opportunities for less interpreter overhead, and to optimize complex, but repetative instruction sequences with a single hardwired call Approach: called explicitly to outline a module (or symbol) Impact: medium

Garbage Collector Goal: to release resources as quickly as possible Rationale: BATs referenced from a MAL program keep resources locked. Approach: In cooperation with a resource scheduler we should identify those that can be released quickly. It requires a forced gargabe collection call at the end of the BAT's lifespan. Impact: large Status: Implemented. Algorithm based on end-of-life-span analysis.

Foreign Key replacements Goal: to improve multi-attribute joins over foreign key constraints Rationale: the code produced by the SQL frontend involves foreign key constraints, which provides many opportunities for speedy code using a join index. Impact: large Status: Implemented in the SQL strategic optimizer.

5.1.1 Optimizer Dependencies

The optimizers are highly targeted to a particular problem. Aside from the resources available to invest in plan optimization, optimizers are partly dependent and may interfere.

To aid selection of the components of interest, we have grouped them in a preferred order of deployment.

Group A:	Code Inliner.
	Constant Expression Evaluator.
	Relational Expression Evaluator.
	Strength Reduction.
Group B:	Common Term Optimizer.
	Query Evaluation Maps.

Group C:	Join Path Optimizer.
	Ranges Propagation.
	Operator Cost Reduction.
	Operator Sort.
	Foreign Key handling.
	Aggregate Groups.
	Data Cube optimizer.
	Heuristic Rule Rewrite.
group D:	Code Parallizer.
	Accumulator Evaluations.
	Result Cacher.
	Replication Manager.
group E:	MAL Compiler.
	Dynamic Query Scheduler.
	Vector Execution.
	Staged Execution.
group F:	Alias Removal.
	Dead Code Removal.
	Garbage Collector.

The interaction also excludes combinations. For example, the Accumulator should be used after the Partition optimizer.

5.1.2 Optimizer Building Blocks

Some instructions are independent of the execution context. In particular, expressions over side-effect free functions with constant parameters could be evaluated before the program block is considered further.

A major task for an optimizer is to select instruction (sequences) which can and should be replaced with cheaper ones. The cost model underlying this decision depends on the processing stage and the overall objective. For example, based on a symbolic analysis their may exist better implementations within the interpreter to perform the job (e.g. hashjoin vs mergejoin). Alternative, expensive intermediates may be cached for later use.

Plan enumeration is often implemented as a Memo structure, which designates alternative sub-plans based on a cost metric. Perhaps we can combine these memo structures into a large table for all possible combinations encountered for a user.

The MAL language does not imply a specific optimizer to be used. Its programs are merely a sequence of specifications, which is interpreted by an engine specific to a given task. Activation of the engine is controlled by a scenario, which currently includes two hooks for optimization; a strategic optimizer and a tactical optimizer. Both engines take a MAL program and produce a (new/modified) MAL program for execution by the lower layers.

MAL programs end-up in the symbol table linked to a user session. An optimizer has the freedom to change the code, provided it is known that the plan derived is invariant to changes in the environment. All others lead to alternative plans, which should be collected as a trail of MAL program blocks. These trails can be inspected for a posteriori analysis, at least in terms of some statistics on the properties of the MAL program structures automatically. Alternatively, the trail may be pruned and re-optimized when appropriate from changes in the environment.

The rule applied for all optimizers is to not-return before checking the state of the MAL program, and to assure the dataflow and variable scopes are properly set. It costs some performance, but the difficulties that arise from optimizer interference are very hard to debug. One of the easiest pitfalls is to derive an optimized version of a MAL function while it is already referenced by or when polymorphic typechecking is required afterwards.

5.1.3 Building Your Own Optimizer

Implementation of your own MAL-MAL optimizer can best be started from refinement of one of the examples included in the code base. Beware that only those used in the critical path of SQL execution are thorouhly tested. The others are developed up to the point that the concept and approach can be demonstrated.

The general structure of most optimizers is to actively copy a MAL block into a new program structure. At each step we determine the action taken, e.g. replace the instruction or inject instructions to achieve the desired goal.

A tally on major events should be retained, because it gives valuable insight in the effectiveness of your optimizer. The effects of all optimizers is collected in a system catalog.

Each optimizer ends with a strong defense line, `optimizerCheck()` It performs a complete type and data flow analysis before returning. Moreover, if you are in debug mode, it will keep a copy of the plan produced for inspection. Studying the differences between optimizer steps provide valuable information to improve your code.

The functionality of the optimizer should be clearly delineated. The guiding policy is that it is always safe to not apply an optimizer step. This helps to keep the optimizers as independent as possible.

It really helps if you start with a few tiny examples to test your optimizer. They should be added to the Tests directory and administered in Tests/All.

Breaking up the optimizer into different components and grouping them together in arbitrary sequences calls for careful programming.

One of the major hurdles is to test interference of the optimizer. The test set is a good starting point, but does not garantee that all cases have been covered.

In principle, any subset of optimizers should work flawlessly. With a few tens of optimizers this amounts to potential millions of runs. Adherence to a partial order reduces the problem, but still is likely to be too resource consumptive to test continously.

5.1.4 Optimizer framework

The large number of query transformers calls for a flexible scheme for the deploy them. The approach taken is to make all optimizers visible at the language level as a signature `optimizer.F()` and `optimizer.F(mod,fcn)`. The latter designates a target function to be inspected by the optimizer `F()`. Then (semantic) optimizer merely inspects a MAL block for their occurrences and activitates it.

The optimizer routines have access to the client context, the MAL block, and the program counter where the optimizer call was found. Each optimizer should remove itself from the MAL block.

The optimizer repeatedly runs through the program until no optimizer call is found.

Note, all optimizer instructions are executed only once. This means that the instruction can be removed from further consideration. However, in the case that a designated function is selected for optimization (e.g., commonTerms(user,qry)) the pc is assumed 0. The first instruction always denotes the signature and can not be removed.

To safeguard against incomplete optimizer implementations it is advisable to perform an optimizerCheck at the end. It takes as arguments the number of optimizer actions taken and the total cpu time spent. The body performs a full flow and type check and re-initializes the lifespan administration. In debugging mode also a copy of the new block is retained for inspection.

5.1.5 Lifespan analysis

Optimizers may be interested in the characteristic of the barrier blocks for making a decision. The variables have a lifespan in the code blocks, denoted by properties beginLifespan,endLifespan. The beginLifespan denotes the intruction where it receives its first value, the endLifespan the last instruction in which it was used as operand or target.

If, however, the last use lies within a BARRIER block, we can not be sure about its end of life status, because a block redo may implictly revive it. For these situations we associate the endLifespan with the block exit.

In many cases, we have to determine if the lifespan interferes with a optimization decision being prepared. The lifespan is calculated once at the beginning of the optimizer sequence. It should either be maintained to reflect the most accurate situation while optimizing the code base. In particular, it means that any move/remove/addition of a MAL instruction calls for either a recalculation or further propagation. Unclear what will be the best strategy. For the time being we just recalc.

See is all arguments mentioned in the instruction at point pc are still visible at instruction qc and have not been updated in the mean time. Take into account that variables may be declared inside a block. This can be calculated using the BARRIER/CATCH and EXIT pairs.

The safety property should be relatively easy to determine for each MAL function. This calls for accessing the function MAL block and to inspect the arguments of the signature.

Any instruction may block identification of a common subexpression. It suffices to stumble upon an unsafe function whose parameter lists has a non-empty intersection with the targeted instruction. To illustrate, consider the sequence

```
L1 := f(A,B,C);
...
G1 := g(D,E,F);
...
l2:= f(A,B,C);
...
L2:= h()
```

The instruction G1:=g(D,E,F) is blocking if G1 is an alias for `{A,B,C}`. Alternatively, function g() may be unsafe and `{D,E,F}` has a non-empty intersection with `{A,B,C}`. An alias can only be used later on for readonly (and not be used for a function with side effects).

5.1.6 Flow analysis

In many optimization rules, the data flow dependency between statements is of crucial importance. The MAL language encodes a multi-source, multi-sink dataflow network. Optimizers typically extract part of the workflow and use the language properties to enumerate semantic equivalent solutions, which under a given cost model turns out to result in better performance.

The flow graph plays a crucial role in many optimization steps. It is unclear as yet what primitives and what storage structure is most adequate. For the time being we introduce the operations needed and evaluate them directly against the program

For each variable we should determine its scope of stability. End-points in the flow graph are illustrative as dead-code, that do not produce persistent data. It can be removed when you know there are no side-effect.

Side-effect free evaluation is a property that should be known upfront. For the time being, we assume it for all operations known to the system. The property `"unsafe"` is reserved to identify cases where this does not hold. Typically, a bun-insert operation is unsafe, as it changes one of the parameters.

5.2 Optimizer Toolkit

In this section, we introduce the collection of MAL optimizers included in the code base. The tool kit is incrementally built, triggered by experimentation and curiousity. Several optimizers require further development to cope with the many features making up the MonetDB system. Such limitations on the implementation are indicated where appropriate.

Experience shows that construction and debugging of a front-end specific optimizer is simplified when you retain information on the origin of the MAL code produced as long as possible. For example, the snippet `sql.insert(col, 12@0, "hello")` can be the target of simple SQL rewrites using the module name as the discriminator.

Pipeline validation. The pipelines used to optimize MAL programs contain dependencies. For example, it does not make much sense to call the deadcode optimizer too early in the pipeline, although it is not an error. Moreover, some optimizers are merely examples of the direction to take, others are critical for proper functioning for e.g. SQL.

5.2.1 Access mode optimization

The routine OPTIMIZER.ACCESSMODE() reduces the number of read/write mode changes of variables to a minimum. Especially setting a BAT to write mode is expensive, because it often implies creation of a private copy first.

A full implementation is delayed until really needed.

5.2.2 Accumulator Evaluations

Bulk arithmetic calculations are pretty expensive, because new BATs are created for each expression. This memory hunger can be reduced by detecting opportunities for accummulator processing, i.e. where a (temporary) variable is overwritten. For example, consider the program snippet

```
    t3:= batcalc.*(64,t2);
    t4:= batcalc,+(t1,t3);
    optimizer.accumulators();
```

If variable t2 is a temporary variable and not used any further in the program block, we can re-use its storage space and propagate its alias through the remainder of the code.

```
    batcalc.*(t2,64,t2);
    t4:= batcalc.+(t2,t1,t2);
```

The implementation is straight forward. It only deals with the arithmetic operations available in BATCALC right now. This set will be gradually be extended. The key decision is to determine whether we may overwrite any of the arguments. This is hard to detect at compile time, e.g. the argument may be the result of a binding operation or represent a view over a persistent BAT. Therefore, the compiler injects the call ALGEBRA.REUSE(), which avoids overwriting persistent BATs by taking a copy.

5.2.3 Alias Removal

The routine OPTIMIZER.ALIASREMOVAL() walks through the program looking for simple assignment statements, e.g. V:=W. It replaces all subsequent occurrences of V by W, provided V is assigned a value once and W does not change in the remainder of the code. Special care should be taken for iterator blocks as illustrated in the case below:

```
        i:=0;
        b:= "done";
barrier go:= true;
        c:=i+1;
        d:="step";
        v:=d;
        io.print(v);
        i:=c;
redo go:= i<2;
exit go;
        io.print(b);
        optimizer.aliasRemoval();
```

The constant strings are propagated to the PRINT() routine, while the initial assigment I:=0 should be retained. The code block becomes:

```
        i:=0;
barrier go:= true;
        c:=i+1;
        io.print("step");
        i:=c;
redo go:= i<2;
exit go;
        io.print("done");
```

A special case is backward propagation of constants. The following snippet is the result of the JITO emptyset. It can be further reduced to avoid useless assignments.

```
    _53 := sql.bind("sys","_tables","type",0);
    (_54,_56,_58,_60) := bat.partition(_53);
    _53 := nil;
    _67 := _54;
    _54 := nil;
    _75 := _67;
    _67 := nil;
    _83 := _75;
    _75 := nil;
```

5.2.4 Code Factorization

In most real-life situations queries are repeatedly called with only slight changes in their parameters. This situation can be captured by the query compilers by keeping a cache of recent query plans. In MonetDB context such queries are represented as parameterized MAL programs.

To further optimize the cached functions it might help to split the query plan into two sections. One section with those actions that do not depend on the arguments given and another section that contains the heart of the query using all information. Such a program can be represented by a MAL factory, which is a re-entrend query plan.

An example of how factorize changes the code is shown below:

```
function test(s:str):lng;
    b:= bat.new(:int,:str);
    bat.insert(b,1,"hello");
    z:= algebra.select(b,s,s);
    i:= aggr.count(z);
    return i;
end test;
optimizer.factorize("user","test");
```

which translates into the following block:

```
factory user.test(s:str):lng;
    b := bat.new(:int,:str);
    bat.insert(b,1,"hello");
barrier always := true;
    z := algebra.select(b,s,s);
    i := aggr.count(z);
    yield i;
    redo always;
exit always;
end test;
```

The factorizer included is a prototype implementation of MAL factorization. The approach taken is to split the program into two pieces and wrap it as a MAL factory. The optimization assumes that the database is not changed on tables accessed only once during the factory lifetime. Such changes should be detected from the outside and followed by re-starting the factory.

A refined scheme where the user can identify the 'frozen' parameters is left for the future. As the mapping of a query to any of the possible available factories to deal with the request. For the time being we simple reorganize the plan for all parameters

The factorize operation interferes with OPTIMIZER.EXPRESSIONACCUMULATION() because that may overwrite the arguments. For the time being, this is captured in a local routine.

5.2.5 Coercion Removal

A simple optimizer that removes coercions that are not needed. They may result from a sloppy code-generator or function call resolution decision. For example:

```
        v:= calc.int(23);
```

becomes a single assignment without function call.

The primary role is a small illustration of coding an optimizer algorithm.

5.2.6 Common Subexpression Elimination

Common subexpression elimination merely involves a scan through the program block to detect re-curring statements. The key problem to be addressed is to make sure that the parameters involved in the repeatative instruction are invariant.

The analysis of OPTIMIZER.COMMONTERMS() is rather crude. All functions with possible side-effects on their arguments should have been marked as 'unsafe'. Their use within a MAL block breaks the dataflow graph for all objects involved (BATs, everything kept in boxes).

The common subexpression optimizer locates backwards the identical instructions. It stops as soon as it has found an identical one. Before we can replace the expression with the variable(s) of the previous one, we should assure that we haven;t passed a non-empty barrier block.

```
    b:= bat.new(:int,:int);
    c:= bat.new(:int,:int);
    d:= algebra.select(b,0,100);
    e:= algebra.select(b,0,100);
    k1:= 24;
    k2:= 27;
    l:= k1+k2;
    l2:= k1+k2;
    l3:= l2+k1;
    optimizer.commonTerms();
```

is translated into the code block where the first two instructions are not common, because they have side effects.

```
    b := bat.new(:int,:int);
    c := bat.new(:int,:int);
    d := algebra.select(b,0,100);
    e := d;
    l := calc.+(24,27);
    l3 := calc.+(l,24);
```

5.2.7 Constant Expression Evaluation

Expressions produced by compilers involving only constant arguments can be evaluated once. It is particular relevant in functions that are repeatably called. One time queries would not benefit from this extra step.

Consider the following snippet, which contains recursive use of constant arguments

```
a:= 1+1;        io.print(a);
b:= 2;           io.print(b);
c:= 3*b;        io.print(c);
d:= calc.flt(c);io.print(d);
e:= mmath.sin(d);io.print(e);
optimizer.aliasRemoval();
optimizer.evaluate();
```

The code produced by the optimizer would be

```
io.print(2);
io.print(2);
io.print(6);
io.print(6);
io.print(-0.279415488);
```

Likewise we attempt to catch barrier blocks based on constants.

5.2.8 Costmodel Approach

Cost models form the basis for many optimization decisions. The cost parameters are typically the size of the (intermediate) results and response time. Alternatively, they are running aggregates, e.g. max memory and total execution time, obtained from a simulated run. The current implementation contains a framework and an example for building your own cost-based optimized.

The optimizer.costModel() works its way through a MAL program in search for relational operators and estimates their result size. The estimated size is left behind as the property rows.

```
r{rows=100} := bat.new(:oid,:int);
s{rows=1000}:= bat.new(:oid,:int);
rs:= algebra.select(s,1,1);
rr:= bat.reverse(r);
j:= algebra.join(rs,rr);
optimizer.costModel();
```

changes the properties of the instructions as follows:

```
r{rows=100} := bat.new(:oid,:int);
s{rows=1000} := bat.new(:oid,:int);
rs{rows=501} := algebra.select(s,1,1);
rr{rows=100} := bat.reverse(r);
j{rows=100} := algebra.join(rs,rr);
```

The cost estimation does not use any statistics on the actual data distribution yet. It relies on the rows property provided by the front-end or other optimizers. It just applies a few heuristic cost estimators. However, it ensures that empty results are only tagged with ROWS=0 if the estimate is accurate, otherwise it assumes at least one result row. This property makes it possible to safely pass the result of the cost estimation to the EMPTYSET optimizer for code reduction.

5.2.9 The dataflow optimizer

MAL programs are largely logical descriptions of an execution plan. At least as it concerns side-effect free operations. For these sub-plans the order of execution needs not to be a priori fixed and a dataflow driven evaluation is possible. Even using multiple cores to work their way through the dataflow graph.

The dataflow optimizer analyses the code and wraps all instructions eligible for dataflow driven execution with a guarded block. Ofcourse, this is only necessary if you can upfront determine there are multiple threads of execution possible.

Upon execution, the interpreter instantiates multiple threads based on an the number of processor cores available. Subsequently, the eligible instructions are queued and consumed by the interpreter threads.

Dataflow blocks may not be nested. Therefore, any dataflow block produced for inlined code is removed first.

Initial experiments on e.g. the RDF benchmark showed a speed up of 20\% on average, with a peak of a factor 2 for individual queries. The main reason for this limited gain stems from the little opportunities for parallel execution in the SQL code plans

5.2.10 Dead Code Removal

Dead code fragments are recognized by assignments to variables whose value is not consumed any more. It can be detected by marking all variables used as arguments as being relevant. In parallel, we built a list of instructions that should appear in the final result. The new code block is than built in one scan, discarding the superflous instructions.

Instructions that produce side effects to the environment, e.g., printing and BAT updates, should be taken into account. Such (possibly recursive) functions should be marked with a property (UNSAFE). For now we recognize a few important ones Likewise, instructions marked as control flow instructions should be retained.

An illustrative example is the following MAL snippet:

```
V7 := bat.new(:oid,:int);
V10 := bat.new(:int,:oid);
V16 := algebra.markH(V7);
V17 := algebra.join(V16,V7);
V19 := bat.new(:oid,:int);
V22 := bat.new(:oid,:int);
V23 := algebra.join(V16,V22);
io.print("done");
optimizer.deadCodeRemoval();
```

The dead code removal trims this program to the following short block:

```
io.print("done");
```

A refinement of the dead code comes from using arguments that ceased to exist due to actions taken by an optimizer. For example, in the snippet below the PUSHRANGES optimizer may conclude that variable V31 becomes empty and simply injects a 'dead' variable by dropping the assignment statement. This makes other code dead as well.

```
V30 := algebra.select( V7, 10,100);
V31 := algebra.select(V30,-1,5);
V32 := aggr.sum(V31);
io.print(V32);
```

[implementation pending]

5.2.11 Emptyset Reduction

One of the key decisions during MAL optimization is to estimate the size of the BATs produced and consumed. Two cases are of interest for symbolic processing. Namely, when a BAT is known to contain no tuples and those that have precisely one element. Such information may come from application domain knowledge or as a side effect from symbolic evaluation. It is associated with the program under inspection as properties.

The empty set property is used by the reduction algorithm presented here. Any empty set is propagated through the program to arrive at a smaller and therefore faster evaluation.

For example, consider the following MAL test:

```
    V1 := bat.new(:oid,:int);
    V7 := bat.new(:oid,:int);
    V10{rows=0} := bat.new(:int,:oid);
    V11 := bat.reverse(V10);
    V12 := algebra.kdifference(V7,V11);
    V16 := algebra.markH(V12);
    V17 := algebra.join(V16,V7);
    bat.append(V1,V17);
optimizer.costModel();
optimizer.emptySet();
```

Calling the optimizers replaces this program by the following code snippet.

```
    V1 := bat.new(:oid,:int);
    V7 := bat.new(:oid,:int);
    V10{rows=0} := bat.new(:int,:oid);
    V11{rows=0} := bat.new(:oid,:int);
    V12 := V7;
    V16 := algebra.markH(V12);
    V17 := algebra.join(V16,V7);
    bat.append(V1,V17);
```

This block can be further optimized using alias propagation and dead code removal. The final block becomes:

```
    V1 := bat.new(:oid,:int);
    V7 := bat.new(:oid,:int);
    V16 := algebra.markH(V7);
    V17 := algebra.join(V16,V7);
    bat.append(V1,V17);
```

During empty set propagation, new candidates may appear. For example, taking the intersection with an empty set creates a target variable that is empty too. It becomes an immediate target for optimization. The current implementation is conservative. A limited set of instructions is considered. Any addition to the MonetDB instruction set would call for assessment on their effect.

5.2.12 SQL specifics

The `bind` operations of SQL requires special care, because they refer to containers that might initially be empty, but aren't upon a second call. This calls for a defensive approach, where a constraint check is left behind to detect a plan whose conditions are not met anymore. Of course, we can drop the constraint if we know that a plan is used onlye once (and not recursively). This can be marked by the SQL compiler, who is in control over the query cache.

5.2.13 Garbage Collection

Garbage collection of temporary variables, such as strings and BATs, takes place upon returning from a function call. Especially for BATs this may keep sizable resources locked longer than strictly necessary. Although the programmer can influence their lifespan by assignment of the NIL, thereby triggering the garbage collector, it is more appropriate to rely on an optimizer to inject these statements. For, it keeps the program smaller and a better target for code-optimizations.

The operation OPTIMIZER.GARBAGECOLLECTOR() removes all BAT references that are at their end of life to make room for new ones. It is typically called as one of the last optimizer steps. A snippet of a the effect of the garbage collector:

```
t1 := bat.new(:oid,:int);
t2 := array.grid(132000,8,1,0);
t3 := array.grid(1,100,10560,0);
t4 := array.grid(1,100,10560,0,8);
t5 := batcalc.+(t2,t4);
t6 := batcalc.oid(t5);
t7 := algebra.join(t6,t1);
optimizer.garbageCollector();
```

is translated into the following code block:

```
t1 := bat.new(:oid,:int);
t2 := array.grid(132000,8,1,0);
t3 := array.grid(1,100,10560,0);
t4 := array.grid(1,100,10560,0,8);
t5 := batcalc.+(t2,t4);
bat.setGarbage(t2);
bat.setGarbage(t4);
t6 := batcalc.oid(t5);
bat.setGarbage(t5);
t7 := algebra.join(t6,t1);
bat.setGarbage(t6);
bat.setGarbage(t1);
```

The current algorithm is straight forward. After each instruction, we check whether its BAT arguments are needed in the future. If not, we inject a garbage collection statement to release them, provided there are no other reasons to retain it. This should be done carefully, because the instruction may be part of a loop. If the variable is defined inside the loop, we can safely remove it.

5.2.14 Heuristic rewrites rules

One of the oldest optimizer tricks in relational query processing is to apply heuristic rules to reduce the processing cost. For example, a selection predicate is pushed through another operator to reduce the number of tuples to consider. Heuristic rewrites are relatively easy to apply in a context where the expression is still close to a relational algebra tree. Therefore, many of the standard rewrite rules are already applied by the SQL front-end as part of its strategic optimization decisions.

Finding rewrite opportunities within a linear MAL program may be more difficult. For example, the pattern should respect the flow of control that may already be introduced. The last resort for the optimizer builder is to write a C-function to look for a pattern of interest and transform it. The code base contains an example how to built such user specific optimizer routines. It translates the pattern:

```
y:= reverse(R);
z:= select(y,l,h);
```

into the statement:

```
z:= selectHead(x,R,l,h)
```

5.2.15 Join Paths

The routine OPTIMIZER.JOINPATH() walks through the program looking for join operations and cascades them into multiple join paths. To illustrate, consider

```
a:= bat.new(:oid,:oid);
b:= bat.new(:oid,:oid);
c:= bat.new(:oid,:str);
j1:= algebra.join(a,b);
j2:= algebra.join(j1,c);
j3:= algebra.join(b,b);
j4:= algebra.join(b,j3);
```

The optimizer will first replace all arguments by their join sequence. The underlying instructions are left behind for the deadcode optimizer to be removed.

```
a:= bat.new(:oid,:oid);
j1:= algebra.join(a,b);
j2:= algebra.joinPath(a,b,c);
j3:= algebra.join(b,b);
j4:= algebra.joinPath(b,b,b);
```

In principle, the joinpaths may contain common subpaths, whose materialization would improve performance. The SQL front-end produces often produces snippets of the following structure

```
t1:= algebra.join(b,c);
z1:= algebra.join(a,t1);
```

```
        ...
    t2:= algebra.join(b,d);
    z2:= algebra.join(a,t2);
```

The joinpath would merge them into

```
    z1:= algebra.joinPath(a,b,c);
        ...
    z2:= algebra.joinPath(a,b,d);
```

which are handle by a heuristic looking at the first two argments and re-uses a materialized join.

```
    _13:= algebra.join(a,b);
    z1:= algebra.join(_13,c);
        ...
    z2:= algebra.join(_13,d);
```

An alternative is to make recognition of the common re-useable paths an integral part of the joinPath body.

```
    x3:= algebra.join(a,b);
        r3:= bat.reverse(x3);
        j1:= join(c,r3);
        rb:= bat.reverse(b);
        ra:= bat.reverse(a);
        j1:= algebra.joinpath(c,rb,ra);
```

As a final step in the speed up of the joinpath we consider clustering large operands if that is expected to improve IO behavior.

5.2.16 Macro and Orcam Processing

The optimizers form the basis for replacing code fragments. Two optimizers are focused on code expansion and contraction. The former involves replacing individual instructions by a block of MAL code, i.e. a MACRO call. The latter depicts the inverse operation, a group of instructions is replaced by a single MAL assignment statement, i.e. a ORCAM call.

The macro facility is limited to type-correct MAL functions, which means that replacement is not essential from a semantic point of view. They could have been called, or the block need not be compressed. It is equivalent to inline code expansion.

The MACRO and ORCAM transformations provide a basis to develop front-end specific code generation templates. The prototypical test case is the following template:

```
function user.joinPath( a:bat[:any_1,:any_2],
                b:bat[:any_2,:any_3],
                c:bat[:any_3,:any_4]):bat[:any_1,:any_4]
address fastjoinpath;
    z:= join(a,b);
    zz:= join(z,c);
    return zz;
end user.joinPath;
```

The call OPTIMIZER.MACRO("USER", "JOINPATH") hunts for occurrences of the instruction call in the block in which it is called and replaces it with the body, i.e. it in-lines the code. Conversely, the optimizer.orcam("user", "joinPath") attempts to localize a block of two join operations and, when found, it is replaced by the direct call to joinPath. In turn, type resolution then directs execution to a built-in function fastjoinpath.

The current implementation is limited to finding a consecutive sequence, ending in a return-statement. The latter is needed to properly embed the result in the enclosed environment. It may be extended in the future to consider the flow of control as well.

5.2.17 Known issues

The functions subject to expansion or contraction should be checked on 'proper' behavior.

The current implementation is extremely limited. The macro optimizer does not recognize use of intermediate results outside the block being contracted. This should be checked and it should block the replacement, unless the intermediates are part of the return list. Likewise, we assume here that the block has a single return statement, which is also the last one to be executed.

The macro optimizer can only expand functions. Factories already carry a significant complex flow of control that is hard to simulate in the nested flow structure of an arbitrary function.

The orcam optimizer can not deal with calls controlled by a barrier. It would often require a rewrite of several other statements as well.

```
pattern optimizer.macro(targetmod:str,targetfcn:str):void
address OPTmacro
comment "Inline the code of the target function.";
pattern optimizer.macro(mod:str,fcn:str,targetmod:str,targetfcn:str):void
address OPTmacro
comment "Inline a target function used in a specific function.";

pattern optimizer.orcam(targetmod:str,targetfcn:str):void
address OPTorcam
comment "Inverse macro processor for current function";
pattern optimizer.orcam(mod:str,fcn:str,targetmod:str,targetfcn:str):void
address OPTorcam
comment "Inverse macro, find pattern and replace with a function call.";
```

5.3 Memo-based Query Execution

Modern cost-based query optimizers use a memo structure to organize the search space for an efficient query execution plan. For example, consider an oid join path 'A.B.C.D'. We can start the evaluation at any point in this path.

Its memo structure can be represented by a (large) MAL program. The memo levels are encapsulated with a choice operator. The arguments of the second dictate which instructions to consider for cost evaluation.

```
...
scheduler.choice("getVolume");
T1:= algebra.join(A,B);
T2:= algebra.join(B,C);
```

```
T3:= algebra.join(C,D);
scheduler.choice("getVolume",T1,T2,T3);
T4:= algebra.join(T1,C);
T5:= algebra.join(A,T2);
T6:= algebra.join(T2,D);
T7:= algebra.join(B,T3);
T8:= algebra.join(C,D);
scheduler.choice("getVolume",T4,T5,T6,T7,T8);
T9:= algebra.join(T4,D);
T10:= algebra.join(T5,D);
T11:= algebra.join(A,T6);
T12:= algebra.join(A,T7);
T13:= algebra.join(T1,T8);
scheduler.choice("getVolume",T9,T10,T11,T12,T13);
answer:= scheduler.pick(T9, T10, T11, T12, T13);
```

The `scheduler.choice()` operator calls a builtin `getVolume` for each target variable and expects an integer-valued cost. In this case it returns the total number of bytes uses as arguments.

The target variable with the lowest cost is chosen for execution and remaining variables are turned into a temporary NOOP operation.(You may want to re-use the memo) They are skipped by the interpreter, but also in subsequent calls to the scheduler. It reduces the alternatives as we proceed in the plan.

A built-in naive cost function is used. It would be nice if the user could provide a private cost function defined as a `pattern` with a polymorphic argument for the target and a `:lng` result. Its implementation can use the complete context information to make a decision. For example, it can trace the potential use of the target variable in subsequent statements to determine a total cost when this step is taken towards the final result.

A complete plan likely includes other expressions to prepare or use the target variables before reaching the next choice point. It is the task of the choice operator to avoid any superfluous operation.

The MAL block should be privately owned by the caller, which can be assured with `scheduler.isolation()`.

A refinement of the scheme is to make cost analysis part of the plan as well. Then you don't have to include a hardwired cost function.

```
Acost:= aggr.count(A);
Bcost:= aggr.count(B);
Ccost:= aggr.count(C);
T1cost:= Acost+Bcost;
T2cost:= Bcost+Ccost;
T3cost:= Ccost+Dcost;
scheduler.choice(T1cost,T1, T2cost,T2, T3cost,T3);
T1:= algebra.join(A,B);
T2:= algebra.join(B,C);
T3:= algebra.join(C,D);
...
```

5.3.1 Merge Tables

A merge association table (MAT) descriptor defines an ordered collection of type compatible BATs, whose union represents a single (virtual) BAT. The MAT may represent a partitioned BAT (see BPM), but could also be an arbitrary collection of temporary BATs within a program fragment.

The MAT definition lives within the scope of a single block. The MAT optimizer simply expands the plan to deal with its components on an instruction basis. Only when a blocking operator is encounted, the underlying BAT is materialized.

The MAT object cannot be passed as an argument to any function without first being materialized. Simply because the MAT is not known by the type system and none of the lower level operations are aware of its existence.

In the first approach of the MAT optimizer we assume that the first BAT in the MAT sequence is used as an accumulator. Furthermore, no semantic knowledge is used to reduce the possible superflous (semi)joins. Instead, we limit expansion to a single argument. This is changed at a later stage when a cost-based evaluation can be used to decide different.

To illustrate, consider:

```
m0:= bat.new(:oid,:int);
m1:= bat.new(:oid,:int);
m2:= bat.new(:oid,:int);
b := mat.new(m0,m1,m2);
s := algebra.select(b,1,3);
i := aggr.count(s);
io.print(s);
io.print(i);
c0 := bat.new(:int,:int);
c1 := bat.new(:int,:int);
c := mat.new(c0,c1);
j := algebra.join(b,c);
io.print(j);
```

The selection and aggregate operations can simply be rewritten using a MAT:

```
_33 := algebra.select(m0,1,3);
_34 := algebra.select(m1,1,3);
_35 := algebra.select(m2,1,3);
    s := mat.new(_33,_34,_35);
    i := 0:int;
    _36 := aggr.count(_33);
    i := calc.+(i,_36);
    _37 := aggr.count(_34);
    i := calc.+(i,_37);
    _38 := aggr.count(_35);
    i := calc.+(i,_38);
    io.print(i);
```

The print operation does not have MAT semantics yet. It requires a function that does not produce the header with each call. Instead, we can also pack the elements before printing.

```
          s := mat.pack(_33,_34,_35);
          io.print(s);
```

For the join we have to generate all possible combinations, not knowing anything about the properties of the components. The current heuristic is to limit expansion to a single argument. This leads to

```
          b := mat.pack(m0,m1,m2);
          _39 := algebra.join(b,c0);
          _40 := algebra.join(b,c1);
          j := mat.new(_39,_40);
```

The drawback of the scheme is the potential explosion in MAL statements. A challenge of the optimizer is to find the minimum by inspection of the properties of the MAT elements. For example, it might attempt to partially pack elements before proceding. This would be a runtime scheduling decision.

Alternatively, the system could use MAT iterators to avoid packing at the cost of more complex program analysis afterwards.

```
      ji:= bat.new(:oid,:int);
      barrier b:= mat.newIterator(m0,m1,m2);
      barrier c:= mat.newIterator(c0,c1);
      ji := algebra.join(b,c);
      bat.insert(j,ji);
      redo c:= mat.newIterator(c0,c1);
      redo b:= mat.newIterator(m0,m1,m2);
      exit c;
      exit b;
```

5.3.2 Multiplex Compilation

The MonetDB operator multiplex concept has been pivotal to easily apply any scalar function to elements in a containers. Any operator CMD came with its multiplex variant [CMD]. Given the signature CMD(T1,..,TN) : TR, it could be applied also as [CMD](BAT[:ANY_1,:T1],...,BAT[ANY_1,TN]) :BAT[ANY_1,TR].

The semantics of the multiplex is to perform the positional join on all bat-valued parameters, and to execute the CMD for each combination of matching tuples. All results are collected in a result BAT. All but one argument may be replaced by a scalar value.

The generic solution to the multiplex operators is to translate them to a MAL loop. A snippet of its behaviour:

```
     b:= bat.new(:int,:int);
     bat.insert(b,1,1);
     c:bat[:int,:int]:= mal.multiplex("calc.+",b,1);
         optimizer.multiplex();
```

The current implementation requires the target type to be mentioned explicitly. The result of the optimizer is:

```
     b := bat.new(:int,:int);
     bat.insert(b,1,1);
     _8 := bat.new(:int,:int);
```

```
barrier (_11,_12,_13):= bat.newIterator(b);
    _15 := calc.+(_13,1);
    bat.insert(_8,_12,_15);
    redo (_11,_12,_13):= bat.hasMoreElements(b);
exit (_11,_12,_13);
    c := _8;
```

5.3.3 BAT Partitions

Limitations on the addressing space in older PCs and the need for distributed storage makes that BATs ideally should be looked upon as a union of smaller BATs which are processed within the (memory) resource limitations given.

The PARTITION() optimizer with the supportive bat partition library `bpm` addresses the issue with an adaptive database segmentation algorithm. It is designed incrementally with a focus on supporting the SQL front-end. In particularly, the operators considered is a limited subset of MAL. Occurrence of an operator outside this set terminates the optimizer activities.

The operation OPTIMIZER.PARTITIONS() hunts for bindings of SQL column BATs and prepare code for using partitioned versions instead.

We use two implementations. The first one attempts to find segments of linear dependent data and builds an iterator around it. This approach is tricky, because you have to take care of special cases. In particular, the semantics of the operators on the sequence construction posed quite some problems.

The naive() approach simply looks at individual operations and surround them with an iterator. An alias table is kept around for re-use and detect already partitioned operands. The drawback is that potentially a partitioned BAT is read multiple times [it depends on the re-use of variables, which can be calculated] and write+read of intermediates. Experiments should demonstrate the optimal one.

5.3.4 Peephole optimization

Recursive descend query compilers easily miss opportunities for better code generation, because limited context is retained or lookahead available. The peephole optimizer is built around such recurring patterns and compensates for the compilers 'mistakes'. The collection of peephole patterns should grow over time and front-end specific variations are foreseen.

The SQL frontend heavily relies on a pivot table, which is a generated oid sequence. Unfortunately, this is not seen and the pattern '$i := calc.oid(0@0); $j:= algebra.markT($k,$i);' occurs often. This can be replaced with '$j:= algebra.markT($k)';

Another example of a 2-way instruction sequence produced is then '$j:= algebra.markT($k); $l:= bat.reverse($j);', which can be replaced by '$l:= algebra.markH($k);'.

The reverse-reverse operation also falls into this category. Reversal pairs may result from the processing scheme of a front-end compiler or from a side-effect from other optimization steps. Such reversal pairs should be removed as quickly as possible, so as to reduce the complexity of finding alternative optimization opportunities. As in all cases we should ensure that the intermediates dropped are not used for other purposes as well.

```
r:bat[:int,:int]:= bat.new(:int,:int);
o:= calc.oid(0@0);
```

```
z:= algebra.markT(r,o);
rr:= bat.reverse(z);
s := bat.reverse(r);
t := bat.reverse(s);
io.print(t);
optimizer.peephole();
```

which is translated by the peephole optimizer into:

```
r:bat[:int,:int] := bat.new(:int,:int);
rr := algebra.markH(r);
io.print(r);
```

Another example is the combination of a BAT partition operation followed by a reconstruction without using the partitions individually.

5.3.5 Query Execution Plans

A commonly used data structure to represent and manipulate a query is a tree (or graph). Its nodes represent the operators and the leaves the operands. Such a view comes in handy when you have to re-organize whole sections of code or to built-up an optimized plan bottom up, e.g. using a memo structure.

The MAL optimizer toolkit provides functions to overlay the body of any MAL block with a tree (graph) structure and to linearize them back into a MAL block. The linearization order is determined by a recursive descend tree walk from the anchor points in the source program.

To illustrate, consider the code block:

```
#T1:= bat.new(:int,:int);
#T2:= bat.new(:int,:int);
#T3:= bat.new(:int,:int);
#T4:= bat.new(:int,:int);
a:= algebra.select(T1,1,3);
b:= algebra.select(T2,1,3);
c:= algebra.select(T3,0,5);
d:= algebra.select(T4,0,5);
e:= algebra.join(a,c);
f:= algebra.join(b,d);
h:= algebra.join(e,f);
optimizer.dumpQEP();
```

which produces an indented structure of the query plan.

```
h := algebra.join(e,f);
    e := algebra.join(a,c);
        a := algebra.select(T1,1,3);
            T1 := bat.new(:int,:int);
        c := algebra.select(T3,0,5);
            T3 := bat.new(:int,:int);
    f := algebra.join(b,d);
        b := algebra.select(T2,1,3);
```

```
                T2 := bat.new(:int,:int);
            d := algebra.select(T4,0,5);
                T4 := bat.new(:int,:int);
```

Any valid MAL routine can be overlayed with a tree (graph) view based on the flow dependencies, but not all MAL programs can be derived from a simple tree. For example, the code snippet above when interpreted as a linear sequence can not be represented unless the execution order itself becomes an operator node itself.

However, since we haven't added or changed the original MAL program, the routine `qep.propagate` produces the orginial program, where the linear order has priority. If, however, we had entered new instructions into the tree, they would have been placed in close proximity of the other tree nodes.

Special care is given to the flow-of-control blocks, because to produce a query plan section that can not easily be moved around. [give dot examples]

5.3.6 Range Propagation

Almost all queries are interested in a few slices of the table. If applied to a view, the query plans often contain multiple selections over the same column. They may also have fixed range arguments comming from fragmentation criteria.

The purpose of the PUSHRANGES optimizer is to minimize the number of table scans by cascading the range terms as much as possible. Useless instructions are removed from the plan.

```
b := bat.new(:oid,:int);
s1:= algebra.select(b,1,100);
s2:= algebra.select(s1,5,95);
s3:= algebra.select(s2,50,nil);
s4:= algebra.select(s3,nil,75);
optimizer.pushranges();
```

This lengthly sequence can be compressed into a single one:

```
b := bat.new(:oid,:int);
s1:= algebra.select(b,50,75);
```

A union over two range selections from a single source could also be a target.

```
t1:= algebra.select(b,1,10);
t2:= algebra.select(b,0,5);
t3:= algebra.union(t1,t2);
```

would become

```
t3:= algebra.select(0,10);
```

5.3.7 The recycler

Query optimization and processing in off-the-shelf database systems is often still focused on individual queries. The queries are analyzed in isolation and ran against a kernel regardless opportunities offered by concurrent or previous invocations.

This approach is far from optimal and two directions to improve are explored: materialized views and (partial) result-set reuse. Materialized views are derived from query logs. They represent common sub-queries, whose materialization improves subsequent processing times. Re-use of (partial) results is used in those cases where a zooming-in or navigational application is at stake.

The Recycler optimizer and module extends this with a middle out approach. They exploit the materialize-all-intermediate approach of MonetDB by deciding to keep a hold on them as long as deemed beneficial.

The approach taken is to mark the instructions in a MAL program using the recycler optimizer call, such that their result is retained in a global recycle cache hardwired in the MAL interpreter. Instructions become subject to the Recycler if at least one of its arguments is a BAT and all others are either constants or variables already known in the Recycler.

Upon execution, the recycler is called from the inner loop of the MAL interpreter to first check for an up-to-date result to be picked up at no cost. Otherwise, it evaluates the instruction and calls upon policy functions to decide if it is worthwhile to keep.

The Recycler comes with a few policy controlling operators to experiment with its effect in concrete settings. The retain policy controls when to keep results around, the reuse policy looks after exact duplicate instructions or uses semantical knowledge on MAL instructions to detect potential reuse gain (e.g. reuse select results). And finally, the cache policy looks after the storage space for the intermediate result pool. The details are described in the recycle module.

```
pattern optimizer.recycle():str
address OPTrecycle;
pattern optimizer.recycle(mod:str, fcn:str):str
address OPTrecycle
comment "Replicator code injection";
```

The number of overloaded instructions is kept to a minimum.

```
#ifndef _OPT_RECYCLER_
#define _OPT_RECYCLER_
#include "opt_prelude.h"
#include "opt_support.h"
#include "mal_recycle.h"

/* #define DEBUG_OPT_RECYCLER */
```

The variables are all checked for being eligible as a variable subject to recycling control. A variable may only be assigned a value once. The function is a sql.bind(-,-,-,0) or all arguments are already recycle enabled or constant.

The arguments of the function cannot be recycled. They change with each call. This does not mean that the instructions using them can not be a target of recycling.

Just looking at a kept result target is not good enough. You have to sure that the arguments are also the same. This rules out function arguments.

The recycler is targeted towards a query only database. The best effect is obtained for a single-user mode (sql_debug=32) when the delta-bats are not processed which allows longer instruction chains to be recycled. Update statements are not recycled. They trigger cleaning of the recycle cache at the end of the query. Only intermediates derived from the updated columns are invalidated. Separate update instructions in queries, such as bat.append implementing 'OR', are monitored and also trigger cleaning the cache.

```
#include "mal_config.h"
#include "opt_recycler.h"
#include "mal_instruction.h"

static int
OPTrecycleImplementation(Client cntxt, MalBlkPtr mb, MalStkPtr stk, InstrPtr p)
{
 int i, j, cnt, actions = 0;
 Lifespan span;
 InstrPtr *old, q;
 int limit, updstmt = 0;
 char *recycled;
 short app_sc = -1,app_tbl = -1;

 (void) cntxt;
 (void) stk;
 /* watch out, instructions may introduce new variables */
 limit= mb->stop;
 old = mb->stmt;

 for (i = 1; i<limit; i++) {
  p = old[i];
  if (getModuleId(p)==sqlRef &&
            (getFunctionId(p) == affectedRowsRef ||
              getFunctionId(p) == exportOperationRef ||
              getFunctionId(p) == appendRef ||
              getFunctionId(p) == updateRef ||
              getFunctionId(p) == deleteRef) )
   updstmt = 1;
 }

 span = setLifespan(mb);
 if ( span == NULL)
  return 0;

 recycled= GDKzalloc(sizeof(char)*mb->vtop*2);
 if ( recycled == NULL)
  return 0;
 newMalBlkStmt(mb, mb->ssize);
 pushInstruction(mb,old[0]);

 /* create a handle for recycler */
 q= newFcnCall(mb,"recycle","prelude");
 for (i = 1; i<limit; i++) {
```

```
p = old[i];
if (hasSideEffects(p,TRUE) || isUnsafeFunction(p)){
 if( getModuleId(p)== recycleRef ){ /*don't inline recycle instr. */
  freeInstruction(p);
  continue;
 }
 pushInstruction(mb,p);
  /*  update instructions are not recycled but monitored*/
 if( isUpdateInstruction(p)){
  if (getModuleId(p) == batRef &&
   (getArgType(mb,p,1)==TYPE_bat
   || isaBatType(getArgType(mb, p,1)))){
   recycled[getArg(p,1)]= 0;
   q= newFcnCall(mb,"recycle","reset");
   pushArgument(mb,q, getArg(p,1));
   actions++;
  }
  if (getModuleId(p) == sqlRef){
   if (getFunctionId(p) == appendRef){
    app_sc = getArg(p,1);
    app_tbl = getArg(p,2);
   } else {
    q= newFcnCall(mb,"recycle","reset");
    pushArgument(mb,q, getArg(p,1));
    pushArgument(mb,q, getArg(p,2));
    if (getFunctionId(p) == updateRef)
     pushArgument(mb,q, getArg(p,3));
   }
   actions++;
  }
 }
 continue;
}
if (p->barrier  && p->token != CMDcall){
 /* never save a barrier unless it is a command and side-effect free */
 pushInstruction(mb,p);
 continue;
}

if( p->token== ENDsymbol){
 if ( updstmt && app_sc >= 0 ){
   q= newFcnCall(mb,"recycle","reset");
   pushArgument(mb,q, app_sc);
   pushArgument(mb,q, app_tbl);
 }
 (void) newFcnCall(mb,"recycle","epilogue");
 pushInstruction(mb,p);
```

```
   continue;
  }

  /* don’t change instructions in update statements */
  if( updstmt){
   pushInstruction(mb,p);
   continue;
  }

  /* skip simple assignments */
  if( p->token == ASSIGNsymbol){
   pushInstruction(mb,p);
   continue;
  }

  /* general rule: all arguments are constants or recycled */
  cnt = 0;
  for (j=p->retc; j<p->argc; j++)
   if(recycled[getArg(p,j)] || isVarConstant(mb, getArg(p,j)) )
    cnt++;
  if (cnt == p->argc-p->retc) {
#ifdef DEBUG_OPT_RECYCLER
   stream_printf(cntxt->fdout,"recycle instruction n");
   printInstruction(cntxt->fdout,mb, 0, p,LIST_MAL_ALL);
#endif
   actions ++;
   p->recycle = REC_MAX_INTEREST; /* this instruction is to be monitored */
   for (j= 0; j < p->retc; j++)
    if (getLastUpdate(span, getArg(p,j)) == i)
     recycled[getArg(p,j)] = 1;
  }
```

The expected gain is largest if we can re-use selections on the base tables in SQL. These, however, are marked as uselect() calls, which only produce the oid head. For cheap types we preselect using select() and re-map uselect() back over this temporary. For the time being for all possible selects encountered are marked for re-use.

```
  /* take care of semantic driven recyling */
  /* for selections check the bat argument only
  the range is often template parameter*/
  if(( getFunctionId(p)== selectRef ||
       getFunctionId(p)== antiuselectRef ||
       getFunctionId(p)== likeselectRef ||
       getFunctionId(p)== putName("like",4) ||
              getFunctionId(p)== putName("thetaselect",11) ||
              getFunctionId(p)== putName("thetauselect",12) ) &&
         recycled[getArg(p,1)] ){
   p->recycle = REC_MAX_INTEREST;
```

```
 actions ++;
 if (getLastUpdate(span, getArg(p,0)) == i)
  recycled[getArg(p,0)] = 1;
}
if( getFunctionId(p)== uselectRef && recycled[getArg(p,1)]) {
 if (!ATOMvarsized( getGDKType( getArgType(mb,p,2)))) {
  q = copyInstruction(p);
  getArg(q,0)= newTmpVariable(mb,TYPE_any);
  setFunctionId(q, selectRef);
  q->recycle = REC_MAX_INTEREST;
  recycled[getArg(q,0)] = 1;
  pushInstruction(mb,q);
  getArg(p,1) = getArg(q,0);
  setFunctionId(p,markTRef);
  p->argc = 2;
 }
 p->recycle = REC_MAX_INTEREST;
 actions ++;
 if (getLastUpdate(span, getArg(p,0)) == i)
  recycled[getArg(p,0)] = 1;
}

if(getModuleId(p) == pcreRef) {
 if (( getFunctionId(p)== selectRef && recycled[getArg(p,2)]) ||
     ( getFunctionId(p)== uselectRef && recycled[getArg(p,2)])){
  p->recycle = REC_MAX_INTEREST;
  actions ++;
  if (getLastUpdate(span, getArg(p,0)) == i)
  recycled[getArg(p,0)] = 1;
 }
 else if( getFunctionId(p)== likeuselectRef && recycled[getArg(p,1)]) {
  q = copyInstruction(p);
  getArg(q,0)= newTmpVariable(mb,TYPE_any);
  setFunctionId(q, likeselectRef);
  q->recycle = REC_MAX_INTEREST;
  recycled[getArg(q,0)] = 1;
  pushInstruction(mb,q);
  getArg(p,1) = getArg(q,0);
  setFunctionId(p,markTRef);
  setModuleId(p,algebraRef);
  p->argc = 2;
  p->recycle = REC_MAX_INTEREST;
  actions ++;
  if (getLastUpdate(span, getArg(p,0)) == i)
   recycled[getArg(p,0)] = 1;
 }
}
```

The sql.bind instructions should be handled carefully The delete and update BATs should not be recycled, because they may lead to view dependencies that later interfere with the transaction commits.

```
  if (getModuleId(p)== sqlRef &&
   (((getFunctionId(p)==bindRef || getFunctionId(p) == putName("bind_idxbat",11)) &&
    getVarConstant(mb, getArg(p,4)).val.ival != 0) ||
    getFunctionId(p)== binddbatRef) ) {
    recycled[getArg(p,0)]=0;
    p->recycle = REC_NO_INTEREST; /* this instruction is not monitored */
   }

  if (getModuleId(p) == octopusRef &&
    getFunctionId(p) == bindRef ) {
    recycled[getArg(p,0)] = 1;
    p->recycle = REC_MAX_INTEREST;
  }

  pushInstruction(mb,p);
 }
 GDKfree(span);
 GDKfree(old);
 GDKfree(recycled);
 mb->recycle = actions > 0;
 return actions;
}
```

5.3.8 Optimizer code wrapper

The optimizer wrapper code is the interface to the MAL optimizer calls. It prepares the environment for the optimizers to do their work and removes the call itself to avoid endless recursions.

Before an optimizer is finished, it should leave a clean state behind. Moreover, the information of the optimization step is saved for debugging and analysis.

The wrapper expects the optimizers to return the number of actions taken, i.e. number of succesful changes to the code.

```
exportOptimizer[?](recycle)
#endif

#include "opt_statistics.h"
wrapOptimizer[?](recycle,OPT_CHECK_TYPES)
```

5.3.9 Remote Queries

MAL variables may live at a different site from where they are used. In particular, the SQL front-end uses portions of remote BATs as replication views. Each time such a view is needed, the corresponding BAT is fetched and added to the local cache.

Consider the following snippet produced by a query compiler,

```
mid:= mapi.reconnect("s0_0","localhost",50000,"monetdb","monetdb","mal");
b:bat[:oid,:int] := mapi.bind(mid,"rvar");
c:=algebra.select(b,0,12);
io.print(c);
d:=algebra.select(b,5,10);
low:= 5+1;
e:=algebra.select(d,low,7);
i:=aggr.count(e);
io.printf(" count %d\n",i);
io.print(d);
```

which uses a BAT RVAR stored at the remote site DB1.

There are several options to execute this query. The remote BAT can be fetched as soon as the bind operation is executed, or a portion can be fetched after a remote select, or the output for the user could be retrieved. An optimal solution depends on the actual resources available at both ends and the time to ship the BAT.

The remote query optimizer assumes that the remote site has sufficient resources to handle the instructions. For each remote query it creates a private connection. It is re-used in subsequent calls .

The remote environment is used to execute the statements. The objects are retrieved just before they are locally needed.

```
mid:= mapi.reconnect("s0_0","localhost",50000,"monetdb","monetdb","mal");
mapi.rpc(mid,"b:bat[:oid,:int] :=bbp.bind(\"rvar\");");
mapi.rpc(mid,"c:=algebra.select(b,0,12);");
c:bat[:oid,:int]:= mapi.rpc(mid, "io.print(c);");
io.print(c);
mapi.rpc(mid,"d:=algebra.select(b,5,10);");
low:= 5+1;
mapi.put(mid,"low",low);
mapi.rpc(mid,"e:=algebra.select(d,low,7);");
mapi.rpc(mid,"i:=aggr.count(d);");
i:= mapi.rpc(mid,"io.print(i);");
io.printf(" count %d\n",i);
io.print(d);
```

To reduce the number of interprocess communications this code can be further improved by glueing the instructions together when until the first result is needed.

5.3.10 Singleton Set Reduction

Application semantics and precise cost analysis may identify the result of an operation to produce a BAT with a single element. Such variables can be tagged with the property

SINGLETON, whereafter the operation OPTIMIZER.SINGLETON() derives an MAL program using a symbolic evaluation as far as possible.

