Volume Title
ASP Conference Series, Vol. **Volume Number**
Author

Towards Dynamic Catalogues

Bart Scheers[1,2], Fabian Groffen[1], and the TKP Team[1,3]

[1]*Astronomical Institute "Anton Pannekoek", University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

[2]*Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands*

[3]*ASTRON, P.O.Box 2, 7990 AA Dwingeloo, The Netherlands*

Abstract. The International LOFAR Telescope is designed to carry out unique science in the spatial, spectral, polarisation and temporal domains.

The Transients Key Science Project aims to study all transient and variable sources detected by LOFAR. One of its products will be an up-to-date catalogue of all sources detected by LOFAR, i.e. a spectral light-curve database, with real-time capabilities, and expected to grow gradually with $50-100$ TB/yr. The response time to transient and variable events depends strongly on the query execution plans of the algorithms that search the LOFAR light-curve database for previous (non-)detections in the spatial, spectral, polarisation and temporal domains.

Here we show how the Transients Key Science Project of LOFAR approaches these challenges by using column-stores, sharded databases and implementing the new array query language SciQL (pronounced as 'cycle').

1. Multiple-Domain Coverage of LOFAR

LOFAR, the next-generation radio telescope, currently in its commissioning phase, is sensitive in the unexplored low-frequency regime of $30-240$ MHz and designed to carry out unique science: (1) high-speed all-sky surveys, (2) searching for fast and slow transient and variable sources, and (3) cataloguing the millions of sources and their millions of measurements. As a consequence LOFAR is going to produce tens of terabytes per day. High-cadence data rates of tens of gigabits per second are neither exceptional. Storing these huge volumes of scientific data requires unique database management systems that are, moreover, able to query the data scientifically with acceptable response times. Here we show how the Transients Key Science Project (Fender 2007) of LOFAR approaches these challenges by using column-stores, sharded databases and implementing the new array query language SciQL (pronounced as 'cycle').

2. Catching Transient and Variable Souces

MonetDB, the open source column-store, is fundamentally different from design than the classical row-store relational database management systems (RDBMSs), e.g., MySQL or Postgres, but all are accessed by the same Structured Query Language (SQL). Direct