During its evaluation, more singleton sets can be created, leading to a ripple effect through the code. A non-optimizable instruction leads to a construction of a new table with the single instance.

```
        b:= bat.new(:int,:int);
        bat.insert(b,1,2);
        c{singleton}:= algebra.select(b,0,4);
        d:= algebra.markH(c);
        io.print(d);
        optimizer.singleton();
```

is translated by into the code block

```
        b := bat.new(:int,:int);
        bat.insert(b,1,2);
        c{singleton} := algebra.select(b,0,4);
        (_15,_16):= bat.unpack(c{singleton});
        d := bat.pack(nil,_16);
        io.print(d);
```

5.3.11 Stack Reduction

The compilers producing MAL may generate an abundance of temporary variables to hold the result of expressions. This leads to a polution of the runtime stack space, because space should be allocated and garbage collection tests should be performed.

The routine OPTIMIZER.REDUCE() reduces the number of scratch variables to a minimum. All scratch variables of the same underlying type share the storage space. The result of this optimizer can be seen using the MonetDB debugger, which marks unused variables explicitly. Experience with the SQL front-end shows, that this optimization step easily reduces the stack consumption by over 20%.

This optimizer needs further testing. Furthermore, the other optimizers should be careful in setting the isused property, or this property can again be easily derived.

5.3.12 Strength Reduction

An effective optimization technique in compiler construction is to move invariant statements out of the loops. The equivalent strategy can be applied to the guarded blocks in MAL programs. Any variable introduced in a block and assigned a value using a side-effect free operation is a candidate to be moved. Furthermore, it may not be used outside the block and the expression may not depend on variables assigned a value within the same block.

```
        j:= "hello world";
barrier go:=true;
        i:= 23;
        j:= "not moved";
        k:= j;
        io.print(i);
        redo go:= false;
exit go;
```

```
        z:= j;
optimizer.strengthReduction();
```

which is translated into the following code:

```
        j := "hello world";
        i := 23;
barrier go := true;
        j := "not moved";
        k := j;
        io.print(i);
        redo go:= false;
exit go;
        z:= j;
```

Application is only applicable to loops and not to guarded blocks in general, because execution of a statement outside the guarded block consumes processing resources which may have been prohibited by the block condition.

For example, it doesn't make sense to move creation of objects outside the barrier.

6 The MAL Debugger

In practice it is hard to write a correct MAL program the first time around. Instead, it is more often constructed by trial-and-error. As long as there are syntax and semantic errors the MAL compiler provides a sufficient handle to proceed. Once it passes the compiler we have to resort to a debugger to assess its behavior.

Note, the MAL debugger described here can be used in conjunction with the textual interface client *mclient* only. The JDBC protocol does not permit passing through information that 'violates' the protocol.

6.1 Program Debugging

To ease debugging and performance monitoring, the MAL interpreter comes with a gdb-like debugger. An illustrative session elicits the functionality offered.

```
mal>function test(i:int):str;
mal> io.print(i);
mal> i:= i*2;
mal> b:= bat.new(:int,:int);
mal> bat.insert(b,1,i);
mal> io.print(b);
mal> return test:= "ok";
mal>end test;
mal>user.test(1);
[ 1 ]
#-----------------#
# h     t         # name
# int   int       # type
#-----------------#
[ 1,      2       ]
```

The debugger can be entered at any time using the call mdb.start(). An overview of the available commands is readily available.

```
mal>mdb.start();
#mdb !end main;
mdb>help
next              -- Advance to next statement
continue          -- Continue program being debugged
catch             -- Catch the next exception
break [<var>]     -- set breakpoint on current instruction or <var>
delete [<var>]    -- remove break/trace point <var>
debug <int>       -- set kernel debugging mask
dot [<int>] [<file>]  -- generate the dependency graph
step              -- advance to next MAL instruction
module            -- display a module signatures
atom              -- show atom list
finish            -- finish current call
exit              -- terminate executionr
```

```
quit             -- turn off debugging
list <obj>       -- list current program block
List <obj>       -- list with type information
span             -- list the life span of variables
var  <obj>       -- print symbol table for module
optimizer <obj>  -- display optimizer steps
print <var>      -- display value of a variable
print <var> <cnt>[<first>] -- display BAT chunk
info <var>       -- display bat variable properties
run              -- restart current procedure
where            -- print stack trace
down             -- go down the stack
up               -- go up the stack
trace <var>      -- trace assignment to variables
trap <mod>.<fcn> -- catch MAL function call in console
set {timer,thread,flow,io,memory,bigfoot} -- set trace switches
unset            -- turn off switches
help             -- this message
mdb>
```

The term <OBJ> is an abbreviation for a MAL operation <MOD>.<FCN>, optionally extended with a version number, i.e. [<NR>]. The VAR denotes a variable in the current stack frame. Debugger commands may be abbreviated.

We walk our way through a debugging session, highlighting the effects of the debugger commands. The call to mdb.start() has been encapsulated in a complete MAL function, as shown by issuing the list command. A more detailed listing shows the binding to the C-routine and the result of type resolution.

```
mal>mdb.start();
#end main;
mdb>l
function user.main():int;
mdb.start();
end main;
mdb>L
function user.main():int;       # 0  (main:int)
mdb.start();        # 1 MDBstart (_1:void)
end main;       # 2
```

The user module is the default place for function defined at the console. The modules loaded can be shown typeing the command 'module' (or 'm' for short). The function signatures become visible using the module and optionally the function name.

```
mdb>m alarm
#command alarm.alarm(secs:int,action:str):void address ALARMsetalarm;
#command alarm.ctime():str address ALARMctime;
#command alarm.epilogue():void address ALARMepilogue;
#command alarm.epoch():int address ALARMepoch;
#command alarm.prelude():void address ALARMprelude;
#command alarm.sleep(secs:int):void address ALARMsleep;
```

```
#command alarm.time():int address ALARMtime;
#command alarm.timers():bat[:str,:str] address ALARMtimers;
#command alarm.usec():lng address ALARMusec;
mdb>m alarm.sleep
#command alarm.sleep(secs:int):void address ALARMsleep;
mdb>
```

The debugger mode is left with a <return>. Any subsequent MAL instruction re-activates the debugger to await for commands. The default operation is to step through the execution using the 'next' ('n') or 'step' ('s) commands, as shown below.

```
mal>user.test(1);
#    user.test(1);
mdb>n
#    io.print(i);
mdb>
[ 1 ]
#    i := calc.*(i,2);
mdb>
#    b := bat.new(:int,:int);
mdb>
```

The last instruction shown is next to be executed. The result can be shown using a print statement, which contains the location of the variable on the stack frame, its name, its value and type. The complete stack frame becomes visible with 'values' ('v') command:

```
#    bat.insert(b,1,i);
mdb>
#    io.print(b);
mdb>v
#Stack for 'test' size=32 top=11
#[0] test        = nil:str
#[1] i   = 4:int
#[2] _2  = 0:int   unused
#[3] _3  = 2:int  constant
#[4] b   = <tmp_1226>:bat[:int,:int]   count=1 lrefs=1 refs=0
#[5] _5  = 0:int   type variable
#[6] _6  = nil:bat[:int,:int]   unused
#[7] _7  = 1:int  constant
#[8] _8  = 0:int   unused
#[9] _9  = "ok":str  constant
```

The variables marked 'unused' have been introduced as temporary variables, but which are not referenced in the remainder of the program. It also illustrates basic BAT properties, a complete description of which can be obtained using the 'info' ('i') command. A sample of the BAT content can be printed passing tuple indices, e.g. 'print b 10 10' prints the second batch of ten tuples.

6.2 Handling Breakpoints

A powerful mechanism for debugging a program is to set breakpoints during the debugging session. The breakpoints are designated by a target variable name, a [module.]function name, or a MAL line number (#<number>).

The snippet below illustrates the reaction to set a break point on assignment to variable 'i'.

```
mal>mdb.start();
#end main;
mdb>
mal>user.test(1);
#    user.test(1);
mdb>break i
breakpoint on 'i' not set
mdb>n
#    io.print(i);
mdb>break i
mdb>c
[ 1 ]
#    i := calc.*(i,2);
mdb>
```

The breakpoints remain in effect over multiple function calls. They can be removed with the DELETE statement. A list of all remaining breakpoints is obtained with BREAKPOINTS.

The interpreter can be instructed to call the debugger as soon as an exception is raised. Simply add the instruction MDB.SETCATCH(TRUE).

6.3 Profile Switches

Switches control the level of detail output shown while debugging or tracing program execution. They are toggled with the SET and UNSET command. The following switches are currently supported:

TIMER	activates a listing of all instructions being executed. It is measured in wall-clock time.
FLOW	shows the total byte size of all BAT target results and input arguments. It is a good indicator on the amount of data being processed.
MEMORY	keeps track on growing memory needs.
IO	keeps track on the amount of physical IO and is used to detect operators consuming excessive amounts of space.
BIGFOOT	keeps track of the current and maximum virtual memory footprint of the BATs.[incomplete]

The snippet below shows setting the MEMORY and TIMER switch. The switches take effect at the next instruction.

```
mdb>set timer
mdb>set flow
mdb>c
[ 3 ]
#    26 usec#   0  0#    io.print(i=3)
#     6 usec#   0  0#    i := calc.*(i=6, _3=2)
#    10 usec#   0  0#    b := bat.new(_5=0, _6=0)
#     7 usec#   0  8#    bat.insert(b=<tmp_167>bat[:int,:int]{1}, _8=1, i=6)
#-----------------#
# h     t         # name
# int   int       # type
#-----------------#
[ 1,      6       ]
#    41 usec#   0  8#    io.print(b=<tmp_167>bat[:int,:int]{1})
#     7 usec#   0  0#    return test := "ok";
#   211 usec#   0  0#    user.test(_2=3)
```

6.4 Program Inspection

The debugger commands available for inspection of the program and symbol tables are:

LIST (LIST) [<MOD>.<FCN>['['<NR>']']]
: A listing of the current MAL block, or one designated by the <mod>.<fcn> is produced. The [<NR>] extension provides access to an element in the MAL block history. The alternative name 'List' also produces the type information.

OPTIMIZER [<MOD>.<FCN>['['<NR>']']]
: Gives an overview of the optimizer actions in the history of the MAL block. Intermediate results can be accessed using the list command.

ATOMS
: Lists the atoms currently known

MODULES [<MOD>]
: Lists the modules currently known. An optional <mod> argument produces a list of all signatures within the module identified.

DOT [<MOD>.<FCN>['['<NR>']']] [<FILE>]
: A dataflow diagram can be produced using the DOT command. It expects a function identifier with an optional history index and produces a file for the Linux program DOT, which can produce a nice, multi-page graph to illustrate plan complexity.

```
mdb>dot user.test
```

This example produces the USER-TST.DOT in the current working directory. The program call

```
dot -Tps user-tst-0.dot -o user-tst-0.ps
```

creates a postscript file with the graphs. With the Adobe reader professional you can break it up into multiple pages. An alternative is the program available from HTTP://WWW.TUG.ORG/TEX-ARCHIVE/SUPPORT/POSTER/POSTER.C The result is shown in the figure below:

Since the flow graphs become rather complex, an optional variable list limits its size.[TODO]

6.5 Runtime Inspection and Reflection

Part of the debugger functionality can also be used directly with MAL instructions. The execution trace of a snippet of code can be visualized encapsulation with MDB.SETTRACE(TRUE) and MDB.SETTRACE(FALSE). Likewise, the performance can be monitored with the command MDB.SETTIMER(ON/OFF). Using a boolean argument makes it easy to control the (performance) trace at run time. The following snippet shows the effect of patching the test case.

```
mal>function test(i:int):str;
mal> mdb.setTrace(true);
mal> io.print(i);
mal> i:= i*2;
mal> b:= bat.new(:int,:int);
mal> bat.insert(b,1,i);
mal> io.print(b);
mal> mdb.setTrace(false);
mal> return test:= "ok";
mal>end test;
mal>user.test(1);
#    mdb.setTrace(_3=true)
[ 1 ]
#    io.print(i=1)
#    i := calc.*(i=2, _5=2)
#    b := bat.new(_7=0, _8=0)
#    bat.insert(b=<tmp_1226>, _10=1, i=2)
#-----------------#
# h     t         # name
# int   int       # type
#-----------------#
[ 1,      2       ]
#    io.print(b=<tmp_1226>)
#   261 usec!    user.test(_2=1)
mal>
```

The command MDB.SETTIMER() toggles the performance traceing flag. The argument is a boolen to designate its state. The primary output of the timer switch is statistics in micro-seconds, the memory tracer shows the arena increment, and the IO tracer shows in- and out-blocks. The time spent on preparing the trace information is excluded from the report. For more detailed timing information the Linux tool *valgrind* may be of help.

The routines MDB.SETFLOW(), MDB.SETMEMORY(), and MDB.SETIO() (de-)activate the other switches.

```
mal>function test(i:int):str;
mal> mdb.setTimer(true);
mal> io.print(i);
mal> i:= i*2;
mal> b:= bat.new(:int,:int);
mal> bat.insert(b,1,i);
mal> io.print(b);
mal> mdb.setTimer(false);
mal> return test:= "ok";
mal>end test;
mal>user.test(1);
#     6 usec#    mdb.setTimer(_3=true)
[ 1 ]
#    43 usec#    io.print(i=1)
#     5 usec#    i := calc.*(i=2, _5=2)
#    24 usec#    b := bat.new(_7=0, _8=0)
#    10 usec#    bat.insert(b=<tmp_1226>, _10=1, i=2)
#-----------------#
# h     t         # name
# int   int       # type
#-----------------#
[ 1,      2       ]
#   172 usec#    io.print(b=<tmp_1226>)
#   261 usec#    user.test(_2=1)
```

It is also possible to activate the debugger from within a program using MDB.START(). It remains in this mode until you either issue a quit command, or the command mdb.stop() instruction is encountered. The debugger is only activated when the user can direct its execution from the client interface. Otherwise, there is no proper input channel and the debugger will run in trace mode.

The program listing functionality of the debugger is also captured in the MAL debugger module. The current code block can be listed using MDB.LIST() and MDB.LIST(). An arbitrary code block can be shown with MDB.LIST(*module,function*) and MDB.LIST(*module,function*). A BAT representation of the current function is return by MDB.GETDEFINITION().

The symbol table and stack content, if available, can be shown with the operations MDB.VAR() and MDB.LIST(*module,function*) Access to the stack frames may be helpful in the context of exception handling. The operation MDB.GETSTACKDEPTH() gives the depth and individual elements can be accessed as BATs using MDB.GETSTACKFRAME(*n*). The top stack frame is accessed using MDB.GETSTACKFRAME().

6.6 Debugger Attachment

Debugging a running MAL process is simplified with a few hooks in the kernel. It is illustrated with a short example.

First open a client connection using MAL as preferred language. Then the state of the system can be inspected, in particular, the clients active can be looked up.

```
mal>b:= clients.getLogins();
mal>c:= clients.getUsers();
mal>io.print(b,c);
#--------------------------------------------------#
# client        login                          users     # name
# int   str                             str       # type
```

```
#-------------------------------------------------#
[ 0,      "Thu Feb  7 15:57:08 2008",     "0"     ]
[ 1,      "Thu Feb  7 15:57:11 2008",     "0"     ]
```

Locate the process you are interested in and obtain its identifier, say N (the first column in the list above). The next step is to gracefully put the running process into debugging mode without jeopardizing the application running.

```
mal> mdb.setTrap(1);
#process 1 put to sleep
mal> mdb.grab();
```

As soon as the next MAL instruction of process N starts the target process is put to sleep and you can access the context for debugging. The control ends when you leave the debugger with a 'quit' command.

7 The MAL Profiler

A key issue in the road towards a high performance implementation is to understand where resources are being spent. This information can be obtained using different tools and at different levels of abstraction. A coarse grain insight for a particular application can be obtained using injection of the necessary performance capturing statements in the instruction sequence. Fine-grain, platform specific information can be obtained using existing profilers, like valgrind (http://www.valgrind.org), or hardware performance counters.

The MAL profiler collects detailed performance information, such as cpu, memory and statement information. It is optionally extended with IO activity, which is needed for coarse grain profiling only, and estimated bytes read/written by an instruction.

The execution profiler is supported by hooks in the MAL interpreter. The default strategy is to ship an event record immediately over a stream to a separate performance monitor, formatted as a tuple. An alternative strategy is preparation for off-line performance analysis.

Reflective performance analysis is supported by an event cache, the event log becomes available as a series of BATs.

7.1 Event Filtering

The profiler supports selective retrieval of performance information by tagging the instructions of interest. This means that a profiler call has a global effect, all concurrent users are affected by the performance overhead. Therefore, it is of primary interest to single user sessions.

The example below illustrates how the different performance counter groups are activated, instructions are filtered for tracking, and where the profile information is retained for a posteriori analysis.

```
#profiler.activate("event");
#profiler.activate("pc");
#profiler.activate("operation");
profiler.activate("time");
profiler.activate("ticks");
#profiler.activate("cpu");
#profiler.activate("memory");
#profiler.activate("io");
#profiler.activate("bytes");
#profiler.activate("diskspace");
profiler.activate("statement");
profiler.setFilter("*","insert");
profiler.setFilter("*","print");

profiler.openStream("/tmp/MonetDBevents");
profiler.start();
b:= bbp.new(:int,:int);
bat.insert(b,1,15);
bat.insert(b,2,4);
```

```
     bat.insert(b,3,9);
     io.print(b);
     profiler.stop();
     profiler.closeStream();
```

In this example, we are interested in all functions name INSERT and PRINT. A wildcard can be used to signify any name, e.g. no constraints are put on the module in which the operations are defined. Several profiler components are ignored, shown by commenting out the code line.

Execution of the sample leads to the creation of a file with the following content. The ticks are measured in micro-seconds.

```
# time, ticks,  stmt  # name
[ "15:17:56",   12,   "_27 := bat.insert(<tmp_15>{3},1,15);" ]
[ "15:17:56",   2,    "_30 := bat.insert(<tmp_15>{3},2,4);"  ]
[ "15:17:56",   2,    "_33 := bat.insert(<tmp_15>{3},3,9);"  ]
[ "15:17:56",   245,  "_36 := io.print(<tmp_15>{3});",   ]
```

7.2 Event Caching

Aside from shipping events to a separate process, the profiler can keep the events in a local BAT group. It is the default when no target file has been opened to collect the information.

Ofcourse, every measurement scheme does not come for free and may even obscure performance measurements obtained through e.g. valgrind. The separate event caches can be accessed using the operator PROFILER.GETTRACE(*name*). The current implementation only supports access to TIME,TICKS,PC,STATEMENT. The event cache can be cleared with PROFILER.CLEARTRACE().

Consider the following MAL program snippet:

```
     profiler.setAll();
     profiler.start();
     b:= bbp.new(:int,:int);
     bat.insert(b,1,15);
     io.print(b);
     profiler.stop();
     s:= profiler.getTrace("statement");
     t:= profiler.getTrace("ticks");
     io.print(s,t);
```

The performance result of the program execution becomes:

```
#-----------------------------------------------------------#
# h     t                                       t         # name
# int   str                                     int       # type
#-----------------------------------------------------------#
[ 1,      "b := bbp.new(0,0);",                   51      ]
[ 2,      "$6 := bat.insert(<tmp_22>,1,15);",     16      ]
[ 3,      "$9 := io.print(<tmp_22>);",            189     ]
```

7.3 Monitoring Variables

The easiest scheme to obtain performance data is to retrieve the performance properties of an instruction directly after it has been executed using getEvent(). It reads the profiling stack maintained, provided you have started monitoring.

```
profiler setFilter(b);
profiler.start();
....
b:= algebra.select(a,0 1000); # some expensive operation
(clk, memread, memwrite):= profiler.getEvent();
...
profiler.stop();
```

7.4 The Stethoscope

The performance profiler infrastructure provides precisely control through annotation of a MAL program. Often, however, inclusion of profiling statements is an afterthought.

The program `stethoscope` addresses this situation by providing a simple application that can attach itself to a running server and extracts the profiler events from concurrent running queries.

The arguments to `stethoscope` are the profiler properties to be traced and the applicable filter expressions. For example,

```
stethoscope -t bat.insert algebra.join
```

tracks the microsecond ticks of two specific MAL instructions. A synopsis of the calling conventions:

```
        stethoscope [options] +[aefoTtcmibds]
        -d | --dbname=<database_name>
        -u | --user=<user>
        -P | --password=<password>
        -p | --port=<portnr>
        -g | --gnuplot=<boolean>
        -h | --host=<hostname>

Event selector:
        a =aggregates
    e =event
    f =function
    o =operation called
    T =time
    t =ticks
    c =cpu statistics
    m =memory resources
    i =io resources
    b =bytes read/written
    d =diskspace needed
    s =statement
```

```
        p =pgfaults,cntxtswitches
```

Ideally, the stream of events should be piped into a 2D graphical tool, like xosview (Linux). A short term solution is to generate a gnuplot script to display the numerics organized as time lines. With a backup of the event lists give you all the information needed for a descent post-mortem analysis.

A convenient way to watch most of the SQL interaction you may use the command: stethoscope +tis algebra.* bat.* group.* sql.* aggr.*

8 The MAL Modules

This section contains a synopsis of the modules being shipped and which use knowledge of the MAL runtime context. They are sorted by module name and repetitive reading may be required to understand all details.

batExtensions	Extensions to the kernel/bat module.
BBP	BAT buffer pool interface.
Box	Box variable interface.
Chopper	Break a collection into chunks.
Clients	Client record inspection.
Constants	Global system defined values.
Factory	Factory management interface.
Inspect	Inspect the runtime symbol table(s).
I/O	The input/output interface.
Language	MAL language extension features.
MDB	MAL debugger interface.
Manual	Online manual material.
Mserver	MonetDB server interface.
MAT	Multiple association tables.
PBM	Dealing with partitioned BATS.
Profiler	Performance profiler.
PCRE	Regular expression handling over strings.
Statistics	A server-side statistics catalog.
Table	Table output interface.
Transactions	Transaction interface

8.1 Module Loading

The server is bootstrapped by processing a MAL script with module definitions or extensions. For each module file encountered, the object library lib_<modulename>.so is searched for in the location identified by monet_mod_path=exec_prefix/lib/MonetDB5:exec_prefix/lib/MonetDB5/lib:exec_prefix/lib/MonetDB5/bin.

The corresponding signature are defined in .../lib(64)/<modulename>.mal.

The default bootstrap script is called .../lib/MonetDB5/mal_init.mal and it is designated in the configuration file as the mal_init property. The rationale for this set-up is that database administrators can extend/overload the bootstrap procedure without affecting the software package being distributed. It merely requires a different direction for the mal_init property. The scheme also isolates the functionality embedded in modules from inadvertise use on non-compliant databases.

Unlike previous versions of MonetDB, modules can not be unloaded. Dynamic libraries are always global and, therefore, it is best to load them as part of the server initialization phase.

8.2 Module file loading

The default location to search for the module is in monet_mod_path unless an absolute path is given. Loading further relies on the Linux policy to search for the module location in the following order: 1) the colon-separated list of directories in the user's LD_LIBRARY_PATH, 2) the libraries specified in /etc/ld.so.cache and 3) /usr/lib followed by /lib. If the module contains a routine _init, then that code is executed before the loader returns. Likewise the routine _fini is called just before the module is unloaded.

A module loading conflict emerges if a function is redefined. A duplicate load is simply ignored by keeping track of modules already loaded.

8.3 BAT Extensions

The kernel libraries are unaware of the MAL runtime semantics. This calls for declaring some operations in the MAL module section and register them in the kernel modules explicitly.

A good example of this borderline case are BAT creation operations, which require a mapping of the type identifier to the underlying implementation type.

Another example concerns the (un)pack operations, which direct access the runtime stack to (push)pull the values needed.

`pattern bat.new(ht:any_1, tt:any_2, b:bat[:any_3,:any_4]) :bat[:any_1,:any_2]`
  address CMDBATclone comment "Creates a new empty transient BAT by cloning another.";

`pattern bat.new(ht:any_1, tt:any_2) :bat[:any_1,:any_2]`
  address CMDBATnew comment "Creates a new empty transient BAT, with head- and tail-types as indicated.";

`pattern bat.new(ht:any_1, tt:any_2, size:int) :bat[:any_1,:any_2]`
  address CMDBATnewint comment "Creates a new BAT with sufficient space.";

`pattern bat.new(ht:any_1, tt:any_2, size:lng) :bat[:any_1,:any_2]`
  address CMDBATnew comment "Creates a new BAT and allocate space.";

`pattern bat.new(ht:oid, tt:any_2, size:int) :bat[:oid,:any_2]`
  address CMDBATnewint;

`pattern bat.new(ht:oid, tt:any_2, size:lng) :bat[:oid,:any_2]`
  address CMDBATnew;

`pattern bat.new(b:bat[:any_1,:any_2] ) :bat[:any_1,:any_2]`
  address CMDBATnewDerived;

`pattern bat.new(b:bat[:any_1,:any_2], size:lng) :bat[:any_1,:any_2]`
address CMDBATnewDerived;

`command bat.new(nme:str):bat[:any_1,:any_2]`
address CMDBATderivedByName comment "Localize a bat by name and produce a clone.";

`command bat.reduce(b:bat[:any_1,:any_2]):bat[:any_1,:any_2]`
address CMDBATreduce comment "Drop auxillary BAT structures.";

`command bat.flush(b:bat[:any_1,:any_2]):void`
address CMDBATflush comment "Designate a BAT as not needed anymore.";

`pattern bat.setGarbage(b:bat[:any_1,:any_2]):void`
address CMDBATsetGarbage comment "Designate a BAT as garbage.";

`pattern bat.partition(b:bat[:any_1,:any_2]):bat[:any_1,:any_2]...`
address CMDbatpartition comment "Create a series of cheap slices over the first argument. The BUNs are distributed evenly.";

`pattern`
`bat.partition(b:bat[:any_1,:any_2],pieces:int,part:int):bat[:any_1,:any_2]`
address CMDbatpartition2 comment "Create a series of cheap slices over the first argument. The BUNs are distributed evenly.";

`pattern bat.unpack(b:bat[:any_1,:any_2])(h:any_1,t:any_2)`
address CMDbatunpack comment "Extract the first tuple from a BAT.";

`pattern bat.pack(h:any_1,t:any_2):bat[:any_1,:any_2]`
address CMDbatpack comment "Pack a pair of values into a BAT.";

`pattern bat.setBase(b:bat[:any_1,:any_2],c:bat[:any_1,:any_2]...):void`
address CMDsetBase comment "Give the non-empty BATs consecutive oid bases.";

8.4 BAT Buffer Pool

The BBP module implements a box interface over the BAT buffer pool. It is primarilly meant to ease inspection of the BAT collection managed by the server.

The two predominant approaches to use bbp is to access the BBP with either *bind* or *take*. The former merely maps the BAT name to the object in the bat buffer pool. A more controlled scheme is to *deposit*, *take*, *release* and *discard* elements. Any BAT B created can be brought under this scheme with the name N. The association N->B is only maintained in the box administration and not reflected in the BAT descriptor. In particular, taking a BAT object out of the box leads to a private copy to isolate the user from concurrent updates on the underlying store. Upon releasing it, the updates are merged with the master copy [todo].

The remainder of this module contains operations that rely on the MAL runtime setting, but logically belong to the kernel/bat module.

module bbp;

`command open():void`
address CMDbbpopen comment "Locate the bbp box and open it.";

`command close():void`
address CMDbbpclose comment "Close the bbp box.";

`command destroy():void`
address CMDbbpdestroy comment "Destroy the box";

`pattern take(name:str) :bat[:any_1,:any_2]`
address CMDbbptake comment "Load a particular bat.";

`pattern deposit(name:str,v:bat[:any_1,:any_2]) :void`
address CMDbbpdeposit comment "Enter a new bat into the bbp box.";

`pattern deposit(name:str,loc:str) :bat[:any_1,:any_2]`
address CMDbbpbindDefinition comment "Relate a logical name to a physical BAT in the buffer pool.";

`pattern commit():void`
address CMDbbpReleaseAll comment "Commit updates for this client.";

`pattern releaseAll():void`
address CMDbbpReleaseAll comment "Commit updates for this client.";

`pattern release(name:str,val:bat[:any_1,:any_2]) :void`
address CMDbbprelease comment "Commit updates and release this BAT.";

`pattern release(b:bat[:any_1,:any_2]):void`
address CMDbbpreleaseBAT comment "Remove the BAT from further consideration";

`pattern destroy(b:bat[:any_1,:any_2]):void`
address CMDbbpdestroyBAT1 comment "Schedule a BAT for removal at session end.";

`pattern destroy(b:bat[:any_1,:any_2],immediate:bit)`
address CMDbbpdestroyBAT comment "Schedule a BAT for removal at session end or immediately.";

`pattern toString(name:str):str`
address CMDbbptoStr comment "Get the string representation of an element in the box.";

`pattern discard(name:str):void`
address CMDbbpdiscard comment "Remove the BAT from the box.";

`pattern iterator(nme:str):lng`
address CMDbbpiterator comment "Locate the next element in the box.";

`pattern prelude():void`
address CMDbbpprelude comment "Initialize the bbp box.";

`pattern bind(name:str):bat[:any_1,:any_2]`
address CMDbbpbind comment "Locate the BAT using its logical name";

`pattern bind(head:str,tail:str):bat[:any_1,:any_2]`
address CMDbbpbind2 comment "Locate the BAT using the head and tail names in the BAT buffer pool");

`pattern bind(idx:int):bat[:any_1,:any_2]`
address CMDbbpbindindex comment "Locate the BAT using its BBP index in the BAT buffer pool";

`pattern getObjects():bat[:int,:str]`
address CMDbbpGetObjects comment "View of the box content.";

`command getHeadType() :bat[:int,:str]`
address CMDbbpHeadType comment "Map a BAT into its head type";

`command getTailType() :bat[:int,:str]`
address CMDbbpTailType comment "Map a BAT into its tail type";

`command getNames() :bat[:int,:str]`
address CMDbbpNames comment "Map BAT into its bbp name";

`command getRNames() :bat[:int,:str]`
address CMDbbpRNames comment "Map a BAT into its bbp physical name";

`command getName( b:bat[:any_1,:any_2]):str`
address CMDbbpName comment "Map a BAT into its internal name";

`command getCount() :bat[:int,:lng]`
address CMDbbpCount comment "Create a BAT with the cardinalities of all known BATs";

`command getRefCount() :bat[:int,:int]`
address CMDbbpRefCount comment "Create a BAT with the (hard) reference counts";

`command getLRefCount() :bat[:int,:int]`
address CMDbbpLRefCount comment "Create a BAT with the logical reference counts";

`command getLocation() :bat[:int,:str]`
address CMDbbpLocation comment "Create a BAT with their disk locations";

`command getHeat() :bat[:int,:int]`
address CMDbbpHeat comment "Create a BAT with the heat values";

`command getDirty() :bat[:int,:str]`
address CMDbbpDirty comment "Create a BAT with the dirty/ diffs/clean status";

`command getStatus() :bat[:int,:str]`
address CMDbbpStatus comment "Create a BAT with the disk/load status";

`command getKind():bat[:int,:str]`
address CMDbbpKind comment "Create a BAT with the persistency status";

`command getRefCount(b:bat[:any_1,:any_2]) :int`
address CMDgetBATrefcnt comment "Utility for debugging MAL interpreter";

`command getLRefCount(b:bat[:any_1,:any_2]) :int`
address CMDgetBATlrefcnt comment "Utility for debugging MAL interpreter";

`command getDiskSpace() :lng`
: address CMDbbpDiskSpace comment "Estimate the amount of disk space occupied by dbfarm";

`command getPageSize():int`
: address CMDgetPageSize comment "Obtain the memory page size";

8.5 Constants

The `const` module provides a box abstraction store for global constants. Between sessions, the value of the constants is saved on disk in the form of a simple MAL program, which is scanned and made available by opening the box. A future implementation should provide transaction support over the box, which would permit multiple clients to exchange (scalar) information easily.

The default constant box is initialized with session variables, such as 'user','dbname', 'dbfarm', and 'dbdir'. These actions are encapsulated in the prelude routine called.

A box should be opened before being used. It is typically used to set-up the list of current users and to perform authorization. The constant box is protected with a simple authorization scheme, prohibiting all updates unless issued by the system administrator.

module const;

`pattern open():void`
: address CSTopen comment "Locate and open the constant box.";

`pattern close():void`
: address CSTclose comment "Close the constant box.";

`pattern destroy():void`
: address CSTdestroy comment "Destroy the box.";

`pattern take(name:str):any_1`
: address CSTtake comment "Take a variable out of the box.";

`pattern deposit(name:str,val:any_1) :void`
: address CSTdeposit comment "Add a variable to the box.";

`pattern releaseAll():void`
: address CSTreleaseAll comment "Release all variables in the box.";

`pattern release(name:str) :void`
: address CSTrelease comment "Release a constant value.";

`pattern release(name:any_1):void`
: address CSTrelease comment "Release a constant value.";

`pattern toString(name:any_1):str`
: address CSTtoString comment "Get the string representation of an element in the box.";

`pattern discard(name:any_1) :void`
: address CSTdiscard comment "Release the const from the box.";

`pattern newIterator()(:lng,:str)`
: address CSTnewIterator comment "Locate next element in the box.";

`pattern hasMoreElements()(:lng,:str)`
  address CSThasMoreElements comment "Locate next element in the box.";

8.6 BAT Iterators

Many low level algorithms rely on an iterator to break a collection into smaller pieces. Each piece is subsequently processed by a block.

For very large BATs it may make sense to break it into chunks and process them separately to solve a query. An iterator pair is provided to chop a BAT into fixed size elements. Each chunk is made available as a BATview. It provides read-only access to an underlying BAT. Adjusting the bounds is cheap, once the BATview descriptor has been constructed.

The smallest granularity is a single BUN, which can be used to realize an iterator over the individual BAT elements. For larger sized chunks, the operators return a BATview.

All iterators require storage space to administer the location of the next element. The BAT iterator module uses a simple lng variable, which also acts as a cursor for barrier statements.

The larger chunks produced are currently static, i.e. their size is a parameter of the call. Dynamic chunk sizes are interesting for time-series query processing. (See another module)

`command bat.newIterator(b:bat[:any_1,:any_2], size:lng)`
  (:lng,:bat[:any_1,:any_2]) address CHPnewChunkIterator comment "Create an iterator with fixed granule size. The result is a view.";

`command bat.hasMoreElements(b:bat[:any_1,:any_2], size:lng)`
  (:lng, :bat[:any_1,:any_2]) address CHPhasMoreElements comment "Produce the next chunk for processing.";

`pattern bat.newIterator(b:bat[:any_1,:any_2]) (:lng, h:any_1, t:any_2)`
  address CHPbunIterator comment "Process the buns one by one extracted from a void table.";

`pattern bat.newIterator(b:bat[:any_1,:bat]) (:lng, h:any_1, t:any_2)`
  address CHPbunIterator comment "Process the buns one by one extracted from a void table.";

`pattern bat.hasMoreElements(b:bat[:any_1,:any_2]) (:lng, h:any_1, t:any_2)`
  address CHPbunHasMoreElements;

`pattern bat.hasMoreElements(b:bat[:oid,:any_2]) (:lng, h:oid, t:any_2)`
  address CHPbunHasMoreElements comment "Produce the next bun for processing.";

`pattern bat.hasMoreElements(b:bat[:any_1,:bat]) (:lng, h:any_1, t:any_2)`
  address CHPbunHasMoreElements comment "Produce the next bun for processing.";

The head and tail values can also be extracted using the cursor. It points to the first bun in the chunk under consideration. It is often more effective due to use the iterator with automatic extraction of head and tail value; the overhead involved is much less.

`pattern bat.getHead(b:bat[:any_1,:any],i:lng):any_1`
  address CHPgetHead comment "return the BUN head value using the cursor.";

`pattern bat.getTail(b:bat[:any_2,:any_1],i:lng):any_1`
: address CHPgetTail comment "return the BUN tail value using the cursor.";

8.7 Box definitions

This module shows the behavior of a simple box of objects. Objects are stored into the box using *deposit* and taken out with *take*. Once you are done, elements can be removed by name or reference using *discard*.

A box should be opened before being used. It is typically used to set-up the list of current users and to perform authorization.

module box;

`pattern open(nme:str):any_1`
: address BOXopen comment "Locate the box and open it.";

`pattern close(bname:str):void`
: address BOXclose comment "Close the box.";

`pattern destroy(bname:str):void`
: address BOXdestroy comment "Destroy the box.";

`pattern take(bnme:str, vnme:str):any_1`
: address BOXtake comment "Locate the typed value in the box.";

`pattern deposit(bname:str,name:str,v:any_1):void`
: address BOXdeposit comment "Enter a new value into the box.";

`pattern releaseAll(bname:str) :void`
: address BOXreleaseAll comment "Release all objects for this client.";

`pattern release(bname:str,nme:str,val:any_1):void`
: address BOXrelease comment "Release the BAT from the client pool.";

`pattern toString(bname:str,name:str) :str`
: address BOXtoString comment "Get the string representation of the i-th element in the box.";

`pattern discard(bname:str,name:str) :void`
: address BOXdiscard comment "Release the BAT from the client pool.";

`pattern iterator(nme:str):lng`
: address BOXiterator comment "Locates the next element in the box.";

`command getBoxNames():bat[:int,:str]`
: address BOXgetBoxNames comment "Retrieve the names of all boxes.";

8.8 Client Management

Each online client is represented with an entry in the clients table. The client may inspect his record at run-time and partially change its properties. The administrator sees all client records and has the right to adjust global properties.

module clients;

`pattern setListing(flag:int):int`
address CLTsetListing comment "Turn on/off echo of MAL instructions: 2 - show mal instruction, 4 - show details of type resolutoin, 8 - show binding information.";

`pattern setHistory(s:str)`
address CLTsetHistory comment "Designate console history file for readline.";

`pattern getId():int`
address CLTgetClientId comment "Return a number that uniquely represents the current client.";

`pattern getInfo( ):bat[:str,:str]`
address CLTInfo comment "Pseudo bat with client attributes.";

`pattern getScenario():str`
address CLTgetScenario comment "Retrieve current scenario name.";

`pattern setScenario(msg:str):str`
address CLTsetScenario comment "Switch to other scenario handler, return previous one.";

`pattern quit():void`
address CLTquit comment "Terminate the client session.";

`pattern quit(idx:int):void`
address CLTquit comment "Terminate the session for a single client using a soft error. It is the privilige of the console user.";

Administrator operations

`command getLogins( ):bat[:int,:str]`
address CLTLogin comment "Pseudo bat of client login time.";

`command getLastCommand( ):bat[:int,:str]`
address CLTLastCommand comment "Pseudo bat of client's last command time.";

`command getActions( ):bat[:int,:int]`
address CLTActions comment "Pseudo bat of client's command counts.";

`command getTime( ):bat[:int,:lng]`
address CLTTime comment "Pseudo bat of client's total time usage(in usec).";

`command getUsers( ):bat[:int,:str]`
address CLTusers comment "Pseudo bat of users logged in.";

`pattern stop(id:int)`
address CLTstop comment "Stop the query execution at the next eligble statement.";

`pattern suspend(id:int):void`
address CLTsuspend comment "Put a client process to sleep for some time. It will simple sleep for a second at a time, until the awake bit has been set in its descriptor";

`command wakeup(id:int):void`
: address CLTwakeup comment "Wakeup a client process";

`pattern setTimeout(q:int,s:int):void`
: address CLTsetTimeout comment "Abort a query after q seconds (q=0 means run undisturbed). The session timeout aborts the connection after spending too many seconds on query processing.";

`pattern getTimeout()(q:int,s:int)`
: address CLTgetTimeout comment "A query is aborted after q seconds (q=0 means run undisturbed). The session timeout aborts the connection after spending too many seconds on query processing.";

`command shutdown(forced:bit):void`
: address CLTshutdown comment "Close all client connections. If forced=false the clients are moved into FINISHING mode, which means that the process stops at the next cycle of the scenario. If forced=true all client processes are immediately killed";

8.9 Factory management

The factory infrastructure can be inspected and steered with the commands provided here.

module factories;

`command getPlants()(mod:bat[:oid,:str], fcn:bat[:oid,:str])`
: address FCTgetPlants comment "Retrieve the names for all active factories.";

`command getCaller():int`
: address FCTgetCaller comment "Retrieve the unique identity of the factory caller.";

`command getOwners():bat[:oid,:str]`
: address FCTgetOwners comment "Retrieve the factory owners table.";

`command getArrival():bat[:oid,:timestamp]`
: address FCTgetArrival comment "Retrieve the time stamp the last call was made.";

`command getDeparture():bat[:oid,:timestamp]`
: address FCTgetDeparture comment "Retrieve the time stamp the last answer was returned.";

`pattern shutdown(m:str, f:str):void`
: address FCTshutdown comment "Close a factory.";

8.10 Inspection

This module introduces a series of commands that provide access to information stored within the interpreter data structures. It's primary use is debugging. In all cases, the pseudo BAT operation is returned that should be garbage collected after being used.

The main performance drain would be to use a pseudo BAT directly to successively access it components. This can be avoided by first assigning the pseudo BAT to a variable.

module inspect;

`command getWelcome():str`
: address INSPECTgetWelcome comment "Return the server message of the day string";

`pattern getDefinition(mod:str,fcn:str) :bat[:str,:str]`
: address INSPECTgetDefinition comment "Returns a string representation of a specific function.";

`pattern getSignature(mod:str,fcn:str) :bat[:str,:str]`
: address INSPECTgetSignature comment "Returns the function signature(s).";

`pattern getAddress(mod:str,fcn:str) :bat[:str,:str]`
: address INSPECTgetAddress comment "Returns the function signature(s).";

`pattern getComment(mod:str,fcn:str) :bat[:str,:str]`
: address INSPECTgetComment comment "Returns the function help information.";

`pattern getSource(mod:str,fcn:str):str`
: address INSPECTgetSource comment "Return the original input for a function.";

`pattern getKind():bat[:oid,:str]`
: address INSPECTgetkind comment "Obtain the instruction kind.";

`pattern getModule():bat[:oid,:str]`
: address INSPECTgetAllModules comment "Obtain the function name.";

`pattern getFunction():bat[:oid,:str]`
: address INSPECTgetAllFunctions comment "Obtain the function name.";

`pattern getSignatures():bat[:oid,:str]`
: address INSPECTgetAllSignatures comment "Obtain the function signatures.";

`pattern getAddresses():bat[:oid,:str]`
: address INSPECTgetAllAddresses comment "Obtain the function address.";

`pattern getSize():lng`
: address INSPECTgetSize comment "Return the storage size for the current function (in bytes).";

`pattern getSize(mod:str):bat[:str,:lng]`
: address INSPECTgetModuleSize comment "Return the storage size for a module (in bytes).";

`pattern getSize(mod:str,fcn:str):lng`
: address INSPECTgetFunctionSize comment "Return the storage size for a function (in bytes).";

`pattern getType(v:bat[:any_1,:any_2]) (ht:str, tt:str)`
: address INSPECTtypeName comment "Return the concrete type of a variable (expression).";

`pattern getType(v:any_1) :str`
: address INSPECTtypeName comment "Return the concrete type of a variable (expression).";

`command getTypeName(v:int):str`
: address INSPECTtypename comment "Get the type name associated with a type id.";

`pattern getTypeIndex(v:bat[:any_1,:any_2]) (ht:int, tt:int)`
: address INSPECTtypeIndex comment "Return the type index of a BAT head and tail.";

`pattern getTypeIndex(v:any_1):int`
: address INSPECTtypeIndex comment "Return the type index of a variable. For BATs, return the type index for its tail.";

`pattern equalType(l:any, r:any):bit`
: address INSPECTequalType comment "Return true if both operands are of the same type";

`command getAtomNames():bat[:int,:str]`
: address INSPECTatom_names comment "Collect a BAT with the atom names.";

`command getAtomSuper():bat[:int,:str]`
: address INSPECTatom_sup_names comment "Collect a BAT with the atom names.";

`command getAtomSizes():bat[:int,:int]`
: address INSPECTatom_sizes comment "Collect a BAT with the atom sizes.";

`command getEnvironment():bat[:str,:str]`
: address INSPECTgetEnvironment comment "Collect the environment variables.";

8.11 Input/Output module

The IO module provides simple ASCII-IO rendering options. It is modeled after the tuple formats, but does not attempt to outline the results. Instead, it is geared at speed, which also means that some functionality regarding the built-in types is duplicated from the atoms definitions.

A functional limited form of formatted printf is also provided. It accepts at most one variable. A more complete approach is the tablet module.

The commands to load and save a BAT from/to an ASCII dump are efficient, but work only for binary tables.

module io;

`pattern stdin():bstream`
: address io_stdin comment "return the input stream to the database client";

`pattern stderr():streams`
: address io_stderr comment "return the error stream for the database console";

`pattern stdout():streams`
address io_stdout comment "return the output stream for the database client";

`pattern print(val:any_1,lst:any...):void`
address IOprint_val comment "Print a MAL value tuple .";

`pattern print(b1:bat[:any_1,:any]...):void`
address IOtable comment "BATs are printed with '#' for legend lines, and the BUNs on seperate lines between brackets, containing each to comma separated values (head and tail). If multiple BATs are passed for printing, print() performs an implicit natural join, producing a multi attribute table.";

`pattern ftable( filep:streams, b1:bat[:any_1,:any], b:bat[:any_1,:any]... ):void`
address IOftable comment "Print an n-ary table to a file.";

`pattern print(order:int,b:bat[:any_1,:any], b2:bat[:any_1,:any]...):void`
address IOotable comment "The same as normal table print, but enforces to use the order of BAT number [1..argc] to do the printing.";

`pattern table(b1:bat[:any_1,:any], b2:bat[:any_1,:any]...):void`
address IOttable comment "Print an n-ary table. Like print, but does not print oid column";

`pattern table(order:int, b1:bat[:any_1,:any], b2:bat[:any_1,:any]...):void`
address IOtotable comment "Print an n-ary table.";

`pattern ftable(fp:streams, order:int, b1:bat[:any_1,:any], b:bat[:any_1,:any]...):void`
address IOfotable comment "Print an n-ary table to a file.";

`pattern print(val:any_1):void`
address IOprint_val comment "Print a MAL value tuple .";

`pattern print(val:bat[:any_1,:any_2]):void`
address IOprint_val comment "Print a MAL value tuple .";

`pattern prompt(val:any_1):void`
address IOprompt_val comment "Print a MAL value without brackets.";

`pattern printf(fmt:str,val:any...):void`
address IOprintf comment "Select default format ";

`pattern printf(fmt:str):void`
address IOprintf comment "Select default format ";

`pattern printf(filep:streams,fmt:str,val:any...):void`
address IOprintfStream comment "Select default format ";

`pattern printf(filep:streams,fmt:str):void`
address IOprintfStream comment "Select default format ";

`command data(fname:str):str`
address IOdatafile comment "Signals receipt of tuples in a file fname. It returns the name of the file, if it still exists.";

`command export(b:bat[:any_1,:any_2], filepath:str):bit`
: address IOexport comment "Export a BAT as ASCII to a file. If the 'filepath' is not absolute, it is put into the .../dbfarm/$DB directory. Success of failure is indicated.";

`command import(b:bat[:any_1,:any_2], filepath:str) :bat[:any_1,:any_2]`
: address IOimport comment "Import a BAT from an ASCII dump. The new tuples are *inserted* into the parameter BAT. You have to create it! Its signature must match the dump, else parsing errors will occur and FALSE is returned.";

8.12 Language Extensions

Iterators over scalar ranges are often needed, also at the MAL level. The barrier and control primitives are sufficient to mimic them directly.

The modules located in the kernel directory should not rely on the MAL datastructures. That's why we have to deal with some bat operations here and delegate the signature to the proper module upon loading.

Running a script is typically used to initialize a context. Therefore we need access to the runtime context. For the call variants we have to determine an easy way to exchange the parameter/return values.

module language;

`command newRange(v:oid)(:bit,:oid)`
: address RNGnewRange_oid;

`command newRange(v:sht)(:bit,:sht)`
: address RNGnewRange_sht;

`command newRange(v:int)(:bit,:int)`
: address RNGnewRange_int;

`command newRange(v:lng)(:bit,:lng)`
: address RNGnewRange_lng;

`command newRange(v:flt)(:bit,:flt)`
: address RNGnewRange_flt;

`command newRange(v:dbl)(:bit,:dbl)`
: address RNGnewRange_dbl comment "This routine introduces an iterator over a scalar domain.";

`command nextElement(step:oid,last:oid)(:bit,:oid)`
: address RNGnextElement_oid;

`command nextElement(step:sht,last:sht)(:bit,:sht)`
: address RNGnextElement_sht;

`command nextElement(step:int,last:int)(:bit,:int)`
: address RNGnextElement_int;

`command nextElement(step:lng,last:lng)(:bit,:lng)`
: address RNGnextElement_lng;

`command nextElement(step:flt,last:flt)(:bit,:flt)`
  address RNGnextElement_flt;

`command nextElement(step:dbl,last:dbl)(:bit,:dbl)`
  address RNGnextElement_dbl comment "Advances the iterator with a fixed value until it becomes >= last.";

`command raise(msg:str) :str`
  address CMDraise comment "Raise an exception labeled with a specific message.";

`command assert(v:bit,term:str):void`
  address MALassertBit;

`command assert(v:sht,term:str):void`
  address MALassertSht;

`command assert(v:int,term:str):void`
  address MALassertInt;

`command assert(v:lng,term:str):void`
  address MALassertLng;

`command assert(v:str,term:str):void`
  address MALassertStr;

`command assert(v:oid,term:str):void`
  address MALassertOid;

`pattern assert(v:any_1,pname:str,oper:str,val:any_2):void`
  address MALassertTriple comment "Assertion test.";

`pattern assertSpace(depth:int)`
  address safeguardStack comment "Ensures that the current call does not consume more than depth*vtop elements on the stack.";

`pattern dataflow():int`
  address MALstartDataflow comment "The current guarded block is executed using dataflow control. ";

`pattern register(m:str,f:str,code:str,help:str):void`
  address CMDregisterFunction comment"Compile the code string and register it as a MAL function.";

`pattern setMemoryTrace(flg:bit):void`
  address CMDsetMemoryTrace comment "Set the flag to trace the memory footprint";

`pattern setThreadTrace(flg:bit):void`
  address CMDsetThreadTrace comment "Set the flag to trace the interpreter threads";

`pattern setTimerTrace(flg:bit):void`
  address CMDsetTimerTrace comment "Set the flag to trace the execution time";

pattern `setIOTrace(flg:bit):void`
address CMDsetIOTrace comment "Set the flag to trace the IO";

pattern `call(s:str):void`
address CMDcallString comment "Evaluate a MAL string program.";

pattern `call(s:bat[:oid,:str]):void`
address CMDcallBAT comment "Evaluate a program stored in a BAT.";

pattern `source(f:str):void`
address CMDevalFile comment "Merge the instructions stored in the file with the current program.";

8.13 MAL debugger interface

This module provides access to the functionality offered by the MonetDB debugger and interpreter status. It is primarilly used in interactive sessions to activate the debugger at a given point. Furthermore, the instructions provide the necessary handle to generate information for post-mortum analysis.

To enable ease of debugging and performance monitoring, the MAL interpreter comes with a hardwired gdb-like text-based debugger. A limited set of instructions can be included in the programs themselves, but beware that debugging has a global effect. Any concurrent user will be affected by breakpoints being set.

The prime scheme to inspect the MAL interpreter status is to use the MAL debugger directly. However, in case of automatic exception handling it helps to be able to obtain BAT versions of the critical information, such as stack frame table, stack trace, and the instruction(s) where an exception occurred. The inspection typically occurs in the exception handling part of the MAL block.

Beware, a large class of internal errors can not easily captured this way. For example, bus-errors and segmentation faults lead to premature termination of the process. Similar, creation of the post-mortum information may fail due to an inconsistent state or insufficient resources.

module mdb;

pattern `start():void`
address MDBstart comment "Start interactive debugger";

pattern `start(clientid:int):void`
address MDBstart comment "Start interactive debugger on a client";

pattern `start(mod:str,fcn:str):void`
address MDBstartFactory comment "Start interactive debugger on a running factory";

pattern `stop():void`
address MDBstop comment "Stop the interactive debugger";

pattern `inspect(mod:str,fcn:str):void`
address MDBinspect comment "Run the debugger on a specific function";

command `modules():bat[:int,:str]`
address CMDmodules comment "List available modules";

pattern `setTrap(mod:str, fcn:str, b:bit):void`
: address MDBtrapFunction comment "Suspend upon a call to the MAL function.";

pattern `setTrap(idx:int):void`
: address mdbTrapClient comment "Call debugger for a specific process.";

pattern `setTrace(b:bit):void`
: address MDBsetTrace comment "Turn on/off tracing of current routine";

pattern `setTrace(b:str):void`
: address MDBsetVarTrace comment "Turn on/off tracing of a variable ";

pattern `setCatch(b:bit):void`
: address MDBsetCatch comment "Turn on/off catching exceptions";

pattern `setThread(b:bit):void`
: address MDBsetThread comment "Turn on/off thread identity for debugger";

pattern `setTimer(b:bit):void`
: address MDBsetTimer comment "Turn on/off performance timer for debugger";

pattern `setMemoryTrace(b:bit):void`
: address MDBsetBigfoot comment "Turn on/off memory foot print tracer for debugger";

pattern `setFlow(b:bit):void`
: address MDBsetFlow comment "Turn on/off memory flow debugger";

pattern `setMemory(b:bit):void`
: address MDBsetMemory comment "Turn on/off memory statistics tracing.";

pattern `setIO(b:bit):void`
: address MDBsetIO comment "Turn on/off io statistics tracing";

pattern `setCount(b:bit):void`
: address MDBsetCount comment "Turn on/off bat count statistics tracing";

command `getDebug():int`
: address MDBgetDebug comment "Get the kernel debugging bit-set. See the MonetDB configuration file for details";

command `setDebug(flg:str):int`
: address MDBsetDebugStr comment "Set the kernel debugging bit-set and return its previous value. The recognized options are: threads, memory, properties, io, transactions, modules, algorithms, estimates, xproperties";

command `setDebug(flg:int):int`
: address MDBsetDebug comment "Set the kernel debugging bit-set and return its previous value.";

command `getException(s:str):str`
: address MDBgetExceptionVariable comment "Extract the variable name from the exception message";

`command getReason(s:str):str`
: address MDBgetExceptionReason comment "Extract the reason from the exception message";

`command getContext(s:str):str`
: address MDBgetExceptionContext comment "Extract the context string from the exception message";

`pattern list():void`
: address MDBlist comment "Dump the current routine on standard out.";

`pattern listMapi():void`
: address MDBlistMapi comment "Dump the current routine on standard out with Mapi prefix.";

`pattern list(M:str,F:str):void`
: address MDBlist3 comment "Dump the routine M.F on standard out.";

`pattern List():void`
: address MDBlistDetail comment "Dump the current routine on standard out.";

`pattern List(M:str,F:str):void`
: address MDBlist3Detail comment "Dump the routine M.F on standard out.";

`pattern var():void`
: address MDBvar comment "Dump the symboltable of current routine on standard out.";

`pattern var(M:str,F:str):void`
: address MDBvar3 comment "Dump the symboltable of routine M.F on standard out.";

`pattern lifespan(M:str,F:str):void`
: address MDBlifespan comment "Dump the current routine lifespan information on standard out.";

`pattern grab():void`
: address mdbGrab comment "Call debugger for a suspended process.";

`pattern trap():void`
: address mdbTrap comment "A suspended process for debugging.";

`pattern dot(M:str,F:str,s:str):void`
: address MDBshowFlowGraph comment "Dump the data flow of the function M.F in a format recognizable by the command 'dot' on the file s";

`pattern getStackDepth():int`
: address MDBStkDepth comment "Return the depth of the calling stack.";

`pattern getStackFrame(i:int):bat[:str,:str]`
: address MDBgetStackFrameN;

`pattern getStackFrame():bat[:str,:str]`
: address MDBgetStackFrame comment "Collect variable binding of current (n-th) stack frame.";

`pattern getStackTrace():bat[:void,:str]`
address MDBStkTrace;

`pattern dump()`
address MDBdump comment "Dump instruction, stacktrace, and stack";

`pattern getDefinition():bat[:void,:str]`
address MDBgetDefinition comment "Returns a string representation of the current function with typing information attached";

8.14 Manual Inspection

This module introduces a series of commands that provide access to the help information stored in the runtime environment.

The manual bulk operations ease offline inspection of all function definitions. It includes an XML organized file, because we expect external tools to massage it further for presentation.

module manual;

`pattern help(text:str)`
address MANUALhelp comment "Produces a list of all <module>.<function> that match the text pattern. The wildcard '*' can be used for <module> and <function>. Using the '(' asks for signature information and using ')' asks for the complete help record.";

`pattern search(text:str)`
address MANUALsearch comment "Search the manual for command descriptions that match the regular expression 'text'";

`pattern createXML(mod:str):void`
address MANUALcreate1 comment "Generate a synopsis of a module";

`pattern createXML():void`
address MANUALcreate0 comment "Produces a XML-formatted manual over all modules loaded.";

`pattern section(mod:str):void`
address MANUALcreateSection comment "Generate a synopsis of a module for the reference manual";

`pattern index():void`
address MANUALcreateIndex comment "Produces an overview of all names grouped by module.";

`pattern summary():void`
address MANUALcreateSummary comment "Produces a manual with help lines grouped by module.";

`pattern completion(pat:str):bat[:int,:str]`
address MANUALcompletion comment "Produces the wordcompletion table.";

8.15 MAPI interface

The complete Mapi library is available to setup communication with another Mserver.

Clients may initialize a private listener to implement specific services. For example, in an OLTP environment it may make sense to have a listener for each transaction type, which simply parses a sequence of transaction parameters.

Authorization of access to the server is handled as part of the client record initialization phase.

This library internally uses pointer handles, which we replace with an index in a locally maintained table. It provides a handle to easily detect havoc clients.

A cleaner and simplier interface for distributed processing is available in the module remote.

module mapi;

`command listen():int`
: address SERVERlisten_default comment "Start a Mapi server with the default settings.";

`command listen(port:int):int`
: address SERVERlisten_port comment "Start a Mapi listener on the port given.";

`command listen(port:int, maxusers:int):int`
: address SERVERlisten2 comment "Start a Mapi listener.";

`command listen(port:int, maxusers:int, cmd:str):int`
: address SERVERlisten3 comment "Start the Mapi listener on <port> for <maxusers>. For a new client connection MAL procedure <cmd>(Stream s_in, Stream s_out) is called.If no <cmd> is specified a new client thread is forked.";

`command stop():void`
: address SERVERstop comment "Terminate connection listeners.";

`command suspend():void`
: address SERVERsuspend comment "Suspend accepting connections.";

`command resume():void`
: address SERVERresume comment "Resume connection listeners.";

`command malclient(in:streams, out:streams):void`
: address SERVERclient comment "Start a Mapi client for a particular stream pair.";

`command trace(mid:int,flag:int):void`
: address SERVERtrace comment "Toggle the Mapi library debug tracer.";

`pattern reconnect(host:str, port:int, usr:str, passwd:str,lang:str):int`
: address SERVERreconnectWithoutAlias comment "Re-establish connection with a remote mserver.";

`pattern reconnect(host:str, port:int, db_alias:str, usr:str, passwd:str,lang:str):int`
: address SERVERreconnectAlias comment "Re-establish connection with a remote mserver.";

`command reconnect(mid:int):void`
address SERVERreconnect comment "Re-establish a connection.";

`pattern connect(host:str, port:int, usr:str, passwd:str,lang:str):int`
address SERVERconnect comment "Establish connection with a remote mserver.";

`command disconnect(dbalias:str):int`
address SERVERdisconnectWithAlias comment "Close connection with a remote Mserver.";

`command disconnect():int`
address SERVERdisconnectALL comment "Close connections with all remote Mserver.";

`command setAlias(dbalias:str)`
address SERVERsetAlias comment "Give the channel a logical name.";

`command lookup(dbalias:str):int`
address SERVERlookup comment "Retrieve the connection identifier.";

`command disconnect(mid:int):void`
address SERVERdisconnect comment "Terminate the session.";

`command destroy(mid:int):void`
address SERVERdestroy comment "Destroy the handle for an Mserver.";

`command ping(mid:int):int`
address SERVERping comment "Test availability of an Mserver.";

`command query(mid:int, qry:str):int`
address SERVERquery comment "Sent the query for execution";

`command query_handle(mid:int, qry:str):int`
address SERVERquery_handle comment "Sent the query for execution.";

`pattern query_array(mid:int, qry:str, arg:str...):int`
address SERVERquery_array comment "Sent the query for execution replacing '?' by arguments.";

`command prepare(mid:int, qry:str):int`
address SERVERprepare comment "Prepare a query for execution.";

`command finish(hdl:int):int`
address SERVERfinish comment "Remove all remaining answers.";

`command get_field_count(hdl:int):int`
address SERVERget_field_count comment "Return number of fields.";

`command get_row_count(hdl:int):lng`
address SERVERget_row_count comment "Return number of rows.";

`command rows_affected(hdl:int):lng`
address SERVERrows_affected comment "Return number of affected rows.";

`command fetch_row(hdl:int):int`
address SERVERfetch_row comment "Retrieve the next row for analysis.";

command `fetch_all_rows(hdl:int):lng`
    address SERVERfetch_all_rows comment "Retrieve all rows into the cache.";

command `fetch_field(hdl:int,fnr:int):str`
    address SERVERfetch_field_str comment "Retrieve a single field.";

command `fetch_field(hdl:int,fnr:int):int`
    address SERVERfetch_field_int comment "Retrieve a single int field.";

command `fetch_field(hdl:int,fnr:int):lng`
    address SERVERfetch_field_lng comment "Retrieve a single lng field.";

command `fetch_field(hdl:int,fnr:int):sht`
    address SERVERfetch_field_sht comment "Retrieve a single sht field.";

command `fetch_field(hdl:int,fnr:int):void`
    address SERVERfetch_field_void comment "Retrieve a single void field.";

command `fetch_field(hdl:int,fnr:int):oid`
    address SERVERfetch_field_oid comment "Retrieve a single void field.";

command `fetch_field(hdl:int,fnr:int):chr`
    address SERVERfetch_field_chr comment "Retrieve a single chr field.";

command `fetch_field_array(hdl:int):bat[:int,:str]`
    address SERVERfetch_field_bat comment "Retrieve all fields for a row.";

command `fetch_line(hdl:int):str`
    address SERVERfetch_line comment "Retrieve a complete line.";

command `fetch_reset(hdl:int):int`
    address SERVERfetch_reset comment "Reset the cache read line.";

command `next_result(hdl:int):int`
    address SERVERnext_result comment "Go to next result set.";

command `error(mid:int):int`
    address SERVERerror comment "Check for an error in the communication.";

command `getError(mid:int):str`
    address SERVERgetError comment "Get error message.";

command `explain(mid:int):str`
    address SERVERexplain comment "Turn the error seen into a string.";

pattern `put(mid:int, nme:str, val:any_1):void`
    address SERVERput comment "Send a value to a remote site.";

pattern `put(nme:str, val:any_1):str`
    address SERVERputLocal comment "Prepare sending a value to a remote site.";

pattern `rpc(key:int,qry:str...):any`
    address SERVERmapi_rpc_single_row comment "Sent a simple query for execution and fetch result.";

pattern `rpc(key:int,qry:str):bat[:any_1,:any_2]`
    address SERVERmapi_rpc_bat;

`command rpc(key:int,qry:str):void`
    address SERVERquery comment "Sent a simple query for execution.";

`pattern bind(key:int,rschema:str,rtable:str,rcolumn:str,i:int):bat[:any_1,:any_2]`
    address SERVERbindBAT comment "Bind a remote variable to a local one.";

`pattern bind(key:int,rschema:str,rtable:str,i:int):bat[:any_1,:any_2]`
    address SERVERbindBAT comment "Bind a remote variable to a local one.";

`pattern bind(key:int,remoteName:str):bat[:any_1,:any_2]`
    address SERVERbindBAT comment "Bind a remote variable to a local one.";

    mapi.listen();

8.16 Multiple association tables

A MAT is a convenient way to deal represent horizontal fragmented tables. It combines the definitions of several, type compatible BATs under a single name. It is produced by the mitosis optimizer and the operations are the target of the mergetable optimizer.

The MAT is materialized when the operations can not deal with the components individually, or the incremental operation is not supported. Normally all mat.new() operations are removed by the mergetable optimizer. In case a mat.new() is retained in the code, then it will behaves as a mat.pack();

The primitives below are chosen to accomodate the SQL front-end to produce reasonable efficient code.

    module mat;

`pattern new(b:bat[:any_1,:any_2]...):bat[:any_1,:any_2]`
    address MATpack comment "Define a Merge Association Table (MAT). Faal back to the pack operation when this is called ";

`pattern pack(:any_2...):bat[:void,:any_2]`
    address MATpackValues comment "Materialize the MAT (of values) into a BAT";

`pattern pack(b:bat[:any_1,:any_2]...):bat[:any_1,:any_2]`
    address MATpack comment "Materialize the MAT into a BAT";

`pattern pack2(b:bat[:any_1,:any_2]...):bat[:any_1,:any_2]`
    address MATpack2 comment "Materialize the MAT into a BAT (by an append all)";

`pattern print(b:bat[:any_1,:any_2]...):void`
    address MATprint;

`pattern newIterator(grp:bat[:any_1,:any_2]...):bat[:any_1,:any_2]`
    address MATnewIterator comment "Create an iterator over a MAT";

`pattern hasMoreElements(grp:bat[:any_1,:any_2]...):bat[:any_1,:any_2]`
    address MAThasMoreElements comment "Find the next element in the merge table";

`command info(g:str, e:str):bat[:any_1,:any_2]`
address MATinfo comment "retrieve the definition from the partition catalogue";

8.17 BAT Partition Manager

In real-life database applications the BATs tend to grow beyond the memory size. This leads to a heavy IO dominated behavior, which can partly be avoided by breaking up the query into a sequence of subqueries using a map-reduce strategy. The BAT partition manager (BPM) module is designed to support this strategy using range- and hash-partitioning.

Consider we want to reorganize R:bat[:oid,:int] into two partitions, based on splitting by tail value. The following MAL program illustrates the snippet of actions needed:

bpm.open(); Ralias:= bpm.deposit("myversion",R:bat[:oid,:int]); bpm.rangePartition(Ralias,nil:int,100); bpm.rangePartition(Ralias,101,200); bpm.close();

The command BPM.DEPOSIT registers a BAT as one for which a partitioned copy is required. The first partition call breaks the orginal BAT into two pieces: (nil:int,100) and (101,nil:int). The second call breaks the latter into (101,200) and (201,nil:int). The BAT partitions share the persistency properties. Partitioning on the head simple calls for a reverse operation on the source BAT first.

The partition manager also supports hash-based partitioning. Its argument is the number of hash bucket bits.

bpm.open(); Rev:= bat.reverse(R:bat[:oid,:int]); Ralias:= bpm.deposit("myHashVersion",Rev); # creates side effects bpm.hashPartition(Ralias,2); bpm.close();

This example creates a hash-partition based on the head.

The design is based on the assumption that partitions are reasonably large. This helps to limit plan explosion. (or a scheduler should step in)

8.17.1 Derived partitioning

A relational front-end would benefit from derived horizontal fragmentation. It would enable grouping together related fragments on the same site. Assume a relation R(A,B) which is already partitioned on A the derived fragmentation on the head is enforced with

bpm.derivePartition(B,A);

8.17.2 Using partitions

The partitioned BAT can be used in two ways. A query plan can be rewritten into a generator over the partitions, or it can be used by optimizers to derived all subqueries first for symbolic evaluation.

The former is illustrated with the snippet to select part of a partitioned BAT. In this example we collect the partial results in the accumulator BAT U.

bpm.open(); Ralias:bat[:oid,:int]:= bpm.take("myversion"); U:= bat.new(:oid,:int); barrier Rp:= bpm.newIterator(Ralias); ... t:= algebra.select(Rp,0,100); U:= algebra.union(tu,t); ... redo Rp:= bpm.hasMoreElements(Ralias); exit Rp; bpm.close();

The properties of the partitioned BATs are particularly useful during query optimization. However, it only works if the BAT identifier can be determined at compile time. For SQL it can be simply looked up in the catalog as part of a preparatory optimizer step.

To illustrate, the same problem handled by an optimizer that produces the plan based on a known number of partitions:

```
bpm.open(); R:bat[:oid,:int]:= bpm.take("myversion"); # get the partition
alias optimizer.mergetable(); T:= algebra.select(R,0,100);
```

is translated to the plan:

```
bpm.open(); R:bat[:oid,:int]:= bpm.take("myversion"); # get the
partition alias R0:bat[:oid,:int]:= bpm.take(R,0, nil:oid,nil:oid, 0,100);
R1:bat[:oid,:int]:= bpm.take(R,1, nil:oid,nil:oid, 101,200); R2:bat[:oid,:int]:=
bpm.take(R,2, nil:oid,nil:oid, 201,nil:int); R:= mat.new(R0,R1,R2); T:=
algebra.select(R,0,100); optimizer.multitable();
```

In this translation Ri also gets the properties of the BATs. It is now up to the mat optimizer to decide about further plan expansion or an iterator approach.

8.17.3 Partition updates

The content of the partitions is preferrable updated in bulk. This calls for accumulation of insertions/deletions in pending update BATs, as already performed in the SQL code generator. Once the transaction is commited, the updates are propagated (in parallel) to all partitions.

```
bpm.open(); Ralias:bat[:oid,:int] := bpm.take("myversion"); bpm.insert(Ralias,
Rinsert);# handle pending inserts bpm.delete(Ralias, Rdelete);# handle
pending deletes bpm.replace(Ralias, Rold, Rnew);# handle pending updates
bpm.close();
```

The replace operator works on the assumption that the head of Rold and Rnew is unique.

It remains possible to retrieve a partition and directly insert elements, but then it is up to the compiler to ensure that the boundery conditions are met.

8.17.4 Partitioned results

In many situations, you would like to keep the partial results as a partitioned BAT again. The easiest solution is to create a partitioned BAT, whose partitions are empty. Subsequently, we insert the temporary results. Depending on the fragmentation criteria, pieces may align with the pieces known, or lead to a redistribution of the buns to the correct bats.

In the previous plan for this becomes

```
bpm.open(); Tmp := bpm.deposit("tmp",:bat[:oid,:int]); bpm.rangePartition(tmp,nil:int,100);
bpm.rangePartition(tmp,101,nil:int);

Ralias:bat[:oid,:int]:= bpm.take("myversion"); # get the partition alias
R0:bat[:oid,:int]:= bpm.take("myversion", nil:oid,nil:oid, 0,100); T0:=
algebra.select(R0,0,100); bpm.insert(Tmp,T0);

R1:bat[:oid,:int]:= bpm.take("myversion", nil:oid,nil:oid, 101,200); T1:= alge-
bra.select(R1,0,100); bpm.insert(Tmp,T1);
```

R2:bat[:oid,:int]:= bpm.take("myversion", nil:oid,nil:oid, 201,nil:int); T2:= algebra.select(R2,0,100); bpm.insert(Tmp,T2);

Note that a symbolic optimizer can reduce this plan to a small snippet.

The rationale for the update approach is that re-distribution of temporary results are hidden behind the BPM.INSERT() interface. The only decision that should be taken by the optimizer is the fragmentation criteria for the temporary results.

For temporary results the range bounds need not be stored in the BPM catalog. Instead, the mat approach could be used to reduce the plan size.

bpm.open(); Ralias:= bpm.take("myversion",:bat[:oid,:int]); # get the partition alias R0:= bpm.take(Ralias, 0); T0:=algebra.select(R0,0,100);

R1:= bpm.take(Ralias, 1); T1:= algebra.select(R1,0,100);

R2:= bpm.take(Ralias, 2); R:= mat.new(T0,T1,T2); T2:=algebra.select(R2,0,100);

8.17.5 Partition iterators

The default strategy for an optimizer is to replace a reference to a partitioned BAT by an iterator.

l:= bpm.new(); barrier Elm:bat[:oid,:int]:= bpm.newIterator(Ralias); t:= algebra.select(Elm,0,20); bpm.addPartition(l,t); redo Elm:bat[:oid,:int]:= bpm.newIterator(Ralias); exit Elm;

Variations on this theme are iterators that search for partitions overlapping a range or those that are not empty.

8.17.6 Partition selection

Partition aware relational operators further reduce the performance overhead and at the same time avoid cluttering the MAL plans with too much flow of control constructs. A few operators relevant for the SQL environment will be added.

The select operation can be overloaded in the BPM to improve processing further. For example, the operation

t := bpm.select(Ralias,0,100);

extracts portions of all three partitions and creates a non-partitioned result BAT. If the partition bounds align with the selection criteria this operation becomes cheap. It can be used to convey information on the bounds to optimizers.

The lifetime of a partitioned table is inherited from its components. How to detect that a temporary BAT is removed from the BBP? Currently we have to explicitly call the bpm.garbage() on those partitioned BATs.

At the end of a query plan we have to garbage collect any of the left-over partitioned temporary tables.

8.18 Performance profiler

A key issue in developing fast programs using the Monet database back-end requires a keen eye on where performance is lost. Although performance tracking and measurements are highly application dependent, a simple to use tool makes life a lot easier.

Activation of the performance monitor has a global effect, i.e. all concurrent actions on the kernel are traced, but the events are only sent to the client initiated the profiler thread.

8.18.1 Monet Event Logger

The Monet Event Logger generates records of each event of interest indicated by a log filter, i.e. a pattern over module and function names.

The log record contents is derived from counters being (de-)activated. A complete list of recognized counters is shown below.

8.18.2 Execution tracing

Tracing is a special kind of profiling, where the information gathered is not sent to a remote system, but stored in the database itself. Each profile event is given a separate BAT

```
# thread and time since start
profiler.activate("tick");
# cpu time in nano-seconds
profiler.activate("cpu");
# memory allocation information
profiler.activate("memory");
# IO activity
profiler.activate("io");
# Module,function,program counter
profiler.activate("pc");
# actual MAL instruction executed
profiler.activate("statement");
```

The profiler event can be handled in several ways. The default strategy is to ship the event record immediately over a stream to a performance monitor. An alternative strategy is preparation of off-line performance analysis.

To reduce the interference of performance measurement with the experiments, the user can use an event cache, which is emptied explicitly upon need.

module profiler;

`command activate(name:str):void`
address CMDactivateProfiler comment "Make the specified counter active.";

`command deactivate(name:str):void`
address CMDdeactivateProfiler comment "Deactivate the counter";

`pattern openStream():void`
address CMDopenProfilerStream comment "Sent the events to output stream";

`pattern openStream(fnme:str):void`
address CMDsetProfilerFile comment "Send the log events to a file, stdout or console";

`pattern openStream(host:str, port:int):void`
address CMDsetProfilerStream comment "Send the log events to a stream ";

`command closeStream():void`
address CMDcloseProfilerStream comment "Stop sending the event records";

`pattern setAll():void`
address CMDsetAllProfiler comment "Short cut for setFilter(*,*).";

`pattern setNone():void`
address CMDsetNoneProfiler comment "Short cut for clrFilter(*,*).";

`pattern setFilter(mod:str,fcn:str):void`
address CMDsetFilterProfiler comment "Generate an event record for all function calls that satisfy the regular expression mod.fcn. A wildcard (*) can be used as name to identify all";

`pattern setFilter(v:any):void`
address CMDsetFilterVariable comment "Generate an event record for every instruction where v is used.";

`pattern clrFilter(mod:str,fcn:str):void`
address CMDclrFilterProfiler comment "Clear the performance trace bit of the selected functions.";

`pattern clrFilter(v:any):void`
address CMDsetFilterVariable comment "Stop tracing the variable" ;

`pattern setStartPoint(mod:str,fcn:str):void`
address CMDstartPointProfiler comment "Start performance tracing at mod.fcn";

`pattern setEndPoint(mod:str,fcn:str)`
address CMDendPointProfiler comment "End performance tracing after mod.fcn";

`pattern start():void`
address CMDstartProfiler comment "Start performance tracing";

`command noop():void`
address CMDnoopProfiler comment "Fetch any pending performance events";

`pattern stop():void`
address CMDstopProfiler comment "Stop performance tracing";

`command reset():void`
address CMDclearTrace comment "Clear the profiler traces";

`command dumpTrace():void`
address CMDdumpTrace comment "List the events collected";

`command getTrace(e:str):bat[:int,:any_1]`
address CMDgetTrace comment "Get the trace details of a specific event";

`pattern getEvent()(:lng,:lng,:lng)`
address CMDgetEvent comment "Retrieve the performance indicators of the previous instruction";

`command cleanup():void`
address CMDcleanup comment "Remove the temporary tables for profiling";

`command getDiskReads():lng`
address CMDgetDiskReads comment "Obtain the number of physical reads";

`command getDiskWrites():lng`
address CMDgetDiskWrites comment "Obtain the number of physical reads";

`command getUserTime():lng`
address CMDgetUserTime comment "Obtain the user timing information.";

`command getSystemTime():lng`
address CMDgetSystemTime comment "Obtain the user timing information.";

`pattern getFootprint():lng`
address CMDgetFootprint comment "Get the memory footprint and reset it";

`pattern getMemory():lng`
address CMDgetMemory comment "Get the amount of memory claimed and reset it";

8.19 PCRE library interface

The PCRE library is a set of functions that implement regular expression pattern matching using the same syntax and semantics as Perl, with just a few differences. The current implementation of PCRE (release 4.x) corresponds approximately with Perl 5.8, including support for UTF-8 encoded strings. However, this support has to be explicitly enabled; it is not the default.

8.20 Remote querying functionality

Communication with other mservers at the MAL level is a delicate task. However, it is indispensable for any distributed functionality. This module provides an abstract way to store and retrieve objects on a remote site. Additionally, functions on a remote site can be executed using objects available in the remote session context. This yields in four primitive functions that form the basis for distribution methods: get, put, register and exec.

The get method simply retrieves a copy of a remote object. Objects can be simple values, strings or BATs. The same holds for the put method, but the other way around. A local object can be stored on a remote site. Upon a successful store, the put method returns the remote identifier for the stored object. With this identifier the object can be addressed, e.g. using the get method to retrieve the object that was stored using put.

The get and put methods are symmetric. Performing a get on an identifier that was returned by put, results in an object with the same value and type as the one that was put. The result of such an operation is equivalent to making an (expensive) copy of the original object.

The register function takes a local MAL function and makes it known at a remote site. It ensures that it does not overload an already known operation remotely, which could create a semantic conflict. Deregister a function is forbidden, because it would allow for taking over the remote site completely. C-implemented functions, such as io.print() cannot be remotely stored. It would require even more complicated (byte?) code shipping and remote compilation to make it work. Currently, the remote procedure may only returns a single value.

The choice to let exec only execute functions was made to avoid problems to decide what should be returned to the caller. With a function it is clear and simple to return that what the function signature prescribes. Any side effect (e.g. io.print calls) may cause havoc in the system, but are currently ignored.

This leads to the final contract of this module. The methods should be used correctly, by obeying their contract. Failing to do so will result in errors and possibly undefined behaviour.

The resolve() function can be used to query Merovingian. It returns one or more databases discovered in its vincinity matching the given pattern.

module remote;

# module loading and unloading funcs

command `prelude():void`
: address RMTprelude comment "Initialise the remote module.";

command `epilogue():void`
: address RMTepilogue comment "Release the resources held by the remote module.";

# global connection resolve function

command `resolve(pattern:str):bat[:oid,:str]`
: address RMTresolve comment "resolve a pattern against Merovingian and return the URIs";

# session local connection instantiation functions

command `connect(uri:str, user:str, passwd:str):str`
: address RMTconnect comment "returns a newly created connection for uri, using user name and password";

command `connect(uri:str, user:str, passwd:str, scen:str):str`
: address RMTconnectScen comment "returns a newly created connection for uri, using user name, password and scenario";

command `disconnect(conn:str):void`
: address RMTdisconnect comment "disconnects the connection pointed to by handle (received from a call to connect()";

# core transfer functions

pattern `get(conn:str, ident:str):any`
: address RMTget comment "retrieves a copy of remote object ident";

pattern `put(conn:str, object:any):str`
: address RMTput comment "copies object to the remote site and returns its identifier";

pattern `register(conn:str, mod:str, fcn:str):void`
: address RMTregister comment "register <mod>.<fcn> at the remote site";

`pattern exec(conn:str, mod:str, func:str):str`
address RMTexec comment "remotely executes <mod>.<func> and returns the handle to its result";

`pattern exec(conn:str, mod:str, func:str)(:str, :str)`
address RMTexec comment "remotely executes <mod>.<func> and returns the handle to its result";

`pattern exec(conn:str, mod:str, func:str, :str...):str`
address RMTexec comment "remotely executes <mod>.<func> using the argument list of remote objects and returns the handle to its result";

`pattern exec(conn:str, mod:str, func:str, :str...)(:str, :str)`
address RMTexec comment "remotely executes <mod>.<func> using the argument list of remote objects and returns the handle to its result";

8.21 Statistics box.

Most optimizers need easy access to key information for proper plan generation. Amongst others, this volatile information consists of the tuple count, size, min- and max-value, the null-density, and a histogram of the value distribution.

The statistics are management by a Box, which gives a controlled environment to manage a collection of BATs and system variables.

BATs have to be deposit into the statistics box separately, because the costs attached maintaining them are high. The consistency of the statistics box is partly the responsibility of the upper layers. There is no automatic triggering when the BATs referenced are heavily modified or are being destroyed. They disappear from the statistics box the first time an invalid access is attempted or during system reboot.

The staleness of the information can be controlled in several ways. The easiest, and most expensive, is to assure that the statistics are updated when you start the server. Alternative, you can set a expiration interval, which will update the information only when it is considered expired. This test will be triggered either at server restart or your explicit call to update the statistics tables. The statistics table is commited each time you change it.

A forced update can be called upon when the front-end expects the situation to be changed drastically.

The statistics table is mostly used internally, but once in a while you need a dump for closed inspection. in your MAL program for inspection. Just use the BBP bind operation to locate them in the buffer pool.

module statistics;

`pattern open():void`
address STATopen comment "Locate and open the statistics box";

`pattern close():void`
address STATclose comment "Close the statistics box ";

`pattern destroy():void`
address STATdestroy comment "Destroy the statistics box";

pattern `take(name:any_1):any_2`
  address STATtake comment "Take a variable out of the statistics box";

pattern `deposit(name:str) :void`
  address STATdepositStr comment "Enter a new BAT into the statistics box";

pattern `deposit(name:bat[:any_1,:any_2]) :void`
  address STATdeposit comment "Enter a new BAT into the statistics box";

pattern `releaseAll():void`
  address STATreleaseAll comment "Release all variables in the box";

pattern `release(name:str) :void`
  address STATreleaseStr comment "Release a single BAT from the box";

pattern `release(name:bat[:any_1,:any_2]):void`
  address STATrelease comment "Release a single BAT from the box";

pattern `toString(name:any_1):str`
  address STATtoString comment "Get the string representation of an element in the box";

pattern `discard(name:str) :void`
  address STATdiscard comment "Release a BAT by name from the box";

pattern `discard(name:bat[:any_1,:any_2]) :void`
  address STATdiscard2 comment "Release a BAT variable from the box";

pattern `newIterator()(:lng,:str)`
  address STATnewIterator comment "Locate next element in the box";

pattern `hasMoreElements()(:lng,:str)`
  address STAThasMoreElements comment "Locate next element in the box";

command `update()`
  address STATupdate comment "Check for stale information";

command `forceUpdate()`
  address STATforceUpdateAll comment "Bring all information up to date";

command `forceUpdate(bnme:str)`
  address STATforceUpdate comment "Bring the statistics up to date for one BAT";

command `prelude() :void`
  address STATprelude comment "Initialize the statistics package";

command `epilogue() :void`
  address STATepilogue comment "Release the resources of the statistics package";

pattern `dump() :void`
  address STATdump comment "Display the statistics table";

command `getObjects():bat[:int,:str]`
  address STATgetObjects comment "Return a table with BAT names managed";

`pattern getHotset():bat[:int,:str]`
: address STATgetHotset comment "Return a table with BAT names that have been touched since the start of the session";

`pattern getCount(nme:str):lng`
: address STATgetCount comment "Return latest stored count information";

`pattern getSize(nme:str):lng`
: address STATgetSize comment "Return latest stored count information";

`pattern getMin(nme:str):lng`
: address STATgetMin comment "Return latest stored minimum information";

`pattern getMax(nme:str):lng`
: address STATgetMax comment "Return latest stored maximum information";

`pattern getHistogram(nme:str):bat[:any_1,:any_2]`
: address STATgetHistogram comment "Return the latest histogram");

8.22 The table interface

A database cannot live without ASCII tabular print/dump/load operations. It is needed to produce reasonable listings, to exchange answers with a client, and to keep a database version for backup. This is precisely where the tablet module comes in handy. [This module should replace all other table dump/load functions]

We start with a simple example to illustrate the plain ASCII representation and the features provided. Consider the relational table answer(name:str, age:int, sex:chr, address:str, dob:date) obtained by calling the routine tablet.page(B1,...,Bn) where the Bi represent BATS.

```
[ "John Doe",           25,     'M',    "Parklane 5",   "25-12-1978" ]
[ "Maril Streep",       23,     'F',    "Church 5",     "12-07-1980" ]
[ "Mr. Smith",          53,     'M',    "Church 1",     "03-01-1950" ]
```

The lines contain the representation of a list in Monet tuple format. This format has been chosen to ease parsing by any front-end. The scalar values are represented according to their type. For visual display, the columns are aligned by placing enough tabs between columns based on sampling the underlying bat to determine a maximal column width. (Note,actual commas are superfluous).

The arguments to the command can be any sequence of BATs, but which are assumed to be aligned. That is, they all should have the same number of tuples and the j-th tuple tail of Bi is printed along-side the j-th tuple tail of Bi+1.

Printing both columns of a single bat is handled by tablet as a print of two columns. This slight inconvenience is catch-ed by the io.print(b) command, which resolves most back-ward compatibility issues.

In many cases, this output would suffice for communication with a front-end. However, for visual inspection the user should be provided also some meta information derived from the database schema. Likewise, when reading a table this information is needed to prepare a first approximation of the schema namings. This information is produced by the command tablet.header(B1,...,Bn), which lists the column role name. If no role name is give, a default is generated based on the BAT name, e.g. B1_tail.

```
#-------------------------------------------------------#
# name,           age, sex, address,       dob          #
#-------------------------------------------------------#
[ "John Doe",      25, ’M’, "Parklane 5", "25-12-1978" ]
[ "Maril Streep",  23, ’F’, "Church 5",   "12-07-1980" ]
[ "Mr. Smith",     53, ’M’, "Church 1",   "03-01-1950" ]
```

The command tablet.display(B1,...,Bn) is a contraction of tablet.header(); tablet.page().

In many cases, the `tablet` produced may be too long to consume completely by the front end. In that case, the user needs page size control, much like the more/less utilities under Linux. However, no guarantee is given for arbitrarily going back and forth. [but works as long as we materialize results first]. A portion of the tablet can be printed by identifying the rows of interest as the first parameter(s) in the page command, e.g.

```
tablet.page(nil,10,B1,...,Bn);  #prints first 10 rows
tablet.page(10,20,B1,...,Bn);   #prints next 10 rows
tablet.page(100,nil,B1,...,Bn); #starts printing at tuple 100 until end
```

A paging system also provides the commands tablet.firstPage(), tablet.nextPage(), tablet.prevPage(), and tablet.lastPage() using a user controlled tablet size tablet.setPagesize(L).

The tablet display operations use a client (thread) specific formatting structure. This structure is initialized using either tablet.setFormat(B1,...,Bn) or tablet.setFormat(S1,...,Sn) (Bi is a BAT, Si a scalar). Subsequently, some additional properties can be set/modified, column width and brackets. After printing/paging the BAT resources should be freed using the command tablet.finish().

Any access outside the page-range leads to removal of the report structure. Subsequent access will generate an error. To illustrate, the following code fragment would be generated by the SQL compiler

```
        tablet.setFormat(B1,B2);
        tablet.setDelimiters("|","\t","|\n");
        tablet.setName(0, "Name");
        tablet.setNull(0, "?");
        tablet.setWidth(0, 15);
        tablet.setBracket(0, " ", ",");
        tablet.setName(1, "Age");
        tablet.setNull(1, "-");
        tablet.setDecimal(1, 9,2);
        tablet.SQLtitle("Query: select * from tables");
        tablet.page();
        tablet.SQLfooter(count(B1),cpuTicks);
```

This table is printed with tab separator(s) between elements and the bar (`|`) to mark begin and end of the string. The column parameters give a new title, a null replacement value, and the preferred column width. Each column value is optionally surrounded by brackets. Note, scale and precision can be applied to integer values only. A negative scale leads to a right adjusted value.

The title and footer operations are SQL specific routines to decorate the output.

Another example involves printing a two column table in XML format. [Alternative, tablet.XMLformat(B1,B2) is a shorthand for the following:]

```
        tablet.setFormat(B1,B2);
        tablet.setTableBracket("<rowset>","</rowset>");
        tablet.setRowBracket("<row>","</row>");
        tablet.setBracket(0, "<name>", "</name>");
        tablet.setBracket(1, "<age>", "</age>");
        tablet.page();
```

8.22.1 Tablet properties

More detailed header information can be obtained with the command tablet.setProperties(S), where S is a comma separated list of properties of interest, followed by the tablet.header(). The properties to choose from are: bat, name, type, width, sorted, dense, key, base, min, max, card,....

```
#---------------------------------------#
# B1,   B2,     B3,     B4,     B5      # BAT
# str,  int,    chr,    str,    date    # type
# true, false,  false,  false,  false   # sorted
# true, true,   false,  false,  false   # key
# ,     23,     'F',    ,               # min
# ,     53,     'M',    ,               # max
# 4,     4,     4,      4,      4       # count
# 4,i    3,     2,      2,      3       # card
# name, age,    sex,   address, dob     # name
#---------------------------------------#
```

8.22.2 Scalar tablets

In line with the 10-year experience of Monet, printing scalar values follow the tuple layout structure. This means that the header() command is also applicable. For example, the sequence "i:=0.2;v:=sin(i); tablet.display(i,v);" produces the answer:

```
#-----------------#
# i,    v         #
#-----------------#
[ 0.2,  0.198669 ]
#-----------------#
```

All other formatted printing should be done with the printf() operations contained in the module IO.

8.22.3 Tablet dump/restore

Dump and restore operations are abstractions over sequence of tablet commands. The command tablet.dump(stream,B1,...,Bn) is a contraction of the sequence tablet.setStream(stream); tablet.setProperties("name,type,dense,sorted,key,min,max"); tablet.header(B1,..,Bn); tablet.page(B1,..,Bn). The result can be read by tablet.load(stream,B1,..,Bn) command. If loading is successful, e.g. no parsing errors occurred, the tuples are appended to the corresponding BATs.

8.22.4 Front-end extension

A general bulk loading of foreign tables, e.g. CSV-files and fixed position records, is not provided. Instead, we extend the list upon need. Currently, the routines tablet.SQLload(stream,delim1,delim2, B1,..,Bn) reads the files using the Oracle(?) storage. The counterpart for dumping is tablet.SQLdump(stream,delim1,delim2);

8.22.5 The commands

The load operation is for bulk loading a table, each column will be loaded into its own bat. The arguments are void-aligned bats describing the input, ie the name of the column, the tuple separator and the type. The nr argument can be -1 (The input (datafile) is read until the end) or a maximum.

The dump operation is for dumping a set of bats, which are aligned. Again with void-aligned arguments, with name (currently not used), tuple separator (the last is the record separator) and bat to be dumped. With the nr argument the dump can be limited (-1 for unlimited).

The output operation is for ordered output. A bat (possibly form the collection) gives the order. For each element in the order bat the values in the bats are searched, if all are found they are output in the datafile, with the given separators.

The scripts from the tablet.mil file are all there too for backward compatibility with the old Mload format files.

The load_format loads the format file, since the old format file was in a table format it can be loaded with the load command.

The result from load_format can be used with load_data to load the data into a set of new bats.

These bats can be made persistent with the make_persistent script or merge with existing bats with the merge_data script.

The dump_format scripts dump a format file for a given set of to be dumped bats. These bats can be dumped with dump_data.

module tablet;

`command load( names:bat[:oid,:str], seps:bat[:oid,:str],`
types:bat[:oid,:str], datafile:str, nr:int) :bat[:str,:bat] address CMDtablet_load comment "Load a bat using specific format.";

`command input( names:bat[:oid,:str], seps:bat[:oid,:str],`
types:bat[:oid,:str], s:streams, nr:int) :bat[:str,:bat] address CMDtablet_input comment "Load a bat using specific format.";

`command dump(names:bat[:oid,:str], seps:bat[:oid,:str],`
bats:bat[:oid,:bat], datafile:str, nr:int) :void address CMDtablet_dump comment "Dump the bat in ASCII format";

`command output(order:bat[:any_1,:any_2], seps:bat[:oid,:str],`
bats:bat[:oid,:bat], s:streams) :void address CMDtablet_output comment "Send the bat to an output stream.";

`pattern display(v:any...):int`
address TABdisplayRow comment "Display a formatted row";

`pattern display(v:bat[:any_1,:any]...):int`
address TABdisplayTable comment "Display a formatted table";

`pattern page(b:bat[:any_1,:any]...):int`
address TABpage comment "Display all pages at once without header";

`pattern header(b:any...):int`
address TABheader comment "Display the minimal header for the table";

`pattern setProperties(prop:str):int`
address TABsetProperties comment "Define the set of properties";

`pattern dump(s:streams,b:bat[:any,:any]...):int`
address TABdump comment "Print all pages with header to a stream";

`pattern setFormat(b:any...):void`
address TABsetFormat comment "Initialize a new reporting structure.";

`pattern finish():void`
address TABfinishReport comment "Free the storage space of the report descriptor";

`pattern setStream(s:streams):void`
address TABsetStream comment "Redirect the output to a stream.";

`pattern setPivot(b:bat[:void,:oid]) :void`
address TABsetPivot comment "The pivot bat identifies the tuples of interest. The only requirement is that all keys mentioned in the pivot tail exist in all BAT parameters of the print comment. The pivot also provides control over the order in which the tuples are produced.";

`pattern setDelimiter(sep:str):void`
address TABsetDelimiter comment "Set the column separator.";

`pattern setTableBracket(lbrk:str,rbrk:str)`
address TABsetTableBracket comment "Format the brackets around a table";

`pattern setRowBracket(lbrk:str,rbrk:str)`
address TABsetRowBracket comment "Format the brackets around a row";

Set the column properties

`pattern setColumn(idx:int, v:any_1)`
address TABsetColumn comment "Bind i-th output column to a variable";

`pattern setName(idx:int, nme:str)`
address TABsetColumnName comment "Set the display name for a given column";

`pattern setBracket(idx:int,lbrk:str,rbrk:str)`
address TABsetColumnBracket comment "Format the brackets around a field";

`pattern setNull(idx:int, fmt:str)`
address TABsetColumnNull comment "Set the display format for a null value for a given column";

`pattern setWidth(idx:int, maxwidth:int)`
: address TABsetColumnWidth comment "Set the maximal display witdh for a given column. All values exceeding the length are simple shortened without any notice.";

`pattern setPosition(idx:int,f:int,i:int)`
: address TABsetColumnPosition comment "Set the character position to use for this field when loading according to fixed (punch-card) layout.";

`pattern setDecimal(idx:int,s:int,p:int)`
: address TABsetColumnDecimal comment "Set the scale and precision for numeric values";

`pattern setTryAll()`
: address TABsetTryAll comment "Skip error lines and assemble an error report";

`pattern setComplaints(b:bat[:oid,:str]) :void`
: address TABsetComplaints comment "The comlaints bat identifies all erroneous lines encountered ";

`command firstPage():void`
: address TABfirstPage comment "Produce the first page of output";

`command lastPage():void`
: address TABlastPage comment "Produce the last page of output";

`command nextPage():void`
: address TABnextPage comment "Produce the next page of output";

`command prevPage():void`
: address TABprevPage comment "Produce the prev page of output";

`command getPageCnt():void`
: address TABgetPageCnt comment "Return the size in number of pages";

`command getPage(i:int):void`
: address TABgetPage comment "Produce the i-th page of output";

8.22.6 Raw Load

Front-ends can bypass most of the overhead in loading the BATs by preparing the corresponding files directly and replace those created by e.g. the SQL frontend. This strategy is only advisable for cases where we have very large files >200GB and/or are created by a well debugged code.

To experiment with this approach, the code base responds on negative number of cores by dumping the data directly in BAT storage format into a collections of files on disk. It reports on the actions to be taken to replace BATs. This technique is initially only supported for fixed-sized columns.

8.23 Transaction management

In the philosophy of Monet, transaction management overhead should only be paid when necessary. Transaction management is for this purpose implemented as a module. This code base is largely absolute and should be re-considered when serious OLTP is being supported. Note, however, the SQL front-end obeys transaction semantics.

module transaction;

`command sync():bit`
address TRNglobal_sync comment "Save all persistent BATs";

`command commit():bit`
address TRNglobal_commit comment "Global commit on all BATs";

`command abort():bit`
address TRNglobal_abort comment "Global abort on all BATs";

`command subcommit(b:bat[:any_1,:str]):bit`
address TRNsubcommit comment "commit only a set of BATnames, passed in the tail (to which you must have exclusive access!)";

`pattern commit(c:any...)`
address TRNtrans_commit comment "Commit changes in certain BATs.";

`pattern abort(c:any...)`
address TRNtrans_abort comment "Abort changes in certain BATs.";

`pattern clean(c:any...)`
address TRNtrans_clean comment "Declare a BAT clean without flushing to disk.";

`command prev(b:bat[:any_1,:any_2]):bat[:any_1,:any_2]`
address TRNtrans_prev comment "The previous stae of this BAT";

`command alpha(b:bat[:any_1,:any_2]) :bat[:any_1,:any_2]`
address TRNtrans_alpha comment "List insertions since last commit.";

`command delta(b:bat[:any_1,:any_2]) :bat[:any_1,:any_2]`
address TRNtrans_delta comment "List deletions since last commit.";

8.24 The Inner Core

The innermost library of the MonetDB database system is formed by the library called GDK, an abbreviation of Goblin Database Kernel. Its development was originally rooted in the design of a pure active-object-oriented programming language, before development was shifted towards a re-usable database kernel engine.

GDK is a C library that provides ACID properties on a DSM model [*Copeland85*] , using main-memory database algorithms [*Garcia-Molina92*] built on virtual-memory OS primitives and multi-threaded parallelism. Its implementation has undergone various changes over its decade of development, many of which were driven by external needs to obtain a robust and fast database system.

The coding scheme explored in GDK has also laid a foundation to communicate over time experiences and to provide (hopefully) helpful advice near to the place where the code-reader needs it. Of course, over such a long time the documentation diverges from reality. Especially in areas where the environment of this package is being described. Consider such deviations as historic landmarks, e.g. crystallization of brave ideas and mistakes rectified at a later stage.

8.25 Short Outline

The facilities provided in this implementation are:

- GDK or Goblin Database Kernel routines for session management
- BAT routines that define the primitive operations on the database tables (BATs).
- BBP routines to manage the BAT Buffer Pool (BBP).
- ATOM routines to manipulate primitive types, define new types using an ADT interface.
- HEAP routines for manipulating heaps: linear spaces of memory that are GDK's vehicle of mass storage (on which BATs are built).
- DELTA routines to access inserted/deleted elements within a transaction.
- HASH routines for manipulating GDK's built-in linear-chained hash tables, for accelerating lookup searches on BATs.
- TM routines that provide basic transaction management primitives.
- TRG routines that provided active database support. [DEPRECATED]
- ALIGN routines that implement BAT alignment management.

The Binary Association Table (BAT) is the lowest level of storage considered in the Goblin runtime system [*Goblin*] . A BAT is a self-descriptive main-memory structure that represents the **binary relationship** between two atomic types. The association can be defined over:

`void:`	virtual-OIDs: a densely ascending column of OIDs (takes zero-storage).
`bit:`	Booleans, implemented as one byte values.
`chr:`	A single character (8 bits **integer**s). DEPRECATED for storing text (Unicode not supported).
`bte:`	Tiny (1-byte) integers (8-bit **integer**s).
`sht:`	Short integers (16-bit **integer**s).
`int:`	This is the C **int** type (32-bit).
`oid:`	Unique **long int** values uses as object identifier. Highest bit cleared always. Thus, oids-s are 31-bit numbers on 32-bit systems, and 63-bit numbers on 64-bit systems.
`wrd:`	Machine-word sized integers (32-bit on 32-bit systems, 64-bit on 64-bit systems).
`ptr:`	Memory pointer values. DEPRECATED. Can only be stored in transient BATs.
`flt:`	The IEEE **float** type.
`dbl:`	The IEEE **double** type.
`lng:`	Longs: the C **long long** type (64-bit integers).
`str:`	UTF-8 strings (Unicode). A zero-terminated byte sequence.
`bat:`	Bat descriptor. This allows for recursive adminstered tables, but severely complicates transaction management. Therefore, they CAN ONLY BE STORED IN TRANSIENT BATs.

This model can be used as a back-end model underlying other -higher level- models, in order to achieve **better performance** and **data independence** in one go. The relational model and the object-oriented model can be mapped on BATs by vertically splitting every table (or class) for each attribute. Each such a column is then stored in a BAT with type **bat**[**oid,attribute**], where the unique object identifiers link tuples in the different BATs. Relationship attributes in the object-oriented model hence are mapped to **bat**[**oid,oid**] tables, being equivalent to the concept of *join indexes* [*Valduriez87*] .

The set of built-in types can be extended with user-defined types through an ADT interface. They are linked with the kernel to obtain an enhanced library, or they are dynamically loaded upon request.

Types can be derived from other types. They represent something different than that from which they are derived, but their internal storage management is equal. This feature facilitates the work of extension programmers, by enabling reuse of implementation code, but is also used to keep the GDK code portable from 32-bits to 64-bits machines: the **oid** and **ptr** types are derived from **int** on 32-bits machines, but is derived from **lng** on 64 bits machines. This requires changes in only two lines of code each.

To accelerate lookup and search in BATs, GDK supports one built-in search accelerator: hash tables. We choose an implementation efficient for main-memory: bucket chained hash [*LehCar86,Analyti92*] . Alternatively, when the table is sorted, it will resort to merge-scan operations or binary lookups.

BATs are built on the concept of heaps, which are large pieces of main memory. They can also consist of virtual memory, in case the working set exceeds main-memory. In this case, GDK supports operations that cluster the heaps of a BAT, in order to improve performance of its main-memory.

8.25.1 Rationale

The rationale for choosing a BAT as the building block for both relational and object-oriented system is based on the following observations:

- - Given the fact that CPU speed and main-memory increase in current workstation hardware for the last years has been exceeding IO access speed increase, traditional disk-page oriented algorithms do no longer take best advantage of hardware, in most database operations.

  Instead of having a disk-block oriented kernel with a large memory cache, we choose to build a main-memory kernel, that only under large data volumes slowly degrades to IO-bound performance, comparable to traditional systems [*boncz95,boncz96*] .
- - Traditional (disk-based) relational systems move too much data around to save on (main-memory) join operations.

  The fully decomposed store (DSM [*Copeland85*)] assures that only those attributes of a relation that are needed, will have to be accessed.
- - The data management issues for a binary association is much easier to deal with than traditional *struct*-based approaches encountered in relational systems.
- - Object-oriented systems often maintain a double cache, one with the disk-based representation and a C pointer-based main-memory structure. This causes expensive conversions and replicated storage management. GDK does not do such ‘pointer swizzling’.

It used virtual-memory (**mmap()**) and buffer management advice (**madvise()**) OS primitives to cache only once. Tables take the same form in memory as on disk, making the use of this technique transparent [*oo7*] .

A RDBMS or OODBMS based on BATs strongly depends on our ability to efficiently support tuples and to handle small joins, respectively.

The remainder of this document describes the Goblin Database kernel implementation at greater detail. It is organized as follows:

GDK Interface:
It describes the global interface with which GDK sessions can be started and ended, and environment variables used.

Binary Association Tables:
As already mentioned, these are the primary data structure of GDK. This chapter describes the kernel operations for creation, destruction and basic manipulation of BATs and BUNs (i.e. tuples: Binary UNits).

BAT Buffer Pool:
All BATs are registered in the BAT Buffer Pool. This directory is used to guide swapping in and out of BATs. Here we find routines that guide this swapping process.

GDK Extensibility:
Atoms can be defined using a unified ADT interface. There is also an interface to extend the GDK library with dynamically linked object code.

GDK Utilities:
Memory allocation and error handling primitives are provided. Layers built on top of GDK should use them, for proper system monitoring. Thread management is also included here.

Transaction Management:
For the time being, we just provide BAT-grained concurrency and global transactions. Work is needed here.

BAT Alignment:
Due to the mapping of multi-ary datamodels onto the BAT model, we expect many correspondences among BATs, e.g. *bat(oid,attr1),.. bat(oid,attrN)* vertical decompositions. Frequent activities will be to jump from one attribute to the other ('bunhopping'). If the head columns are equal lists in two BATs, merge or even array lookups can be used instead of hash lookups. The alignment interface makes these relations explicitly manageable.

In GDK, complex data models are mapped with DSM on binary tables. Usually, one decomposes N-ary relations into N BATs with an **oid** in the head column, and the attribute in the tail column. There may well be groups of tables that have the same sets of **oid**s, equally ordered. The alignment interface is intended to make this explicit. Implementations can use this interface to detect this situation, and use cheaper algorithms (like merge-join, or even array lookup) instead.

BAT Iterators:

Iterators are C macros that generally encapsulate a complex for-loop. They would be the equivalent of cursors in the SQL model. The macro interface (instead of a function call interface) is chosen to achieve speed when iterating main-memory tables.

Common BAT Operations:

These are much used operations on BATs, such as aggregate functions and relational operators. They are implemented in terms of BAT- and BUN-manipulation GDK primitives.

8.26 Interface Files

In this section we summarize the user interface to the GDK library. It consist of a header file (gdk.h) and an object library (gdklib.a), which implements the required functionality. The header file must be included in any program that uses the library. The library must be linked with such a program.

8.26.1 Database Context

The MonetDB environment settings are collected in a configuration file. Amongst others it contains the location of the database directory. First, the database directory is closed for other servers running at the same time. Second, performance enhancements may take effect, such as locking the code into memory (if the OS permits) and preloading the data dictionary. An error at this stage normally lead to an abort.

8.26.2 GDK session handling

int	GDKinit (char *db, char *dbfarm, int allocmap)
int	GDKexit (int status)

The session is bracketed by *GDKinit* and *GDKexit*. Initialization involves setting up the administration for database access, such as memory allocation for the database buffer pool. During the exit phase any pending transaction is aborted and the database is freed for access by other users. A zero is returned upon encountering an erroneous situation.

8.27 Binary Association Tables

Having gone to the previous preliminary definitions, we will now introduce the structure of Binary Association Tables (BATs) in detail. They are the basic storage unit on which GDK is modelled.

The BAT holds an unlimited number of binary associations, called BUNs (**Binary UNits**). The two attributes of a BUN are called **head** (left) and **tail** (right) in the remainder of this document.

The above figure shows what a BAT looks like. It consists of two columns, called head and tail, such that we have always binary tuples (BUNs). The overlooking structure is the **BAT record**. It points to a heap structure called the **BUN heap**. This heap contains the atomic values inside the two columns. If they are fixed-sized atoms, these atoms reside directly in the BUN heap. If they are variable-sized atoms (such as string or polygon), however, the columns has an extra heap for storing those (such **variable-sized atom heaps** are then referred to as **Head Heap**s and **Tail Heap**s). The BUN heap then contains integer byte-offsets (fixed-sized, of course) into a head- or tail-heap.

The BUN heap contains a contiguous range of BUNs. It starts after the **first** pointer, and finishes at the end in the **free** area of the BUN. All BUNs after the **inserted** pointer have been added in the last transaction (and will be deleted on a transaction abort). All BUNs between the **deleted** pointer and the **first** have been deleted in this transaction (and will be reinserted at a transaction abort).

The location of a certain BUN in a BAT may change between successive library routine invocations. Therefore, one should avoid keeping references into the BAT storage area for long periods.

Passing values between the library routines and the enclosing C program is primarily through value pointers of type *ptr*. Pointers into the BAT storage area should only be used for retrieval. Direct updates of data stored in a BAT is forbidden. The user should adhere to the interface conventions to guarantee the integrity rules and to maintain the (hidden) auxiliary search structures.

8.27.1 GDK variant record type

When manipulating values, MonetDB puts them into value records. The built-in types have a direct entry in the union. Others should be represented as a pointer of memory in pval or as a string, which is basically the same. In such cases the *len* field indicates the size of this piece of memory.

8.27.2 The BAT record

The elements of the BAT structure are introduced in the remainder. Instead of using the underlying types hidden beneath it, one should use a *BAT* type that is supposed to look like this:

```
typedef struct {
          /* static BAT properties */
          bat    batCacheid;       /* bat id: index in BBPcache */
          int    batPersistence;   /* persistence mode */
          bit    batCopiedtodisk;  /* BAT is saved on disk? */
          bit    batSet;           /* all tuples in the BAT are unique? */
          /* dynamic BAT properties */
          int    batHeat;          /* heat of BAT in the BBP */
          sht    batDirty;         /* BAT modified after last commit? */
          bit    batDirtydesc;     /* BAT descriptor specific dirty flag */
          Heap*  batBuns;          /* Heap where the buns are stored */
          /* DELTA status */
          BUN    batDeleted;       /* first deleted BUN */
          BUN    batFirst;         /* empty BUN before the first alive BUN */
          BUN    batInserted;      /* first inserted BUN */
          BUN    batCount;         /* Tuple count */
          /* Head properties */
          int    htype;            /* Head type number */
          str    hident;           /* name for head column */
          bit    hkey;             /* head values should be unique? */
          bit    hsorted;          /* are head values currently ordered? */
          bit    hvarsized;        /* for speed: head type is varsized? */
          bit    hnonil;                   /* head has no nils */
          oid    halign;           /* alignment OID for head. */
          /* Head storage */
          int    hloc;             /* byte-offset in BUN for head elements */
          Heap   *hheap;           /* heap for varsized head values */
          Hash   *hhash;           /* linear chained hash table on head */
          /* Tail properties */
          int    ttype;            /* Tail type number */
          str    tident;           /* name for tail column */
          bit    tkey;             /* tail values should be unique? */
          bit    tnonil;           /* tail has no nils */
          bit    tsorted;          /* are tail values currently ordered? */
          bit    tvarsized;        /* for speed: tail type is varsized? */
          oid    talign;           /* alignment OID for head. */
          /* Tail storage */
          int    tloc;             /* byte-offset in BUN for tail elements */
          Heap   theap;            /* heap for varsized tail values */
          Hash   thash;            /* linear chained hash table on tail */
 } BAT;
```

The internal structure of the **BAT** record is in fact much more complex, but GDK programmers should refrain of making use of that.

The reason for this complex structure is to allow for a BAT to exist in two incarnations at the time: the *normal view* and the *reversed view*. Each bat *b* has a BATmirror(*b*) which has the negative **cacheid** of b in the BBP.

Since we don't want to pay cost to keep both views in line with each other under BAT updates, we work with shared pieces of memory between the two views. An update to one will thus automatically update the other. In the same line, we allow **synchronized BATs** (BATs with identical head columns, and marked as such in the **BAT Alignment** interface) now to be clustered horizontally.

8.27.3 Heap Management

Heaps are the low-level entities of mass storage in BATs. Currently, they can either be stored on disk, loaded into memory, or memory mapped.

int	HEAPalloc (Heap *h, size_t nitems, size_t itemsize);
int	HEAPfree (Heap *h);
int	HEAPextend (Heap *h, size_t size);
int	HEAPload (Heap *h, str nme,ext, int trunc);
int	HEAPsave (Heap *h, str nme,ext);
int	HEAPcopy (Heap *dst,*src);
int	HEAPdelete (Heap *dst, str o, str ext);
int	HEAPwarm (Heap *h);

These routines should be used to alloc free or extend heaps; they isolate you from the different ways heaps can be accessed.

8.27.4 Internal HEAP Chunk Management

Heaps are used in BATs to store data for variable-size atoms. The implementor must manage malloc()/free() functionality for atoms in this heap. A standard implementation is provided here.

`void`	HEAP_initialize (Heap* h, size_t nbytes, size_t nprivate, int align)
`void`	HEAP_destroy (Heap* h)
`var_t`	HEAP_malloc (Heap* heap, size_t nbytes)
`void`	HEAP_free (Heap *heap, var_t block)
`int`	HEAP_private (Heap* h)
`void`	HEAP_printstatus (Heap* h)
`void`	HEAP_check (Heap* h)

The heap space starts with a private space that is left untouched by the normal chunk allocation. You can use this private space e.g. to store the root of an rtree *HEAP_malloc* allocates a chunk of memory on the heap, and returns an index to it. *HEAP_free* frees a previously allocated chunk *HEAP_private* returns an integer index to private space.

8.27.5 BAT construction

`BAT*`	BATnew (int headtype, int tailtype, BUN cap)
`BAT*`	BATextend (BAT *b, BUN newcap)

A temporary BAT is instantiated using *BATnew* with the type aliases of the required binary association. The aliases include the built-in types, such as *TYPE_int....TYPE_ptr*, and the atomic types introduced by the user. The initial capacity to be accommodated within a BAT is indicated by *cap*. Their extend is automatically incremented upon storage overflow. Failure to create the BAT results in a NULL pointer.

The routine *BATclone* creates an empty BAT storage area with the properties inherited from its argument.

8.27.6 BUN manipulation

BAT*	BATins (BAT *b, BAT *c, bit force)
BAT*	BATappend (BAT *b, BAT *c, bit force)
BAT*	BATdel (BAT *b, BAT *c, bit force)
BAT*	BUNins (BAT *b, ptr left, ptr right, bit force)
BAT*	BUNappend (BAT *b, ptr right, bit force)
BAT*	BUNreplace (BAT *b, ptr left, ptr right, bit force)
int	BUNdel (BAT *b, ptr left, ptr right, bit force)
int	BUNdelHead (BAT *b, ptr left, bit force)
BUN	BUNfnd (BAT *b, ptr head)
void	BUNfndOID (BUN result, BATiter bi, oid *head)
void	BUNfndSTD (BUN result, BATiter bi, ptr head)

BUN	BUNlocate (BAT *b, ptr head, ptr tail)
ptr	BUNhead (BAT *b, BUN p)
ptr	BUNtail (BAT *b, BUN p)

The BATs contain a number of fixed-sized slots to store the binary associations. These slots are called BUNs or BAT units. A BUN variable is a pointer into the storage area of the BAT, but it has limited validity. After a BAT modification, previously obtained BUNs may no longer reside at the same location.

The association list does not contain holes. This density permits users to quickly access successive elements without the need to test the items for validity. Moreover, it simplifies transport to disk and other systems. The negative effect is that the user should be aware of the evolving nature of the sequence, which may require copying the BAT first.

The update operations come in three flavors. Element-wise updates can use *BUNins*, *BUNappend*, *BUNreplace*, *BUNdel*, and *BUNdelHead*. The batch update operations are *BATins*, *BATappend* and *BATdel*.

Only experts interested in speed may use *BUNfastins*, since it skips most consistency checks, does not update search accelerators, and does not maintain properties such as the *hsorted* and *tsorted* flags. Beware!

The routine *BUNfnd* provides fast access to a single BUN providing a value for the head of the binary association. A very fast shortcut for *BUNfnd* if the selection type is known to be integer or OID, is provided in the form of the macro *BUNfndOID*.

To select on a tail, one should use the reverse view obtained by *BATmirror*.

The routines *BUNhead* and *BUNtail* return a pointer to the first and second value in an association, respectively. To guard against side effects on the BAT, one should normally copy this value into a scratch variable for further processing.

Behind the interface we use several macros to access the BUN fixed part and the variable part. The BUN operators always require a BAT pointer and BUN identifier.

- BAThtype(b) and BATttype(b) find out the head and tail type of a BAT.
- BUNfirst(b) returns a BUN pointer to the first BUN as a BAT.
- BUNlast(b) returns the BUN pointer directly after the last BUN in the BAT.
- BUNhead(b, p) and BUNtail(b, p) return pointers to the head-value and tail-value in a given BUN.
- BUNhloc(b, p) and BUNtloc(b, p) do the same thing, but knowing in advance that the head-atom resp. tail-atom of a BAT is fixed size.
- BUNhvar(b, p) and BUNtvar(b, p) do the same thing, but knowing in advance that the head-atom resp. tail-atom of a BAT is variable sized.

8.27.7 BAT properties

BUN	BATcount (BAT *b)
void	BATsetcapacity (BAT *b, BUN cnt)
void	BATsetcount (BAT *b, BUN cnt)
BUN	BATbuncount (BAT *b)
str	BATrename (BAT *b, str nme)
BAT *	BATkey (BAT *b, int onoff)
BAT *	BATset (BAT *b, int onoff)

BAT *	BATmode (BAT *b, int mode)
BAT *	BATsetaccess (BAT *b, int mode)
int	BATdirty (BAT *b)
int	BATgetaccess (BAT *b)

The function *BATcount* returns the number of associations stored in the BAT.

The function *BATbuncount* returns the space that is occupied in associations in the BAT. This is not the same as *BATcount*, since the first N associations may be unused or delta data.

The BAT is given a new logical name using *BATrename*.

The integrity properties to be maintained for the BAT are controlled separately. A key property indicates that duplicates in the association dimension are not permitted. The BAT is turned into a set of associations using *BATset*. Key and set properties are orthogonal integrity constraints. The strongest reduction is obtained by making the BAT a set with key restrictions on both dimensions.

The persistency indicator tells the retention period of BATs. The system support three modes: PERSISTENT, TRANSIENT, and SESSION. The PERSISTENT BATs are automatically saved upon session boundary or transaction commit. TRANSIENT BATs are removed upon transaction boundary. SESSION BATs are removed at the end of a session. They are normally used to maintain temporary results. All BATs are initially TRANSIENT unless their mode is changed using the routine *BATmode*.

The BAT properties may be changed at any time using *BATkey*, *BATset*, and *BATmode*.

Valid BAT access properties can be set with *BATsetaccess* and *BATgetaccess*: BAT_READ, BAT_APPEND, and BAT_WRITE. BATs can be designated to be read-only. In this case some memory optimizations may be made (slice and fragment bats can point to stable subsets of a parent bat). A special mode is append-only. It is then allowed to insert BUNs at the end of the BAT, but not to modify anything that already was in there.

8.27.8 BAT manipulation

BAT *	BATclear (BAT *b)
BAT *	BATcopy (BAT *b, int ht, int tt, int writeable)
BAT *	BATmark (BAT *b, oid base)
BAT *	BATmark_grp (BAT *b, BAT *g, oid *s)
BAT *	BATnumber (BAT *b)
BAT *	BATmirror (BAT *b)
BAT *	BATreset (BAT *b)

The routine *BATclear* removes the binary associations, leading to an empty, but (re-)initialized BAT. Its properties are retained. A temporary copy is obtained with *BATcopy*. The new BAT has an unique name. The routine *BATmark* creates a binary association that introduces a new tail column of fresh densely ascending OIDs. The base OID can be given explicitly, or if oid_nil is passed, is chosen as a new unique range by the system. A similar routine is *BATnumber*, which copies the heads and assigns an integer index to the tail. It plays a crucial role in administration of query results.

The routine *BATmirror* returns the mirror image BAT (where tail is head and head is tail) of that same BAT. This does not involve a state change in the BAT (as previously): both views on the BAT exist at the same time.

8.27.9 BAT Input/Output

BAT *	BATload (str name)
BAT *	BATsave (BAT *b)
int	BATmmap (BAT *b, int hb, int tb, int hh, int th)
int	BATmadvise (BAT *b, int hb, int tb, int hh, int th)
int	BATmmap_pin (BAT *b)
int	BATmmap_unpin (BAT *b)
int	BATdelete (BAT *b)

A BAT created by *BATnew* is considered temporary until one calls the routine *BATsave* or *BATmode*. This routine reserves disk space and checks for name clashes in the BAT directory. It also makes the BAT persistent. The empty BAT is initially marked as ordered on both columns. Failure to read or write the BAT results in a NULL, otherwise it returns the BAT pointer.

MonetDB now has a mmap trim thread that takes care of flushing the memory mapped regions when MonetDB starts to consume too much main memory. Heaps (that are randomly accessed) can be excluded from this mechanism, by pinning them. BATmmap_pin/unpin do this for all heaps of a BAT.

8.27.10 Heap Storage Modes

The discriminative storage modes are memory-mapped, compressed, or loaded in memory. The **BATmmap()** changes the storage mode of each heap associated to a BAT. As can be seen in the bat record, each BAT has one BUN-heap (*bn*), and possibly two heaps (*hh* and *th*) for variable-sized atoms. The *BATmadvise* call works in the same way. Using the *madvise()* system call it issues buffer management advise to the OS kernel, as for the expected usage pattern of the memory in a heap.

8.27.11 Printing

int	BATprintf (stream *f, BAT *b)
int	BATmultiprintf (stream *f, int argc, BAT *b[], int printoid, int order, int printorderby)

The functions to convert BATs into ASCII and the reverse use internally defined formats. They are primarily meant for ease of debugging and to a lesser extent for output processing. Printing a BAT is done essentially by looping through its components, printing each association. If an index is available, it will be used. The *BATmultiprintf* command assumes a set of BATs with corresponding oid-s in the head columns. It performs the multijoin over them, and prints the multi-column result on the file.

8.27.12 BAT clustering

BAT *	BATsort (BAT *b)
BAT *	BATsort_rev (BAT *b)
BAT *	BATorder (BAT *b)
BAT *	BATorder_rev (BAT *b)
BAT *	BATrevert (BAT *b)
int	BATordered (BAT *b)

When working in a main-memory situation, clustering of data on disk-pages is not important. Whenever mmap()-ed data is used intensively, reducing the number of page faults is a hot issue.

The above functions rearrange data in MonetDB heaps (used for storing BUNs var-sized atoms, or accelerators). Applying these clusterings will allow that MonetDB's main-memory oriented algorithms work efficiently also in a disk-oriented context.

The *BATsort* functions return a copy of the input BAT, sorted in ascending order on the head column. *BATordered* starts a check on the head values to see if they are ordered. The result is returned and stored in the *hsorted* field of the BAT. *BATorder* is similar to *BATsort*, but sorts the BAT itself, rather than returning a copy (BEWARE: this operation destroys the delta information. TODO:fix). The *BATrevert* puts all the live BUNs of a BAT in reverse order. It just reverses the sequence, so this does not necessarily mean that they are sorted in reverse order!

8.28 BAT Buffer Pool

int	BBPfix (bat bi)
int	BBPunfix (bat bi)
int	BBPincref (bat bi, int logical)
int	BBPdecref (bat bi, int logical)
void	BBPhot (bat bi)
void	BBPcold (bat bi)
str	BBPname (bat bi)
bat	BBPindex (str nme)
BAT*	BATdescriptor (bat bi)
bat	BBPcacheid (BAT *b)

The BAT Buffer Pool module contains the code to manage the storage location of BATs. It uses two tables *BBPlogical* and *BBphysical* to relate the BAT name with its corresponding file system name. This information is retained in an ASCII file within the database home directory for ease of inspection. It is loaded upon restart of the server and saved upon transaction commit (if necessary).

The remaining BBP tables contain status information to load, swap and migrate the BATs. The core table is *BBPcache* which contains a pointer to the BAT descriptor with its heaps. A zero entry means that the file resides on disk. Otherwise it has been read or mapped into memory.

BATs loaded into memory are retained in a BAT buffer pool. They retain their position within the cache during their life cycle, which make indexing BATs a stable operation. Their descriptor can be obtained using *BBPcacheid*.

The *BBPindex* routine checks if a BAT with a certain name is registered in the buffer pools. If so, it returns its BAT id. The *BATdescriptor* routine has a BAT id parameter, and returns a pointer to the corresponding BAT record (after incrementing the reference count). The BAT will be loaded into memory, if necessary.

8.29 GDK Extensibility

GDK can be extended with new atoms, search accelerators and storage modes.

8.29.1 Atomic Type Descriptors

The atomic types over which the binary associations are maintained are described by an atom descriptor.

```
void     ATOMproperty (str nme, char *property, int (*fcn)(), int val);
int      ATOMindex (char *nme);
int      ATOMdump ();
void     ATOMdelete (int id);
str      ATOMname (int id);
int      ATOMsize (int id);
int      ATOMalign (int id);
int      ATOMvarsized (int id);
ptr      ATOMnilptr (int id);
int      ATOMfromstr (int id, str s, int* len, ptr* v_dst);
int      ATOMtostr (int id, str s, int* len, ptr* v_dst);
hash_t   ATOMhash (int id, ptr val, in mask);
int      ATOMcmp (int id, ptr val_1, ptr val_2);
int      ATOMconvert (int id, ptr v, int direction);
int      ATOMfix (int id, ptr v);
int      ATOMunfix (int id, ptr v);
int      ATOMheap (int id, Heap *hp, size_t cap);
void     ATOMheapconvert (int id, Heap *hp, int direction);
int      ATOMheapcheck (int id, Heap *hp, HeapRepair *hr);
int      ATOMput (int id, Heap *hp, BUN pos_dst, ptr val_src);
int      ATOMdel (int id, Heap *hp, BUN v_src);
int      ATOMlen (int id, ptr val);
ptr      ATOMnil (int id);
int      ATOMformat (int id, ptr val, char** buf);
int      ATOMprint (int id, ptr val, stream *fd);
ptr      ATOMdup (int id, ptr val );
```

8.29.2 Atom Definition

User defined atomic types can be added to a running system with the following interface:.

- *ATOMproperty()* registers a new atom definition, if there is no atom registered yet under that name. It then installs the attribute of the named property. Valid names are "size", "align", "null", "fromstr", "tostr", "cmp", "hash", "put", "get", "del", "length" and "heap".
- *ATOMdelete()* unregisters an atom definition.
- *ATOMindex()* looks up the atom descriptor with a certain name.

8.29.3 Atom Manipulation

- The *ATOMname()* operation retrieves the name of an atom using its id.
- The *ATOMsize()* operation returns the atoms fixed size.
- The *ATOMalign()* operation returns the atoms minimum alignment. If the alignment info was not specified explicitly during atom install, it assumes the maximum value of {1,2,4,8} smaller than the atom size.

- The *ATOMnilptr()* operation returns a pointer to the nil-value of an atom. We usually take one dedicated value halfway down the negative extreme of the atom range (if such a concept fits), as the nil value.
- The *ATOMnil()* operation returns a copy of the nil value, allocated with GDKmalloc().
- The *ATOMheap()* operation creates a new var-sized atom heap in 'hp' with capacity 'cap'.
- The *ATOMhash()* computes a hash index for a value. 'val' is a direct pointer to the atom value. Its return value should be an hash_t between 0 and 'mask'.
- The *ATOMcmp()* operation computes two atomic values. Its parameters are pointers to atomic values.
- The *ATOMlen()* operation computes the byte length for a value. 'val' is a direct pointer to the atom value. Its return value should be an integer between 0 and 'mask'.
- The *ATOMdel()* operation deletes a var-sized atom from its heap 'hp'. The integer byte-index of this value in the heap is pointed to by 'val_src'.
- The *ATOMput()* operation inserts an atom 'src_val' in a BUN at 'dst_pos'. This involves copying the fixed sized part in the BUN. In case of a var-sized atom, this fixed sized part is an integer byte-index into a heap of var-sized atoms. The atom is then also copied into that heap 'hp'.
- The *ATOMfix()* and *ATOMunfix()* operations do bookkeeping on the number of references that a GDK application maintains to the atom. In MonetDB, we use this to count the number of references directly, or through BATs that have columns of these atoms. The only operator for which this is currently relevant is BAT. The operators return the POST reference count to the atom. BATs with fixable atoms may not be stored persistently.
- The *ATOMfromstr()* parses an atom value from string 's'. The memory allocation policy is the same as in *ATOMget()*. The return value is the number of parsed characters.
- The *ATOMprint()* prints an ASCII description of the atom value pointed to by 'val' on file descriptor 'fd'. The return value is the number of parsed characters.
- The *ATOMformat()* is similar to *ATOMprint()*. It prints an atom on a newly allocated string. It must later be freed with **GDKfree**. The number of characters written is returned. This is minimally the size of the allocated buffer.
- The *ATOMdup()* makes a copy of the given atom. The storage needed for this is allocated and should be removed by the user.

These wrapper functions correspond closely to the interface functions one has to provide for a user-defined atom. They basically (with exception of *ATOMput()*, *ATOMprint()* and *ATOMformat()*) just have the atom id parameter prepended to them.

8.29.4 Unique OIDs

oid	OIDseed (oid seed);
oid	OIDnew (oid inc);

OIDs are special kinds of unsigned integers because the system guarantees uniqueness. For system simplicity and performance, OIDs are now represented as (signed) integers; however this is hidden in the system internals and shouldn't affect semantics.

The *OIDnew(N)* claims a range of N contiguous unique, unused OIDs, and returns the starting value of this range. The highest OIDBITS designate site. [DEPRECATED]

8.29.5 Built-in Accelerator Functions

BAT*	BAThash (BAT *b, BUN masksize)
BAT *	BAThashsplit (BAT *b, BUN n, int unary)
BAT *	BATrangesplit (BAT *b, int n)

The current BAT implementation supports one search accelerator: hashing. The routine *BAThash* makes sure that a hash accelerator on the head of the BAT exists. A zero is returned upon failure to create the supportive structures.

The hash data structures are currently maintained during update operations.

A *BAT* can be redistributed over n buckets using a hash function with *BAThashsplit*. The return value is a list of BAT pointers. Similarly, a range partitioning based is supported.

8.29.6 Multilevel Storage Modes

We should bring in the compressed mode as the first, maybe built-in, mode. We could than add for instance HTTP remote storage, SQL storage, and READONLY (cd-rom) storage.

8.30 GDK Utilities

Interfaces for memory management, error handling, thread management and system information.

8.30.1 GDK memory management

void*	GDKmalloc (size_t size)
void*	GDKzalloc (size_t size)
void*	GDKmallocmax (size_t size, size_t *maxsize, int emergency)
void*	GDKrealloc (void* pold, size_t size)
void*	GDKreallocmax (void* pold, size_t size, size_t *maxsize, int emergency)
void	GDKfree (void* blk)
str	GDKstrdup (str s)
void*	GDKvmalloc (size_t size, size_t *maxsize, int emergency)
void*	GDKvmrealloc (void* pold, size_t oldsize, size_t newsize, size_t oldmax, size_t *maxsize, int emergency)
void	GDKvmfree (void* blk, size_t size, size_t maxsize)

These utilities are primarily used to maintain control over critical interfaces to the C library. Moreover, the statistic routines help in identifying performance and bottlenecks in the current implementation.

Compiled with -DMEMLEAKS the GDK memory management log their activities, and are checked on inconsistent frees and memory leaks.

8.30.2 GDK error handling

str	GDKmessage
bit	GDKsilent

int	GDKfatal(str msg)
int	GDKwarning(str msg)
int	GDKerror (str msg)
int	GDKgoterrors ()
int	GDKsyserror (str msg)
str	GDKerrbuf
	GDKsetbuf (str buf)

The error handling mechanism is not sophisticated yet. Experience should show if this mechanism is sufficient. Most routines return a pointer with zero to indicate an error.

The error messages are also copied to standard output unless *GDKsilent* is set to a non-zero value. The last error message is kept around in a global variable.

Error messages can also be collected in a user-provided buffer, instead of being echoed to a stream. This is a thread-specific issue; you want to decide on the error mechanism on a thread-specific basis. This effect is established with *GDKsetbuf*. The memory (de)allocation of this buffer, that must at least be 1024 chars long, is entirely by the user. A pointer to this buffer is kept in the pseudo-variable *GDKerrbuf*. Normally, this is a NULL pointer.

The GDKembedded variable is a property set in the configuration file to indicate that the kernel is only allowed to run as a single process. This can be used to remove all locking overhead. The actual state of affairs is maintained in GDKprotected, which is set when locking is required, e.g. when multiple threads become active.

The kernel maintains a central table of all active threads. They are indexed by their tid. The structure contains information on the input/output file descriptors, which should be set before a database operation is started. It ensures that output is delivered to the proper client. The Thread structure should be ideally made directly accessible to each thread. This speeds up access to tid and file descriptors.

8.31 Transaction Management

int	TMcommit ()
int	TMabort ()
int	TMsubcommit ()

MonetDB by default offers a global transaction environment. The global transaction involves all activities on all persistent BATs by all threads. Each global transaction ends with either *TMabort* or *TMcommit*, and immediately starts a new transaction. *TMcommit* implements atomic commit to disk on the collection of all persistent BATs. For all persistent BATs, the global commit also flushes the delta status for these BATs (see *BATcommit/BATabort*). This allows to perform *TMabort* quickly in memory (without re-reading all disk images from disk). The collection of which BATs is persistent is also part of the global transaction state. All BATs that where persistent at the last commit, but were made transient since then, are made persistent again by *TMabort*. In other words, BATs that are deleted, are only physically deleted at *TMcommit* time. Until that time, rollback (*TMabort*) is possible.

Use of *TMabort* is currently NOT RECOMMENDED due to two bugs:

- *TMabort after a failed* %TMcommit does not bring us back to the previous committed state; but to the state at the failed *TMcommit*.

- At runtime, *TMabort* does not undo BAT name changes, whereas a cold MonetDB restart does.

In effect, the problems with *TMabort* reduce the functionality of the global transaction mechanism to consistent checkpointing at each *TMcommit*. For many applications, consistent checkpointingis enough.

Extension modules exist that provide fine grained locking (lock module) and Write Ahead Logging (sqlserver). Applications that need more fine-grained transactions, should build this on top of these extension primitives.

TMsubcommit is intended to quickly add or remove BATs from the persistent set. In both cases, rollback is not necessary, such that the commit protocol can be accelerated. It comes down to writing a new BBP.dir.

Its parameter is a BAT-of-BATs (in the tail); the persistence status of that BAT is committed. We assume here that the calling thread has exclusive access to these bats. An error is reported if you try to partially commit an already committed persistent BAT (it needs the rollback mechanism).

8.31.1 Delta Management

BAT *	BATcommit (BAT *b)
BAT *	BATfakeCommit (BAT *b)
BAT *	BATundo (BAT *b)
BAT *	BATprev (BAT *b)
BAT *	BATalpha (BAT *b)
BAT *	BATdelta (BAT *b)

The BAT keeps track of updates with respect to a 'previous state'. Do not confuse 'previous state' with 'stable' or 'commited-on-disk', because these concepts are not always the same. In particular, they diverge when BATcommit, BATfakecommit, and BATundo are called explictly, bypassing the normal global *TMcommit* protocol (some applications need that flexibility).

BATcommit make the current BAT state the new 'stable state'. This happens inside the global *TMcommit* on all persistent BATs previous to writing all bats to persistent storage using a BBPsync[?].

EXPERT USE ONLY: The routine *BATfakeCommit* updates the delta information on BATs and clears the dirty bit. This avoids any copying to disk. Expert usage only, as it bypasses the global commit protocol, and changes may be lost after quitting or crashing MonetDB.

BATabort undo-s all changes since the previous state. The global TMabort[?] achieves a rollback to the previously committed state by doing BATabort on all persistent bats.

BUG: after a failed *TMcommit*, *TMabort* does not do anything because *TMcommit* does the *BATcommit*s *before* attempting to sync to disk instead of AFTER doing this.

The previous state can also be queried. *BATprev* is a view on the current BAT as it was in the previous state. *BATalpha* shows only the BUNs inserted since the previous state, and *BATdelta* the deleted buns.

CAVEAT: *BATprev*, *BATalpha* and *BATdelta* only return views if the underlying BATs are read-only (often not the case when BATs are being updated). Otherwise, copies must be made anyway.

8.32 BAT Alignment and BAT views

int	ALIGNsynced (BAT* b1, BAT* b2)
int	ALIGNsync (BAT *b1, BAT *b2)
int	ALIGNrelated (BAT *b1, BAT *b2)
int	ALIGNsetH ((BAT *dst, BAT *src)
BAT *	BATpropcheck (BAT *b, int mode)
BAT*	VIEWcreate (BAT *h, BAT *t)
int	isVIEW (BAT *b)
bat	VIEWhparent (BAT *b)
bat	VIEWtparent (BAT *b)
BAT*	VIEWhead (BAT *b)
BAT*	VIEWcombine (BAT *b)
BAT*	VIEWreset (BAT *b)
BAT*	BATmaterialize (BAT *b)

Alignments of two columns of a BAT means that the system knows whether these two columns are exactly equal. Relatedness of two BATs means that one pair of columns (either head or tail) of both BATs is aligned. The first property is checked by *ALIGNsynced*, the latter by *ALIGNrelated*.

The *BATpropcheck* examines a BAT and tries to set all applicable properties (key,sorted,align,dense).

All algebraic BAT commands propagate the properties - including alignment properly on their results.

VIEW BATs are BATs that lend their storage from a parent BAT. They are just a descriptor that points to the data in this parent BAT. A view is created with *VIEWcreate*. The cache id of the parent (if any) is returned by *VIEWhparent* and *VIEWtparent* (otherwise it returns 0).

VIEW bats are read-only!!

The *VIEWcombine* gives a view on a BAT that has two head columns of the parent. The *VIEWhead* constructs a BAT view that has the same head column as the parent, but has a void column with seqbase=nil in the tail. *VIEWreset* creates a normal BAT with the same contents as its view parameter (it converts void columns with seqbase!=nil to materialized oid columns).

The *BATmaterialize* materializes a VIEW (TODO) or void bat inplace. This is useful as materialization is usually needed for updates.

8.33 BAT Iterators

BATloop	(BAT *b; BUN p, BUN q)
BATloopDEL	(BAT *b; BUN p; BUN q; int dummy)
DELloop	(BAT *b; BUN p, BUN q, int dummy)
HASHloop	(BAT *b; Hash *h, size_t dummy; ptr value)
HASHloop_bit	(BAT *b; Hash *h, size_t idx; bit *value, BUN w)
HASHloop_chr	(BAT *b; Hash *h, size_t idx; char *value, BUN w)

HASHloop_bte	(BAT *b; Hash *h, size_t idx; bte *value, BUN w)
HASHloop_sht	(BAT *b; Hash *h, size_t idx; sht *value, BUN w)
HASHloop_bat	(BAT *b; Hash *h, size_t idx; bat *value, BUN w)
HASHloop_ptr	(BAT *b; Hash *h, size_t idx; ptr *value, BUN w)
HASHloop_int	(BAT *b; Hash *h, size_t idx; int *value, BUN w)
HASHloop_oid	(BAT *b; Hash *h, size_t idx; oid *value, BUN w)
HASHloop_wrd	(BAT *b; Hash *h, size_t idx; wrd *value, BUN w)
HASHloop_flt	(BAT *b; Hash *h, size_t idx; flt *value, BUN w)
HASHloop_lng	(BAT *b; Hash *h, size_t idx; lng *value, BUN w)
HASHloop_dbl	(BAT *b; Hash *h, size_t idx; dbl *value, BUN w)
HASHloop_str	(BAT *b; Hash *h, size_t idx; str value, BUN w)
HASHlooploc	(BAT *b; Hash *h, size_t idx; ptr value, BUN w)
HASHloopvar	(BAT *b; Hash *h, size_t idx; ptr value, BUN w)
SORTloop	(BAT *b,p,q,tl,th,s)

The *BATloop()* looks like a function call, but is actually a macro. The following example gives an indication of how they are to be used:

```
void
print_a_bat(BAT *b)
{
        BATiter bi = bat_iterator(b);
        BUN p, q;

        BATloop(b, p, q)
                printf("Element %3d has value %d\n",
                           *(int*) BUNhead(bi, p), *(int*) BUNtail(bi, p));
}
```

8.33.1 simple sequential scan

The first parameter is a BAT, the p and q are BUN pointers, where p is the iteration variable.

```
#define BATloop(r, p, q)
 for(q = BUNlast(r), p = BUNfirst(r);p < q; p++)
```

8.33.2 batloop where the current element can be deleted/updated

Normally it is *strictly forbidden* to update the BAT over which is being iterated, or delete the current element. This can only be done with the specialized batloop below. When doing a delete, do not forget to update the current pointer with a p = *BUNdelete*(b,p) (the delete may modify the current pointer p). After the delete/update has taken place, the pointer p is in an inconsistent state till the next iteration of the batloop starts.

```
#define BATloopDEL(r, p, q)
 for(p = BUNfirst(r), q = BUNlast(r); p < q;
     q = MIN(q,BUNlast(r)), p++)
```

8.33.3 sequential scan over deleted BUNs

Stable BUNS that were deleted, are conserved to transaction end. You may inspect these data items. Again, the b is a BAT, p and q are BUNs, where p is the iteration variable.

```
#define DELloop(b, p, q)
 for (q = (b)->batFirst, p = (b)->batDeleted; p < q; p++)
```

8.33.4 hash-table supported loop over BUNs

The first parameter 'b' is a BAT, the second ('h') should point to 'b->H->hash', and 'v' a pointer to an atomic value (corresponding to the head column of 'b'). The 'hb' is an integer index, pointing out the 'hb'-th BUN.

```
#define GDK_STREQ(l,r) (*(char*) (l) == *(char*) (r) && !strcmp(l,r))

#define HASHloop(bi, h, hb, v)
 for (hb = h->hash[HASHprobe(h, v)]; hb != BUN_NONE; hb = h->link[hb])
  if (ATOMcmp(h->type, v, BUNhead(bi, hb)) == 0)
#define HASHloop_str(bi, h, hb, v)
 for (hb = h->hash[strHash(v)&h->mask]; hb != BUN_NONE; hb = h->link[hb])
  if (GDK_STREQ(v, BUNhvar(bi, hb)))
```

For string search, we can optimize if the string heap has eliminated all doubles. This is the case when not too many different strings are stored in the heap. You can check this with the macro *strElimDoubles()* If so, we can just compare integer index numbers instead of strings:

```
#define HASHloop_fstr(bi, h, hb, idx, v)
 for (hb = h->hash[strHash(v)&h->mask], idx = strLocate((bi.b)->hheap,v);
      hb != BUN_NONE; hb = h->link[hb])
       if (*(var_t*) BUNhloc(bi, hb) == idx)
```

The following example shows how the hashloop is used:

```
void
print_books(BAT *author_books, str author)
{
        BAT *b = author_books;
        BUN i;

        printf("%s\n==================\n", author);
        HASHloop(b, (b)->H->hash, i, author)
                        printf("%s\n", ((str) BUNtail(b, i));
}
```

Note that for optimization purposes, we could have used a *HASHloop_str* instead, and also a *BUNtvar* instead of a *BUNtail* (since we know the tail-type of *author_books* is string, hence variable-sized). However, this would make the code less general.

8.33.5 specialized hashloops

HASHloops come in various flavors, from the general *HASHloop*, as above, to specialized versions (for speed) where the type is known (e.g. *HASHloop_int*), or the fact that the atom is fixed-sized (*HASHlooploc*) or variable-sized (*HASHloopvar*).

```
#define HASHlooploc(bi, h, hb, v)
 for (hb = h->hash[HASHprobe(h, v)]; hb != BUN_NONE; hb = h->link[hb])
  if (ATOMcmp(h->type, v, BUNhloc(bi, hb)) == 0)
#define HASHloopvar(bi, h, hb, v)
 for (hb = h->hash[HASHprobe(h, v)]; hb != BUN_NONE; hb = h->link[hb])
  if (ATOMcmp(h->type, v, BUNhvar(bi, hb)) == 0)
```

8.33.6 loop over a BAT with ordered tail

Here we loop over a BAT with an ordered tail column (see for instance *BATsort*). Again, 'p' and 'q' are iteration variables, where 'p' points at the current BUN. 'tl' and 'th' are pointers to atom corresponding to the minimum (included) and maximum (included) bound in the selected range of BUNs. A nil-value means that there is no bound. The 's' finally is an integer denoting the bunsize, used for speed.

```
#define SORTloop(b,p,q,tl,th)
 if (!(BATtordered(b)&1)) GDKerror("SORTloop: BAT not sorted. n");
 else for (p = (ATOMcmp((b)->ttype,tl,ATOMnilptr((b)->ttype))?
       SORTfndfirst(b,tl):BUNfirst(b)),
    q = (ATOMcmp((b)->ttype,th,ATOMnilptr((b)->ttype))?
       SORTfndlast(b,th):BUNlast(b)); p < q; p++)


/* OIDDEPEND */
#if SIZEOF_OID == SIZEOF_INT
#define SORTfnd_oid(b,v) SORTfnd_int(b,v)
#define SORTfndfirst_oid(b,v) SORTfndfirst_int(b,v)
#define SORTfndlast_oid(b,v) SORTfndlast_int(b,v)
sortloop[?.10](oid,int,oid,simple,&oid_nil)
#else
#define SORTfnd_oid(b,v) SORTfnd_lng(b,v)
#define SORTfndfirst_oid(b,v) SORTfndfirst_lng(b,v)
#define SORTfndlast_oid(b,v) SORTfndlast_lng(b,v)
sortloop[?.10](oid,lng,oid,simple,&oid_nil)
#endif
#if SIZEOF_WRD == SIZEOF_INT
#define SORTfnd_wrd(b,v) SORTfnd_int(b,v)
#define SORTfndfirst_wrd(b,v) SORTfndfirst_int(b,v)
#define SORTfndlast_wrd(b,v) SORTfndlast_int(b,v)
sortloop[?.10](wrd,int,wrd,simple,&wrd_nil)
#else
#define SORTfnd_wrd(b,v) SORTfnd_lng(b,v)
#define SORTfndfirst_wrd(b,v) SORTfndfirst_lng(b,v)
#define SORTfndlast_wrd(b,v) SORTfndlast_lng(b,v)
sortloop[?.10](wrd,lng,wrd,simple,&wrd_nil)
```

```
#endif
#define SORTloop_bit(b,p,q,tl,th) SORTloop_chr(b,p,q,tl,th)
```

8.34 Common BAT Operations

Much used, but not necessarily kernel-operations on BATs.

8.34.1 BAT aggregates

BAT*	BAThistogram(BAT *b)
BAT*	BATsample(BAT* b,BUN n)

The routine *BAThistogram* produces a new BAT with a frequency distribution of the tail of its operand.

The routine *BATsample* returns a random sample on n BUNs of a BAT.

For each BAT we maintain its dimensions as separately accessible properties. They can be used to improve query processing at higher levels.

8.34.2 Alignment transformations

Some classes of algebraic operators transform a sequence in an input BAT always in the same way in the output result. An example are the () function (including histogram(b), which is identical to (b.reverse)). That is to say, if synced(b2,b2) => synced((b1),(b2))

Another example is b.fetch(position-bat). If synced(b2,b2) and the same position-bat is fetched with, the results will again be synced. This can be mimicked by transforming the *alignment-id* of the input BAT with a one-way function onto the result.

We use **output->halign = NOID_AGGR(input->halign)** for the **output = (input)** case, and **output->align = NOID_MULT(input1->align,input2->halign)** for the fetch.

8.34.3 BAT relational operators

BAT *	BATjoin (BAT *l, BAT *r, BUN estimate)
BAT *	BATouterjoin (BAT *l, BAT *r, BUN estimate)
BAT *	BATthetajoin (BAT *l, BAT *r, int mode, BUN estimate)
BAT *	BATsemijoin (BAT *l, BAT *r)
BAT *	BATselect (BAT *b, ptr tl, ptr th)
BAT *	BATfragment (BAT *b, ptr l, ptr h, ptr L, ptr H)
BAT *	BATsunique (BAT *b)
BAT *	BATkunique (BAT *b)
BAT *	BATsunion (BAT *b, BAT *c)
BAT *	BATkunion (BAT *b, BAT *c)
BAT *	BATsintersect (BAT *b, BAT *c)
BAT *	BATkintersect (BAT *b, BAT *c)
BAT *	BATsdiff (BAT *b, BAT *c)
BAT *	BATkdiff (BAT *b, BAT *c)

The BAT library comes with a full-fledged collection of relational operators. The two selection operators *BATselect* and *BATfragment* produce a partial copy of the BAT. The former performs a search on the tail; the latter considers both dimensions. The *BATselect* operation takes two inclusive ranges as search arguments. Interpretation of a NULL argument depends on the position, i.e. a domain lower or upper bound.

The operation *BATsort* sorts the BAT on the header and produces a new BAT. A side effect is the clustering of the BAT store on the sort key.

The *BATjoin* over R[A, B] and S[C, D] performs an equi-join over B and C. It results in a BAT over A and D. The *BATouterjoin* implements a left outerjoin over the BATs involved. The *BATsemijoin* over R[A, B] and S[C, D] produces the subset of R[A, B] that satisfies the semijoin over A and C.

The full-materialization policy intermediate results in MonetDB means that a join can produce an arbitrarily large result and choke the system. The Data Distilleries tool therefore first computes the join result size before the actual join (better waste time than crash the server). To exploit that perfect result size knowledge, an result-size estimate parameter was added to all equi-join implementations. TODO: add this for semijoin/select/unique/diff/intersect

The routine *BATsunique* considers both dimensions in the double elimination it performs; it produces a set. The routine *BATtunique* considers only the head column, and produces a unique head column.

BATs that satisfy the set property can be further processed with the set operations *BATsunion*, *BATsintersect*, and *BATsdiff*. The same operations are also available in versions that only look at the head column:*BATkunion*, *BATkdiff*, and *BATkintersect* (which shares its implementation with *BATsemijoin*).

The kernel code modules are encapsulated with MAL wrappers. A synopsis of their functionality is described below. The signature details can be found in the appendix.

8.35 Aggregates Module

This module contains some efficient aggregate functions that compute their result in one scan, rather than in the iterative manner of the generic MIL aggregrate implementations.

The implementation code is derived from the original 'aggr' module. It uses a complete type-specific code expansion to avoid any type-checking in the inner-most loops. Where feasible, it replaced (expansive) hash-lookup by significantly cheaper positional void-lookups (if the head-column of the group-extend BAT (`"e"`) is `"void"`) or at least by (also positional) array lookups (in case the group-ids span a reasonably small range);

In addition to the 2-parameter

8.36 Timers and Timed Interrupts

This module handles various signaling/timer functionalities. The Monet interface supports two timer commands: *alarm* and *sleep*. Their argument is the number of seconds to wait before the timer goes off. The *sleep* command blocks till the alarm goes off. The *alarm* command continues directly, executes off a MIL string when it goes off. The parameterless routines *time* and *ctime* provide access to the cpu clock.They return an integer and string, respectively.

8.37 BAT Algebra

This modules contains the most common algebraic BAT manipulation commands. We call them *algebra*, because all operations take values as parameters, and produce new result values, but *do not modify their parameters*. Unlike the previous Monet versions, we reduce the number of functions returning a BAT reference. This was previously needed to simplify recursive bat-expression and manage reference counts. In the current version we return only a BAT identifier when a new bat is being created.

All parameters to the modules are passed by reference. In particular, this means that string values are passed to the module layer as (str *) and we have to de-reference them before entering the gdk library. This calls for knowlegde on the underlying BAT typs's

8.38 Basic array support

The array support library constructs the index arrays essential for the Relational Algebra Model language. The grid filler operation assumes that there is enough space. The shift variant multiplies all elements with a constant factor. It is a recurring operation for the RAM front-end and will save an additional copying.

The optimization is captured in a contraction macro.

8.39 Binary Association Tables

This module contains the commands and patterns to manage Binary Association Tables (BATs). The relational operations you can execute on BATs have the form of a neat algebra, described in algebra.mx

But a database system needs more that just this algebra, since often it is crucial to do table-updates (this would not be permitted in a strict algebra).

All commands needed for BAT updates, property management, basic I/O, persistence, and storage options can be found in this module.

All parameters to the modules are passed by reference. In particular, this means that string values are passed to the module layer as (str *) and we have to de-reference them before entering the gdk library. (Actual a design error in gdk to differentiate passing int/str) This calls for knowledge on the underlying BAT types's

8.39.1 Wrapping

The remainder contains the wrapper code over the version 4

8.40 InformationFunctions

In most cases we pass a BAT identifier, which should be unified with a BAT descriptor. Upon failure we can simply abort the function.

The logical head type :oid is mapped to a TYPE_void with sequenceBase. It represents the old fashioned :vid

```
str
BKCnewBAT(int *res, int *ht, int *tt, BUN *cap)
{
 BAT *b;
 if( *ht == TYPE_oid){
  int tpe= TYPE_void;
  if (CMDnew(&b, &tpe, tt, cap) == GDK_SUCCEED) {
   oid o= 0;
   BATseqbase(b, o);
   *res = b->batCacheid;
   BBPkeepref(*res);
   return MAL_SUCCEED;
  }
 } else
 if (CMDnew(&b, ht, tt, cap) == GDK_SUCCEED) {
  *res = b->batCacheid;
  BBPkeepref(*res);
  return MAL_SUCCEED;
 }
 throw(MAL, "bat.new", GDK_EXCEPTION);
}

str
BKCattach(int *ret, int *tt, str *heapfile)
{
 BAT *b;

 if (CMDattach(&b, tt, *heapfile) == GDK_SUCCEED) {
  *ret = b->batCacheid;
  BBPkeepref(*ret);
  return MAL_SUCCEED;
 }
 throw(MAL, "bat.attach", GDK_EXCEPTION);
}

str
BKCdensebat(int *ret, wrd *size)
{
 BAT *b;

 if (CMDdensebat(&b, size) == GDK_SUCCEED) {
  *ret = b->batCacheid;
  BBPkeepref(*ret);
  return MAL_SUCCEED;
 }
 throw(MAL, "bat.densebat", GDK_EXCEPTION);
}

str
BKCreverse(int *ret, int *bid)
{
 BAT *b, *bn = NULL;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.reverse", RUNTIME_OBJECT_MISSING);
 }

 CMDreverse(&bn, b);
 BBPreleaseref(b->batCacheid);
 if (bn) {
  *ret = bn->batCacheid;
  BBPkeepref(bn->batCacheid);
  return MAL_SUCCEED;
 }
 throw(MAL, "bat.reverse", GDK_EXCEPTION);
}

str
BKCmirror(int *ret, int *bid)
{
 BAT *b, *bn = NULL;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.mirror", RUNTIME_OBJECT_MISSING);
 }
 if (CMDmirror(&bn, b) == GDK_SUCCEED) {
  *ret = bn->batCacheid;
  BBPkeepref(*ret);
  BBPreleaseref(b->batCacheid);
  return MAL_SUCCEED;
 }
 *ret = 0;
 BBPreleaseref(b->batCacheid);
 throw(MAL, "bat.mirror", GDK_EXCEPTION);
}

str
BKCrevert(int *ret, int *bid)
{
 BAT *b, *bn;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.revert", RUNTIME_OBJECT_MISSING);
 }
 bn= BATrevert(b);
 if(bn==NULL ){
  BBPkeepref(*ret= b->batCacheid);
  throw(MAL, "bat.revert", GDK_EXCEPTION);
 }
 BBPkeepref(*ret= bn->batCacheid);
 return MAL_SUCCEED;
}

str
BKCorder(int *ret, int *bid)
{
 BAT *b,*bn;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.order", RUNTIME_OBJECT_MISSING);
 }
 bn= BATorder(b);
 if(bn==NULL ){
  BBPkeepref(*ret= b->batCacheid);
  throw(MAL, "bat.order", GDK_EXCEPTION);
 }
 BBPkeepref(*ret= b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCorder_rev(int *ret, int *bid)
{
 BAT *b,*bn;

 (void) ret;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.order_rev", RUNTIME_OBJECT_MISSING);
 }
 bn= BATorder_rev(b);
 if(bn==NULL ){
  BBPkeepref(*ret= b->batCacheid);
  throw(MAL, "bat.order_rev", GDK_EXCEPTION);
 }
 BBPkeepref(*ret= b->batCacheid);
 return MAL_SUCCEED;
}
```

Insertions into the BAT may involve void types (=no storage required) These cases should actually be captured during BUNins, because they may emerge internally as well.

```
void_insertbun ::=
if (b->@1type == TYPE_void && *(oid*) @1 != oid_nil &&
    *(oid*) @1 != (b->@1seqbase + BUNgetpos(b, BUNlast(b))))
{
```

```
printf("val " OIDFMT " seqbase " OIDFMT " pos " BUNFMT " n", *(oid*)@1,
 b->@1seqbase,  BUNgetpos(b, BUNlast(b)) );
      throw(MAL, "bat.insert", OPERATION_FAILED " Insert non-nil values in a void colu
}


char *
BKCinsert_bun(int *r, int *bid, ptr h, ptr t)
{
 BAT *i,*b;
 int param=0;
 (void) r;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.insert", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 derefStr[?.1](b,h,h)
 derefStr[?.1](b,t,t)
 BUNins(b, h, t,FALSE);
 BBPkeepref(*r=b->batCacheid);
 BBPreleaseref(i->batCacheid);
 return MAL_SUCCEED;
}

char *
BKCinsert_bun_force(int *r, int *bid, ptr h, ptr t, bit *force)
{
 BAT *i,*b;
 int param=0;
 (void) r;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.insert", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 derefStr[?.1](b,h,h)
 derefStr[?.1](b,t,t)
 BUNins(b, h, t, *force);
 BBPkeepref(*r=b->batCacheid);
 BBPreleaseref(i->batCacheid);
 return MAL_SUCCEED;
}

str
BKCinsert_bat(int *r, int *bid, int *sid)
{
 BAT *i,*b, *s;
 int param=0;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.insert", RUNTIME_OBJECT_MISSING);
 }
 if ((s = BATdescriptor(*sid)) == NULL) {
  BBPreleaseref(i->batCacheid);
  throw(MAL, "bat.insert", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 if (BATins(b, s,FALSE) == NULL) {
  BBPkeepref(*r=b->batCacheid);
  BBPreleaseref(s->batCacheid);
  BBPreleaseref(i->batCacheid);
  throw(MAL, "bat.insert", GDK_EXCEPTION);
 }
 BBPreleaseref(s->batCacheid);
 BBPkeepref(*r=b->batCacheid);
 BBPreleaseref(i->batCacheid);
 return MAL_SUCCEED;
}

str
BKCinsert_bat_force(int *r, int *bid, int *sid, bit *force)
{
 BAT *i,*b, *s;
 int param=0;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.insert", RUNTIME_OBJECT_MISSING);
 }
 if ((s = BATdescriptor(*sid)) == NULL) {
  BBPreleaseref(i->batCacheid);
  throw(MAL, "bat.insert", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 if (BATins(b, s, *force) == NULL) {
  BBPreleaseref(b->batCacheid);
  BBPreleaseref(s->batCacheid);
  BBPreleaseref(i->batCacheid);
  throw(MAL, "bat.insert", GDK_EXCEPTION);
 }
 BBPkeepref(*r=b->batCacheid);
 BBPreleaseref(s->batCacheid);
 BBPreleaseref(i->batCacheid);
 return MAL_SUCCEED;
```

```
}


str
BKCreplace_bun(int *r, int *bid, ptr h, ptr t)
{
 BAT *i,*b;
 int param=0;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.replace", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 derefStr[?.1](b,h,h)
 derefStr[?.1](b,t,t)
 if (BUNreplace(b, h, t, 0) == NULL) {
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.replace", GDK_EXCEPTION);
 }
 BBPkeepref(*r=b->batCacheid);
 BBPreleaseref(i->batCacheid);
 return MAL_SUCCEED;
}

str
BKCreplace_bat(int *r, int *bid, int *sid)
{
 BAT *i, *b, *bn, *s;
 int param=0;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.replace", RUNTIME_OBJECT_MISSING);
 }
 if ((s = BATdescriptor(*sid)) == NULL) {
  BBPreleaseref(i->batCacheid);
  throw(MAL, "bat.replace", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 bn=BATreplace(b, s, 0);
 if(bn && bn->batCacheid != b->batCacheid){
  BBPreleaseref(i->batCacheid);
  BBPreleaseref(s->batCacheid);
  BBPreleaseref(b->batCacheid);
  if( bn)
   BBPreleaseref(bn->batCacheid);
  throw(MAL, "bat.replace", OPERATION_FAILED "Different BAT returned");
 }
 BBPkeepref(*r=bn->batCacheid);
 BBPreleaseref(i->batCacheid);
 BBPreleaseref(s->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCreplace_bun_force(int *r, int *bid, ptr h, ptr t, bit *force)
{
 BAT *b, *bn;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.replace", RUNTIME_OBJECT_MISSING);
 }
 derefStr[?.1](b,h,h)
 derefStr[?.1](b,t,t)
 bn= BUNreplace(b, h, t, *force);
 BBPreleaseref(b->batCacheid);
 if(bn && bn->batCacheid != b->batCacheid)
  throw(MAL, "bat.replace", OPERATION_FAILED "Different BAT returned");
 BBPkeepref(*r=bn->batCacheid);
 return MAL_SUCCEED;
}

str
BKCreplace_bat_force(int *r, int *bid, int *sid, bit *force)
{
 BAT *b, *bn, *s;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.replace", RUNTIME_OBJECT_MISSING);
 }
 if ((s = BATdescriptor(*sid)) == NULL) {
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.replace", RUNTIME_OBJECT_MISSING);
 }
 bn= BATreplace(b, s, *force);
 BBPreleaseref(s->batCacheid);
 BBPreleaseref(b->batCacheid);
 if(bn && bn->batCacheid != b->batCacheid)
  throw(MAL, "bat.replace_bat", OPERATION_FAILED "Different BAT returned");
 BBPkeepref(*r=bn->batCacheid);
 return MAL_SUCCEED;
}

char *
```

```
BKCdelete_bun(int *r, int *bid, ptr h, ptr t)
{
 BAT *b, *bn;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }
 derefStr[?.1](b,h,h)
 derefStr[?.1](b,t,t)
 bn= BUNdel(b, h, t,FALSE);
 if(bn && bn->batCacheid != b->batCacheid)
  throw(MAL, "bat.delete_bun", OPERATION_FAILED "Different BAT returned");
 BBPkeepref(*r=bn->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

char *
BKCdelete(int *r, int *bid, ptr h)
{
 BAT *b, *bn;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }
 derefStr[?.1](b,h,h)
 bn= BUNdelHead(b, h,FALSE);
 if(bn && bn->batCacheid != b->batCacheid)
  throw(MAL, "bat.delete", OPERATION_FAILED "Different BAT returned");
 BBPkeepref(*r=bn->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCdelete_all(int *r, int *bid)
{
 BAT *b, *bn;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }
 bn=BATclear(b);
 if(bn && bn->batCacheid != b->batCacheid){
  BBPreleaseref(bn->batCacheid);
  throw(MAL, "bat.delete_all", OPERATION_FAILED "Different BAT returned");
 }
 BBPkeepref(*r=b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCdelete_bat_bun(int *r, int *bid, int *sid)
{
 BAT *b, *bn, *s;

 if( *bid == *sid)
  return BKCdelete_all(r,bid);
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }
 if ((s = BATdescriptor(*sid)) == NULL) {
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }

 bn=BATdel(b, s,FALSE);
 BBPreleaseref(s->batCacheid);
 if(bn && bn->batCacheid != b->batCacheid)
  throw(MAL, "bat.delete_bat_buns", OPERATION_FAILED "Different BAT returned");
 BBPkeepref(*r=bn->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCdelete_bat(int *r, int *bid, int *sid)
{
 BAT *i,*b, *s;
 int param=0;

 if ((i = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }
 if ((s = BATdescriptor(*sid)) == NULL) {
  BBPreleaseref(i->batCacheid);
  throw(MAL, "bat.delete", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&b,i,&param);
 if (BATdelHead(b, s,FALSE) != NULL) {
  BBPreleaseref(b->batCacheid);
  BBPreleaseref(s->batCacheid);
  return MAL_SUCCEED;
 }
 BBPkeepref(*r=b->batCacheid);
 BBPreleaseref(s->batCacheid);
 BBPreleaseref(i->batCacheid);
 return MAL_SUCCEED;
}

str
BKCdestroy_bat(bit *r, str *input)
{
 CMDdestroy(r, *input);
 return MAL_SUCCEED;
}

char *
BKCdestroyImmediate(signed char*r, int *bid)
{
 BAT *b;
 char buf[512];

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.destroy", RUNTIME_OBJECT_MISSING);
 }
 BBPlogical(b->batCacheid, buf);
 BBPreleaseref(b->batCacheid);
 CMDdestroy(r, buf);
 return MAL_SUCCEED;
}

char *
BKCdestroy(signed char *r, int *bid)
{
 BAT *b;

 (void) r;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.destroy", RUNTIME_OBJECT_MISSING);
 }
 *bid = 0;
 BATmode(b, TRANSIENT);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

/* The SQL frontend uses void-head bats */
BUN
void_delete_bat(BAT *b, BAT *d, int delta)
{
 BATiter di = bat_iterator(d);
 BUN nr = 0;
 BUN r, s;
 ptr nil = ATOMnilptr(b->ttype);

 if (delta) {
  for (r = d->batInserted; r < BUNlast(d); r++) {
   oid delid = *(oid *) BUNtail(di, r);

   void_inplace5(b, delid, nil, TRUE);
   nr++;
  }
 } else {
  BATloop(d, r, s) {
   oid delid = *(oid *) BUNtail(di, r);

   void_inplace5(b, delid, nil, TRUE);
   nr++;
  }
 }
 return nr;
}

BUN
void_insert_delta(BAT *b, BAT *u)
{
 BATiter ui = bat_iterator(u);
 BUN nr = 0;
 BUN r;

 for (r = u->batInserted; r < BUNlast(u); r++) {
  BUNappend(b, BUNtail(ui, r),FALSE);
  nr++;
 }
 return nr;
}

BUN
void_replace_delta(BAT *b, BAT *u)
{
 BATiter ui = bat_iterator(u);
 BUN nr = 0;
 BUN r;

 for (r = u->batInserted; r < BUNlast(u); r++) {
  oid updid = *(oid *) BUNhead(ui, r);
  ptr val = BUNtail(ui, r);
  void_inplace5(b, updid, val, TRUE);
  nr++;
 }
 return nr;
}

char *
BKCappend_wrap(int *r, int *bid, int *uid)
{
 BAT *b, *i, *u;
 int param=0;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }
 if ((u = BATdescriptor(*uid)) == NULL) {
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&i,b,&param);
 BATappend(i, u,FALSE);
 BBPkeepref(*r=i->batCacheid);
 BBPreleaseref(b->batCacheid);
 BBPreleaseref(u->batCacheid);
 return MAL_SUCCEED;
}

str
BKCappend_val_wrap(int *r, int *bid, ptr u)
{
 BAT *i,*b;
 int param=0;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }

 derefStr[?.1](b,t,u)
 CMDsetaccess(&i,b,&param);
 BUNappend(i, u,FALSE);
 BBPkeepref(*r=i->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
str
BKCappend_reverse_val_wrap(int *r, int *bid, ptr u)
```

```
{
 BAT *i,*b;
 int param=0;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }

 CMDsetaccess(&i,b,&param);
 derefStr[?.1](i,t,u)
 BUNappend(BATmirror(i), u,FALSE);
 BBPkeepref(*r=i->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

char *
BKCappend_force_wrap(int *r, int *bid, int *uid, bit *force)
{
 BAT *b,*i, *u;
 int param=0;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }
 if ((u = BATdescriptor(*uid)) == NULL) {
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&i,b,&param);
 BATappend(i, u, *force);
 BBPkeepref(*r=i->batCacheid);
 BBPreleaseref(u->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCappend_val_force_wrap(int *r, int *bid, ptr u, bit *force)
{
 BAT *b,*i;
 int param=0;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.append", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&i,b,&param);
 derefStr[?.1](i,t,u)
 BUNappend(i, u, *force);
 BBPkeepref(*r=i->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCbun_inplace(int *r, int *bid, oid *id, ptr t)
{
 BAT *o;

 (void) r;
 if ((o = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.inplace", RUNTIME_OBJECT_MISSING);
 }
 void_inplace5(o, *id, t,FALSE);
 BBPreleaseref(o->batCacheid);
 return MAL_SUCCEED;
}

str
BKCbun_inplace_force(int *r, int *bid, oid *id, ptr t, bit *force)
{
 BAT *o;

 (void) r;
 if ((o = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.inplace", RUNTIME_OBJECT_MISSING);
 }
 void_inplace5(o, *id, t, *force);
 BBPreleaseref(o->batCacheid);
 return MAL_SUCCEED;
}

str
BKCbat_inplace(int *r, int *bid, int *rid)
{
 BAT *o, *d;

 (void) r;
 if ((o = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.inplace", RUNTIME_OBJECT_MISSING);
 }
 if ((d = BATdescriptor(*rid)) == NULL) {
  BBPreleaseref(o->batCacheid);
  throw(MAL, "bat.inplace", RUNTIME_OBJECT_MISSING);
 }
 void_replace_bat5(o, d,FALSE);
 BBPreleaseref(o->batCacheid);
 BBPreleaseref(d->batCacheid);
 return MAL_SUCCEED;
}

str
BKCbat_inplace_force(int *r, int *bid, int *rid, bit *force)
{
 BAT *o, *d;

 (void) r;
 if ((o = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.inplace", RUNTIME_OBJECT_MISSING);
 }
 if ((d = BATdescriptor(*rid)) == NULL) {
  BBPreleaseref(o->batCacheid);
  throw(MAL, "bat.inplace", RUNTIME_OBJECT_MISSING);
 }
 void_replace_bat5(o, d, *force);
 BBPreleaseref(o->batCacheid);
 BBPreleaseref(d->batCacheid);
 return MAL_SUCCEED;
}

/*end of SQL enhancement */

char *
BKCgetAlpha(int *r, int *bid)
{
 BAT *b, *c;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getInserted", RUNTIME_OBJECT_MISSING);
 }
 c = BATalpha(b);
 *r = c->batCacheid;
 BBPkeepref(c->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

char *
BKCgetDelta(int *r, int *bid)
{
 BAT *b, *c;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getDeleted", RUNTIME_OBJECT_MISSING);
 }
 c = BATdelta(b);
 *r = c->batCacheid;
 BBPkeepref(c->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCgetCapacity(lng *res, int *bid)
{
 CMDcapacity(res, bid);
 return MAL_SUCCEED;
}

str
BKCgetHeadType(str *res, int *bid)
{
 CMDhead(res, bid);
 return MAL_SUCCEED;
}

str
BKCgetTailType(str *res, int *bid)
{
 CMDtail(res, bid);
 return MAL_SUCCEED;
}

str
BKCgetRole(str *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getType", RUNTIME_OBJECT_MISSING);
 }
 *res = GDKstrdup((*bid > 0) ? b->hident : b->tident);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
```

```
BKCsetkey(int *res, int *bid, bit *param)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setKey", RUNTIME_OBJECT_MISSING);
 }
 BATkey(b, *param ? BOUND2BTRUE :FALSE);
 *res = b->batCacheid;
 BBPkeepref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCsetSet(int *res, int *bid, bit *param)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setSet", RUNTIME_OBJECT_MISSING);
 }
 BATset(b, *param ? BOUND2BTRUE :FALSE);
 *res = b->batCacheid;
 BBPkeepref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCisaSet(int *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isaSet", RUNTIME_OBJECT_MISSING);
 }
 *res = b->batSet;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCsetSorted(bit *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isSorted", RUNTIME_OBJECT_MISSING);
 }
 CMDordered(res, b);
 *res = BATordered(b) ? 1 : 0;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCisSorted(bit *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isSorted", RUNTIME_OBJECT_MISSING);
 }
 *res = BATordered(b) ? 1 : 0;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCisSortedReverse(bit *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isSorted", RUNTIME_OBJECT_MISSING);
 }
 *res = BATordered_rev(b) ? 1 : 0;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
```

We must take care of the special case of a nil column (TYPE_void,seqbase=nil) such nil columns never set hkey (and BUNins will never invalidate it if set) yet a nil column of a BAT with <= 1 entries does not contain doubles => return TRUE.

```
str
BKCgetKey(bit *ret, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setPersistence", RUNTIME_OBJECT_MISSING);
 }
 CMDgetkey(ret, b);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCpersists(int *r, int *bid, bit *flg)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setPersistence", RUNTIME_OBJECT_MISSING);
 }
 BATmode(b, (*flg == TRUE) ? PERSISTENT : (*flg ==FALSE) ? TRANSIENT : SESSION);
 BBPreleaseref(b->batCacheid);
 *r = 0;
 return MAL_SUCCEED;
}

str
BKCsetPersistent(int *r, int *bid)
{
 bit flag= TRUE;
 return BKCpersists(r,bid, &flag);
}

str
BKCisPersistent(bit *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setPersistence", RUNTIME_OBJECT_MISSING);
 }
 *res = (b->batPersistence == PERSISTENT) ? TRUE :FALSE;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCsetTransient(int *r, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setTransient", RUNTIME_OBJECT_MISSING);
 }
 BATmode(b, TRANSIENT);
 *r = 0;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCisTransient(bit *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setTransient", RUNTIME_OBJECT_MISSING);
 }
 *res = b->batPersistence == TRANSIENT;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
```

```
accessMode_export ::=
bat5_export str BKCset@1(int *res, int *bid) ;
bat5_export str BKChas@1(bit *res, int *bid);
```

```
accessMode ::=
str BKCset@1(int *res, int *bid) {
 BAT *b, *bn = NULL;
 int param=@2;
    if( (b= BATdescriptor(*bid)) == NULL ){
        throw(MAL, "bat.set@1", RUNTIME_OBJECT_MISSING);
    }
 CMDsetaccess(&bn,b,&param);
 BBPkeepref(*res=bn->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
str BKChas@1(bit *res, int *bid) {
 BAT *b;
    if( (b= BATdescriptor(*bid)) == NULL ){
        throw(MAL, "bat.set@1", RUNTIME_OBJECT_MISSING);
    }
 *res = BATgetaccess(b)=='@3';
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
```

```
accessMode_export[?.2](WriteMode,0,w)
accessMode_export[?.2](ReadMode,1,r)
accessMode_export[?.2](AppendMode,2,a)
```

```
accessMode[?.2](WriteMode,0,w)
```

```
accessMode[?.2](ReadMode,1,r)
accessMode[?.2](AppendMode,2,a)

str
BKCaccess(int *res, int *bid, int *m)
{
 BAT *b, *bn = NULL;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setAccess", RUNTIME_OBJECT_MISSING);
 }
 CMDsetaccess(&bn, b, m);
 *res = bn->batCacheid;
 BBPkeepref(bn->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCsetAccess(int *res, int *bid, str *param)
{
 BAT *b, *bn = NULL;
 int m;
 int oldid;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setAccess", RUNTIME_OBJECT_MISSING);
 }
 switch (*param[0]) {
 case 'r':
  m = 1;
  break;
 case 'a':
  m = 2;
  break;
 case 'w':
  m = 0;
  break;
 default:
  *res = 0;
  throw(MAL, "bat.setAccess", ILLEGAL_ARGUMENT" Got %c" " expected 'r','a', or 'w'", *
 }
 /* CMDsetaccess(&bn, b, &m);*/

 oldid= b->batCacheid;
 bn = BATsetaccess(b, m);
 if ((bn)->batCacheid == b->batCacheid) {
  BBPkeepref(bn->batCacheid);
 } else {
  BBPreleaseref(oldid);
  BBPfix(bn->batCacheid);
  BBPkeepref(bn->batCacheid);
 }
 *res = bn->batCacheid;
 return MAL_SUCCEED;
}

str
BKCgetAccess(str *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getAccess", RUNTIME_OBJECT_MISSING);
 }
 switch (BATgetaccess(b)) {
 case 1:
  *res = GDKstrdup("read");
  break;
 case 2:
  *res = GDKstrdup("append");
  break;
 case 0:
  *res = GDKstrdup("write");
  break;
 }
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
```

8.40.1 Property management

All property operators should ensure exclusive access to the BAT descriptor. Where necessary use the primary view to access the properties

```
str
BKCinfo(int *retval, int *bid)
{
 BAT *bn = NULL, *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getInfo", RUNTIME_OBJECT_MISSING);
 }
 if (CMDinfo(&bn, b) == GDK_SUCCEED) {
  *retval = bn->batCacheid;
  BBPkeepref(bn->batCacheid);
  BBPreleaseref(*bid);
  return MAL_SUCCEED;
 }
 BBPreleaseref(*bid);
 BBPreleaseref(b->batCacheid);
 throw(MAL, "BKCinfo", GDK_EXCEPTION);
}

str
BKCbatdisksize(lng *tot, int *bid){
 BAT *b;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getDiskSize", RUNTIME_OBJECT_MISSING);
 }
 CMDbatdisksize(tot,b);
 BBPreleaseref(*bid);
 return MAL_SUCCEED;
}

str
BKCbatvmsize(lng *tot, int *bid){
 BAT *b;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getDiskSize", RUNTIME_OBJECT_MISSING);
 }
 CMDbatvmsize(tot,b);
 BBPreleaseref(*bid);
 return MAL_SUCCEED;
}

str
BKCbatsize(lng *tot, int *bid){
 BAT *b;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getDiskSize", RUNTIME_OBJECT_MISSING);
 }
 CMDbatsize(tot,b, FALSE);
 BBPreleaseref(*bid);
 return MAL_SUCCEED;
}

str
BKCgetStorageSize(lng *tot, int *bid)
{
 BAT *b;
 if ((b = BATdescriptor(*bid)) == NULL)
  throw(MAL, "bat.getStorageSize", RUNTIME_OBJECT_MISSING);
 CMDbatsize(tot,b,TRUE);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
str
BKCgetSpaceUsed(lng *tot, int *bid)
{
 BAT *b;
 size_t size = sizeof(BATstore);

 if ((b = BATdescriptor(*bid)) == NULL)
  throw(MAL, "bat.getSpaceUsed", RUNTIME_OBJECT_MISSING);

 if (!isVIEW(b)) {
  BUN cnt = BATcount(b);

  size += headsize(b, cnt);
  size += tailsize(b, cnt);
  /* the upperbound is used for the heaps */
  if (b->hheap)
   size += b->hheap->size;
  if (b->theap)
   size += b->theap->size;
  if (b->H->hash)
   size += sizeof(BUN) * cnt;
  if (b->T->hash)
   size += sizeof(BUN) * cnt;
 }
 *tot = size;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCgetStorageSize_str(lng *tot, str batname)
{
 int bid = BBPindex(batname);

 if (bid == 0)
  throw(MAL, "bat.getStorageSize", RUNTIME_OBJECT_MISSING);
 return BKCgetStorageSize(tot, &bid);
}
```

8.41 Synced BATs

```
str
BKCisSynced(bit *ret, int *bid1, int *bid2)
{
 BAT *b1, *b2;

 if ((b1 = BATdescriptor(*bid1)) == NULL) {
  throw(MAL, "bat.isSynced", RUNTIME_OBJECT_MISSING);
 }
 if ((b2 = BATdescriptor(*bid2)) == NULL) {
  BBPreleaseref(b1->batCacheid);
  throw(MAL, "bat.isSynced", RUNTIME_OBJECT_MISSING);
 }
 CMDsynced(ret, b1, b2);
 BBPreleaseref(b1->batCacheid);
 BBPreleaseref(b2->batCacheid);
 return MAL_SUCCEED;
}
```

8.42 Role Management

```
char *
BKCsetRole(int *r, int *bid, char **hname, char **tname)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setRole", RUNTIME_OBJECT_MISSING);
 }
 if (hname == 0 || *hname == 0 || **hname == 0){
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.setRole", ILLEGAL_ARGUMENT " Head name missing");
 }
 if (tname == 0 || *tname == 0 || **tname == 0){
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.setRole", ILLEGAL_ARGUMENT " Tail name missing");
 }
 BATroles(b, *hname, *tname);
 BBPreleaseref(b->batCacheid);
 *r = 0;
 return MAL_SUCCEED;
}

str
BKCsetColumn(int *r, int *bid, str *tname)
{
 BAT *b;
 str dummy;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setColumn", RUNTIME_OBJECT_MISSING);
 }
 if (tname == 0 || *tname == 0 || **tname == 0){
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.setColumn", ILLEGAL_ARGUMENT " Column name missing");
 }
 /* watch out, hident is freed first */
 dummy= GDKstrdup(b->hident);
 BATroles(b, dummy, *tname);
 GDKfree(dummy);
 BBPreleaseref(b->batCacheid);
 *r =0;
 return MAL_SUCCEED;
}

str
BKCsetColumns(int *r, int *bid, str *hname, str *tname)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setColumns", RUNTIME_OBJECT_MISSING);
 }
 if (hname == 0 || *hname == 0 || **hname == 0){
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.setRole", ILLEGAL_ARGUMENT " Head name missing");
 }
 if (tname == 0 || *tname == 0 || **tname == 0){
  BBPreleaseref(b->batCacheid);
  throw(MAL, "bat.setRole", ILLEGAL_ARGUMENT " Tail name missing");
 }
 BATroles(b, *hname, *tname);
 BBPreleaseref(b->batCacheid);
 *r =0;
 return MAL_SUCCEED;
}


str
BKCsetName(int *r, int *bid, str *s)
{
 BAT *b;
 bit res, *rp = &res;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setName", RUNTIME_OBJECT_MISSING);
 }
 CMDrename(rp, b, *s);
 BBPreleaseref(b->batCacheid);
 *r = 0;
 return MAL_SUCCEED;
}

str
BKCgetBBPname(str *ret, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.getName", RUNTIME_OBJECT_MISSING);
 }
 *ret = GDKstrdup(BBPname(b->batCacheid));
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCunload(bit *res, str *input)
{
 CMDunload(res, *input);
 return MAL_SUCCEED;
}

str
BKCisCached(int *res, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isCached", RUNTIME_OBJECT_MISSING);
 }
 *res = 0;
 BBPreleaseref(b->batCacheid);
 throw(MAL, "bat.isCached", PROGRAM_NYI);
}

str
BKCload(int *res, str *input)
{
 bat bid = BBPindex(*input);
 *res = bid;
 if (bid) {
  BBPincref(bid,TRUE);
  return MAL_SUCCEED;
 }
 throw(MAL, "bat.unload", ILLEGAL_ARGUMENT " File name missing");
}

str
BKChot(int *res, str *input)
{
 (void) res;  /* fool compiler */
 BBPhot(BBPindex(*input));
 return MAL_SUCCEED;
}

str
BKCcold(int *res, str *input)
{
 (void) res;  /* fool compiler */
 BBPcold(BBPindex(*input));
 return MAL_SUCCEED;
}

str
BKCcoldBAT(int *res, int *bid)
{
 BAT *b;

 (void) res;
 (void) bid;  /* fool compiler */
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isCached", RUNTIME_OBJECT_MISSING);
 }
 BBPcold(b->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCheat(int *res, str *input)
{
 int bid = BBPindex(*input);

 if (bid) {
  *res = BBP_lastused(bid) & 0x7fffffff;
 }
 throw(MAL, "bat", PROGRAM_NYI);
}

str
BKChotBAT(int *res, int *bid)
{
 BAT *b;

 (void) res;
 (void) bid;  /* fool compiler */
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.isCached", RUNTIME_OBJECT_MISSING);
 }
 BBPhot(b->batCacheid);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCsave(bit *res, str *input)
{
 CMDsave(res, *input);
 return MAL_SUCCEED;
}

str
BKCsave2(int *r, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.save", RUNTIME_OBJECT_MISSING);
 }

 if (b && BATdirty(b))
  BBPsave(b);
 BBPreleaseref(b->batCacheid);
 *r = 0;
 return MAL_SUCCEED;
}


str
BKCmmap(int *res, int *bid, int *hbns, int *tbns, int *hhp, int *thp)
{
 BAT *b, *bn = NULL;
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.mmap", RUNTIME_OBJECT_MISSING);
 }
 if (CMDmmap(&bn, b, hbns, tbns, hhp, thp) == GDK_SUCCEED) {
  *res = TRUE;
  BBPreleaseref(bn->batCacheid);
  BBPreleaseref(b->batCacheid);
  return MAL_SUCCEED;
 }
 *res =FALSE;
 BBPreleaseref(b->batCacheid);
 throw(MAL, "bat.mmap", GDK_EXCEPTION);
}

str
BKCmmap2(int *res, int *bid, int *mode)
{
 return BKCmmap(res, bid, mode, mode, mode, mode);
}

str
BKCmadvise(int *res, int *bid, int *hbns, int *tbns, int *hhp, int *thp)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.madvice", RUNTIME_OBJECT_MISSING);
 }
 *res = BATmadvise(b, (*hbns == int_nil) ? -1 : *hbns, (*tbns == int_nil) ? -1 : *tbns
 BBPreleaseref(b->batCacheid);
 if (*res)
  throw(MAL, "bat.madvise", GDK_EXCEPTION);
 return MAL_SUCCEED;
}

str
BKCmadvise2(int *res, int *bid, int *mode)
{
 return BKCmadvise(res, bid, mode, mode, mode, mode);
}
```

8.43 Accelerator Control

```
str
BKCaccbuild(int *ret, int *bid, str *acc, ptr *param)
```

```
{
 (void) bid;
 (void) acc;
 (void) param;
 *ret = TRUE;
 throw(MAL, "Accelerator", PROGRAM_NYI);
}

str
BKCaccbuild_std(int *ret, int *bid, int *acc)
{
 (void) bid;
 (void) acc;
 *ret = TRUE;
 throw(MAL, "Accelerator", PROGRAM_NYI);
}


str
BKCsetHash(bit *ret, int *bid, bit *prop)
{
 BAT *b;

 (void) ret;
 (void) prop;  /* fool compiler */
 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setHash", RUNTIME_OBJECT_MISSING);
 }
 BAThash(b, 0);
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}

str
BKCsetSequenceBase(int *r, int *bid, oid *o)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setSequenceBase", RUNTIME_OBJECT_MISSING);
 }
 BATseqbase(b, *o);
 *r = b->batCacheid;
 BBPkeepref(b->batCacheid);
 return MAL_SUCCEED;
}
```

```
str
BKCsetSequenceBaseNil(int *r, int *bid, oid *o)
{
 oid ov = oid_nil;

 (void) o;
 return BKCsetSequenceBase(r, bid, &ov);
}

str
BKCgetSequenceBase(oid *r, int *bid)
{
 BAT *b;

 if ((b = BATdescriptor(*bid)) == NULL) {
  throw(MAL, "bat.setSequenceBase", RUNTIME_OBJECT_MISSING);
 }
 *r = b->hseqbase;
 BBPreleaseref(b->batCacheid);
 return MAL_SUCCEED;
}
```

8.44 BAT calculator

Many applications require extension of the basic calculator and mathematic functions to work on BAT arguments. Although the MAL multiplex optimizer contains a command ('optimizer.multiplex') to generate the necessary code, it is often much more efficient to use one of the dedidacted batcalc functions introduced below.

The operators supported are limited to the built-in fixed length atoms, because they permit ease of storage of the operation result. Variable sized atoms, especially user defined, may require more administrative activities. Furthermore, the operands involved are assumed to be aligned to assure the fastest possible join evaluation.

Optimal processing performance is further obtained when the operators can work as 'accumulators', for then we do not pay the price of space allocation for a new intermediate. It may imply a BATcopy before the accummulator function is being called. A new BAT is of course created when the result of a function does not fit the accumulator.

The implementation does not take into account possible overflows caused by the operators. However, the operators respect the NIL semantics and division by zero produces a NIL.

In addition to arithmetic and comparison operators, casting and mathematical functions are directly supported.

8.45 NULL semantics

The batcalc arithmetic is already constraint to BATs of equal size. Another improvement is obtained when we do not have to check for NULLs in each and every basic operation

+,-,/,* and comparisons. A problem is to propagate the general 'nonil' property, because this depends on the semantics of the operator. Checking the result for a nil value depleats most of the expected gain.

Currently only set for the NOT operation and comparisons operators when neither argument has nils.

8.46 BAT Coercion Routines

The coercion routines over BATs can not easily be speed up using an accumulator approach, because they often require different storage space. Nevertheless, the implementation provided here are much faster compared to the Version 4.* implementation.

The coercion routines are build for speed. They are not protected against overflow.

8.47 BAT if-then-else multiplex expressions.

The assembled code for IF-THEN-ELSE multiplex operations. Again we assume that the BAT arguments are aligned.

8.48 Color multiplexes

[TODO: property propagations and general testing] The collection of routines provided here are map operations for the color string primitives.

In line with the batcalc module, we assume that if two bat operands are provided that they are already aligned on the head. Moreover, the head of the BATs are limited to :oid, which can be cheaply realized using the GRPsplit operation.

8.49 String multiplexes

[TODO: property propagations] The collection of routines provided here are map operations for the atom string primitives.

In line with the batcalc module, we assume that if two bat operands are provided that they are already aligned on the head. Moreover, the head of the BATs are limited to :void, which can be cheaply realized using the GRPsplit operation.

8.50 BAT math calculator

This module contains the multiplex versions of the linked in mathematical functions.

8.51 The math module

This module contains the math commands. The implementation is very simply, the c math library functions are called. See for documentation the ANSI-C/POSIX manuals of the equaly named functions.

NOTE: the operand itself is being modified, rather than that we produce a new BAT. This to save the expensive copying.

8.52 Time/Date multiplexes

[TODO: arithmetic multiplexes] The collection of routines provided here are map operations for the atom time and date primitives.

In line with the batcalc module, we assume that if two bat operands are provided that they are already aligned on the head. Moreover, the head of the BATs are limited to :void, which can be cheaply realized using the GRPsplit operation.

8.53 Basic arithmetic

This module is an extended version of the V4 arithmetic module. It implements the arithmetic operations on the built-in types, *chr*, *bte*, *sht*, *int*, *flt*, *dbl* and *lng*. All combinations are implemented. Limited combinations are implemented for *bit*, *oid* and *str*.

`[binary operators]`
: The implemented operators are first of all all comparison that return a TRUE/FALSE value (*bit* values), i.e. `<=`, `<`, `==`, `!=`, `>=`, and `>=`.

  The module also implements the operators `+`, `-`, `*` and `/`. The rules for the return types operators is as follows. If one of the input types is a floating point the result will be a floating point. The largest type of the input types is taken.

  The *max* and *min* functions return the maximum and minimum of the two input parameters.

`[unary operators]`
: This module also implements the unary *abs*() function, which calculates the absolute value of the given input parameter, as well as the - unary operator.

  The *inv* unary operation calculates the inverse of the input value. An error message is given when the input value is zero.

`[bitwise operators]`
: For integers there are some additional operations. The % operator implements the congruent modulo operation. The `<<` and `>>` are the left and right bit shift. The *or*, *and*, *xor* and *not* for integers are implemented as bitwise boolean operations.

`[boolean operators]`
: The *or*, *and*, *xor* and *not* for the bit atomic type in MIL (this corresponds to what is normally called boolean) are implemented as the logic operations.

`[random numbers]`
: This module also contains the rand and srand functions. The *srand*() function initializes the random number generator using a seed value. The subsequent calls to *rand*() are pseudo random numbers (with the same seed the sequence can be repeated).

The general interpretation for the NIL value is "unknown". This semantics mean that any operation that receives at least one NIL value, will produce a NIL value in the output for sure.

The only exception to this rule are the "==" and "!=" equality test routines (it would otherwise become rather difficult to test whether a value is nil).

8.54 Performance Counters

This is a memory/cpu performance measurement tool for the following processor (families).

- MIPS R10000/R12000 (IP27)
- Sun UltraSparcI/II (sun4u)
- Intel Pentium (i586/P5)
- Intel PentiumPro/PentiumII/PentiumIII/Celeron (i686/P6)
- AMD Athlon (i686/K7)
- Intel Itanium/Itanium2 (ia64)

It uses

- libperfmon libperfex (IRIX) for R10000/R12000,
- (Solaris <= 7) by Richard Enbody, libcpc (Solaris >= 8) for UltraSparcI/II,
- libperfctr (Linux-i?86 >= 2.4), by M. Pettersson for Pentiums & Athlons.
- libpfm (Linux-ia64 >= 2.4), by HP for Itanium[2].

Module counters provides similar interface and facilities as Peter's R10000 perfex module, but it offers no multiplexing of several events; only two events can be monitored at a time. On non-Linux/x86, non-Solaris/UltraSparc, and non-IRIX/R1x000 systems, only the elapsed time in microseconds is measured.

8.55 The group module

This module contains the primitives to construct, derive, and perform statistical operations on BATs representing groups. The default scheme in Monet is to assume the head to represent the group identifier and the tail an element in the group.

Groups play an important role in datamining, where they are used to construct cross-tables. Such cross tables over a single BAT are already supported by the histogram function. This module provides extensions to support identification of groups in a (multi-)dimensional space.

The module implementation has a long history. The first implementation provided several alternatives to produce/derive the grouping. A more complete (and complex) scheme was derived during its extensive use in the context of the Data Distilleries product. The current implementation is partly a cleanup of this code-base, but also enables provides better access to the intermediate structures produced in the process, i.e. the histogram and the sub-group mapping. They can be used for various optimization schemes at the MAL level.

The prime limitation of the current implementation is that an underlying database of `oid->any` BATs is assumed. This enables representation of each group using an oid, and the value representation of the group can be accordingly be retrieved easily. An optimized implementation in which we use positional integer id's (as embodied by Monet's void type) is also available.

This limitation on (v)oid-headers is marginal. The primitive GRPsplit produces for any BAT two copies with both a (v)oid header.

8.55.1 Algorithms

There are several approaches to build a cross table. The one chosen here is aimed at incremental construction, such that re-use of intermediates becomes possible. Starting with the first dimension, a BAT is derived to represent the various groups, called a *GRP BAT* or cross-table BAT.

8.55.2 Cross Table (GRP)

A cross table is an `<oid,oid>` BAT where the first (head) denotes a tuple in the cross table and the second (tail) marks all identical lists. The tail-oids contain group identifiers; that is, *this value is equal* **iff** *two tuples belong to the same group*. The group identifiers are chosen from the domain of the tuple-identifiers. This simplifies getting back to the original tuples, when talking about a group. If the tuple-oid of 'John' is chosen as a group-id, you might view this as saying that each member of the group is 'like John' with respect to the grouping-criterion.

Successively the subgroups can be identified by modifying the GRP BAT or to derive a new GRP BAT for the subgroups. After all groups have been identified this way, a BAT histogram operation can be used to obtain the counts of each data cube. Other aggregation operations using the MIL set aggregate construct `(bat)` can be used as well; note for instance that `histogram == (b.reverse())`.

The Monet interface module specification is shown below. Ideally we should defined stronger type constraints, e.g. command group.new(attr:bat[,:any_1]

The group macro is split along three dimensions:

[type:]	Type specific implementation for selecting the right hash function and data size etc.;
[clustered:]	The select the appropriate algorithm, i.e., with or without taking advantage of an order of values in the parent groups;
[physical properties:]	Values , choosing between a fixed predefined and a custom hashmask. Custom allows the user to determine the size of the hashmask (and indirectly the estimated size of the result). The hashmask is $2^n - 1$ where n is given by the user, or 1023 otherwise, and the derived result size is $4 \ldots 2^n$.

Further research should point out whether fitting a simple statistical model (possibly a simple mixture model) can help choose these parameters automatically; the current idea is that the user (which could be a domain-specific extension of the higher-level language) knows the properties of the data, especially for IR in which the standard grouping settings differ significantly from the original datamining application.

8.56 Lightweight Lock Module

This module provides simple SMP lock and thread functionality as already present in the MonetDB system.

This module provides simple SMP lock and thread functionality as already present in the MonetDB system.

8.57 The Transaction Logger

In the philosophy of MonetDB, transaction management overhead should only be paid when necessary. Transaction management is for this purpose implemented as a separate module and applications are required to obey the transaction policy, e.g. obtaining/releasing locks.

This module is designed to support efficient logging of the SQL database. Once loaded, the SQL compiler will insert the proper calls at transaction commit to include the changes in the log file.

The logger uses a directory to store its log files. One master log file stores information about the version of the logger and the transaction log files. This file is a simple ascii file with the following format: `6DIGIT-VERSION\n[log file number \n]*]*` The transaction log files have a binary format, which stores fixed size logformat headers (flag,nr,bid), where the flag is the type of update logged. The nr field indicates how many changes there were (in case of inserts/deletes). The bid stores the bid identifier.

The key decision to be made by the user is the location of the log file. Ideally, it should be stored in fail-safe environment, or at least the log and databases should be on separate disk columns.

This file system may reside on the same hardware as the database server and therefore the writes are done to the same disk, but could also reside on another system and then the changes are flushed through the network. The logger works under the assumption that it is called to safeguard updates on the database when it has an exclusive lock on the latest version. This lock should be guaranteed by the calling transaction manager first.

Finding the updates applied to a BAT is relatively easy, because each BAT contains a delta structure. On commit these changes are written to the log file and the delta management is reset. Since each commit is written to the same log file, the beginning and end are marked by a log identifier.

A server restart should only (re)process blocks which are completely written to disk. A log replay therefore ends in a commit or abort on the changed bats. Once all logs have been read, the changes to the bats are made persistent, i.e. a bbp sub-commit is done.

8.58 Multi-Attribute Equi-Join

8.59 Priority queues

This module includes functions for accessing and updating a pqueue. A pqueue is an (oid,any) bat. The tail is used as a comparison key. The first element of the pqueue is the smallest one in a min-pqueue or the largest one in a max-pqueue. Each element is larger than (smaller than) or equal to its parent which is defined by (position/2) if position is odd or (position-1)/2 if position is even (positions are from 0 to n-1). The head of the bat is used to keep track of the object-ids which are organized in the heap with respect to their values (tail column).

8.60 System state information

This document introduces a series of bats and operations that provide access to information stored within the Monet Version 5 internal data structures. In all cases, pseudo BAT operation returns a transient BAT that should be garbage collected after being used.

The main performance drain would be to use a pseudo BAT directly to successively access it components. This can be avoided by first assigning the pseudo BAT to a variable.

8.61 Unix standard library calls

The unix module is currently of rather limited size. It should include only those facilities that are UNIX specific, i.e. not portable to other platforms. Similar modules may be defined for Windows platforms.

9 Application Programming Interfaces

MonetDB comes with a complete set of programming libraries. Their basis is the MonetDB application programming interface (Mapi), which describes the protocol understood by the server. The Perl, PHP, and Python libraries are mostly wrappers around the Mapi routines.

The programming interface is based on a client-server architecture, where the client program connects to a server using a TCP/IP connection to exchange commands and receives answers. The underlying protocol uses plain UTF-8 data for ease of use and debugging. This leads to publicly visible information exchange over a network, which may be undesirable. Therefore, a private and secure channel can be set up with the Secure Socket Layer functionality.

A more tightly connection between application logic and database server is described in Section 1.77.6 [Embedded Server], page 39.

9.1 The Mapi Library

The easiest way to extend the functionality of MonetDB is to construct an independent application, which communicates with a running server using a database driver with a simple API and a textual protocol. The effectiveness of such an approach has been demonstrated by the wide use of database API implementations, such as Perl DBI, PHP, ODBC,...

9.1.1 Sample MAPI Application

The database driver implementation given in this document focuses on developing applications in C. The command collection has been chosen to align with common practice, i.e. queries follow a prepare, execute, and fetch_row paradigm. The output is considered a regular table. An example of a mini application below illustrates the main operations.

```
#include <mapilib/Mapi.h>
#include <stdio.h>
#include <stdlib.h>

void die(Mapi dbh, MapiHdl hdl)
{
        if (hdl != NULL) {
                mapi_explain_query(hdl, stderr);
                do {
                        if (mapi_result_error(hdl) != NULL)
                                mapi_explain_result(hdl, stderr);
                } while (mapi_next_result(hdl) == 1);
                mapi_close_handle(hdl);
                mapi_destroy(dbh);
        } else if (dbh != NULL) {
                mapi_explain(dbh, stderr);
                mapi_destroy(dbh);
        } else {
                fprintf(stderr, "command failed\n");
        }
        exit(-1);
}

MapiHdl query(Mapi dbh, char *q)
{
        MapiHdl ret = NULL;
        if ((ret = mapi_query(dbh, q)) == NULL || mapi_error(dbh) != MOK)
                die(dbh, ret);
        return(ret);
}

void update(Mapi dbh, char *q)
{
        MapiHdl ret = query(dbh, q);
        if (mapi_close_handle(ret) != MOK)
                die(dbh, ret);
}

int main(int argc, char *argv[])
{
    Mapi dbh;
    MapiHdl hdl = NULL;
        char *name;
        char *age;

    dbh = mapi_connect("localhost", 50000, "monetdb", "monetdb", "sql", "demo");
    if (mapi_error(dbh))
        die(dbh, hdl);

        update(dbh, "CREATE TABLE emp (name VARCHAR(20), age INT)");
        update(dbh, "INSERT INTO emp VALUES ('John', 23)");
        update(dbh, "INSERT INTO emp VALUES ('Mary', 22)");

        hdl = query(dbh, "SELECT * FROM emp");

    while (mapi_fetch_row(hdl)) {
        name = mapi_fetch_field(hdl, 0);
        age = mapi_fetch_field(hdl, 1);
        printf("%s is %s\n", name, age);
    }

    mapi_close_handle(hdl);
    mapi_destroy(dbh);

    return(0);
}
```

The `mapi_connect()` operation establishes a communication channel with a running server. The query language interface is either "sql", "mil" or "xquery".

Errors on the interaction can be captured using `mapi_error()`, possibly followed by a request to dump a short error message explanation on a standard file location. It has been abstracted away in a macro.

Provided we can establish a connection, the interaction proceeds as in many similar application development packages. Queries are shipped for execution using `mapi_query()` and an answer table can be consumed one row at a time. In many cases these functions suffice.

The Mapi interface provides caching of rows at the client side. `mapi_query()` will load tuples into the cache, after which they can be read repeatedly using `mapi_fetch_row()` or directly accessed (`mapi_seek_row()`). This facility is particularly handy when small, but stable query results are repeatedly used in the client program.

To ease communication between application code and the cache entries, the user can bind the C-variables both for input and output to the query parameters, and output columns, respectively. The query parameters are indicated by '?' and may appear anywhere in the query template.

The Mapi library expects complete lines from the server as answers to query actions. Incomplete lines leads to Mapi waiting forever on the server. Thus formatted printing is discouraged in favor of tabular printing as offered by the `table.print()` commands.

The following action is needed to get a working program. Compilation of the application relies on the *monetdb-config* program shipped with the distribution. It localizes the include files and library directories. Once properly installed, the application can be compiled and linked as follows:

```
cc sample.c `monetdb-clients-config --cflags --libs` -lMapi -o sample
./sample
```

It assumes that the dynamic loadable libraries are in public places. If, however, the system is installed in your private environment then the following option can be used on most ELF platforms.

```
cc sample.c `monetdb-clients-config --cflags --libs` -lMapi -o sample \
`monetdb-clients-config --libs | sed -e's:-L:-R:g'`
./sample
```

The compilation on Windows is slightly more complicated. It requires more attention towards the location of the include files and libraries.

9.1.2 Command Summary

The quick reference guide to the Mapi library is given below. More details on their constraints and defaults are given in the next section.

mapi_bind()	Bind string C-variable to a field
mapi_bind_numeric()	Bind numeric C-variable to field
mapi_bind_var()	Bind typed C-variable to a field
mapi_cache_freeup()	Forcefully shuffle fraction for cache refreshment
mapi_cache_limit()	Set the tuple cache limit
mapi_cache_shuffle()	Set shuffle fraction for cache refreshment

mapi_clear_bindings()	Clear all field bindings
mapi_clear_params()	Clear all parameter bindings
mapi_close_handle()	Close query handle and free resources
mapi_connect()	Connect to a Mserver
mapi_destroy()	Free handle resources
mapi_disconnect()	Disconnect from server
mapi_error()	Test for error occurrence
mapi_execute()	Execute a query
mapi_execute_array()	Execute a query using string arguments
mapi_explain()	Display error message and context on stream
mapi_explain_query()	Display error message and context on stream
mapi_fetch_all_rows()	Fetch all answers from server into cache
mapi_fetch_field()	Fetch a field from the current row
mapi_fetch_field_array()	Fetch all fields from the current row
mapi_fetch_line()	Retrieve the next line
mapi_fetch_reset()	Set the cache reader to the beginning
mapi_fetch_row()	Fetch row of values
mapi_finish()	Terminate the current query
mapi_get_dbname()	Database being served
mapi_get_field_count()	Number of fields in current row
mapi_get_host()	Host name of server
mapi_get_query()	Query being executed
mapi_get_language()	Query language name
mapi_get_mapi_version()	Mapi version name
mapi_get_monet_versionId()	MonetDB version identifier
mapi_get_monet_version()	MonetDB version name
mapi_get_motd()	Get server welcome message
mapi_get_row_count()	Number of rows in cache or -1
mapi_get_last_id()	last inserted id of an auto_increment (or alike) column
mapi_get_trace()	Get trace flag
mapi_get_user()	Current user name
mapi_log()	Keep log of client/server interaction
mapi_next_result()	Go to next result set
mapi_needmore()	Return whether more data is needed
mapi_ping()	Test server for accessibility
mapi_prepare()	Prepare a query for execution
mapi_prepare_array()	Prepare a query for execution using arguments
mapi_query()	Send a query for execution
mapi_query_array()	Send a query for execution with arguments
mapi_query_handle()	Send a query for execution
mapi_quick_query_array()	Send a query for execution with arguments
mapi_quick_query()	Send a query for execution
mapi_quick_response()	Quick pass response to stream
mapi_quote()	Escape characters
mapi_reconnect()	Reconnect with a clean session context
mapi_rows_affected()	Obtain number of rows changed
mapi_seek_row()	Move row reader to specific location in cache

mapi_setAutocommit()	Set auto-commit flag
mapi_setAlgebra()	Use algebra backend
mapi_stream_query()	Send query and prepare for reading tuple stream
mapi_table()	Get current table name
mapi_timeout()	Set timeout for long-running queries[TODO]
mapi_output()	Set output format
mapi_stream_into()	Stream document into server
mapi_profile()	Set profile flag
mapi_trace()	Set trace flag
mapi_virtual_result()	Submit a virtual result set
mapi_unquote()	remove escaped characters

9.1.3 Library Synopsis

The routines to build a MonetDB application are grouped in the library MonetDB Programming Interface, or shorthand Mapi.

The protocol information is stored in a Mapi interface descriptor (mid). This descriptor can be used to ship queries, which return a MapiHdl to represent the query answer. The application can set up several channels with the same or a different `mserver`. It is the programmer's responsibility not to mix the descriptors in retrieving the results.

The application may be multi-threaded as long as the user respects the individual connections represented by the database handlers.

The interface assumes a cautious user, who understands and has experience with the query or programming language model. It should also be clear that references returned by the API point directly into the administrative structures of Mapi. This means that they are valid only for a short period, mostly between successive `mapi_fetch_row()` commands. It also means that it the values are to retained, they have to be copied. A defensive programming style is advised.

Upon an error, the routines `mapi_explain()` and `mapi_explain_query()` give information about the context of the failed call, including the expression shipped and any response received. The side-effect is clearing the error status.

9.1.4 Error Message

Almost every call can fail since the connection with the database server can fail at any time. Functions that return a handle (either `Mapi` or `MapiHdl`) may return NULL on failure, or they may return the handle with the error flag set. If the function returns a non-NULL handle, always check for errors with mapi_error.

Functions that return MapiMsg indicate success and failure with the following codes.

MOK	No error
MERROR	Mapi internal error.
MTIMEOUT	Error communicating with the server.

When these functions return MERROR or MTIMEOUT, an explanation of the error can be had by calling one of the functions `mapi_error_str()`, `mapi_explain()`, or `mapi_explain_query()`.

To check for error messages from the server, call `mapi_result_error()`. This function returns NULL if there was no error, or the error message if there was. A user-friendly message can be printed using `map_explain_result()`. Typical usage is:

```
do {
    if ((error = mapi_result_error(hdl)) != NULL)
        mapi_explain_result(hdl, stderr);
    while ((line = mapi_fetch_line(hdl)) != NULL)
        /* use output */;
} while (mapi_next_result(hdl) == 1);
```

9.1.5 Mapi Function Reference

9.1.6 Connecting and Disconnecting

- Mapi mapi_connect(const char *host, int port, const char *username, const char *password, const char *lang, const char *dbname)

  Setup a connection with a Mserver at a *host*:*port* and login with *username* and *password*. If host == NULL, the local host is accessed. If host starts with a '/' and the system supports it, host is actually the name of a UNIX domain socket, and port is ignored. If port == 0, a default port is used. If username == NULL, the username of the owner of the client application containing the Mapi code is used. If password == NULL, the password is omitted. The preferred query language is any of {sql,mil,mal,xquery }. On success, the function returns a pointer to a structure with administration about the connection.
- MapiMsg mapi_disconnect(Mapi mid)

  Terminate the session described by *mid*. The only possible uses of the handle after this call is *mapi_destroy()* and `mapi_reconnect()`. Other uses lead to failure.
- MapiMsg mapi_destroy(Mapi mid)

  Terminate the session described by *mid* if not already done so, and free all resources. The handle cannot be used anymore.
- MapiMsg mapi_reconnect(Mapi mid)

  Close the current channel (if still open) and re-establish a fresh connection. This will remove all global session variables.
- MapiMsg mapi_ping(Mapi mid)

  Test availability of the server. Returns zero upon success.

9.1.7 Sending Queries

- MapiHdl mapi_query(Mapi mid, const char *Command)

  Send the Command to the database server represented by mid. This function returns a query handle with which the results of the query can be retrieved. The handle should be closed with `mapi_close_handle()`. The command response is buffered for consumption, c.f. mapi_fetch_row().
- MapiMsg mapi_query_handle(MapiHdl hdl, const char *Command)

  Send the Command to the database server represented by hdl, reusing the handle from a previous query. If Command is zero it takes the last query string kept around. The command response is buffered for consumption, e.g. `mapi_fetch_row()`.

- MapiHdl mapi_query_array(Mapi mid, const char *Command, char **argv)

  Send the Command to the database server replacing the placeholders (?) by the string arguments presented.
- MapiHdl mapi_quick_query(Mapi mid, const char *Command, FILE *fd)

  Similar to `mapi_query()`, except that the response of the server is copied immediately to the file indicated.
- MapiHdl mapi_quick_query_array(Mapi mid, const char *Command, char **argv, FILE *fd)

  Similar to `mapi_query_array()`, except that the response of the server is not analyzed, but shipped immediately to the file indicated.
- MapiHdl mapi_stream_query(Mapi mid, const char *Command, int windowsize)

  Send the request for processing and fetch a limited number of tuples (determined by the window size) to assess any erroneous situation. Thereafter, prepare for continual reading of tuples from the stream, until an error occurs. Each time a tuple arrives, the cache is shifted one.
- MapiHdl mapi_prepare(Mapi mid, const char *Command)

  Move the query to a newly allocated query handle (which is returned). Possibly interact with the back-end to prepare the query for execution.
- MapiMsg mapi_execute(MapiHdl hdl)

  Ship a previously prepared command to the backend for execution. A single answer is pre-fetched to detect any runtime error. MOK is returned upon success.
- MapiMsg mapi_execute_array(MapiHdl hdl, char **argv)

  Similar to `mapi\_execute` but replacing the placeholders for the string values provided.
- MapiMsg mapi_finish(MapiHdl hdl)

  Terminate a query. This routine is used in the rare cases that consumption of the tuple stream produced should be prematurely terminated. It is automatically called when a new query using the same query handle is shipped to the database and when the query handle is closed with `mapi_close_handle()`.
- MapiMsg mapi_virtual_result(MapiHdl hdl, int columns, const char **columnnames, const char **columntypes, const int *columnlengths, int tuplecount, const char ***tuples)

  Submit a table of results to the library that can then subsequently be accessed as if it came from the server. columns is the number of columns of the result set and must be greater than zero. columnnames is a list of pointers to strings giving the names of the individual columns. Each pointer may be NULL and columnnames may be NULL if there are no names. tuplecount is the length (number of rows) of the result set. If tuplecount is less than zero, the number of rows is determined by a NULL pointer in the list of tuples pointers. tuples is a list of pointers to row values. Each row value is a list of pointers to strings giving the individual results. If one of these pointers is NULL it indicates a NULL/nil value.

9.1.8 Getting Results

- int mapi_get_field_count(MapiHdl mid)

  Return the number of fields in the current row.

- mapi_int64 mapi_get_row_count(MapiHdl mid)

  If possible, return the number of rows in the last select call. A -1 is returned if this information is not available.
- mapi_int64 mapi_get_last_id(MapiHdl mid)

  If possible, return the last inserted id of auto_increment (or alike) column. A -1 is returned if this information is not available. We restrict this to single row inserts and one auto_increment column per table. If the restrictions do not hold, the result is unspecified.
- mapi_int64 mapi_rows_affected(MapiHdl hdl)

  Return the number of rows affected by a database update command such as SQL's INSERT/DELETE/UPDATE statements.
- int mapi_fetch_row(MapiHdl hdl)

  Retrieve a row from the server. The text retrieved is kept around in a buffer linked with the query handle from which selective fields can be extracted. It returns the number of fields recognized. A zero is returned upon encountering end of sequence or error. This can be analyzed in using `mapi_error()`.
- mapi_int64 mapi_fetch_all_rows(MapiHdl hdl)

  All rows are cached at the client side first. Subsequent calls to `mapi_fetch_row()` will take the row from the cache. The number or rows cached is returned.
- int mapi_quick_response(MapiHdl hdl, FILE *fd)

  Read the answer to a query and pass the results verbatim to a stream. The result is not analyzed or cached.
- MapiMsg mapi_seek_row(MapiHdl hdl, mapi_int64 rownr, int whence)

  Reset the row pointer to the requested row number. If whence is `MAPI_SEEK_SET`, rownr is the absolute row number (0 being the first row); if whence is `MAPI_SEEK_CUR`, rownr is relative to the current row; if whence is `MAPI_SEEK_END`, rownr is relative to the last row.
- MapiMsg mapi_fetch_reset(MapiHdl hdl)

  Reset the row pointer to the first line in the cache. This need not be a tuple. This is mostly used in combination with fetching all tuples at once.
- char **mapi_fetch_field_array(MapiHdl hdl)

  Return an array of string pointers to the individual fields. A zero is returned upon encountering end of sequence or error. This can be analyzed in using `mapi\_error()`.
- char *mapi_fetch_field(MapiHdl hdl, int fnr)

  Return a pointer a C-string representation of the value returned. A zero is returned upon encountering an error or when the database value is NULL; this can be analyzed in using `mapi\_error()`.
- MapiMsg mapi_next_result(MapiHdl hdl)

  Go to the next result set, discarding the rest of the output of the current result set.

9.1.9 Errors

- MapiMsg mapi_error(Mapi mid)

  Return the last error code or 0 if there is no error.

- char *mapi_error_str(Mapi mid)

  Return a pointer to the last error message.
- char *mapi_result_error(MapiHdl hdl)

  Return a pointer to the last error message from the server.
- MapiMsg mapi_explain(Mapi mid, FILE *fd)

  Write the error message obtained from `mserver` to a file.
- MapiMsg mapi_explain_query(MapiHdl hdl, FILE *fd)

  Write the error message obtained from `mserver` to a file.
- MapiMsg mapi_explain_result(MapiHdl hdl, FILE *fd)

  Write the error message obtained from `mserver` to a file.

9.1.10 Parameters

- MapiMsg mapi_bind(MapiHdl hdl, int fldnr, char **val)

  Bind a string variable with a field in the return table. Upon a successful subsequent `mapi\_fetch\_row()` the indicated field is stored in the space pointed to by val. Returns an error if the field identified does not exist.
- MapiMsg mapi_bind_var(MapiHdl hdl, int fldnr, int type, void *val)

  Bind a variable to a field in the return table. Upon a successful subsequent `mapi\_fetch\_row()`, the indicated field is converted to the given type and stored in the space pointed to by val. The types recognized are `{ MAPI\_TINY, MAPI\_UTINY, MAPI\_SHORT, MAPI\_USHORT, MAPI_INT, MAPI_UINT, MAPI_LONG, MAPI_ULONG, MAPI_LONGLONG, MAPI_ULONGLONG, MAPI_CHAR, MAPI_VARCHAR, MAPI_FLOAT, MAPI_DOUBLE, MAPI_DATE, MAPI_TIME, MAPI_DATETIME }`. The binding operations should be performed after the mapi_execute command. Subsequently all rows being fetched also involve delivery of the field values in the C-variables using proper conversion. For variable length strings a pointer is set into the cache.
- MapiMsg mapi_bind_numeric(MapiHdl hdl, int fldnr, int scale, int precision, void *val)

  Bind to a numeric variable, internally represented by MAPI_INT Describe the location of a numeric parameter in a query template.
- MapiMsg mapi_clear_bindings(MapiHdl hdl)

  Clear all field bindings.
- MapiMsg mapi_param(MapiHdl hdl, int fldnr, char **val)

  Bind a string variable with the n-th placeholder in the query template. No conversion takes place.
- MapiMsg mapi_param_type(MapiHdl hdl, int fldnr, int ctype, int sqltype, void *val)

  Bind a variable whose type is described by ctype to a parameter whose type is described by sqltype.
- MapiMsg mapi_param_numeric(MapiHdl hdl, int fldnr, int scale, int precision, void *val)

  Bind to a numeric variable, internally represented by MAPI_INT.

- MapiMsg mapi_param_string(MapiHdl hdl, int fldnr, int sqltype, char *val, int *sizeptr)

  Bind a string variable, internally represented by MAPI_VARCHAR, to a parameter. The sizeptr parameter points to the length of the string pointed to by val. If sizeptr == NULL or *sizeptr == -1, the string is NULL-terminated.
- MapiMsg mapi_clear_params(MapiHdl hdl)

  Clear all parameter bindings.

9.1.11 Miscellaneous

- MapiMsg mapi_setAutocommit(Mapi mid, int autocommit)

  Set the autocommit flag (default is on). This only has an effect when the language is SQL. In that case, the server commits after each statement sent to the server.
- MapiMsg mapi_setAlgebra(Mapi mid, int algebra)

  Tell the backend to use or stop using the algebra-based compiler.
- MapiMsg mapi_cache_limit(Mapi mid, int maxrows)

  A limited number of tuples are pre-fetched after each `execute()`. If maxrows is negative, all rows will be fetched before the application is permitted to continue. Once the cache is filled, a number of tuples are shuffled to make room for new ones, but taking into account non-read elements. Filling the cache quicker than reading leads to an error.
- MapiMsg mapi_cache_shuffle(MapiHdl hdl, int percentage)

  Make room in the cache by shuffling percentage tuples out of the cache. It is sometimes handy to do so, for example, when your application is stream-based and you process each tuple as it arrives and still need a limited look-back. This percentage can be set between 0 to 100. Making shuffle= 100% (default) leads to paging behavior, while shuffle==1 leads to a sliding window over a tuple stream with 1% refreshing.
- MapiMsg mapi_cache_freeup(MapiHdl hdl, int percentage)

  Forcefully shuffle the cache making room for new rows. It ignores the read counter, so rows may be lost.
- char * mapi_quote(const char *str, int size)

  Escape special characters such as `\n`, `\t` in str with backslashes. The returned value is a newly allocated string which should be freed by the caller.
- char * mapi_unquote(const char *name)

  The reverse action of `mapi_quote()`, turning the database representation into a C-representation. The storage space is dynamically created and should be freed after use.
- MapiMsg mapi_output(Mapi mid, char *output)

  Set the output format for results send by the server.
- MapiMsg mapi_stream_into(Mapi mid, char *docname, char *colname, FILE *fp)

  Stream a document into the server. The name of the document is specified in docname, the collection is optionally specified in colname (if NULL, it defaults to docname), and the content of the document comes from fp.
- MapiMsg mapi_profile(Mapi mid, int flag)

  Set the profile flag to time commands send to the server.

- MapiMsg mapi_trace(Mapi mid, int flag)

  Set the trace flag to monitor interaction of the client with the library. It is primarilly used for debugging Mapi applications.
- int mapi_get_trace(Mapi mid)

  Return the current value of the trace flag.
- MapiMsg mapi_log(Mapi mid, const char *fname)

  Log the interaction between the client and server for offline inspection. Beware that the log file overwrites any previous log. For detailed interaction trace with the Mapi library itself use mapi_trace().

The remaining operations are wrappers around the data structures maintained. Note that column properties are derived from the table output returned from the server.

- char *mapi_get_name(MapiHdl hdl, int fnr)
- char *mapi_get_type(MapiHdl hdl, int fnr)
- char *mapi_get_table(MapiHdl hdl, int fnr)
- int mapi_get_len(Mapi mid, int fnr)
- char *mapi_get_dbname(Mapi mid)
- char *mapi_get_host(Mapi mid)
- char *mapi_get_user(Mapi mid)
- char *mapi_get_lang(Mapi mid)
- char *mapi_get_motd(Mapi mid)

9.2 MonetDB Perl Library

Perl is one of the more common scripting languages for which a 'standard' database application programming interface is defined. It is called DBI and it was designed to protect you from the API library details of multiple DBMS vendors. It has a very simple interface to execute SQL queries and for processing the results sent back. DBI doesn't know how to talk to any particular database, but it does know how to locate and load in DBD ('Database Driver') modules. The DBD modules encapsulate the interface library's intricacies and knows how to talk to the real databases.

MonetDB comes with its own DBD module which is included in both the source and binary distribution packages. The module is also available via CPAN.

Two sample Perl applications are included in the source distribution; a MIL session and a simple client to interact with a running server.

For further documentation we refer to the Perl community home page.

9.2.1 A Simple Perl Example

```
use strict;
use warnings;
use DBI();

print "\nStart a simple Monet MIL interaction\n\n";
```

```
# determine the data sources:
my @ds = DBI->data_sources('monetdb');
print "data sources: @ds\n";

# connect to the database:
my $dsn = 'dbi:monetdb:database=test;host=localhost;port=50000;language=mil';
my $dbh = DBI->connect( $dsn,
  undef, undef,  # no authentication in MIL
  { PrintError => 0, RaiseError => 1 }  # turn on exception handling
);
{
  # simple MIL statement:
  my $sth = $dbh->prepare('print(2);');
  $sth->execute;
  my @row = $sth->fetchrow_array;
  print "field[0]: $row[0], last index: $#row\n";
}
{
  my $sth = $dbh->prepare('print(3);');
  $sth->execute;
  my @row = $sth->fetchrow_array;
  print "field[0]: $row[0], last index: $#row\n";
}
{
  # deliberately executing a wrong MIL statement:
  my $sth = $dbh->prepare('( xyz 1);');
  eval { $sth->execute }; print "ERROR REPORTED: $@" if $@;
}
$dbh->do('var b:=new(int,str);');
$dbh->do('insert(b,3,"three");');
{
  # variable binding stuff:
  my $sth = $dbh->prepare('insert(b,?,?);');
  $sth->bind_param( 1,     7 , DBI::SQL_INTEGER() );
  $sth->bind_param( 2,'seven' );
  $sth->execute;
}
{
  my $sth = $dbh->prepare('print(b);');
  # get all rows one at a time:
  $sth->execute;
  while ( my $row = $sth->fetch ) {
    print "bun: $row->[0], $row->[1]\n";
  }
  # get all rows at once:
  $sth->execute;
  my $t = $sth->fetchall_arrayref;
  my $r = @$t;         # row count
  my $f = @{$t->[0]};  # field count
  print "rows: $r, fields: $f\n";
  for my $i ( 0 .. $r-1 ) {
    for my $j ( 0 .. $f-1 ) {
      print "field[$i,$j]: $t->[$i][$j]\n";
    }
  }
}
{
  # get values of the first column from each row:
  my $row = $dbh->selectcol_arrayref('print(b);');
  print "head[$_]: $row->[$_]\n" for 0 .. 1;
}
{
  my @row = $dbh->selectrow_array('print(b);');
  print "field[0]: $row[0]\n";
  print "field[1]: $row[1]\n";
}
{
  my $row = $dbh->selectrow_arrayref('print(b);');
  print "field[0]: $row->[0]\n";
  print "field[1]: $row->[1]\n";
}
$dbh->disconnect;
print "\nFinished\n";
```

9.3 MonetDB PHP Library

The MonetDB distribution comes with a MAPI based PHP interface For general compilation of MonetDB see the howto's for Unix and Linux. The unix configure process normally tries to detect if you have PHP including developer packages installed and builds the PHP module only if you have it. With the –WITH-PHP option you could tell 'configure' where to find the PHP installation.

When the build process is complete you should have a PHP extension_dir under your MonetDB prefix directory. Usually this is PREFIX/LIB(64)/PHP5. It contains the loadable php module, *lib/php5/monetdb.dll*.

Depending on your local setup you could now use these files by coping them into the system extension_dir or with a private webserver you could simply reset the environment variables, include_path and extension_dir. For example you could have a php.ini file which has the following php section.

```
[PHP]
safe_mode = Off
safe_mode_gid = Off
extension_dir = /opt/MonetDB-4.6/lib(64)/php4
```

9.3.1 A Simple PHP Example

A tiny example of the use the MonetDB PHP module follows:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml"" xml:lang="en" lang="en">

<head>
        <title>MonetDB Query</title>
</head>

<body>
<?php
        if ( isset($_POST['query']) )
        {
                $db = monetdb_connect('sql', 'localhost', 50000, 'monetdb',
'monetdb')
                                or die('Failed to connect to MonetDB<br>');

                $sql = stripslashes($_POST['query']);
                $res = monetdb_query($sql);
                while ( $row = monetdb_fetch_assoc($res) )
                {
                        print "<pre>\n";
                        print_r($row);
                        print "</pre>\n";
                }
        }

        print "<form method=\"post\" action=\"{$_SERVER['PHP_SELF']}\">\n";
        print "<label for=\"query\">SQL Query:</label>\n";
        print "<input type=\"text\" name=\"query\" id=\"query\"
value=\"{$_POST['query']}\" />\n";
        print "<input type=\"submit\" value=\"Execute\" />\n";
        print "</form>\n";
?>
</body>

</html>
```

More examples can be found in the sources.

The PHP module is aligned with the PostgreSQL implementation. A synopsis of the operations provided:

- proto resource monetdb_connect([string host [, string port [, string username [, string password [, string language]]]]]) Open a MonetDB connection
- proto resource monetdb_pconnect([string host [, string port [, string username [, string

password [, string language]]]]]) Open a persistent MonetDB connection

- proto bool monetdb_close([resource connection]) Close a MonetDB connection
- proto string monetdb_dbname([resource connection]) Get the database name
- proto string monetdb_last_error([resource connection]) Get the error message string
- proto string monetdb_host([resource connection]) Returns the host name associated with the connection
- proto array monetdb_version([resource connection]) Returns an array with client, protocol and server version (when available)
- proto bool monetdb_ping([resource connection]) Ping database. If connection is bad, try to reconnect.
- proto resource monetdb_query([resource connection,] string query) Execute a query
- proto resource monetdb_query_params([resource connection,] string query, array params) Execute a query
- proto resource monetdb_prepare([resource connection,] string stmtname, string query) Prepare a query for future execution
- proto resource monetdb_execute([resource connection,] string stmtname, array params) Execute a prepared query
- proto int monetdb_num_rows(resource result) Return the number of rows in the result
- proto int monetdb_num_fields(resource result) Return the number of fields in the result
- proto int monetdb_affected_rows(resource result) Returns the number of affected tuples
- proto string pg_last_notice(resource connection) Returns the last notice set by the back-end
- proto string monetdb_field_name(resource result, int field_number) Returns the name of the field
- proto string monetdb_field_table(resource result, int field_number) Returns the name of the table field belongs to
- proto string monetdb_field_type(resource result, int field_number) Returns the type of the field
- proto int monetdb_field_num(resource result, string field_name) Returns the field number of the named field
- proto mixed monetdb_fetch_result(resource result, [int row_number,] mixed field_name) Returns values from a result identifier
- proto array monetdb_fetch_row(resource result [, int row [, int result_type]]) Get a row as an enumerated array
- proto array monetdb_fetch_assoc(resource result [, int row]) Fetch a row as an assoc array
- proto array monetdb_fetch_array(resource result [, int row [, int result_type]]) Fetch a row as an array
- proto object monetdb_fetch_object(resource result [, int row [, string class_name [, NULL|array ctor_params]]]) Fetch a row as an object
- proto bool monetdb_result_seek(resource result, int offset) Set internal row offset

- proto int monetdb_field_prtlen(resource result, [int row,] mixed field_name_or_number) Returns the printed length
- proto int monetdb_field_is_null(resource result, [int row,] mixed field_name_or_number) Test if a field is NULL
- proto bool monetdb_free_result(resource result) Free result memory
- proto string monetdb_escape_string(string data) Escape string for text/char type
- proto int monetdb_connection_status(resource connnection) Get connection status
- proto bool monetdb_connection_reset(resource connection) Reset connection (reconnect)
- proto bool monetdb_put_line([resource connection,] string query) Send null-terminated string to back-end server
- proto bool monetdb_end_copy([resource connection]) Sync with back-end. Completes the Copy command
- proto array monetdb_copy_to(resource connection, string table_name [, string delimiter [, string null_as]]) Copy table to array
- proto bool monetdb_copy_from(resource connection, string table_name , array rows [, string delimiter [, string null_as]]) Copy table from array
- proto bool monetdb_connection_busy(resource connection) Get connection is busy or not
- proto bool monetdb_send_query(resource connection, string query) Send asynchronous query
- proto bool monetdb_send_query_params(resource connection, string query) Send asynchronous parameterized query
- proto bool monetdb_send_prepare(resource connection, string stmtname, string query) Asynchronously prepare a query for future execution
- proto bool monetdb_send_execute(resource connection, string stmtname, array params) Executes prevriously prepared stmtname asynchronously
- proto resource monetdb_get_result(resource connection) Get asynchronous query result
- proto mixed monetdb_result_status(resource result[, long result_type]) Get status of query result
- proto array monetdb_meta_data(resource db, string table) Get meta_data
- proto array monetdb_convert(resource db, string table, array values[, int options]) Check and convert values for MonetDB SQL statement
- proto mixed monetdb_insert(resource db, string table, array values[, int options]) Insert values (filed=>value) to table
- proto mixed monetdb_update(resource db, string table, array fields, array ids[, int options]) Update table using values (field=>value) and ids (id=>value)
- proto mixed monetdb_delete(resource db, string table, array ids[, int options]) Delete records has ids (id=>value)
- proto mixed monetdb_select(resource db, string table, array ids[, int options]) Select records that has ids (id=>value)

9.4 The MonetDB MAPI and SQL client python API

9.5 Introduction

This is the new native python client API. This API is cross-platform, and doesn't depend on any monetdb libraries. It has support for python 2.5, 2.6 and 3.0 and is Python DBAPI 2.0 compatible.

9.6 Changes

A number of things are different compared to the old version that uses the mapi library:

- No dependecies on MonetDB libraries anymore
- MAPI protocol is now implemented in pure python
- Added unit tests for the SQL API
- The MAPI module is now named monetdb.mapi
- The SQL module is now named monetdb.sql
- Small changes in argument names for functions
- Type conversion is working (for example a monetdb int becomes a python int)
- If you want to use the mapi module with python3.* you should use the mapi3.py module (import monetdb.mapi3 as mapi). This is done automatically for the sql module
- Dropped support for the dictionary based cursor

9.7 Installation

To install the MonetDB python API run the following command from the python source directory:

```
# python setup.py install
```

That's all, now you are ready to start using the API.

9.8 Documentation

The python code is well documented, so if you need to find documentation you should have a look at the source code. Below is an interactive example on how to use the monetdb SQL API which should get you started quite fast.

9.9 Examples

There are some examples in the 'examples' folder, but here are is a line by line example of the SQL API:

```
> # import the SQL module
> import monetdb.sql
>
> # set up a connection. arguments below are the defaults
> connection = monetdb.sql.connect(username="monetdb", password="monetdb", hostname="localh
>
> # create a cursor
> cursor = connection.cursor()
>
> # increase the rows fetched to increase performance (optional)
> cursor.arraysize = 100
>
> # execute a query (return the number of rows to fetch)
> cursor.execute('SELECT * FROM tables')
26
>
> # fetch only one row
> cursor.fetchone()
[1062, 'schemas', 1061, None, 0, True, 0, 0]
>
> # fetch the remaining rows
> cursor.fetchall()
[[1067, 'types', 1061, None, 0, True, 0, 0],
 [1076, 'functions', 1061, None, 0, True, 0, 0],
 [1085, 'args', 1061, None, 0, True, 0, 0],
 [1093, 'sequences', 1061, None, 0, True, 0, 0],
 [1103, 'dependencies', 1061, None, 0, True, 0, 0],
 [1107, 'connections', 1061, None, 0, True, 0, 0],
 [1116, '_tables', 1061, None, 0, True, 0, 0],
 ...
 [4141, 'user_role', 1061, None, 0, True, 0, 0],
 [4144, 'auths', 1061, None, 0, True, 0, 0],
 [4148, 'privileges', 1061, None, 0, True, 0, 0]]
>
> # Show the table meta data
> cursor.description
[('id', 'int', 4, 4, None, None, None),
 ('name', 'varchar', 12, 12, None, None, None),
 ('schema_id', 'int', 4, 4, None, None, None),
 ('query', 'varchar', 168, 168, None, None, None),
 ('type', 'smallint', 1, 1, None, None, None),
 ('system', 'boolean', 5, 5, None, None, None),
 ('commit_action', 'smallint', 1, 1, None, None, None),
 ('temporary', 'tinyint', 1, 1, None, None, None)]
```

If you would like to communicate with the database at a lower level you can use the MAPI library:

```
> # If you use python 2.6, python 3.0 or higher:
> from monetdb import mapi
>
> # If you use python2.5
> from monetdb import mapi25 as mapi
>
> server = mapi.Server()
> server.connect(hostname="localhost", port=50000, username="monetdb", password="monetdb",
> server.cmd("sSELECT * FROM tables;")
...
```

9.10 MonetDB JDBC Driver

The most obvious way to connect to a data source using the Java programming language is by making use of the in Java defined JDBC framework. MonetDB has a native Java JDBC driver type 4 which allows use of the MonetDB database in a Java alike way.

It is quite difficult to have a fully complete JDBC implementation. Also this driver isn't complete in that sense. However, it is believed that the most prominent parts of the JDBC interface are implemented, and in such a way that they adhere to the specifications. If you make extensive use of JDBC semantics and rely on many of its features, please read the release notes which are to be found in the SRC/JDBC directory of the sql CVS tree.

This document gives a short description how to use the MonetDB JDBC driver in Java applications. A familiarity with the Java JDBC API is required to fully understand this document. Please note that you can find the complete JDBC API on Sun's web site HTTP://JAVA.SUN.COM/.

In order to use the MonetDB JDBC driver in Java applications you need (of course) a running MonetDB/SQL instance, preferably via merovingian.

9.10.1 Getting the driver Jar

The easiest way to acquire the driver is to download it from our download page. You will find a file called called MONETDB-X.Y-JDBC.JAR where X and Y are major and minor version numbers. The current release as of this writing is 1.11.

9.10.2 Compiling the driver (using ant, optional)

If you prefer to build the driver yourself, make sure you acquire the MonetDB Java repository, e.g. as part of the Super Source Tarball. The Java sources are built using Apache's Ant tool. Simply issuing the command ANT DISTJDBC should be sufficient to build the driver jar-archive in the subdirectory JARS. See the ANT web site for more documentation on the ant build-tool: HTTP://ANT.APACHE.ORG/. The Java sources require at least a Java 2 platform 1.4 compatible compiler. The JDBC driver, however, currently cannot be compiled with a Java 1.6 or up compiler.

9.10.3 Testing the driver using the JdbcClient utility

Before you start developing your programs which use the MonetDB JDBC driver it is generally a good idea to check if the driver actually works in your environment. JdbcClient is a no longer distributed, but when compling from sources, it is still built and put in the JARS directory. Follow the steps below to assure your setup is complete:

1. start merovingian
2. create a database using MONETDB CREATE MYTEST
3. run the JdbcClient utility using JAVA -JAR {PATH/TO/JDBCCLIENT.JAR} -DMYTEST -UMONETDB (with password monetdb)

The last step should give you something like this:

```
% java -jar jars/jdbcclient.jar -umonetdb
password:

Welcome to the MonetDB interactive JDBC terminal!
Database: MonetDB 5.0.0
Driver: MonetDB Native Driver 1.5 (Steadfast_pre4 20061124)
Type \q to quit, \h for a list of available commands
auto commit mode: on
monetdb->
```

From here you can execute a simple query to assure yourself everything is setup to work correctly. If the connection fails, observe the error messages from JdbcClient and the merovingian logs for clues.

9.10.4 Using the driver in your Java programs

To use the MonetDB JDBC driver, the MONETDB-X.Y-JDBC.JAR jar-archive has to be in the Java classpath. Make sure this is actually the case.

Loading the driver in your Java program requires two lines of code:

```
// make sure the ClassLoader has the MonetDB JDBC driver loaded
Class.forName("nl.cwi.monetdb.jdbc.MonetDriver");
// request a Connection to a MonetDB server running on 'localhost'
Connection con = DriverManager.getConnection("jdbc:monetdb://localhost/database", "monetdb"
```

The first line makes sure the Java ClassLoader has initialised (and loaded) the Driver class of the MonetDB JDBC package, so it is registered with the DriverManager. The second line requests a Connection object from the DriverManager which is suitable for MonetDB.

The string passed to the "GETCONNECTION()"method is defined as "JDBC:MONETDB://<HOST>[:<PORT>]/<DATABASE>" where elements between "<" and ">" are required and elements between "[" and "]" are optional.

9.10.5 A sample Java program

```
import java.sql.*;

/**
 * This example assumes there exist tables a and b filled with some data.
 * On these tables some queries are executed and the JDBC driver is tested
 * on it's accuracy and robustness against 'users'.
 *
 * @author Fabian Groffen
 */
public class MJDBCTest {
    public static void main(String[] args) throws Exception {
        // make sure the driver is loaded
        Class.forName("nl.cwi.monetdb.jdbc.MonetDriver");
        Connection con = DriverManager.getConnection("jdbc:monetdb://localhost/database", "
        Statement st = con.createStatement();
        ResultSet rs;
```

```
rs = st.executeQuery("SELECT a.var1, COUNT(b.id) as total FROM a, b WHERE a.var1 =
// get meta data and print columns with their type
ResultSetMetaData md = rs.getMetaData();
for (int i = 1; i <= md.getColumnCount(); i++) {
    System.out.print(md.getColumnName(i) + ":" +
        md.getColumnTypeName(i) + "\t");
}
System.out.println("");
// print the data: only the first 5 rows, while there probably are
// a lot more. This shouldn't cause any problems afterwards since the
// result should get properly discarded on the next query
for (int i = 0; rs.next() && i < 5; i++) {
    for (int j = 1; j <= md.getColumnCount(); j++) {
        System.out.print(rs.getString(j) + "\t");
    }
    System.out.println("");
}

// tell the driver to only return 5 rows, it can optimize on this
// value, and will not fetch any more than 5 rows.
st.setMaxRows(5);
// we ask the database for 22 rows, while we set the JDBC driver to
// 5 rows, this shouldn't be a problem at all...
rs = st.executeQuery("select * from a limit 22");
// read till the driver says there are no rows left
for (int i = 0; rs.next(); i++) {
    System.out.print("[" + rs.getString("var1") + "]");
    System.out.print("[" + rs.getString("var2") + "]");
    System.out.print("[" + rs.getInt("var3") + "]");
    System.out.println("[" + rs.getString("var4") + "]");
}

// this close is not needed, should be done by next execute(Query) call
// however if there can be some time between this point and the next
// execute call, it is from a resource perspective better to close it.
//rs.close();

// unset the row limit; 0 means as much as the database sends us
st.setMaxRows(0);
// we only ask 10 rows
rs = st.executeQuery("select * from b limit 10;");
// and simply print them
while (rs.next()) {
    System.out.print(rs.getInt("rowid") + ", ");
    System.out.print(rs.getString("id") + ", ");
    System.out.print(rs.getInt("var1") + ", ");
```

```
            System.out.print(rs.getInt("var2") + ", ");
            System.out.print(rs.getString("var3") + ", ");
            System.out.println(rs.getString("var4"));
        }

        // this close is not needed, as the Statement will close the last
        // ResultSet around when it's closed
        // again, if that can take some time, it's nicer to close immediately
        // the reason why these closes are commented out here, is to test if
        // the driver really cleans up it's mess like it should
        //rs.close();

        // perform a ResultSet-less query (with no trailing ; since that should
        // be possible as well and is JDBC standard)
        // Note that this method should return the number of updated rows. This
        // method however always returns -1, since Monet currently doesn't
        // support returning the affected rows.
        st.executeUpdate("delete from a where var1 = 'zzzz'");

        // closing the connection should take care of closing all generated
        // statements from it...
        // don't forget to do it yourself if the connection is reused or much
        // longer alive, since the Statement object contains a lot of things
        // you probably want to reclaim if you don't need them anymore.
        //st.close();
        con.close();
    }
}
```

9.11 MonetDB ODBC Driver

Short for Open DataBase Connectivity, a standard database access method developed by the SQL Access group in 1992. The goal of ODBC is to make it possible to access any data from any application, regardless of which database management system (DBMS) is handling the data. ODBC manages this by inserting a middle layer, called a database driver, between an application and the DBMS. The purpose of this layer is to translate the application's data queries into commands that the DBMS understands. For this to work, both the application and the DBMS must be ODBC-compliant – that is, the application must be capable of issuing ODBC commands and the DBMS must be capable of responding to them.

The ODBC driver for MonetDB is included in the Windows installer and Linux RPMs. The source can be found in the *SQL* CVS tree.

To help you setup your system to use the ODBC driver with MonetDB, two how-tos are available, one for Windows users and one for Linux/UNIX users.

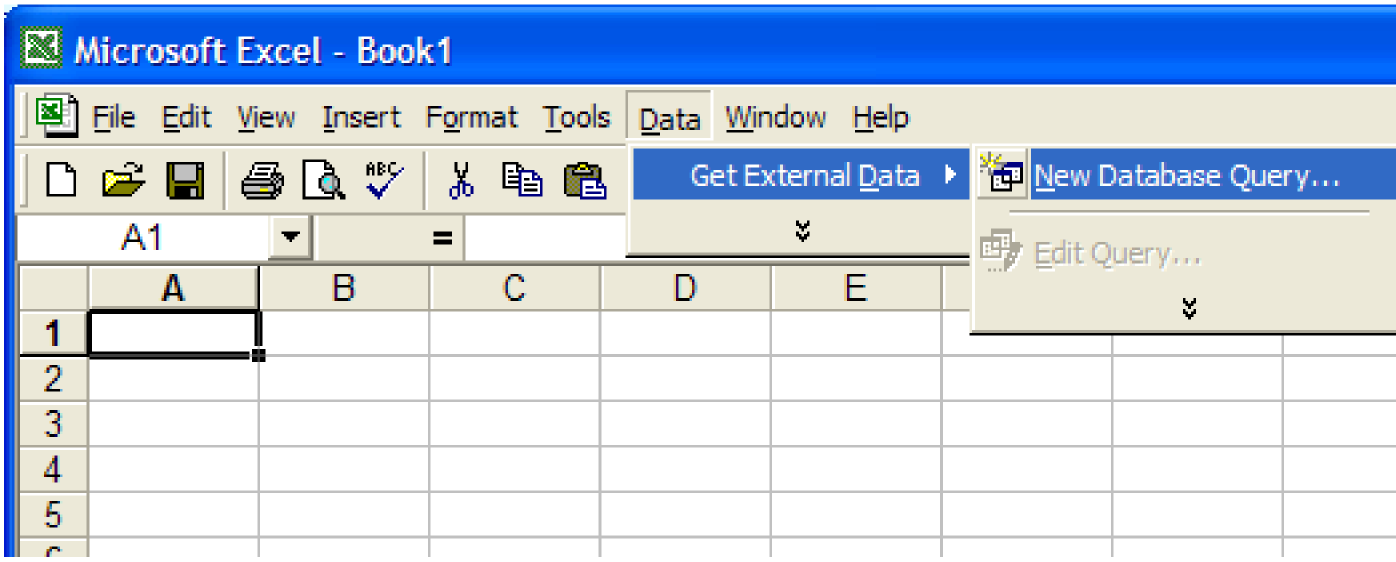
Microsoft Excel demo

A little demo showing how to import data from a MonetDB server into Microsoft Excel.

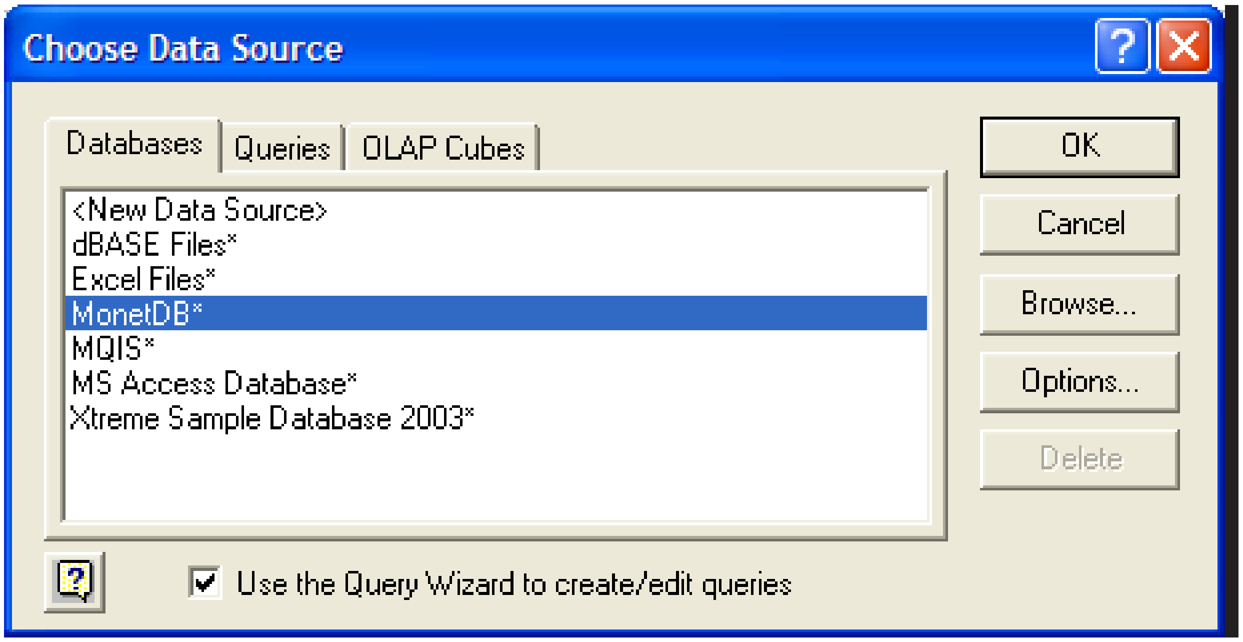
Using Excel with the MonetDB ODBC Driver

Start up the MonetDB SQL Server and Excel.

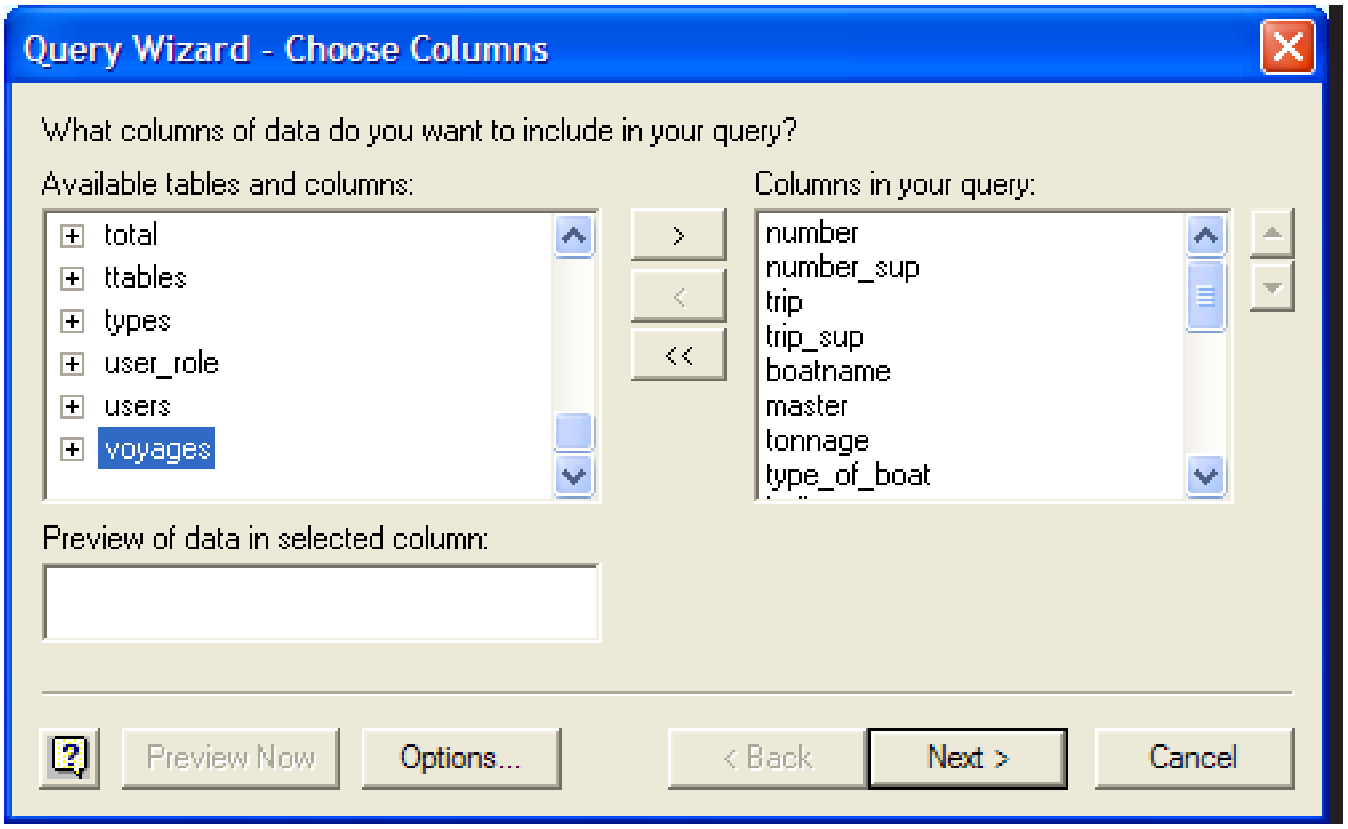
In Excel, select from the drop down menu, first *Data*, then *Get External Data*, and finally *New Database Query...*



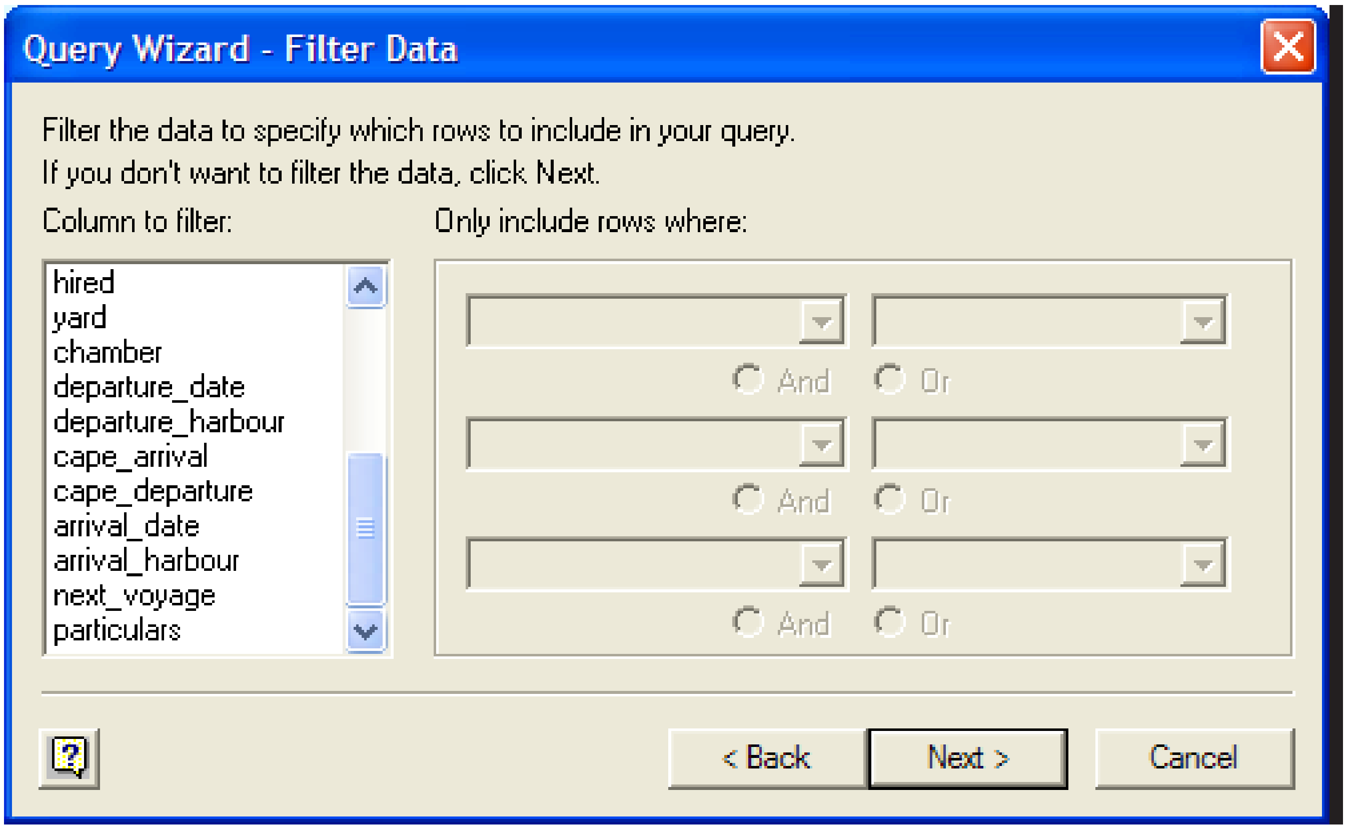
If MonetDB was installed correctly, there should be an entry *MonetDB* in the dialog box that opens. Select it and click on *OK*.



In the wizard that opens, scroll down in the list on the left hand side and select *voyages*. Then click on the button labeled > and then on *Next* >.



In the next page of the wizard, click on *Next* >.



Query Wizard - Filter Data

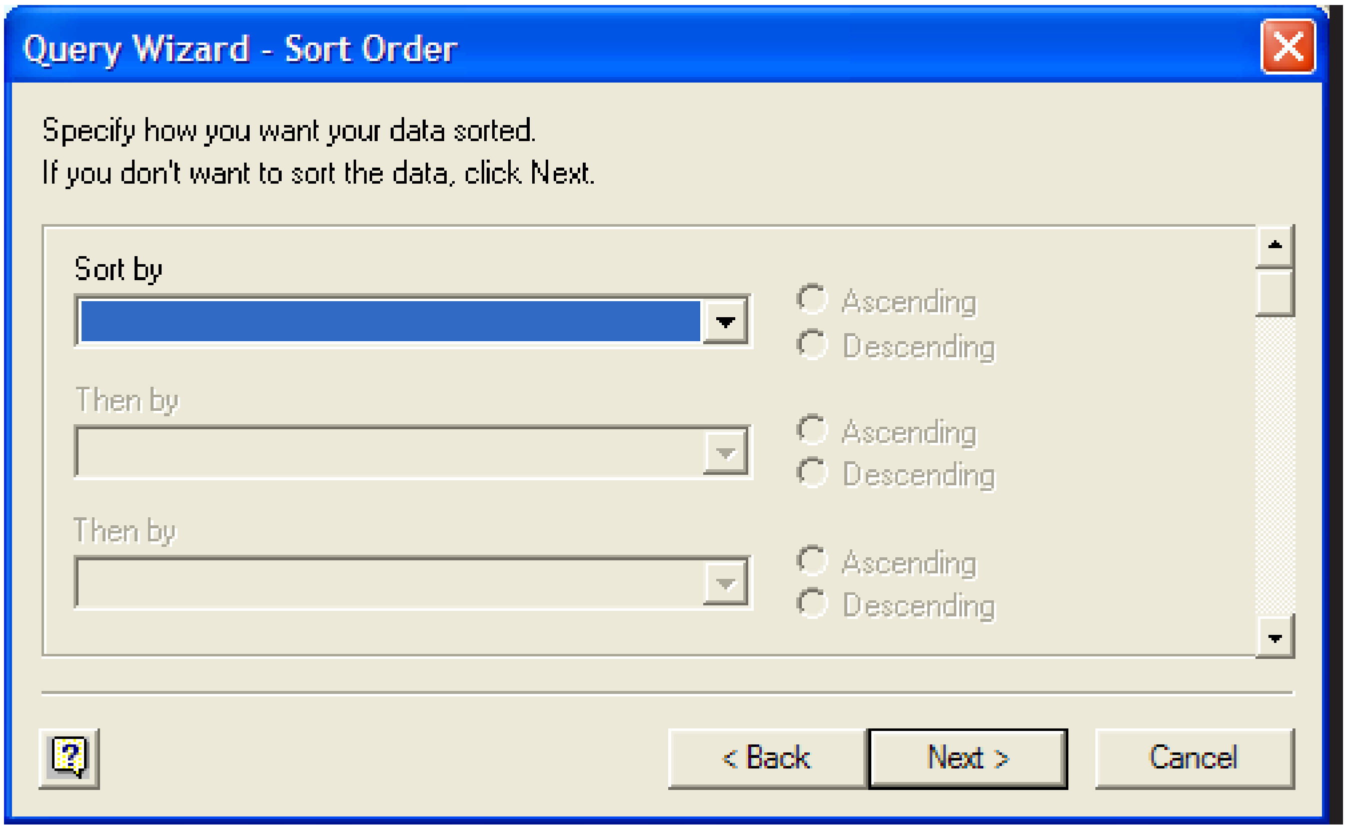
Filter the data to specify which rows to include in your query.
If you don't want to filter the data, click Next.

Column to filter:

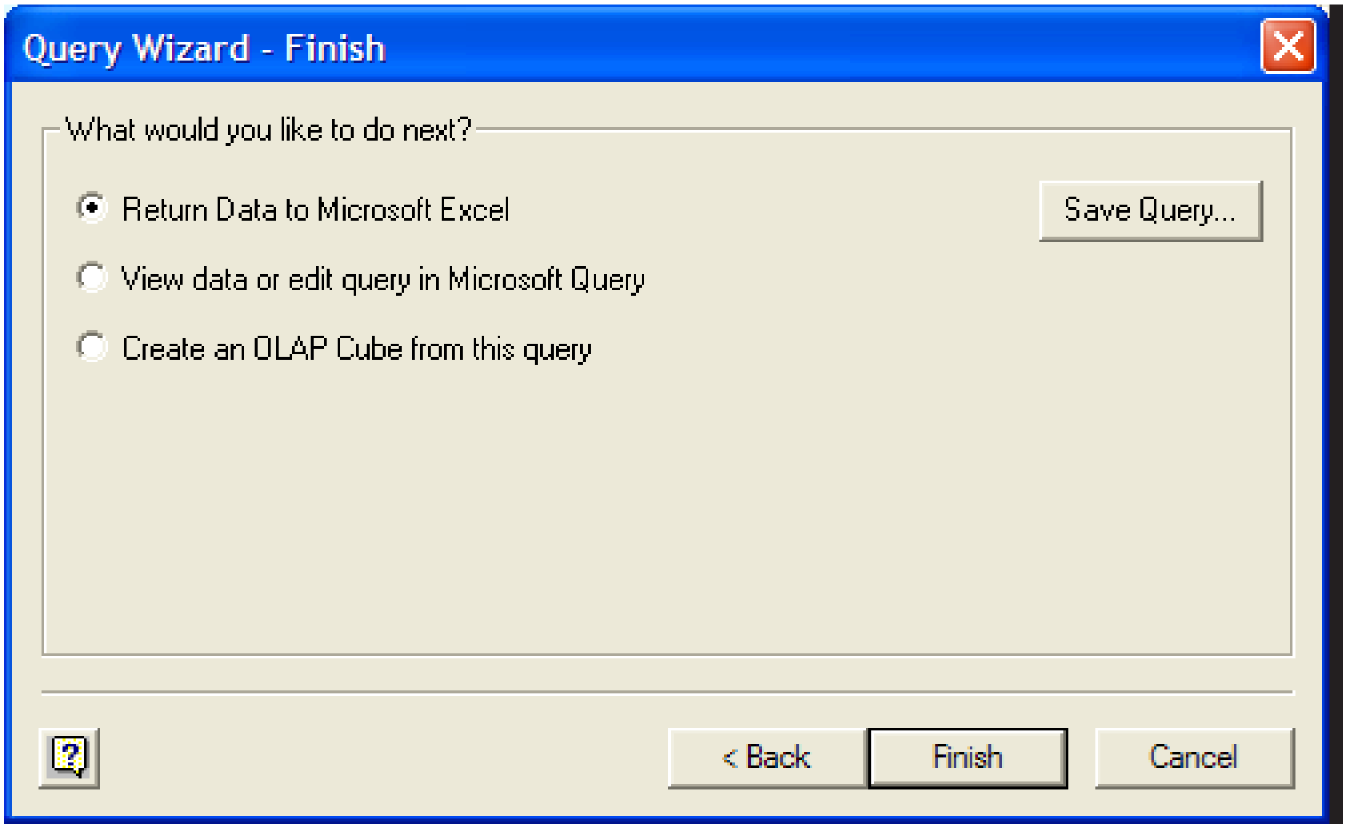
hired
yard
chamber
departure_date
departure_harbour
cape_arrival
cape_departure
arrival_date
arrival_harbour
next_voyage
particulars

Only include rows where:

In the next page of the wizard, click on *Next* >.


Query Wizard - Sort Order
Specify how you want your data sorted.
If you don't want to sort the data, click Next.
Sort by
Ascending
Descending
Then by
Ascending
Descending
Then by
Ascending
Descending
< Back
Next >
Cancel

In the final page of the wizard, click on *Finish*.

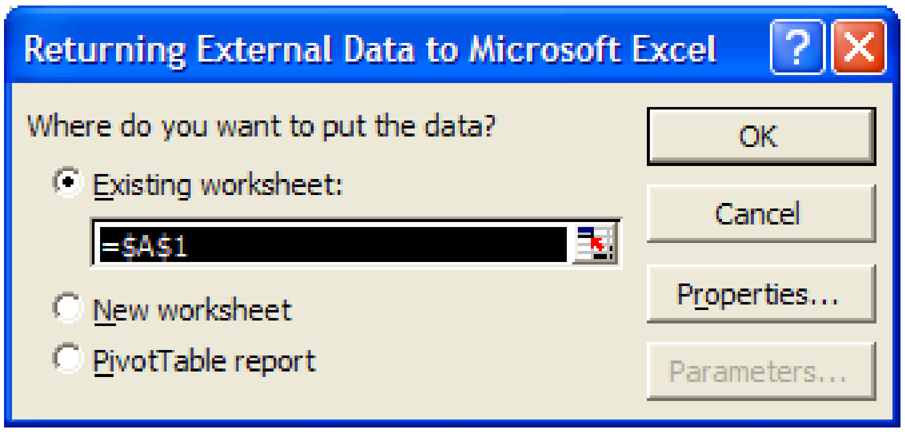


Query Wizard - Finish

What would you like to do next?

(● Return Data to Microsoft Excel) ○ View data or edit query in Microsoft Query ○ Create an OLAP Cube from this query

A new dialog window opens. Click on *OK* to insert the data into the current Excel worksheet.



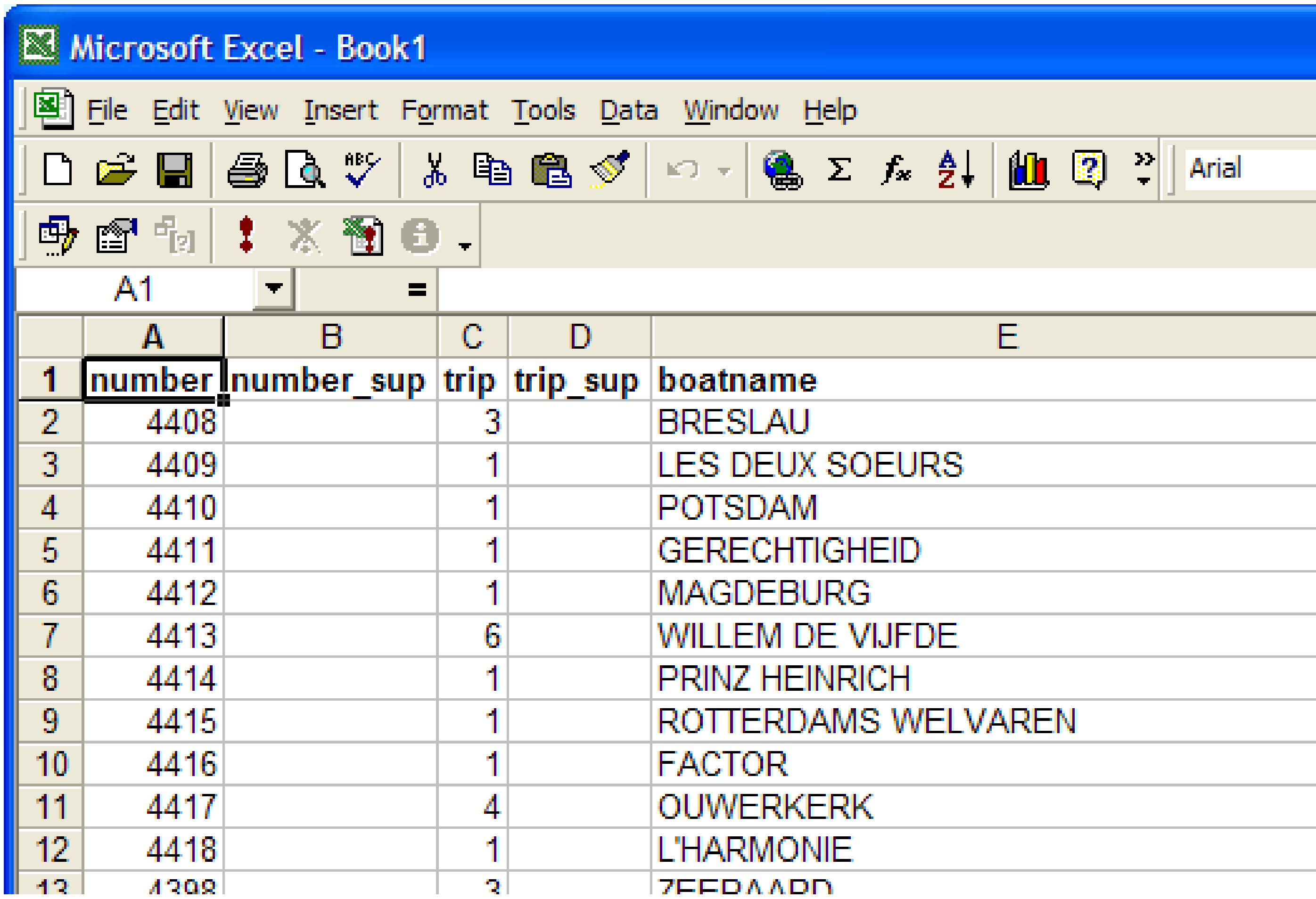
Returning External Data to Microsoft Excel

Where do you want to put the data?

(● Existing worksheet:) ○ New worksheet ○ PivotTable report

=A1

That's all.

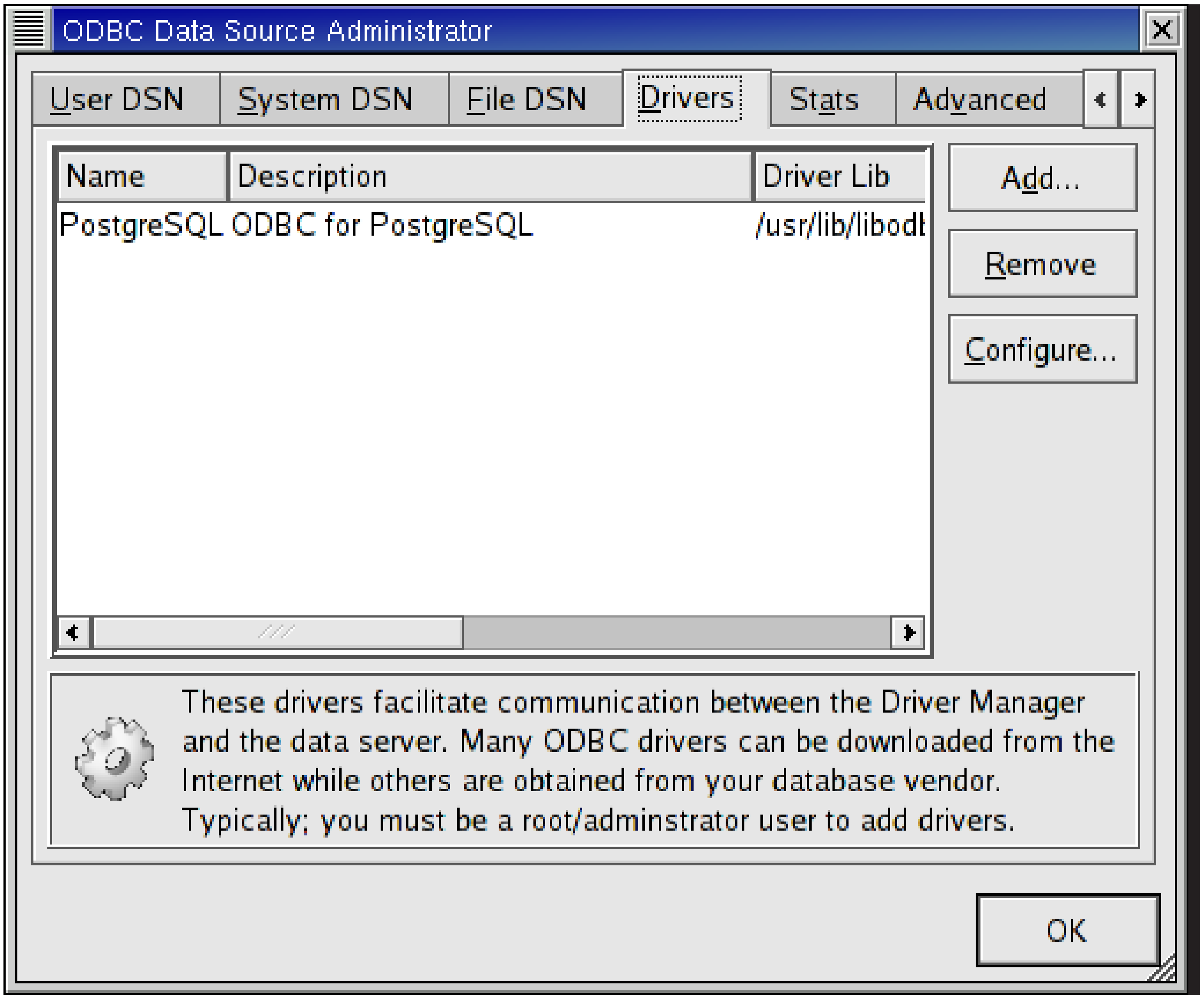


	A	B	C	D	E
1	number	number_sup	trip	trip_sup	boatname
2	4408		3		BRESLAU
3	4409		1		LES DEUX SOEURS
4	4410		1		POTSDAM
5	4411		1		GERECHTIGHEID
6	4412		1		MAGDEBURG
7	4413		6		WILLEM DE VIJFDE
8	4414		1		PRINZ HEINRICH
9	4415		1		ROTTERDAMS WELVAREN
10	4416		1		FACTOR
11	4417		4		OUWERKERK
12	4418		1		L'HARMONIE
13	4398		3		ZEEPAARD

Installing the MonetDB ODBC Driver for unixODBC

Configuring the MonetDB Driver

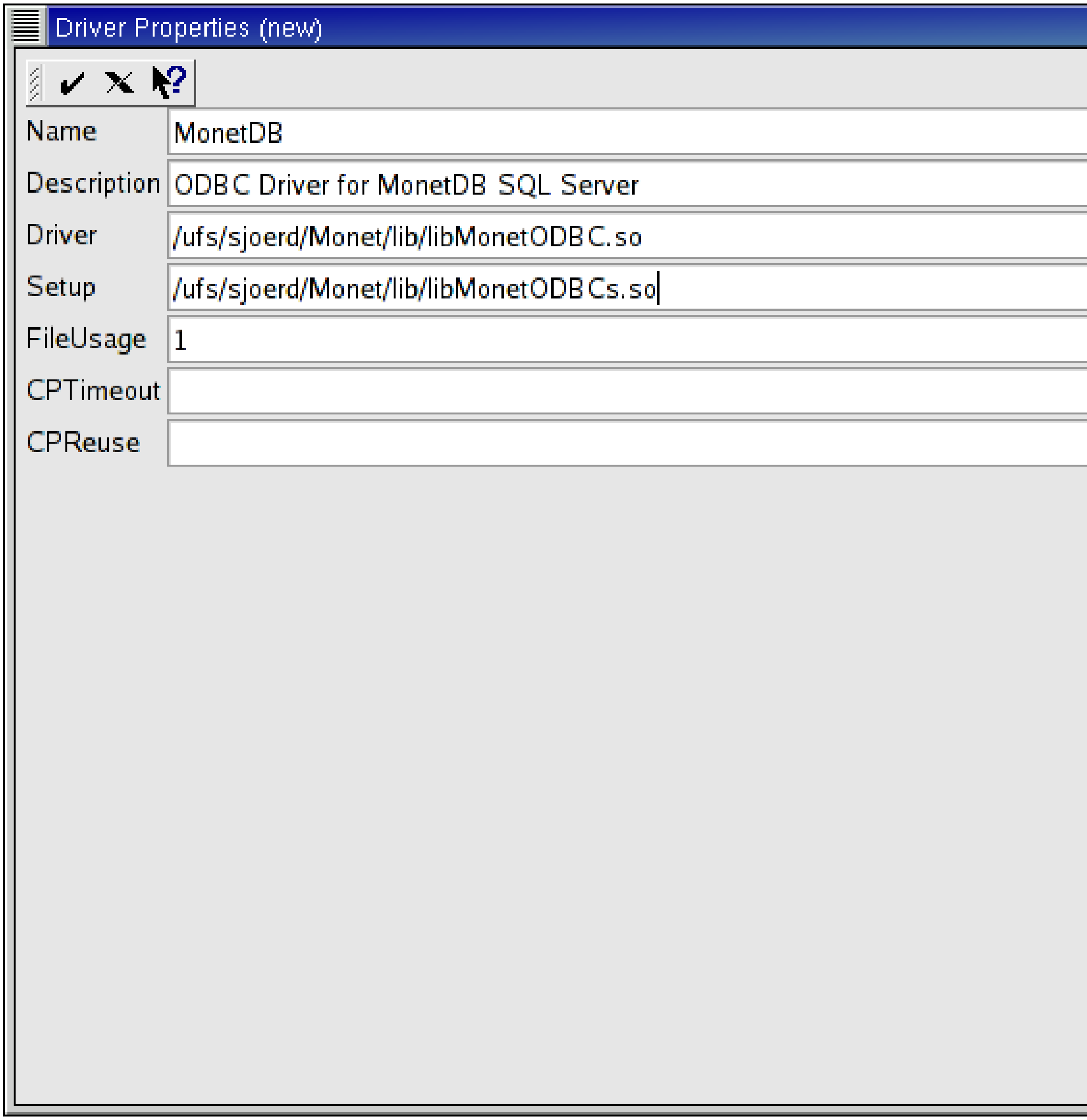
As Superuser, start the unixODBC configuration program *ODBCConfig* and select the *Drivers* tab.



Name	Description	Driver Lib
PostgreSQL	ODBC for PostgreSQL	/usr/lib/libodb

These drivers facilitate communication between the Driver Manager and the data server. Many ODBC drivers can be downloaded from the Internet while others are obtained from your database vendor. Typically; you must be a root/adminstrator user to add drivers.

On this tab, click on the button labeled *Add...* and fill in the fields as follows.



Name	MonetDB
Description	ODBC Driver for MonetDB SQL Server
Driver	/ufs/sjoerd/Monet/lib/libMonetODBC.so
Setup	/ufs/sjoerd/Monet/lib/libMonetODBCs.so
FileUsage	1
CPTimeout	
CPReuse	

`Name` MonetDB

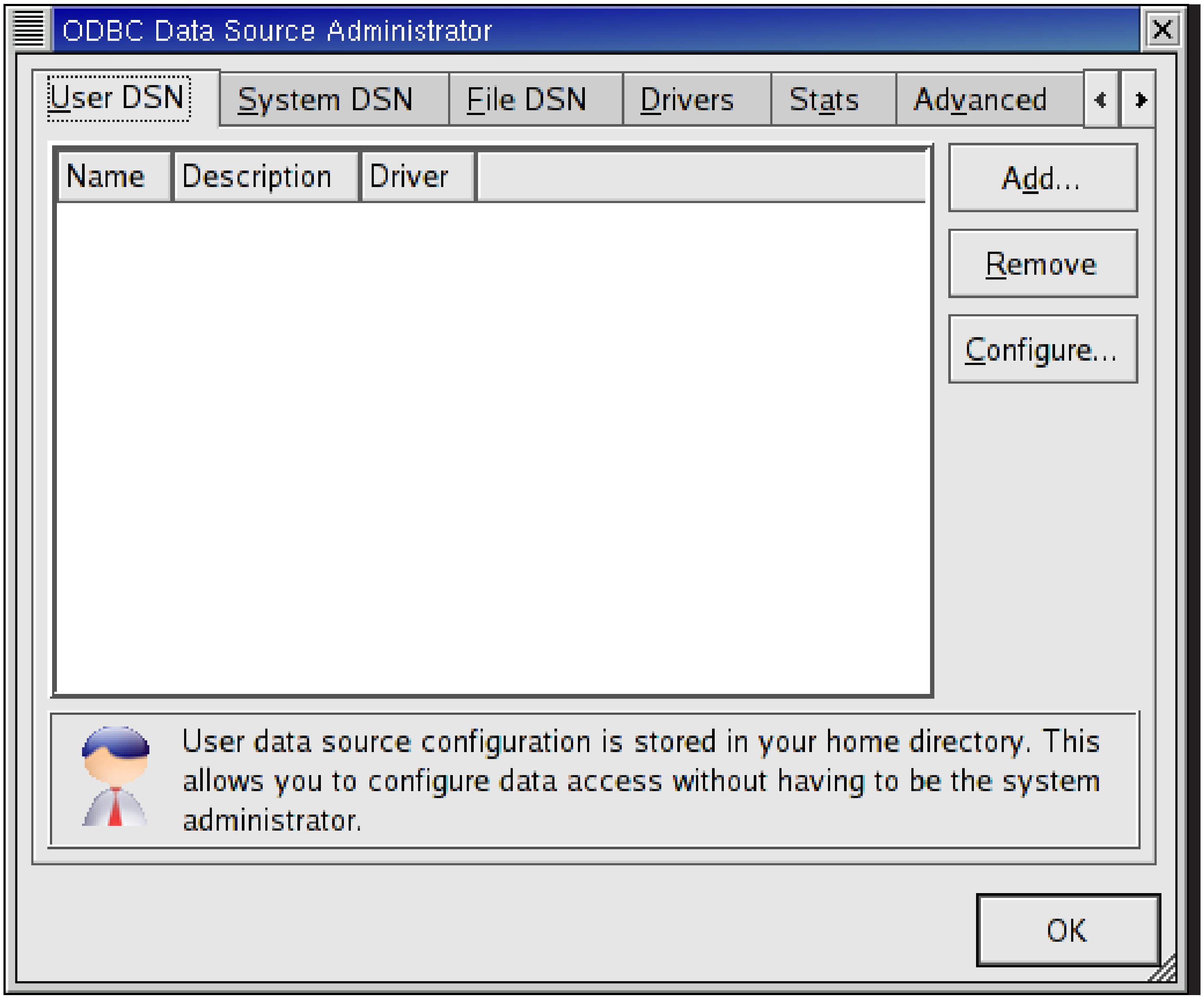
`Description`
ODBC Driver for MonetDB SQL Server

`Driver`	<*path-to-MonetDB*>/lib(64)/libMonetODBC.so
`Setup`	<*path-to-MonetDB*>/lib(64)/libMonetODBCs.so

Don't change the other fields. When done, click on the check mark in the top left corner of the window. The first window should now contain an entry for MonetDB. Click on *OK*

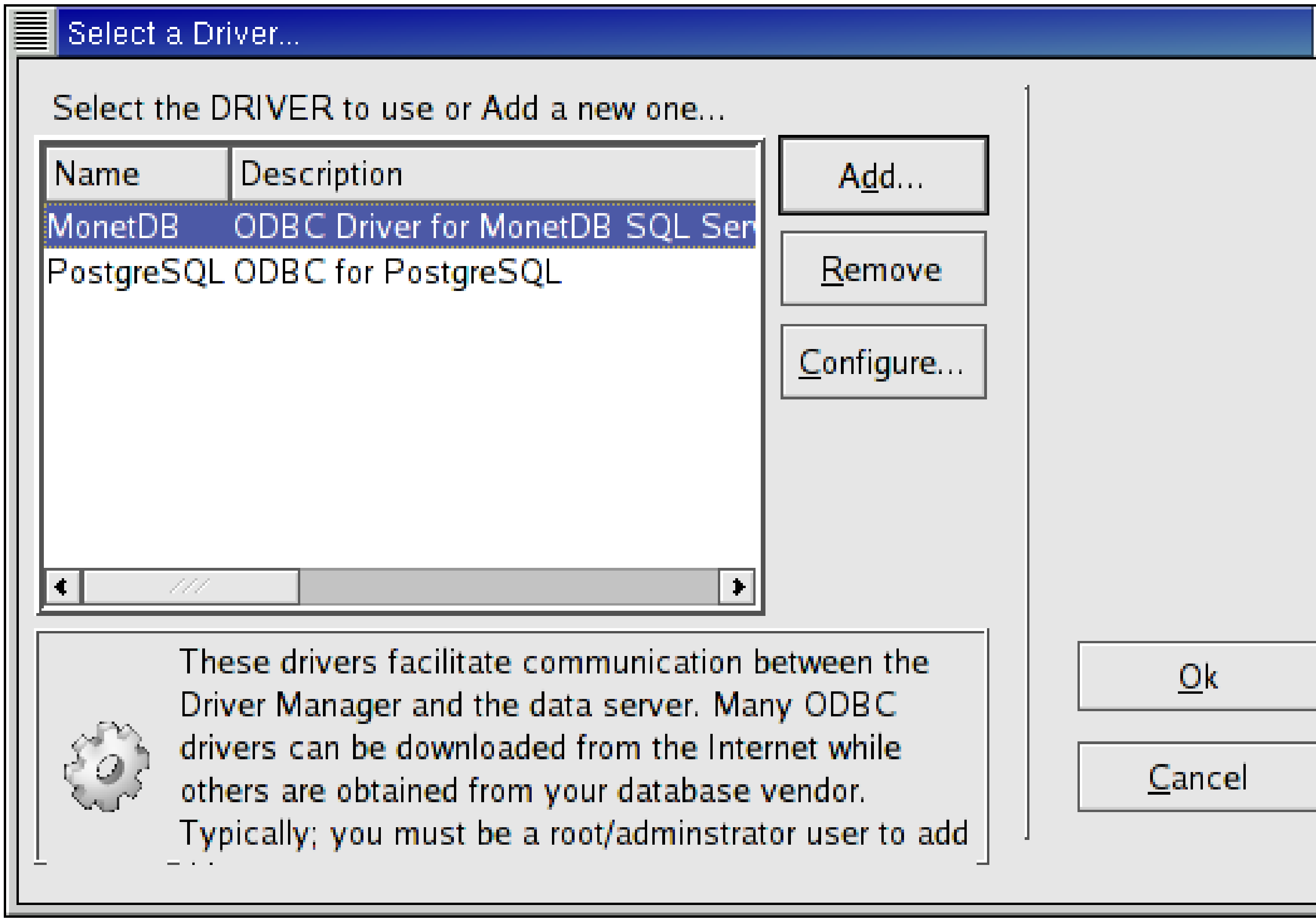
<h3>Configuring a Data Source</h3>

Now as normal user start *ODBCConfig* again.



User data source configuration is stored in your home directory. This allows you to configure data access without having to be the system administrator.

On the *User DSN* tab click on the *Add...* button. A new window pops up in which you have to select the ODBC driver. Click on the entry for *MonetDB* and click on *OK*.

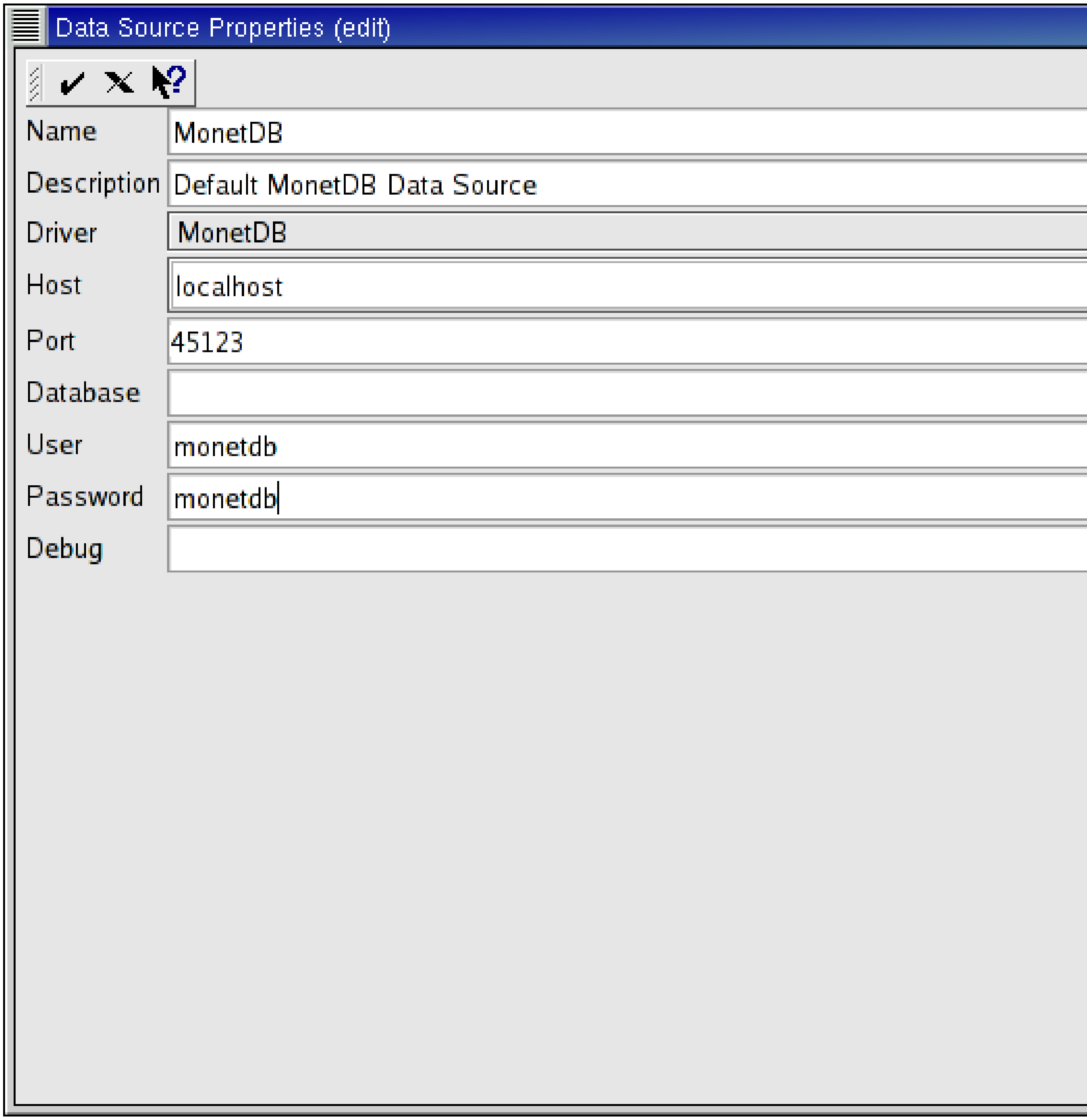


Select the DRIVER to use or Add a new one...

Name	Description
MonetDB	ODBC Driver for MonetDB SQL Ser
PostgreSQL	ODBC for PostgreSQL

These drivers facilitate communication between the Driver Manager and the data server. Many ODBC drivers can be downloaded from the Internet while others are obtained from your database vendor. Typically; you must be a root/adminstrator user to add

A new window pops up. Fill in the fields as follows.



Name	MonetDB
Description	Default MonetDB Data Source
Driver	MonetDB
Host	localhost
Port	45123
Database	
User	monetdb
Password	monetdb
Debug	

`Name` MonetDB

`Description`
Default MonetDB Data Source

`Host`	localhost
`Port`	50000
`User`	monetdb
`Password`	monetdb

Don't change the other fields. When done, click on the check mark in the top left corner of the window. The first window should now contain an entry for MonetDB. Click on *OK*

Appendix A Instruction Summary

The table below gives a condensed overview of the operations defined in each of the modules.

aggr.avg	aggr.count_no_nil	aggr.min	aggr.size
aggr.cardinality	aggr.histogram	aggr.prod	aggr.sum
aggr.count	aggr.max	aggr.product	
alarm.alarm	alarm.epoch	alarm.time	
alarm.ctime	alarm.prelude	alarm.timers	
alarm.epilogue	alarm.sleep	alarm.usec	
algebra.antijoin	algebra.joinPath	algebra.project	algebra.sortTH
algebra.antiuselect	algebra.kdifference	algebra.rangesplit	algebra.sortTail
algebra.bandjoin	algebra.kintersect	algebra.reuse	algebra.split
algebra.copy	algebra.kunion	algebra.revert	algebra.ssort
algebra.crossproduct	algebra.kunique	algebra.sample	algebra.ssort_rev
algebra.difference	algebra.leftfetchjoin		algebra.sunion
algebra.exist	algebra.leftjoin	algebra.sdifference	algebra.sunique
algebra.fetch	algebra.like	algebra.select	algebra.thetajoin
algebra.fetchjoin	algebra.markH	algebra.selectH	algebra.thetaselect
algebra.find	algebra.markT	algebra.selectNotNil	algebra.thetauselect
algebra.fragment	algebra.mark_grp	algebra.semijoin	algebra.topN
algebra.groupby	algebra.materialize	algebra.sintersect	algebra.tunique
algebra.hashjoin	algebra.merge	algebra.slice	algebra.uhashsplit
algebra.hashsplit	algebra.mergejoin	algebra.sort	algebra.union
algebra.indexjoin	algebra.number	algebra.sortHT	algebra.unique
algebra.intersect	algebra.outerjoin	algebra.sortReverse	algebra.urangesplit
algebra.join	algebra.position	algebra.sortReverseTail	
array.grid	array.product	array.project	
bat.append	bat.getSpaceUsed	bat.isaSet	bat.setColumn
bat.attach	bat.getStorageSize	bat.load	bat.setGarbage
bat.delete	bat.getTail	bat.mirror	bat.setHash
bat.densebat	bat.getTailType	bat.new	bat.setHot
bat.flush	bat.hasAppendMode	bat.newIterator	bat.setKey
bat.getAccess	bat.hasMoreElements	bat.order	bat.setMemoryAdvise
bat.getAlpha	bat.hasReadMode	bat.orderReverse	bat.setMemoryMap
bat.getCapacity	bat.hasWriteMode	bat.pack	bat.setName
bat.getDelta	bat.info	bat.partition	bat.setPersistent
bat.getDiskSize	bat.inplace	bat.reduce	bat.setReadMode
bat.getHead	bat.insert	bat.replace	bat.setRole
bat.getHeadType	bat.isCached	bat.reverse	bat.setSet
bat.getHeat	bat.isPersistent	bat.revert	bat.setSorted
bat.getMemorySize	bat.isSorted	bat.save	bat.setTransient
bat.getName	bat.isSortedReverse	bat.setAccess	bat.setWriteMode
bat.getRole	bat.isSynced	bat.setAppendMode	bat.unload
bat.getSequenceBase	bat.isTransient	bat.setBase	bat.unpack
bat.getSize	bat.isaKey	bat.setCold	
batcalc.!=	batcalc.<	batcalc.bte	batcalc.lng
batcalc.%	batcalc.<=	batcalc.chr	batcalc.not

batcalc.*	batcalc.==	batcalc.dbl	batcalc.oid
batcalc.+	batcalc.>	batcalc.flt	batcalc.or
batcalc.++	batcalc.>=	batcalc.ifthen	batcalc.sht
batcalc.-	batcalc.abs	batcalc.ifthenelse	batcalc.str
batcalc.–	batcalc.and	batcalc.int	batcalc.wrd
batcalc./	batcalc.bit	batcalc.isnil	batcalc.xor
batcolor.blue	batcolor.green	batcolor.red	batcolor.value
batcolor.cb	batcolor.hsv	batcolor.rgb	
batcolor.color	batcolor.hue	batcolor.saturation	
batcolor.cr	batcolor.luminance	batcolor.str	
batmmath.acos	batmmath.cos	batmmath.fmod	batmmath.sinh
batmmath.asin	batmmath.cosh	batmmath.log	batmmath.sqrt
batmmath.atan	batmmath.exp	batmmath.log10	batmmath.tan
batmmath.atan2	batmmath.fabs	batmmath.pow	batmmath.tanh
batmmath.ceil	batmmath.floor	batmmath.sin	
batmtime.day	batmtime.milliseconds		batmtime.year
batmtime.hours	batmtime.month	batmtime.seconds	
batstr.chrAt	batstr.ltrim	batstr.search	batstr.toLower
batstr.endsWith	batstr.nbytes	batstr.startsWith	batstr.toUpper
batstr.length	batstr.r_search	batstr.string	batstr.trim
batstr.like	batstr.replace	batstr.substitute	batstr.unicodeAt
batstr.like_uselect	batstr.rtrim	batstr.substring	
bbp.bind	bbp.getDiskSpace	bbp.getObjects	bbp.prelude
bbp.close	bbp.getHeadType	bbp.getPageSize	bbp.release
bbp.commit	bbp.getHeat	bbp.getRNames	bbp.releaseAll
bbp.deposit	bbp.getKind	bbp.getRefCount	bbp.take
bbp.destroy	bbp.getLRefCount	bbp.getStatus	bbp.toString
bbp.discard	bbp.getLocation	bbp.getTailType	
bbp.getCount	bbp.getName	bbp.iterator	
bbp.getDirty	bbp.getNames	bbp.open	
blob.blob	blob.prelude	blob.tostring	
blob.nitems	blob.toblob		
box.close	box.discard	box.open	box.take
box.deposit	box.getBoxNames	box.release	box.toString
box.destroy	box.iterator	box.releaseAll	
bpm.adapt	bpm.fold	bpm.mapNxt	bpm.prelude
bpm.addPartition	bpm.garbage	bpm.mapPrv	bpm.rangePartition
bpm.close	bpm.getDimension	bpm.mapThghDbl	bpm.rangePartitionSort
bpm.count	bpm.getNames	bpm.mapThghLng	bpm.replace
bpm.delete	bpm.getNumberOfPartitions		bpm.saveCatalog
bpm.deposit	bpm.hasMoreElements	bpm.mapTlowDbl	bpm.select
bpm.derivePartition	bpm.hashPartition	bpm.mapTlowLng	bpm.sortPartitions
bpm.destroy	bpm.hashPartitions	bpm.new	bpm.sortTail
bpm.discard	bpm.insert	bpm.newIterator	bpm.splitquant
bpm.dump	bpm.mapAlias	bpm.open	bpm.take
bpm.emptySet	bpm.mapBid	bpm.partition	bpm.unfold
bpm.epilogue	bpm.mapName	bpm.pieces	

bstream.create	bstream.destroy	bstream.read	
calc.!=	calc.abs	calc.ifthenelse	calc.not
calc.%	calc.and	calc.inet	calc.oid
calc.*	calc.bat	calc.int	calc.or
calc.+	calc.between	calc.inv	calc.ptr
calc.-	calc.bit	calc.isnil	calc.setoid
calc./	calc.blob	calc.isnotnil	calc.sht
calc.<	calc.bte	calc.length	calc.sign
calc.<<	calc.chr	calc.lng	calc.sizeof
calc.<=	calc.date	calc.max	calc.sqladd
calc.=	calc.daytime	calc.max_no_nil	calc.sqlblob
calc.==	calc.dbl	calc.min	calc.str
calc.>	calc.flt	calc.min_no_nil	calc.timestamp
calc.>=	calc.getBAT	calc.newoid	calc.void
calc.>>	calc.getBATidentifier		calc.wrd
clients.addScenario	clients.getId	clients.getScenario	clients.removeUser
clients.addUser	clients.getInfo	clients.getTime	clients.setHistory
clients.changePassword		clients.getUsername	clients.setListing
clients.changeUsername		clients.getUsers	clients.setPassword
clients.checkPermission		clients.quit	clients.setScenario
clients.exit	clients.getLastCommand		clients.shutdown
clients.getActions	clients.getLogins	clients.removeScenario	
cluster.column	cluster.map	cluster.table	
cluster.key	cluster.new		
color.blue	color.green	color.red	color.value
color.cb	color.hsv	color.rgb	color.ycc
color.color	color.hue	color.saturation	
color.cr	color.luminance	color.str	
const.close	const.epiloque	const.prelude	const.take
const.deposit	const.hasMoreElements		const.toString
const.destroy	const.newIterator	const.release	
const.discard	const.open	const.releaseAll	
constraints.emptySet			
crackers.DeleteMap	crackers.dselect	crackers.getCrackerBAT	
crackers.InsertAVLIndex		crackers.getMap	crackers.materializeHead
crackers.activeCacheConsciousCrackHashJoin		crackers.getTotalStorage	
crackers.alignJoin	crackers.extendCrackerBAT		crackers.pmaddReference
crackers.alignedJoin	crackers.extendCrackerMap		crackers.pmclearReferences
crackers.bandJoin	crackers.fmaddReference		crackers.pmjoinselect
crackers.buildAVLIndex		crackers.hselect	crackers.pmmaxTail
crackers.cacheConsciousCrackHashJoin		crackers.insert	crackers.pmproject
crackers.cacheConsciousCrackHashJoinAlignOnly		crackers.insertionsBForce	
crackers.crackHashJoin		crackers.insertionsBOnNeed	
crackers.crackJoin	crackers.fmclearReferences		crackers.pmselect
crackers.crackOrdered		crackers.insertionsBOnNeedGradually	
crackers.crackOrdered_validate		crackers.insertionsBOnNeedGraduallyRipple	
crackers.crackUnordered_validate		crackers.insertionsForget	

crackers.deleteAVL	crackers.fmcreateMap	crackers.insertionsPartiallyForget	
crackers.deletionsOnNeed		crackers.joinselect	crackers.pmtselect
crackers.deletionsOnNeedGradually		crackers.joinuselect	crackers.positionproject
crackers.deletionsOnNeedGraduallyRipple		crackers.mapCount	crackers.printAVLTree_int
crackers.djoinselect	crackers.fmremoveMap	crackers.markedproject	
crackers.dproject	crackers.fullAlignment		crackers.printCrackerBAT
date.!=	date.<=	date.>	date.date
date.<	date.==	date.>=	date.isnil
daytime.!=	daytime.<=	daytime.>	daytime.isnil
daytime.<	daytime.==	daytime.>=	
factories.getArrival	factories.getDeparture		factories.shutdown
factories.getCaller	factories.getOwners	factories.getPlants	
group.avg	group.max	group.prelude	group.size
group.count	group.min	group.refine	group.sum
group.derive	group.new	group.refine_reverse	group.variance
identifier.identifier		identifier.prelude	
inet.!=	inet.>	inet.host	inet.new
inet.<	inet.>=	inet.hostmask	inet.setmasklen
inet.<<	inet.>>	inet.isnil	inet.text
inet.<<=	inet.>>=	inet.masklen	
inet.<=	inet.abbrev	inet.netmask	
inet.=	inet.broadcast	inet.network	
inspect.equalType	inspect.getComment	inspect.getSignature	inspect.getStatistics
inspect.getAddress	inspect.getDefinition		inspect.getType
inspect.getAddresses	inspect.getEnvironment		inspect.getTypeIndex
inspect.getAtomNames	inspect.getFunction	inspect.getSignatures	
inspect.getAtomSizes	inspect.getKind	inspect.getSize	inspect.getTypeName
inspect.getAtomSuper	inspect.getModule	inspect.getSource	inspect.getWelcome
io.data	io.import	io.prompt	io.stdout
io.export	io.print	io.stderr	io.table
io.ftable	io.printf	io.stdin	
language.assert	language.newRange	language.setIOTrace	language.source
language.assertSpace	language.nextElement	language.setMemoryTrace	
language.call	language.raise	language.setThreadTrace	
language.dataflow	language.register	language.setTimerTrace	
lock.create	lock.set	lock.try	
lock.destroy	lock.tostr	lock.unset	
mal.multiplex			
manual.completion	manual.help	manual.search	manual.summary
manual.createXML	manual.index	manual.section	
mapi.bind	mapi.fetch_field_array		mapi.query_array
mapi.connect	mapi.fetch_line	mapi.listen_ssl	mapi.query_handle
mapi.connect_ssl	mapi.fetch_reset	mapi.lookup	mapi.reconnect
mapi.destroy	mapi.fetch_row	mapi.malclient	mapi.resume
mapi.disconnect	mapi.finish	mapi.next_result	mapi.rpc
mapi.error	mapi.getError	mapi.ping	mapi.setAlias
mapi.explain	mapi.get_field_count	mapi.prepare	mapi.stop

mapi.fetch_all_rows	mapi.get_row_count	mapi.put	mapi.suspend
mapi.fetch_field	mapi.listen	mapi.query	mapi.trace
mat.hasMoreElements	mat.new	mat.pack	
mat.info	mat.newIterator	mat.print	
mdb.List	mdb.getReason	mdb.listMapi	mdb.setMemoryTrace
mdb.collect	mdb.getStackDepth	mdb.modules	mdb.setThread
mdb.dot	mdb.getStackFrame	mdb.setCatch	mdb.setTimer
mdb.dump	mdb.getStackTrace	mdb.setCount	mdb.setTrace
mdb.getContext	mdb.grab	mdb.setDebug	mdb.start
mdb.getDebug	mdb.inspect	mdb.setFlow	mdb.stop
mdb.getDefinition	mdb.lifespan	mdb.setIO	mdb.var
mdb.getException	mdb.list	mdb.setMemory	
mkey.bulk_rotate_xor_hash		mkey.hash	mkey.rotate
mmath.acos	mmath.cot	mmath.isnan	mmath.sin
mmath.asin	mmath.exp	mmath.log	mmath.sinh
mmath.atan	mmath.fabs	mmath.log10	mmath.sqrt
mmath.atan2	mmath.finite	mmath.pi	mmath.srand
mmath.ceil	mmath.floor	mmath.pow	mmath.tan
mmath.cos	mmath.fmod	mmath.rand	mmath.tanh
mmath.cosh	mmath.isinf	mmath.round	
mtime.add	mtime.dayname	mtime.local_timezone	mtime.start_dst
mtime.adddays	mtime.daynum	mtime.milliseconds	mtime.str_to_date
mtime.addmonths	mtime.dayofweek	mtime.minutes	mtime.time_add_sec_interval
mtime.addyears	mtime.dayofyear	mtime.month	mtime.time_sub_sec_interval
mtime.compute	mtime.daytime	mtime.monthname	mtime.time_synonyms
mtime.current_date	mtime.diff	mtime.monthnum	mtime.timestamp
mtime.current_time	mtime.dst	mtime.msec	mtime.timestamp_add_month_interval
mtime.current_timestamp		mtime.msecs	mtime.timestamp_add_sec_interval
mtime.date	mtime.end_dst	mtime.olddate	mtime.timestamp_sub_month_interval
mtime.date_add_month_interval		mtime.oldduration	mtime.timestamp_sub_sec_interval
mtime.date_add_sec_interval		mtime.prelude	mtime.timezone
mtime.date_sub_sec_interval		mtime.rule	mtime.timezone_local
mtime.date_to_str	mtime.epilogue	mtime.seconds	mtime.weekday
mtime.day	mtime.hours	mtime.setTimezone	mtime.weekofyear
optimizer.accessmode	optimizer.dumpQEP	optimizer.joinselect	optimizer.prelude
optimizer.accumulators		optimizer.macro	optimizer.pushranges
optimizer.aliases	optimizer.emptySet	optimizer.mergetable	optimizer.recycle
optimizer.clrDebug	optimizer.evaluate	optimizer.mitosis	optimizer.reduce
optimizer.coercions	optimizer.factorize	optimizer.multiplex	optimizer.remap
optimizer.commonTerms		optimizer.octopus	optimizer.remoteQueries
optimizer.constants	optimizer.garbageCollector		optimizer.replicator
optimizer.costModel	optimizer.heuristics	optimizer.optimize	optimizer.setDebug
optimizer.crack	optimizer.history	optimizer.orcam	optimizer.showFlowGraph
optimizer.dataflow	optimizer.inline	optimizer.partitions	optimizer.showPlan
optimizer.deadcode	optimizer.joinPath	optimizer.peephole	optimizer.singleton
pcre.compile	pcre.match	pcre.prelude	pcre.sql2pcre
pcre.index	pcre.patindex	pcre.replace	pcre.uselect

pcre.like_uselect	pcre.pcre_quote	pcre.select	
pqueue.dequeue_max	pqueue.enqueue_min	pqueue.topn_min	
pqueue.dequeue_min	pqueue.init	pqueue.topreplace_max	
pqueue.enqueue_max	pqueue.topn_max	pqueue.topreplace_min	
profiler.activate	profiler.getDiskReads		profiler.noop
profiler.cleanup	profiler.getDiskWrites		profiler.openStream
profiler.closeStream	profiler.getEvent	profiler.getTrace	profiler.reset
profiler.clrFilter	profiler.getFootprint		profiler.setAll
profiler.deactivate	profiler.getMemory	profiler.getUserTime	profiler.setEndPoint
profiler.dumpTrace	profiler.getSystemTime		profiler.setFilter
recycle.dump	recycle.getRetainPolicy		recycle.reset
recycle.dumpQPat	recycle.getReusePolicy		recycle.setCachePolicy
recycle.epilogue	recycle.log	recycle.monitor	recycle.setRetainPolicy
recycle.getCachePolicy		recycle.prelude	recycle.setReusePolicy
remote.connect	remote.disconnect	remote.get	remote.put
remote.create	remote.epilogue	remote.getList	remote.register
remote.destroy	remote.exec	remote.prelude	
replicator.bind	replicator.bind_dbat	replicator.setMaster	replicator.setVersion
sabaoth.epilogue	sabaoth.getLocalConnectionPort		sabaoth.marchScenario
sabaoth.getLocalConnectionHost		sabaoth.marchConnection	
scheduler.choice	scheduler.drop	scheduler.isolation	scheduler.pick
scheduler.costPrediction		scheduler.octopus	scheduler.volumeCost
sema.create	sema.destroy	sema.down	sema.up
sql.forgetPrevious	sql.keepquery	sql.queryId	
sqlblob.sqlblob			
statistics.close	statistics.forceUpdate		statistics.open
statistics.deposit	statistics.getCount	statistics.getObjects	
statistics.destroy	statistics.getHistogram		statistics.prelude
statistics.discard	statistics.getHotset	statistics.getSize	statistics.release
statistics.dump	statistics.getMax	statistics.hasMoreElements	
statistics.epilogue	statistics.getMin	statistics.newIterator	
status.batStatistics	status.getThreads	status.mem_cursize	status.vm_maxsize
status.cpuStatistics	status.ioStatistics	status.mem_maxsize	
status.getDatabases	status.memStatistics	status.vmStatistics	
status.getPorts	status.memUsage	status.vm_cursize	
str.+	str.length	str.rtrim	str.substitute
str.STRepilogue	str.like	str.search	str.substring
str.STRprelude	str.locate	str.space	str.suffix
str.ascii	str.ltrim	str.startsWith	str.toLower
str.chrAt	str.nbytes	str.str	str.toUpper
str.codeset	str.prefix	str.string	str.trim
str.endsWith	str.r_search	str.stringleft	str.unicode
str.iconv	str.repeat	str.stringlength	str.unicodeAt
str.insert	str.replace	str.stringright	
streams.blocked	streams.openReadBytes		streams.socketReadBytes
streams.close	streams.openWrite	streams.readStr	streams.socketWrite
streams.flush	streams.openWriteBytes		streams.socketWriteBytes

streams.openRead	streams.readInt	streams.socketRead	streams.writeInt
tablet.display	tablet.lastPage	tablet.setComplaints	tablet.setProperties
tablet.dump	tablet.load	tablet.setDecimal	tablet.setRowBracket
tablet.finish	tablet.nextPage	tablet.setDelimiter	tablet.setStream
tablet.firstPage	tablet.output	tablet.setFormat	tablet.setTableBracket
tablet.getPage	tablet.page	tablet.setName	tablet.setTryAll
tablet.getPageCnt	tablet.prevPage	tablet.setNull	tablet.setWidth
tablet.header	tablet.setBracket	tablet.setPivot	
tablet.input	tablet.setColumn	tablet.setPosition	
timestamp.!=	timestamp.==	timestamp.epoch	
timestamp.<	timestamp.>	timestamp.isnil	
timestamp.<=	timestamp.>=	timestamp.unix_epoch	
timezone.str	timezone.timestamp		
transaction.abort	transaction.clean	transaction.delta	transaction.subcommit
transaction.alpha	transaction.commit	transaction.prev	transaction.sync
txtsim.editdistance	txtsim.qgramnormalize		txtsim.stringdiff
txtsim.editdistance2	txtsim.qgramselfjoin	txtsim.soundex	
txtsim.levenshtein	txtsim.similarity	txtsim.str2qgrams	
unix.getenv	unix.setenv		
url.getAnchor	url.getDomain	url.getProtocol	url.isaURL
url.getBasename	url.getExtension	url.getQuery	url.new
url.getContent	url.getFile	url.getQueryArg	url.url
url.getContext	url.getHost	url.getRobotURL	
url.getDirectory	url.getPort	url.getUser	
user.main			
zrule.define			

Appendix B Instruction Help

The table below summarizes the commentary lines encountered in the system associated with a MAL kernel modules.

aggr.avg	Grouped tail average on dbl
aggr.cardinality	Return the cardinality of the BAT tail values.
aggr.count	Grouped count
aggr.count_no_nil	Return the number of elements currently in a BAT ignoring BUNs with nil-tail
aggr.histogram	Produce a BAT containing the histogram over the tail values.
aggr.max	Give the highest tail value.
aggr.min	Give the lowest tail value.
aggr.prod	Gives the product of all tail values.
aggr.product	Product over grouped tail on dbl
aggr.size	Grouped count of true values
aggr.sum	Grouped tail sum on dbl
alarm.alarm	execute action in X secs
alarm.ctime	current time as a string
alarm.epilogue	Finalize alarm module
alarm.epoch	current time as unix epoch
alarm.prelude	Initialize alarm module
alarm.sleep	sleep X secs
alarm.time	time in millisecs
alarm.timers	give a list of all active timers
alarm.usec	return cpu microseconds info
algebra.antijoin	Returns the antijoin
algebra.antiuselect	See select() but limited to head values
algebra.bandjoin	This is a join() for which the predicate is that two BUNs match if the left-tail value is within the range [right-head - minus, right-head + plus], depending on (l_in/h_in), the bounds are included. Works only for the builtin numerical types, and their derivates.
algebra.copy	Returns physical copy of a BAT.
algebra.crossproduct	Returns the cross product
algebra.difference	
algebra.exist	Returns true when 'h,t' occurs as a bun in b.
algebra.fetch	Returns a positional selection of b by the oid head values of s
algebra.fetchjoin	Hook directly into the fetch implementation of the join.
algebra.find	Returns the tail value 't' for which some [h,t] BUN exists in b. If no such BUN exists, an error occurs.
algebra.fragment	Select both on head and tail range.
algebra.groupby	Produces a new BAT with groups identified by the head column. The result contains tail times the head value, ie the tail contains the result group sizes.
algebra.hashjoin	Hook directly into the hash implementation of the join.

algebra.hashsplit	Split a BAT on tail column according (hash-value MOD buckets). Returns a recursive BAT, containing the fragments in the tail, their bucket number in the head.
algebra.indexjoin	Hook directly into the index implementation of the join.
algebra.intersect	
algebra.join	Returns all BUNs, consisting of a head-value from 'left' and a tail-value from 'right' for which there are BUNs in 'left' and 'right' with equal tail- resp. head-value (i.e. the join columns are projected out).
algebra.joinPath	internal routine to handle join paths. The type analysis is rather tricky.
algebra.kdifference	Returns the difference taken over only the *head* columns of two BATs. Results in all BUNs of 'left' that are *not* in 'right'. It does *not* do double-elimination over the 'left' BUNs. If you want this, use: 'kdifference(left.kunique,right.kunique)' or: 'kdifference(left,right).kunique'.
algebra.kintersect	Returns the intersection taken over only the *head* columns of two BATs. Results in all BUNs of 'left' that are also in 'right'. Does *not* do double- elimination over the 'left' BUNs. If you want this, use: 'kintersect(kunique(left),kunique(right))' or: 'kunique(kintersect(left,right))'.
algebra.kunion	Returns the union of two BATs; looking at head-columns only. Results in all BUNs of 'left' that are not in 'right', plus all BUNs of 'right'. *no* double-elimination is done. If you want this, do: 'kunion(left.kunique,right.kunique)' or: 'sunion(left,right).kunique'.
algebra.kunique	Select unique tuples from the input BAT. Double elimination is done only looking at the head column. The result is a BAT with property hkeyed() == true.
algebra.leftfetchjoin	Hook directly into the left fetch join implementation.
algebra.leftjoin	
algebra.like	Selects all elements that have 'substr' as in the tail.
algebra.markH	Produces a new BAT with fresh unique dense sequense of OIDs in the head that starts at base (i.e. [base,..base+b.count()-1]).
algebra.markT	Produces a BAT with fresh unique OIDs in the tail starting at `0@0`.
algebra.mark_grp	`"grouped mark"`: Produces a new BAT with per group a locally unique dense ascending sequense of OIDs in the tail. The tail of the first BAT (b) identifies the group that each BUN of b belongs to. The second BAT (g) represents the group extend, i.e., the head is the unique list of group IDs from b's tail. The third argument (s) gives the base value for the new OID sequence of each group.
algebra.materialize	Materialize the void column
algebra.merge	Merge head and tail into a single value
algebra.mergejoin	Hook directly into the merge implementation of the join.
algebra.number	Produces a new BAT with identical head column, and consecutively increasing integers (start at 0) in the tail column.

algebra.outerjoin	Returns all the result of a join, plus the BUNS formed NIL in the tail and the head-values of 'outer' whose tail-value does not match an head-value in 'inner'.
algebra.position	Returns the position of the value pair It returns an error if 'val' does not exist.
algebra.project	Fill the tail column with a constant taken from the aligned BAT.
algebra.rangesplit	Split a BAT on tail column in 'ranges' equally sized consecutive ranges. Returns a recursive BAT, containing the fragments in the tail, the higher-bound of the range in the head. The higher bound of the last range is 'nil'.
algebra.reuse	Reuse a temporary BAT if you can. Otherwise, allocate enough storage to accept result of an operation (not involving the heap)
algebra.revert	Returns a BAT copy with buns in reverse order
algebra.sample	Produce a random selection of size 'num' from the input BAT.
algebra.sdifference	Returns the difference taken over *both* columns of two BATs. Results in all BUNs of 'left' that are *not* in 'right'. Does *not* do double-elimination over the 'left' BUNs. If you want this, use: 'sdifference(left.sunique,right.sunique)' or: 'sdifference(left,right).sunique'.
algebra.select	Select all BUNs of a BAT with a certain tail value. Selection on NIL is also possible (it should be properly casted, e.g.:int(nil)).
algebra.selectH	
algebra.selectNotNil	Select all not-nil values
algebra.semijoin	Returns the intersection taken over only the *head* columns of two BATs. Results in all BUNs of 'left' that are also in 'right'. Does *not* do double-elimination over the 'left' BUNs. If you want this, use: 'kintersect(kunique(left),kunique(right))' or: 'kunique(kintersect(left,right))'.
algebra.sintersect	Returns the intersection taken over *both* columns of two BATs. Results in all BUNs of 'left' that are also in 'right'. Does *not* do double-elimination over the 'left' BUNs, If you want this, use: 'sintersect(sunique(left),sunique(right))' or: 'sunique(sintersect(left,right))'.
algebra.slice	Return the slice with the BUNs at position x till y.
algebra.sort	Returns a BAT copy sorted on the head column.
algebra.sortHT	Returns a lexicographically sorted copy on head,tail.
algebra.sortReverse	Returns a BAT copy reversely sorted on the tail column.
algebra.sortReverseTail	Returns a BAT copy reversely sorted on the tail column.
algebra.sortTH	Returns a lexicographically sorted copy on tail,head.
algebra.sortTail	Returns a BAT copy sorted on the tail column.
algebra.split	Split head into two values
algebra.ssort	Returns copy of a BAT with the BUNs sorted on ascending head values. This is a stable sort.
algebra.ssort_rev	Returns copy of a BAT with the BUNs sorted on descending head values. This is a stable sort.

algebra.sunion	Returns the union of two BATs; looking at both columns of both BATs. Results in all BUNs of 'left' that are not in 'right', plus all BUNs of 'right'. *no* double-elimination is done. If you want this, do: 'sunion(left.sunique,right.sunique)' or: 'sunion(left,right).sunique'.
algebra.sunique	Select unique tuples from the input BAT. Double elimination is done over BUNs as a whole (head and tail). Result is a BAT with real set() semantics.
algebra.thetajoin	Theta join on for 'mode' in { LE, LT, EQ, GT, GE }. JOIN_EQ is just the same as join(). All other options do merge algorithms. Either using the fact that they are ordered() already (left on tail, right on head), or by using/creating binary search trees on the join columns.
algebra.thetaselect	The theta (<=,<,=,>,>=) select()
algebra.thetauselect	The theta (<=,<,=,>,>=) select() limited to head values
algebra.topN	Trim all but the top N tuples.
algebra.tunique	Select unique tuples from the input BAT. Double elimination is done over the BUNs tail. The result is a BAT with property tkeyd()== true
algebra.uhashsplit	Same as hashsplit, but only collect the head values in the fragments
algebra.union	
algebra.unique	
algebra.urangesplit	Same as rangesplit, but only collect the head values in the fragments
algebra.uselect	Value select, but returning only the head values. SEE ALSO:select(bat,val)
array.grid	Fills an index BAT, (grpcount,grpsize,clustersize,offset) and shift all elemenets with a factor s
array.product	Produce an array product
array.project	Fill an array representation with constants
bat.append	append the value u to i
bat.attach	Returns a new BAT with dense head and tail of the given type and uses the given file to initialize the tail. The file will be owned by the server.
bat.delete	Delete from the first BAT all BUNs with a corresponding BUN in the second.
bat.densebat	Creates a new [void,void] BAT of size 'size'.
bat.flush	Designate a BAT as not needed anymore.
bat.getAccess	return the access mode attached to this BAT as a character.
bat.getAlpha	Obtain the list of BUNs added
bat.getCapacity	Returns the current allocation size (in max number of elements) of a BAT.
bat.getDelta	Obtain the list of BUNs deleted
bat.getDiskSize	Approximate size of the (persistent) BAT heaps as stored on disk in pages of 512 bytes. Indices are not included, as they only live temporarily in virtual memory.
bat.getHead	return the BUN head value using the cursor.
bat.getHeadType	Returns the type of the head column of a BAT, as an integer type number.
bat.getHeat	Return the current BBP heat (LRU stamp)

bat.getMemorySize	Calculate the size of the BAT heaps and indices in bytes rounded to the memory page size (see bbp.getPageSize()).
bat.getName	Gives back the logical name of a BAT.
bat.getRole	Returns the rolename of the head column of a BAT.
bat.getSequenceBase	Get the sequence base for the void column of a BAT.
bat.getSize	Calculate the size of the BAT descriptor, heaps and indices in bytes.
bat.getSpaceUsed	Determine the total space (in bytes) occupied by a BAT.
bat.getStorageSize	Determine the total space (in bytes) reserved for a BAT.
bat.getTail	return the BUN tail value using the cursor.
bat.getTailType	Returns the type of the tail column of a BAT, as an integer type number.
bat.hasAppendMode	return true if to this BAT is append only.
bat.hasMoreElements	Produce the next bun for processing.
bat.hasReadMode	return true if to this BAT is read only.
bat.hasWriteMode	return true if to this BAT is read and write.
bat.info	Produce a BAT containing info about a BAT in [attribute,value] format. It contains all properties of the BAT record. See the BAT documentation in GDK for more information.
bat.inplace	inplace replace values on the given locations
bat.insert	Insert one BUN[h,t] in a BAT.
bat.isCached	Bat is stored in main memory.
bat.isPersistent	
bat.isSorted	Returns whether a BAT is ordered on head or not.
bat.isSortedReverse	Returns whether a BAT is ordered on head or not.
bat.isSynced	Tests whether two BATs are synced or not.
bat.isTransient	
bat.isaKey	return whether the head column of a BAT is unique (key).
bat.isaSet	return whether the BAT mode is set to unique.
bat.load	Load a particular BAT from disk
bat.mirror	Returns the head-mirror image of a BAT (two head columns).
bat.new	Localize a bat by name and produce a clone.
bat.newIterator	Process the buns one by one extracted from a void table.
bat.order	Sorts the BAT itself on the head, in place.
bat.orderReverse	Reverse sorts the BAT itself on the head, in place.
bat.pack	Pack a pair of values into a BAT.
bat.partition	Create a series of cheap slices over the first argument
bat.reduce	Drop auxillary BAT structures.
bat.replace	Replace the tail value of one BUN that has some head value.
bat.reverse	Returns the reverse view of a BAT (head is tail and tail is head). BEWARE no copying is involved; input and output refer to the same object!
bat.revert	Puts all BUNs in a BAT in reverse order. (Belongs to the BAT sequence module)
bat.save	Save a BAT to storage, if it was loaded and dirty. Returns whether IO was necessary. Please realize that calling this function violates the atomic commit protocol!!

bat.setAccess	Try to change the update access priviliges to this BAT. Mode: r[ead-only] - allow only read access. a[append-only] - allow reads and update. w[riteable] - allow all operations. BATs are updatable by default. On making a BAT read-only, all subsequent updates fail with an error message.Returns the BAT itself.
bat.setAppendMode	Change access privilige of BAT to append only
bat.setBase	Give the non-empty BATs consecutive oid bases.
bat.setCold	Makes a BAT very cold for the BBP. The chance of being choses for swapout is big, afterwards.
bat.setColumn	Give both columns of a BAT a new name.
bat.setGarbage	Designate a BAT as garbage.
bat.setHash	
bat.setHot	Makes a BAT very hot for the BBP. The chance of being chosen for swapout is small, afterwards.
bat.setKey	Sets the 'key' property of the head column to 'mode'. In 'key' mode, the kernel will silently block insertions that cause a duplicate entries in the head column. KNOWN BUG:when 'key' is set to TRUE, this function does not automatically eliminate duplicates. Use b := b.kunique;
bat.setMemoryAdvise	alias for madvise(b, mode, mode, mode, mode)
bat.setMemoryMap	Alias for mmap(b, mode, mode, mode, mode)
bat.setName	Give a logical name to a BAT.
bat.setPersistent	Make the BAT persistent. Returns boolean which indicates if the BAT administration has indeed changed.
bat.setReadMode	Change access privilige of BAT to read only
bat.setRole	Give a logical name to the columns of a BAT.
bat.setSet	Sets the 'set' property on this BAT to 'mode'. In 'set' mode, the kernel will silently block insertions that cause a duplicate BUN [head,tail] entries in the BAT. KNOWN BUG:when 'set' is set to TRUE, this function does not automatically eliminate duplicates. Use b := b.sunique; Returns the BAT itself.
bat.setSorted	Assure BAT is ordered on the head.
bat.setTransient	Make the BAT transient. Returns boolean which indicates if the BAT administration has indeed changed.
bat.setWriteMode	Change access privilige of BAT to read and write
bat.unload	Swapout a BAT to disk. Transient BATs can also be swapped out. Returns whether the unload indeed happened.
bat.unpack	Extract the first tuple from a BAT.
batcalc.!=	Equate a bat of strings against a singleton
batcalc.%	Binary BAT calculator function with new BAT result
batcalc.*	Binary BAT calculator function with new BAT result
batcalc.+	Concatenate two strings.
batcalc.++	Unary minus over the tail of the bat
batcalc.-	Unary minus over the tail of the bat
batcalc.–	Unary minus over the tail of the bat
batcalc./	Binary BAT calculator function with new BAT result
batcalc.<	Compare a bat of timestamp against a singleton

batcalc.<=	Compare a bat of timestamp against a singleton
batcalc.==	Equate a bat of strings against a singleton
batcalc.>	Compare a bat of timestamp against a singleton
batcalc.>=	Compare a bat of timestamp against a singleton
batcalc.abs	Unary abs over the tail of the bat
batcalc.and	Binary BAT calculator function with new BAT result
batcalc.bit	Coerce an str tail to a bat with bit tail.
batcalc.bte	Coerce an bit tail to a bat with bte tail.
batcalc.chr	
batcalc.dbl	Coerce an flt tail to a bat with dbl tail.
batcalc.flt	Coerce an dbl tail to a bat with flt tail.
batcalc.ifthen	Ifthen operation to assemble a conditional result
batcalc.ifthenelse	If-then-else operation to assemble a conditional result
batcalc.int	Coerce an str tail to a bat with int tail.
batcalc.isnil	Unary check for nil over the tail of the bat
batcalc.lng	Coerce an bit tail to a bat with lng tail.
batcalc.not	Return a BAT with the negated tail
batcalc.oid	Coerce an lng tail to a bat with oid tail.
batcalc.or	Binary BAT calculator function with new BAT result
batcalc.sht	Coerce an bit tail to a bat with sht tail.
batcalc.str	
batcalc.wrd	Coerce an bit tail to a bat with wrd tail.
batcalc.xor	Binary BAT calculator function with new BAT result
batcolor.blue	Extracts blue component from a color atom
batcolor.cb	Extracts Cb(blue color) component from a color atom
batcolor.color	Converts string to color
batcolor.cr	Extracts Cr(red color) component from a color atom
batcolor.green	Extracts green component from a color atom
batcolor.hsv	Converts an HSV triplets to a color atom
batcolor.hue	Extracts hue component from a color atom
batcolor.luminance	Extracts Y(luminance) component from a color atom
batcolor.red	Extracts red component from a color atom
batcolor.rgb	Converts an RGB triplets to a color atom
batcolor.saturation	Extracts saturation component from a color atom
batcolor.str	Identity mapping for string bats
batcolor.value	Extracts value component from a color atom
batmmath.acos	
batmmath.asin	
batmmath.atan	
batmmath.atan2	
batmmath.ceil	
batmmath.cos	
batmmath.cosh	
batmmath.exp	
batmmath.fabs	
batmmath.floor	
batmmath.fmod	

batmmath.log	
batmmath.log10	
batmmath.pow	
batmmath.sin	
batmmath.sinh	
batmmath.sqrt	
batmmath.tan	
batmmath.tanh	
batmtime.day	
batmtime.hours	
batmtime.milliseconds	
batmtime.month	
batmtime.seconds	
batmtime.year	
batstr.chrAt	String array lookup operation.
batstr.endsWith	Suffix check.
batstr.length	Return the length of a string.
batstr.like	
batstr.like_uselect	Perform SQL like operation against a string bat
batstr.ltrim	Strip whitespaces from start of a string.
batstr.nbytes	Return the string length in bytes.
batstr.r_search	Reverse search for a substring. Returns position, -1 if not found.
batstr.replace	Insert a string into another
batstr.rtrim	Strip whitespaces from end of a string.
batstr.search	Search for a substring. Returns position, -1 if not found.
batstr.startsWith	Prefix check.
batstr.string	Return the tail s[offset..n] of a string s[0..n].
batstr.substitute	Substitute first occurrence of 'src' by 'dst'. Iff repeated = true this is repeated while 'src' can be found in the result string. In order to prevent recursion and result strings of unlimited size, repeating is only done iff src is not a substring of dst.
batstr.substring	Substring extraction using [start,start+length]
batstr.toLower	Convert a string to lower case.
batstr.toUpper	Convert a string to upper case.
batstr.trim	Strip whitespaces around a string.
batstr.unicodeAt	get a unicode character (as an int) from a string position.
bbp.bind	Locate the BAT using its BBP index in the BAT buffer pool
bbp.close	Close the bbp box.
bbp.commit	Commit updates for this client.
bbp.deposit	Relate a logical name to a physical BAT in the buffer pool.
bbp.destroy	Schedule a BAT for removal at session end or immediately.
bbp.discard	Remove the BAT from the box.
bbp.getCount	Create a BAT with the cardinalities of all known BATs
bbp.getDirty	Create a BAT with the dirty/ diffs/clean status
bbp.getDiskSpace	Estimate the amount of disk space occupied by dbfarm
bbp.getHeadType	Map a BAT into its head type

bbp.getHeat	Create a BAT with the heat values
bbp.getKind	Create a BAT with the persistency status
bbp.getLRefCount	Utility for debugging MAL interpreter
bbp.getLocation	Create a BAT with their disk locations
bbp.getName	Map a BAT into its internal name
bbp.getNames	Map BAT into its bbp name
bbp.getObjects	View of the box content.
bbp.getPageSize	Obtain the memory page size
bbp.getRNames	Map a BAT into its bbp physical name
bbp.getRefCount	Utility for debugging MAL interpreter
bbp.getStatus	Create a BAT with the disk/load status
bbp.getTailType	Map a BAT into its tail type
bbp.iterator	Locate the next element in the box.
bbp.open	Locate the bbp box and open it.
bbp.prelude	Initialize the bbp box.
bbp.release	Remove the BAT from further consideration
bbp.releaseAll	Commit updates for this client.
bbp.take	Load a particular bat.
bbp.toString	Get the string representation of an element in the box.
blob.blob	Noop routine.
blob.nitems	get the number of bytes in this blob.
blob.prelude	
blob.toblob	store a string as a blob.
blob.tostring	get the bytes from blob as a string, starting at byte 'index' till the first 0 byte or the end of the blob.
box.close	Close the box.
box.deposit	Enter a new value into the box.
box.destroy	Destroy the box.
box.discard	Release the BAT from the client pool.
box.getBoxNames	Retrieve the names of all boxes.
box.iterator	Locates the next element in the box.
box.open	Locate the box and open it.
box.release	Release the BAT from the client pool.
box.releaseAll	Release all objects for this client.
box.take	Locate the typed value in the box.
box.toString	Get the string representation of the i-th element in the box.
bpm.adapt	Re-organize segment s using the selection (val1,val2) stored in bat rs.
bpm.addPartition	Add a partition to a fragmented temporary table.
bpm.close	Save and close the BAT partition box.
bpm.count	
bpm.delete	Delete elements from the BAT partitions.
bpm.deposit	Create a new partitioned BAT by name.
bpm.derivePartition	Create a derived fragmentation over the head using src.
bpm.destroy	Destroy the BAT partition box.
bpm.discard	Release all partitioned BATs.
bpm.dump	Give the details of the partition tree

bpm.emptySet	Implement the empty set constraints test efficiently.
bpm.epilogue	
bpm.fold	Collapse the partitioned BAT into a single BAT.
bpm.garbage	Remove a temporary partitioned table.
bpm.getDimension	Obtain the partition boundary values.
bpm.getNames	Retrieve the names of all known partitioned BATs.
bpm.getNumberOfPartitions	Return the number of partitions known
bpm.hasMoreElements	Localize the next partition for processing.
bpm.hashPartition	Create a hash partition on a BAT.
bpm.hashPartitions	Ensure all partitions have a hash in the head.
bpm.insert	Insert elements into the BAT partitions.
bpm.mapAlias	
bpm.mapBid	
bpm.mapName	
bpm.mapNxt	
bpm.mapPrv	
bpm.mapThghDbl	
bpm.mapThghLng	
bpm.mapTlowDbl	
bpm.mapTlowLng	
bpm.new	Create a temporary partitioned table.
bpm.newIterator	Create an iterator over the BAT partitions.
bpm.open	Locate and open the BAT partition box.
bpm.partition	Split all partitions that cover the split value.
bpm.pieces	Count the number of partitions.
bpm.prelude	
bpm.rangePartition	Create the partitions based on a range vector.
bpm.rangePartitionSort	Create the partitions based on a range vector.
bpm.replace	Replace the content of the BAT partitions.
bpm.saveCatalog	
bpm.select	Partitioned based selection
bpm.sortPartitions	Sort all partitions of alias b on the tail.
bpm.sortTail	Implement the sort on tail for partitioned BAT efficiently.
bpm.splitquant	Split all partitions to fit into a memory bound in KB
bpm.take	Retrieve a single component of a partitioned BAT by index.
bpm.unfold	Unfold a BAT into a partitioned one.
bstream.create	create a buffered stream
bstream.destroy	destroy bstream
bstream.read	read at least size bytes into the buffer of s
calc.!=	Inequality of two inets
calc.%	
calc.*	
calc.+	Concatenate two strings
calc.-	negative value
calc./	

calc.<	Whether v is less than w
calc.<<	
calc.<=	Whether v is less than or equal to w
calc.=	Equality of two inets
calc.==	Equality of two timestamps
calc.>	Whether v is greater than w
calc.>=	Whether v is equal to or greater than w
calc.>>	
calc.abs	absolute value
calc.and	
calc.bat	
calc.between	
calc.bit	coercion dbl to bit
calc.blob	
calc.bte	coercion lng to bte
calc.chr	coercion lng to chr
calc.date	
calc.daytime	
calc.dbl	coercion lng to dbl
calc.flt	coercion lng to flt
calc.getBAT	Coerce bat to BAT identifier
calc.getBATidentifier	Coerce bat to BAT identifier
calc.ifthenelse	
calc.inet	Convert a string to an inet
calc.int	coercion dbl to int
calc.inv	inverse value (1/x)
calc.isnil	Nil test for inet value
calc.isnotnil	is a value not equal to nil?
calc.length	
calc.lng	coercion dbl to lng
calc.max	Maximum test for timestamp value
calc.max_no_nil	Maximum test for timestamp value
calc.min	Minimum test for timestamp value
calc.min_no_nil	Minimum test for timestamp value
calc.newoid	Reserves a range of consecutive unique OIDs; returns the lowest in range. equivalent to newoid(0,incr)
calc.not	
calc.oid	coercion dbl to oid
calc.or	
calc.ptr	
calc.setoid	Equivalent to setoid(1:oid).
calc.sht	coercion dbl to sht
calc.sign	Returns +1, 0, -1 based on the sign of the given expression
calc.sizeof	
calc.sqladd	

calc.sqlblob	
calc.str	coercion dbl to str
calc.timestamp	
calc.void	
calc.wrd	coercion dbl to wrd
calc.xor	
clients.addScenario	add the given scenario to the allowed scenarios for the given user
clients.addUser	Allow user with password access to the given scenarios
clients.changePassword	Change the password for the current user
clients.changeUsername	Change the username of the user into the new string
clients.checkPermission	Check permission for a user
clients.exit	Terminate the session for a single client using a soft error.
clients.getActions	Pseudo bat of client's command counts.
clients.getId	Return a number that uniquely represents the current client.
clients.getInfo	Pseudo bat with client attributes.
clients.getLastCommand	Pseudo bat of client's last command time.
clients.getLogins	Pseudo bat of client login time.
clients.getScenario	Retrieve current scenario name.
clients.getTime	Pseudo bat of client's total time usage(in usec).
clients.getUsername	Return the username of the currently logged in user
clients.getUsers	return a BAT with user id and name available in the system with access to the given scenario(s)
clients.quit	Terminate the server. This command can only be initiated from the console.
clients.removeScenario	remove the given scenario from the allowed scenarios for the given user
clients.removeUser	Remove the given user from the system
clients.setHistory	Designate console history file for readline.
clients.setListing	Turn on/off echo of MAL instructions: 2 - show mal instruction, 4 - show details of type resolutoin, 8 - show binding information.
clients.setPassword	Set the password for the given user
clients.setScenario	Switch to other scenario handler, return previous one.
clients.shutdown	Close all client connections. If forced=false the clients are moved into FINISHING mode, which means that the process stops at the next cycle of the scenario. If forced=true all client processes are immediately killed
clients.stop	Stop the query execution at the next eligble statement.
clients.suspend	Put a client process to sleep for some time. It will simple sleep for a second at a time, until the awake bit has been set in its descriptor
clients.wakeup	Wakeup a client process
cluster.column	Reorder tail of the BAT using the cluster map
cluster.key	Create the hash key list

cluster.map	Reorder tail of bat b, using a cluster map
cluster.new	Compute the cluster map for bat b of hash key values. A cluster map is a list of unique (new) BUN positions. The p(refix) sum is a by product which returns the prefix sum of the per masked key frequency.
cluster.table	Cluster the BATs using the first one as reference. Return the oid map used
color.blue	Extracts blue component from a color atom
color.cb	Extracts Cb(blue color) component from a color atom
color.color	Converts string to color
color.cr	Extracts Cr(red color) component from a color atom
color.green	Extracts green component from a color atom
color.hsv	Converts an HSV triplets to a color atom
color.hue	Extracts hue component from a color atom
color.luminance	Extracts Y(luminance) component from a color atom
color.red	Extracts red component from a color atom
color.rgb	Converts an RGB triplets to a color atom
color.saturation	Extracts saturation component from a color atom
color.str	Converts color to string
color.value	Extracts value component from a color atom
color.ycc	Converts an YCC triplets to a color atom
const.close	Close the constant box.
const.deposit	Add a variable to the box.
const.destroy	Destroy the box.
const.discard	Release the const from the box.
const.epiloque	Cleanup the const box
const.hasMoreElements	Locate next element in the box.
const.newIterator	Locate next element in the box.
const.open	Locate and open the constant box.
const.prelude	Initialize the const box
const.release	Release a constant value.
const.releaseAll	Release all variables in the box.
const.take	Take a variable out of the box.
const.toString	Get the string representation of an element in the box.
constraints.emptySet	Check if the BAT is empty.
crackers.DeleteMap	Throw away a certain map
crackers.InsertAVLIndex	Insert u in the AVL tree index of BAT b
crackers.activeCacheConsciousCrackHashJoin	Join two maps based on head values with active cracking. Align the maps to avoid overlapping pieces. Reuse hash tables
crackers.alignJoin	Join and on the fly align a map with an intermediate result bat, i.e., not cracked
crackers.alignedJoin	Join an aligned cracker bat with a map
crackers.bandJoin	Band Join two maps based on head values. Continuously crack the right BAT for each tuple of the left one

crackers.buildAVLIndex	Create an AVL tree index for this BAT
crackers.cacheConsciousCrackHashJoin	Join two maps based on head values. Align the maps to avoid overlapping pieces. Reuse hash tables
crackers.cacheConsciousCrackHashJoinAlignOnly	Join two maps based on head values. Align the maps to avoid overlapping pieces. Reuse hash tables
crackers.crackHashJoin	Join two maps based on head values. Align the maps to avoid overlapping pieces. Reuse hash tables
crackers.crackJoin	Join two maps based on head values. Align the maps to avoid overlapping pieces
crackers.crackOrdered	Break a BAT into three pieces with tail<mid, tail==mid, tail>mid, respectively; maintaining the head-oid order within each piece.
crackers.crackOrdered_validate	Validate whether a BAT is correctly broken into five pieces with tail<low, tail==low, low<tail<hgh, tail==hgh, tail>hgh, respectively; maintaining the head-oid order within each piece.
crackers.crackUnordered_validate	Validate whether a BAT is correctly broken into five pieces with tail<low, tail==low, low<tail<hgh, tail==hgh, tail>hgh, respectively.
crackers.deleteAVL	Delete a collection of values from the index
crackers.deletionsOnNeed	Keep the deletions BAT separatelly and do a complete merge only if a relevant query arrives in the future
crackers.deletionsOnNeedGradually	Keep the deletions BAT separatelly and merge only what is needed if a relevant query arrives in the future
crackers.deletionsOnNeedGraduallyRipple	Keep the deletions BAT separatelly and merge only what is needed using ripple if a relevant query arrives in the future
crackers.djoinselect	Use the pivot. For each tuple in pivot with a 0, check if the respective tuple (in the same position) in the tail of cpair satisfies the range restriction. If yes mark the pivot BUN as 1.
crackers.dproject	Sync the cracking pair and project the tail. Use for disjunctive queries that require a larger bit vector
crackers.dselect	Crack based on dbl and evaluate the dbl disjunctive predicate outside the cracked area. Return a bit vector.
crackers.extendCrackerBAT	Extend the cracker column by P positions
crackers.extendCrackerMap	Extend the cracker map by P positions
crackers.fmaddReference	

	add bp reference to map set of b
crackers.fmclearReferences	clear all references
crackers.fmcreateMap	make new map for debugging
crackers.fmremoveMap	clear all debugging map
crackers.fullAlignment	Align a bat with the cracks on a map
crackers.getCrackerBAT	Get the cracker BAT of b
crackers.getMap	Get a certain map
crackers.getTotalStorage	Get the number of total tuples stored in sideways maps
crackers.hselect	Retrieve the subset head using a cracker index producing preferably a BATview.
crackers.insert	Keep the insertions BAT separatelly and merge in the future on demand with the Ripple
crackers.insertionsBForce	Merge the insertions BAT with the cracker bat and update the cracker index
crackers.insertionsBOnNeed	Keep the insertions BAT separatelly and do a complete merge only if a relevant query arrives in the future
crackers.insertionsBOnNeedGradually	Keep the insertions BAT separatelly and merge only what is needed if a relevant query arrives in the future
crackers.insertionsBOnNeedGraduallyRipple	Keep the insertions BAT separatelly and merge only what is needed using the ripple strategy if a relevant query arrives in the future
crackers.insertionsForget	Append c to the cracked BAT of b and completelly forget the cracker index
crackers.insertionsPartiallyForget	Append c to the cracked BAT of b and partially forget the cracker index, i.e., forget only what is affected
crackers.joinselect	Use the pivot. For each tuple in pivot with a 1, check if the respective tuple (in the same position) in the tail of cpair satisfies the range restriction. If not mark the pivot BUN as 0.
crackers.joinuselect	Join left and right on head-OIDs. From right, only those BUNs qualify that satisfy the range-restriction on the tail. If inPlace is TRUE (and left has an OID head and is not a BAT-view), we operate in-place, overwriting left and returning it as result. Otherwise, the result is a new [:oid,:void] BAT. If isForeignKey is TRUE, we assume that each tuple from left finds a match in right, and hence skip the respective check. (NOTE: This may lead to CRASHES, if isForeignKey is incorrectly passed as TRUE!)
crackers.mapCount	Retrieve the size of the map

crackers.markedproject	Sync the cracking pair and project the tail. The result bat has a marked head
crackers.materializeHead	Materialize the head of BAT b
crackers.pmaddReference	add bp reference to map set of b
crackers.pmclearReferences	clear all references to b
crackers.pmjoinselect	Use the pivot. For each tuple in pivot with a 1, check if the respective tuple (in the same position) in the tail of cpair(collection of pieces) satisfies the range restriction. If not mark the pivot BUN as 0.
crackers.pmmaxTail	Sync/crack the map and get the max of the tail
crackers.pmproject	Sync the map and project the tail based on the pivot
crackers.pmselect	Crack based on dbl and evaluate the dbl conjunctive predicate. Return a bit vector.
crackers.pmtselect	Crack based on dbl and project the dbl tail .
crackers.positionproject	Sync the cracking pair and project the tail. The pivot holds the positions to be projected
crackers.printAVLTree_int	Print the AVL Tree of the cracker index (for debugging purposes)
crackers.printCrackerBAT	Print the cracker BAT of b
crackers.printCrackerDeletions	Print the pending deletions of the cracker BAT of b
crackers.printCrackerIndexBATpart	Print the cracker index of b
crackers.printCrackerInsertions	Print the pending insertions of the cracker BAT of b
crackers.printPendingInsertions	Print the pending insertions
crackers.project	Sync the cracking pair and project the tail
crackers.projectH	Sync the cracking pair and project the head
crackers.select	Retrieve the subset using a cracker index producing preferably a BATview.
crackers.select2	Similar to select but always make sure that we do not create a large piece i.e., bigger than half the size of the cracked piece
crackers.selectAVL	Retrieve the subset using the AVL index
crackers.setStorageThreshold	set the maximum number of total tuples that can be stored in sideways maps
crackers.simpleJoin	Join two maps based on head values by exploiting the already existing partitioning information
crackers.singlePassJoin	

	First partition on separate pieces the left input based on the right index. Then join matching pieces
crackers.sizeCrackerDeletions	Get the size of the pending deletions of the cracker BAT of b
crackers.sizeCrackerInsertions	Get the size of the pending insertions of the cracker BAT of b
crackers.sizePendingInsertions	Get the size of the pending insertions for this map
crackers.sortBandJoin	Band Join two maps based on head values. First sort the right BAT and then continuously binary search the right BAT for each tuple of the left one
crackers.tselect	Retrieve the subset tail using a cracker index producing preferably a BATview.
crackers.uselect	Retrieve the subset using a cracker index producing preferably a BATview.
crackers.verifyCrackerIndex	Check the cracker index and column, whether each value is in the correct chunk
crackers.zcrackOrdered	Break a BAT into three pieces with tail<=low, low<tail<=hgh, tail>hgh, respectively; maintaining the head-oid order within each piece.
crackers.zcrackOrdered_validate	Validate whether a BAT is correctly broken into three pieces with tail<=low, low<tail<=hgh, tail>hgh, respectively; maintaining the head-oid order within each piece.
crackers.zcrackUnordered	Break a BAT into three pieces with tail<=low, low<tail<=hgh, tail>hgh, respectively.
crackers.zcrackUnordered_validate	Validate whether a BAT is correctly broken into three pieces with tail<=low, low<tail<=hgh, tail>hgh, respectively.
date.!=	Equality of two dates
date.<	Equality of two dates
date.<=	Equality of two dates
date.==	Equality of two dates
date.>	Equality of two dates
date.>=	Equality of two dates
date.date	Noop routine.
date.isnil	Nil test for date value
daytime.!=	Equality of two daytimes
daytime.<	Equality of two daytimes
daytime.<=	Equality of two daytimes
daytime.==	Equality of two daytimes
daytime.>	Equality of two daytimes

daytime.>=	Equality of two daytimes
daytime.isnil	Nil test for daytime value
factories.getArrival	Retrieve the time stamp the last call was made.
factories.getCaller	Retrieve the unique identity of the factory caller.
factories.getDeparture	Retrieve the time stamp the last answer was returned.
factories.getOwners	Retrieve the factory owners table.
factories.getPlants	Retrieve the names for all active factories.
factories.shutdown	Close a factory.
group.avg	grouped tail average
group.count	Grouped count
group.derive	Cross tabulation group extension step. Returned head values are identical as in 'ct'. Tail values are from the same domain and indicate further refinement of the groups in 'ct', taking into account also the tail-values in 'attr'.
group.max	Select the minimum element of each group
group.min	Select the minimum element of each group
group.new	Cross tabulation group initialization like GRPgroup, but with user provided #bits in hashmask and #distinct values in range.
group.prelude	
group.refine	refine the ordering of a tail-ordered BAT by sub-ordering on the values of a second bat 'a' (where the heads of a and b match 1-1). The effect of this is similar to (hash-based) GRPderive, with the distinction that the group ids respect the ordering of the group values.
group.refine_reverse	refine the ordering of a tail-ordered BAT by sub-ordering on the values of a second bat 'a' (where the heads of a and b match 1-1). The effect of this is similar to (hash-based) GRPderive, with the distinction that the group ids respect the ordering of the group values.
group.size	Grouped count of true values
group.sum	Tail sum of groups of a sliding window of fixed size
group.variance	grouped tail variance
identifier.identifier	Cast a string to an identifer
identifier.prelude	Initialize the module
inet.!=	Inequality of two inets
inet.<	Whether v is less than w
inet.<<	Whether v is contained within w
inet.<<=	Whether v is contained within or is equal to w
inet.<=	Whether v is less than or equal to w
inet.=	Equality of two inets
inet.>	Whether v is greater than w
inet.>=	Whether v is equal to or greater than w
inet.>>	Whether v contains w
inet.>>=	Whether v contains or is equal to w
inet.abbrev	Abbreviated display format as text
inet.broadcast	Returns the broadcast address for network

inet.host	Extract IP address as text
inet.hostmask	Construct host mask for network
inet.isnil	Nil test for inet value
inet.masklen	Extract netmask length
inet.netmask	Construct netmask for network
inet.network	Extract network part of address
inet.new	Create an inet from a string literal
inet.setmasklen	Set netmask length for inet value
inet.text	Extract IP address and netmask length as text
inspect.equalType	Return true if both operands are of the same type
inspect.getAddress	Returns the function signature(s).
inspect.getAddresses	Obtain the function address.
inspect.getAtomNames	Collect a BAT with the atom names.
inspect.getAtomSizes	Collect a BAT with the atom sizes.
inspect.getAtomSuper	Collect a BAT with the atom names.
inspect.getComment	Returns the function help information.
inspect.getDefinition	Returns a string representation of a specific function.
inspect.getEnvironment	Collect the environment variables.
inspect.getFunction	Obtain the function name.
inspect.getKind	Obtain the instruction kind.
inspect.getModule	Obtain the function name.
inspect.getSignature	Returns the function signature(s).
inspect.getSignatures	Obtain the function signatures.
inspect.getSize	Return the storage size for a function (in bytes).
inspect.getSource	Return the original input for a function.
inspect.getStatistics	Get optimizer property statistics such as #calls, #total actions, #total time
inspect.getType	Return the concrete type of a variable (expression).
inspect.getTypeIndex	Return the type index of a variable. For BATs, return the type index for its tail.
inspect.getTypeName	Get the type name associated with a type id.
inspect.getWelcome	Return the server message of the day string
io.data	Signals receipt of tuples in a file fname. It returns the name of the file, if it still exists.
io.export	Export a BAT as ASCII to a file. If the 'filepath' is not absolute, it is put into the .../dbfarm/$DB directory. Success of failure is indicated.
io.ftable	Print an n-ary table to a file.
io.import	Import a BAT from an ASCII dump. The new tuples are *inserted* into the parameter BAT. You have to create it! Its signature must match the dump, else parsing errors will occur and FALSE is returned.
io.print	Print a MAL value tuple .
io.printf	Select default format

io.prompt	Print a MAL value without brackets.
io.stderr	return the error stream for the database console
io.stdin	return the input stream to the database client
io.stdout	return the output stream for the database client
io.table	Print an n-ary table.
language.assert	Assertion test.
language.assertSpace	Ensures that the current call does not consume more than depth*vtop elements on the stack.
language.call	Evaluate a program stored in a BAT.
language.dataflow	The current guarded block is executed using dataflow control.
language.newRange	This routine introduces an iterator over a scalar domain.
language.nextElement	Advances the iterator with a fixed value until it becomes >= last.
language.raise	Raise an exception labeled with a specific message.
language.register	Compile the code string and register it as a MAL function.
language.setIOTrace	Set the flag to trace the IO
language.setMemoryTrace	Set the flag to trace the memory footprint
language.setThreadTrace	Set the flag to trace the interpreter threads
language.setTimerTrace	Set the flag to trace the execution time
language.source	Merge the instructions stored in the file with the current program.
lock.create	Create an unset lock
lock.destroy	Destroy a lock
lock.set	Try to set a lock. If set, block till it is freed
lock.tostr	Overloaded atom function
lock.try	Try a lock. If free set it, if not return EBUSY
lock.unset	Unset a lock
mal.multiplex	
manual.completion	Produces the wordcompletion table.
manual.createXML	Produces a XML-formatted manual over all modules loaded.
manual.help	Produces a list of all <module>.<function> that match the text pattern. The wildcard '*' can be used for <module> and <function>. Using the '(' asks for signature information and using ')' asks for the complete help record.
manual.index	Produces an overview of all names grouped by module.
manual.search	Search the manual for command descriptions that match the regular expression 'text'
manual.section	Generate a synopsis of a module for the reference manual
manual.summary	Produces a manual with help lines grouped by module.
mapi.bind	Bind a remote variable to a local one.
mapi.connect	Establish connection with a remote mserver.
mapi.connect_ssl	Establish connection with a remote mserver using the secure socket layer.
mapi.destroy	Destroy the handle for an Mserver.
mapi.disconnect	Terminate the session.

mapi.error	Check for an error in the communication.
mapi.explain	Turn the error seen into a string.
mapi.fetch_all_rows	Retrieve all rows into the cache.
mapi.fetch_field	Retrieve a single chr field.
mapi.fetch_field_array	Retrieve all fields for a row.
mapi.fetch_line	Retrieve a complete line.
mapi.fetch_reset	Reset the cache read line.
mapi.fetch_row	Retrieve the next row for analysis.
mapi.finish	Remove all remaining answers.
mapi.getError	Get error message.
mapi.get_field_count	Return number of fields.
mapi.get_row_count	Return number of rows.
mapi.listen	Start the Mapi listener on <port> for <maxusers>. For a new client connection MAL procedure <cmd>(Stream s_in, Stream s_out) is called.If no <cmd> is specified a new client thread is forked.
mapi.listen_ssl	Start the Mapi listener on <port> for <maxusers> using SSL. <keyfile> and <certfile> give the path names for files with the server key and certificates in PEM format. For a new client connection MAL procedure <cmd>(Stream s_in, Stream s_out) is called. If no <cmd> is specified a new client thread is forked.
mapi.lookup	Retrieve the connection identifier.
mapi.malclient	Start a Mapi client for a particular stream pair.
mapi.next_result	Go to next result set.
mapi.ping	Test availability of an Mserver.
mapi.prepare	Prepare a query for execution.
mapi.put	Prepare sending a value to a remote site.
mapi.query	Sent the query for execution
mapi.query_array	Sent the query for execution replacing '?' by arguments.
mapi.query_handle	Sent the query for execution.
mapi.reconnect	Re-establish a connection.
mapi.resume	Resume connection listeners.
mapi.rpc	Sent a simple query for execution.
mapi.setAlias	Give the channel a logical name.
mapi.stop	Terminate connection listeners.
mapi.suspend	Suspend accepting connections.
mapi.trace	Toggle the Mapi library debug tracer.
mat.hasMoreElements	Find the next element in the merge table
mat.info	retrieve the definition from the partition catalogue
mat.new	Define a Merge Association Table (MAT)
mat.newIterator	Create an iterator over a MAT
mat.pack	Materialize the MAT into the first BAT
mat.print	
mdb.List	Dump the routine M.F on standard out.
mdb.collect	Dump the previous instruction to a temporary file

mdb.dot	Dump the data flow of the function M.F in a format recognizable by the command 'dot' on the file s
mdb.dump	Dump instruction, stacktrace, and stack
mdb.getContext	Extract the context string from the exception message
mdb.getDebug	Get the kernel debugging bit-set. See the MonetDB configuration file for details
mdb.getDefinition	Returns a string representation of the current function with typing information attached
mdb.getException	Extract the variable name from the exception message
mdb.getReason	Extract the reason from the exception message
mdb.getStackDepth	Return the depth of the calling stack.
mdb.getStackFrame	Collect variable binding of current (n-th) stack frame.
mdb.getStackTrace	
mdb.grab	Stop and debug another client process.
mdb.inspect	Run the debugger on a specific function
mdb.lifespan	Dump the current routine lifespan information on standard out.
mdb.list	Dump the routine M.F on standard out.
mdb.listMapi	Dump the current routine on standard out with Mapi prefix.
mdb.modules	List available modules
mdb.setCatch	Turn on/off catching exceptions
mdb.setCount	Turn on/off bat count statistics tracing
mdb.setDebug	Set the kernel debugging bit-set and return its previous value.
mdb.setFlow	Turn on/off memory flow debugger
mdb.setIO	Turn on/off io statistics tracing
mdb.setMemory	Turn on/off memory statistics tracing.
mdb.setMemoryTrace	Turn on/off memory foot print tracer for debugger
mdb.setThread	Turn on/off thread identity for debugger
mdb.setTimer	Turn on/off performance timer for debugger
mdb.setTrace	Turn on/off tracing of a variable
mdb.start	Start interactive debugger on a running factory
mdb.stop	Stop the interactive debugger
mdb.var	Dump the symboltable of routine M.F on standard out.
mkey.bulk_rotate_xor_hash	pre: h and b should be synced on head post: [:xor=]([:rotate=](h, nbits), [hash](b))
mkey.hash	compute a hash int number from any value
mkey.rotate	left-rotate an int by nbits
mmath.acos	The acos(x) function calculates the arc cosine of x, that is the value whose cosine is x. The value is returned in radians and is mathematically defined to be between 0 and PI (inclusive).
mmath.asin	The asin(x) function calculates the arc sine of x, that is the value whose sine is x. The value is returned in radians and is mathematically defined to be between -PI/20 and -PI/2 (inclusive).
mmath.atan	The atan(x) function calculates the arc tangent of x, that is the value whose tangent is x. The value is returned in radians and is mathematically defined to be between -PI/2 and PI/2 (inclusive).

mmath.atan2	The atan2(x,y) function calculates the arc tangent of the two variables x and y. It is similar to calculating the arc tangent of y / x, except that the signs of both arguments are used to determine the quadrant of the result. The value is returned in radians and is mathematically defined to be between -PI/2 and PI/2 (inclusive).
mmath.ceil	The ceil(x) function rounds x upwards to the nearest integer.
mmath.cos	The cos(x) function returns the cosine of x, where x is given in radians. The return value is between -1 and 1.
mmath.cosh	The cosh() function returns the hyperbolic cosine of x, which is defined mathematically as (exp(x) + exp(-x)) / 2.
mmath.cot	The cot(x) function returns the Cotangent of x, where x is given in radians
mmath.exp	The exp(x) function returns the value of e (the base of natural logarithms) raised to the power of x.
mmath.fabs	The fabs(x) function returns the absolute value of the floating-point number x.
mmath.finite	The finite(x) function returns true if x is neither infinite nor a 'not-a-number' (NaN) value, and false otherwise.
mmath.floor	The floor(x) function rounds x downwards to the nearest integer.
mmath.fmod	The fmod(x,y) function computes the remainder of dividing x by y. The return value is x - n * y, where n is the quotient of x / y, rounded towards zero to an integer.
mmath.isinf	The isinf(x) function returns -1 if x represents negative infinity, 1 if x represents positive infinity, and 0 otherwise.
mmath.isnan	The isnan(x) function returns true if x is 'not-a-number' (NaN), and false otherwise.
mmath.log	The log(x) function returns the natural logarithm of x.
mmath.log10	The log10(x) function returns the base-10 logarithm of x.
mmath.pi	return an important mathematical value
mmath.pow	The pow(x,y) function returns the value of x raised to the power of y.
mmath.rand	return a random number
mmath.round	The round(n, m) returns n rounded to m places to the right of the decimal point; if m is omitted, to 0 places. m can be negative to round off digits left of the decimal point. m must be an integer.
mmath.sin	The sin(x) function returns the cosine of x, where x is given in radians. The return value is between -1 and 1.
mmath.sinh	The sinh() function returns the hyperbolic sine of x, which is defined mathematically as (exp(x) - exp(-x)) / 2.
mmath.sqrt	The sqrt(x) function returns the non-negative square root of x.
mmath.srand	initialize the rand() function with a seed
mmath.tan	The tan(x) function returns the tangent of x, where x is given in radians
mmath.tanh	The tanh() function returns the hyperbolic tangent of x, which is defined mathematically as sinh(x) / cosh(x).
mtime.add	returns the timestamp that comes 'msecs' (possibly negative) after 'value'.

mtime.adddays	returns the date after a number of days (possibly negative).
mtime.addmonths	returns the date after a number of months (possibly negative).
mtime.addyears	returns the date after a number of years (possibly negative).
mtime.compute	compute the date from a rule in a given year
mtime.current_date	
mtime.current_time	
mtime.current_timestamp	
mtime.date	extracts date from timestamp in a specific timezone.
mtime.date_add_month_interval	Add months to a date
mtime.date_add_sec_interval	Add seconds to a date
mtime.date_sub_sec_interval	Subtract seconds from a date
mtime.date_to_str	create a string from the date, using the specified format (see man strftime)
mtime.day	extract day from rule.
mtime.dayname	Returns day name from a number between [1-7], str(nil) otherwise.
mtime.daynum	Returns number of day [1-7] from a string or nil if does not match any.
mtime.dayofweek	Returns the current day of the week where 1=sunday, .., 7=saturday
mtime.dayofyear	Returns N where d is the Nth day of the year (january 1 returns 1)
mtime.daytime	default time with zeroed components
mtime.diff	returns the number of milliseconds between 'val1' and 'val2'.
mtime.dst	return whether DST holds in the timezone at a certain point of time.
mtime.end_dst	extract rule that determines end of DST from timezone.
mtime.epilogue	
mtime.hours	extracts hour from daytime
mtime.local_timezone	get the local timezone in seconds
mtime.milliseconds	extracts milliseconds from daytime
mtime.minutes	extract minutes from rule.
mtime.month	extract month from rule.
mtime.monthname	Returns month name from a number between [1-12], str(nil) otherwise.
mtime.monthnum	Returns month number [1-12] from a string or nil if does not match any.
mtime.msec	get time of day in msec since 1-1-1970.
mtime.msecs	convert date components to milliseconds
mtime.olddate	create a date from the old instant format.
mtime.oldduration	parse the old duration format and return an (estimated) number of days.
mtime.prelude	
mtime.rule	create a DST start/end date rule.
mtime.seconds	extracts seconds from daytime
mtime.setTimezone	Test and set the timezone.
mtime.start_dst	extract rule that determines start of DST from timezone.
mtime.str_to_date	create a date from the string, using the specified format (see man strptime)

mtime.time_add_sec_interval	Add seconds to a time
mtime.time_sub_sec_interval	Subtract seconds from a time
mtime.time_synonyms	Allow synonyms for the parse format of date/timestamp.
mtime.timestamp	creates a timestamp from (d,00:00:00) parameters (in the local timezone).
mtime.timestamp_add_month_interval	Add months to a timestamp
mtime.timestamp_add_sec_interval	
mtime.timestamp_sub_month_interval	Subtract months from a timestamp
mtime.timestamp_sub_sec_interval	
mtime.timezone	create a timezone as an hour difference from GMT and a DST.
mtime.timezone_local	get the local timezone; which is used for printing timestamps
mtime.weekday	extract weekday from rule.
mtime.weekofyear	Returns the week number in the year.
mtime.year	extracts year from date (nonzero value between -5867411 and +5867411).
optimizer.accessmode	Reduce the number of mode changes.
optimizer.accumulators	Replace calculations with accumulator model
optimizer.aliases	Alias removal optimizer
optimizer.clrDebug	
optimizer.coercions	Handle simple type coercions
optimizer.commonTerms	Common sub-expression optimizer
optimizer.constants	Duplicate constant removal optimizer
optimizer.costModel	Estimate the cost of a relational expression
optimizer.crack	Replace algebra select with crackers select
optimizer.dataflow	Dataflow bracket code injection
optimizer.deadcode	Dead code optimizer
optimizer.dumpQEP	Produce an indented tree visualisation
optimizer.emptySet	Symbolic evaluation of empty BAT expressions
optimizer.evaluate	Evaluate constant expressions once.
optimizer.factorize	Turn function into a factory
optimizer.garbageCollector	Garbage collector optimizer
optimizer.heuristics	Handle simple replacements
optimizer.history	Collect SQL query statistics
optimizer.inline	Expand inline functions
optimizer.joinPath	Join path constructor
optimizer.joinselect	Replace select with join select
optimizer.macro	Inline a target function used in a specific function.
optimizer.mergetable	Resolve the multi-table definitions
optimizer.mitosis	Modify the plan to exploit parallel processing on multiple cores

optimizer.multiplex	Compiler for multiplexed instructions.
optimizer.octopus	Map-execute-reduce parallelism optimizer
optimizer.optimize	Optimize a specific operation
optimizer.orcam	Inverse macro, find pattern and replace with a function call.
optimizer.partitions	Experiment with partitioned databases
optimizer.peephole	Perform local rewrites
optimizer.prelude	Initialize the optimizer
optimizer.pushranges	Push constant range selections through the program
optimizer.recycle	Replicator code injection
optimizer.reduce	Reduce the stack space claims
optimizer.remap	Remapping function calls to a their multiplex variant
optimizer.remoteQueries	Resolve the multi-table definitions
optimizer.replicator	Replication optimizer
optimizer.setDebug	
optimizer.showFlowGraph	Dump the data flow of the function M.F in a format recognizable by the command 'dot' on the file s
optimizer.showPlan	Illustrate the plan derived so far
optimizer.singleton	Perform singleton optimization
optimizer.strengthReduction	Move constant expressions out of the loop
optimizer.trace	Collect trace of a specific operation
pcre.compile	compile a pattern
pcre.index	match a pattern, return matched position (or 0 when not found)
pcre.like_uselect	
pcre.match	POSIX pattern matching against a string
pcre.patindex	Location of the first POSIX pattern matching against a string
pcre.pcre_quote	Return a PCRE pattern string that matches the argument exactly.
pcre.prelude	Initialize pcre
pcre.replace	Replace _all_ matches of "pattern" in "origin_str" with "replacement". Parameter "flags" accept these flags: 'i', 'm', 's', and 'x'. 'e': if present, an empty string is considered to be a valid match 'i': if present, the match operates in case-insensitive mode. Otherwise, in case-sensitive mode. 'm': if present, the match operates in multi-line mode. 's': if present, the match operates in "dot-all" The specifications of the flags can be found in "man pcreapi" The flag letters may be repeated. No other letters than 'e', 'i', 'm', 's' and 'x' are allowed in "flags". Returns the replaced string, or if no matches found, the original string.
pcre.select	Select tuples based on the pattern
pcre.sql2pcre	Convert a SQL like pattern with the given escape character into a PCRE pattern.
pcre.uselect	Select tuples based on the pattern, only returning the head
pqueue.dequeue_max	Removes top element of the max-pqueue and updates it
pqueue.dequeue_min	Removes top element of the min-pqueue and updates it
pqueue.enqueue_max	Inserts element (oid,dbl) in the max-pqueue

pqueue.enqueue_min	Inserts element (oid,dbl) in the min-pqueue
pqueue.init	Creates an empty pqueue of bat a's tailtype with maximum size maxsize
pqueue.topn_max	Return the topn elements of the bat t using a max-pqueue
pqueue.topn_min	Return the topn elements of the bat t using a min-pqueue
pqueue.topreplace_max	Replaces top element with input and updates max-pqueue
pqueue.topreplace_min	Replaces top element with input and updates min-pqueue
profiler.activate	Make the specified counter active.
profiler.cleanup	Remove the temporary tables for profiling
profiler.closeStream	Stop sending the event records
profiler.clrFilter	Stop tracing the variable
profiler.deactivate	Deactivate the counter
profiler.dumpTrace	List the events collected
profiler.getDiskReads	Obtain the number of physical reads
profiler.getDiskWrites	Obtain the number of physical reads
profiler.getEvent	Retrieve the performance indicators of the previous instruction
profiler.getFootprint	Get the memory footprint and reset it
profiler.getMemory	Get the amount of memory claimed and reset it
profiler.getSystemTime	Obtain the user timing information.
profiler.getTrace	Get the trace details of a specific event
profiler.getUserTime	Obtain the user timing information.
profiler.noop	Fetch any pending performance events
profiler.openStream	Send the log events to a stream
profiler.reset	Clear the profiler traces
profiler.setAll	Short cut for setFilter(*,*).
profiler.setEndPoint	End performance tracing after mod.fcn
profiler.setFilter	Generate an event record for every instruction where v is used.
profiler.setNone	Short cut for clrFilter(*,*).
profiler.setStartPoint	Start performance tracing at mod.fcn
profiler.start	Start performance tracing
profiler.stop	Stop performance tracing
recycle.dump	Dump summary of recycle table into a file
recycle.dumpQPat	Dump statistics of query patterns
recycle.epilogue	Called at the start of a recycle controlled function
recycle.getCachePolicy	
recycle.getRetainPolicy	
recycle.getReusePolicy	
recycle.log	Set the name of recycle log file

recycle.monitor	start/stop the monitoring (printing) of the recycler info (storage size used and number of statements retained)
recycle.prelude	Called at the start of a recycle controlled function
recycle.reset	Reset off all recycled variables
recycle.setCachePolicy	Set recycler cache policy with alpha parameter
recycle.setRetainPolicy	Set recycler retainment policy: 0- RETAIN_NONE: baseline, keeps stat, no retain, no reuse 1- RETAIN_ALL: infinite case, retain all 2- RETAIN_CAT: time-based semantics, retain if beneficial 3- RETAIN_ADAPT: adaptive temporal
recycle.setReusePolicy	Set recycler reuse policy
recycle.shutdown	Clear the recycle cache
recycle.start	Initialize recycler for the current block
recycle.stop	Cleans recycler bookkeeping
remote.connect	Returns a newly created connection for dbname, user name and password.
remote.create	Create a user-defined connection to a server.
remote.destroy	Destroy a previously user-defined connection to a server.
remote.disconnect	Disconnects the connection for dbname.
remote.epilogue	Release the resources held by the remote module.
remote.exec	Remotely executes <mod>.<func> using the argument list of remote objects and returns the handle to its result
remote.get	Retrieves a copy of remote object ident.
remote.getList	List available databases with their property for use with connect().
remote.prelude	Initialise the remote module.
remote.put	Copies object to the remote site and returns its identifier.
remote.register	Register <mod>.<fcn> at the remote site.
replicator.bind	Create a named persistent BAT if it was not known
replicator.bind_dbat	Create a named persistent BAT if it was not known
replicator.setMaster	Mark the source of this database
replicator.setVersion	Keep the latest version in the symbol table as a constant
sabaoth.epilogue	Release the resources held by the sabaoth module
sabaoth.getLocalConnectionHost	Returns the hostname this server can be connected to, or nil if none
sabaoth.getLocalConnectionPort	Returns the port this server can be connected to, or 0 if none
sabaoth.marchConnection	Publishes the given host/port as available for connecting to this server
sabaoth.marchScenario	Publishes the given language as available for this server
sabaoth.prelude	Initialise the sabaoth module
sabaoth.retreatScenario	Unpublishes the given language as available for this server

sabaoth.wildRetreat	Unpublishes everything known for this server
scheduler.choice	Select the next step in a query memo plan
scheduler.costPrediction	A sample cost prediction function
scheduler.drop	Remove a worker from the list
scheduler.isolation	Run a private copy of the MAL program
scheduler.octopus	Run the program block in parallel, but don't wait longer then t seconds
scheduler.pick	Pick up the first result
scheduler.volumeCost	A sample cost function based on materialized results
scheduler.worker	Add a worker site to the known list
sema.create	Create an unset sema, with an initial value
sema.destroy	Destroy a semaphore
sema.down	Decrement the semaphpore if >0; else block
sema.up	Increment the semaphore
sql.forgetPrevious	invalidate the previous instruction from future execution
sql.keepquery	
sql.queryId	
sqlblob.sqlblob	Noop routine.
statistics.close	Close the statistics box
statistics.deposit	Enter a new BAT into the statistics box
statistics.destroy	Destroy the statistics box
statistics.discard	Release a BAT variable from the box
statistics.dump	Display the statistics table
statistics.epilogue	Release the resources of the statistics package
statistics.forceUpdate	Bring the statistics up to date for one BAT
statistics.getCount	Return latest stored count information
statistics.getHistogram	Return the latest histogram
statistics.getHotset	Return a table with BAT names that have been touched since the start of the session
statistics.getMax	Return latest stored maximum information
statistics.getMin	Return latest stored minimum information
statistics.getObjects	Return a table with BAT names managed
statistics.getSize	Return latest stored count information
statistics.hasMoreElements	Locate next element in the box
statistics.newIterator	Locate next element in the box
statistics.open	Locate and open the statistics box
statistics.prelude	Initialize the statistics package
statistics.release	Release a single BAT from the box
statistics.releaseAll	Release all variables in the box
statistics.take	Take a variable out of the statistics box

statistics.toString	Get the string representation of an element in the box
statistics.update	Check for stale information
status.batStatistics	Show distribution of bats by kind
status.cpuStatistics	Global cpu usage information
status.getDatabases	Produce a list of known databases in the current dbfarm
status.getPorts	Produce a list of default ports for a specific language
status.getThreads	Produce overview of active threads.
status.ioStatistics	Global IO activity information
status.memStatistics	Global memory usage information
status.memUsage	Get a split-up of how much memory blocks are in use.
status.mem_cursize	The amount of physical swapspace in KB that is currently in use.
status.mem_maxsize	Set the maximum usable amount of physical swapspace in KB.
status.vmStatistics	Get a split-up of how much virtual memory blocks are in use.
status.vm_cursize	the amount of logical VM space in KB that is currently in use
status.vm_maxsize	set the maximum usable amount of physical swapspace in KB
str.+	Concatenate two strings.
str.STRepilogue	
str.STRprelude	
str.ascii	Return unicode of head of string
str.chrAt	String array lookup operation.
str.codeset	Return the locale's codeset
str.endsWith	Suffix check.
str.iconv	String codeset conversion
str.insert	Insert a string into another
str.length	Return the length of a string.
str.like	SQL pattern match function
str.locate	Locate the start position of a string
str.ltrim	Strip whitespaces from start of a string.
str.nbytes	Return the string length in bytes.
str.prefix	Extract the prefix of a given length
str.r_search	Reverse search for a char. Returns position, -1 if not found.
str.repeat	
str.replace	Insert a string into another
str.rtrim	Strip whitespaces from end of a string.
str.search	Search for a character. Returns position, -1 if not found.
str.space	
str.startsWith	Prefix check.
str.str	Noop routine.
str.string	Return substring s[offset..offset+count] of a string s[0..n]
str.stringleft	
str.stringlength	Return the length of a right trimed string (SQL semantics).
str.stringright	
str.substitute	Substitute first occurrence of 'src' by 'dst'. Iff repeated = true this is repeated while 'src' can be found in the result string. In order to prevent recursion and result strings of unlimited size, repeating is only done iff src is not a substring of dst.

str.substring	Extract a substring from str starting at start, for length len
str.suffix	Extract the suffix of a given length
str.toLower	Convert a string to lower case.
str.toUpper	Convert a string to upper case.
str.trim	Strip whitespaces around a string.
str.unicode	convert a unicode to a character.
str.unicodeAt	get a unicode character (as an int) from a string position.
streams.blocked	open a block based stream
streams.close	close and destroy the stream s
streams.flush	flush the stream
streams.openRead	convert an ascii stream to binary
streams.openReadBytes	open a file stream for reading
streams.openWrite	convert an ascii stream to binary
streams.openWriteBytes	open a file stream for writing
streams.readInt	read integer data from the stream
streams.readStr	read string data from the stream
streams.socketRead	open ascii socket stream for reading
streams.socketReadBytes	open a socket stream for reading
streams.socketWrite	open ascii socket stream for writing
streams.socketWriteBytes	open a socket stream for writing
streams.writeInt	write data on the stream
streams.writeStr	write data on the stream
tablet.display	Display a formatted table
tablet.dump	Print all pages with header to a stream
tablet.finish	Free the storage space of the report descriptor
tablet.firstPage	Produce the first page of output
tablet.getPage	Produce the i-th page of output
tablet.getPageCnt	Return the size in number of pages
tablet.header	Display the minimal header for the table
tablet.input	Load a bat using specific format.
tablet.lastPage	Produce the last page of output
tablet.load	Load a bat using specific format.
tablet.nextPage	Produce the next page of output
tablet.output	Send the bat to an output stream.
tablet.page	Display all pages at once without header
tablet.prevPage	Produce the prev page of output
tablet.setBracket	Format the brackets around a field
tablet.setColumn	Bind i-th output column to a variable
tablet.setComplaints	The comlaints bat identifies all erroneous lines encountered
tablet.setDecimal	Set the scale and precision for numeric values
tablet.setDelimiter	Set the column separator.
tablet.setFormat	Initialize a new reporting structure.
tablet.setName	Set the display name for a given column

tablet.setNull	Set the display format for a null value for a given column
tablet.setPivot	The pivot bat identifies the tuples of interest. The only requirement is that all keys mentioned in the pivot tail exist in all BAT parameters of the print comment. The pivot also provides control over the order in which the tuples are produced.
tablet.setPosition	Set the character position to use for this field when loading according to fixed (punch-card) layout.
tablet.setProperties	Define the set of properties
tablet.setRowBracket	Format the brackets around a row
tablet.setStream	Redirect the output to a stream.
tablet.setTableBracket	Format the brackets around a table
tablet.setTryAll	Skip error lines and assemble an error report
tablet.setWidth	Set the maximal display witdh for a given column. All values exceeding the length are simple shortened without any notice.
timestamp.!=	Equality of two timestamps
timestamp.<	Equality of two timestamps
timestamp.<=	Equality of two timestamps
timestamp.==	Equality of two timestamps
timestamp.>	Equality of two timestamps
timestamp.>=	Equality of two timestamps
timestamp.epoch	convert seconds since epoch into a timestamp
timestamp.isnil	Nil test for timestamp value
timestamp.unix_epoch	The Unix epoch time (00:00:00 UTC on January 1, 1970)
timezone.str	
timezone.timestamp	Utility function to create a timestamp from a number of seconds since the Unix epoch
transaction.abort	Abort changes in certain BATs.
transaction.alpha	List insertions since last commit.
transaction.clean	Declare a BAT clean without flushing to disk.
transaction.commit	Commit changes in certain BATs.
transaction.delta	List deletions since last commit.
transaction.prev	The previous stae of this BAT
transaction.subcommit	commit only a set of BATnames, passed in the tail (to which you must have exclusive access!)
transaction.sync	Save all persistent BATs
txtsim.editdistance	Alias for Levenshtein(str,str)
txtsim.editdistance2	Calculates Levenshtein distance (edit distance) between two strings. Cost of transposition is 1 instead of 2
txtsim.levenshtein	Calculates Levenshtein distance (edit distance) between two strings
txtsim.qgramnormalize	'Normalizes' strings (eg. toUpper and replaces non-alphanumerics with one space
txtsim.qgramselfjoin	QGram self-join on ordered(!) qgram tables and sub-ordered q-gram positions

txtsim.similarity	Normalized edit distance between two strings
txtsim.soundex	Soundex function for phonetic matching
txtsim.str2qgrams	
txtsim.stringdiff	calculate the soundexed editdistance
unix.getenv	Get the environment variable string.
unix.setenv	Set the environment variable string.
url.getAnchor	Extract the URL anchor (reference)
url.getBasename	Extract the URL base file name
url.getContent	Get the URL resource in a local file
url.getContext	Get the path context of a URL
url.getDirectory	Extract directory names from the URL
url.getDomain	Extract Internet domain from the URL
url.getExtension	Extract the file extension of the URL
url.getFile	Extract the last file name of the URL
url.getHost	Extract the server name from the URL
url.getPort	Extract the port id from the URL
url.getProtocol	Extract the protocol from the URL
url.getQuery	Extract the query string from the URL
url.getQueryArg	Extract argument mappings from the URL
url.getRobotURL	Extract the location of the robot control file
url.getUser	Extract the user identity from the URL
url.isaURL	Check conformity of the URL syntax
url.new	Construct URL from protocol, host,and file
url.url	Create an URL from a string literal
user.main	
zrule.define	Introduce a synomym timezone rule.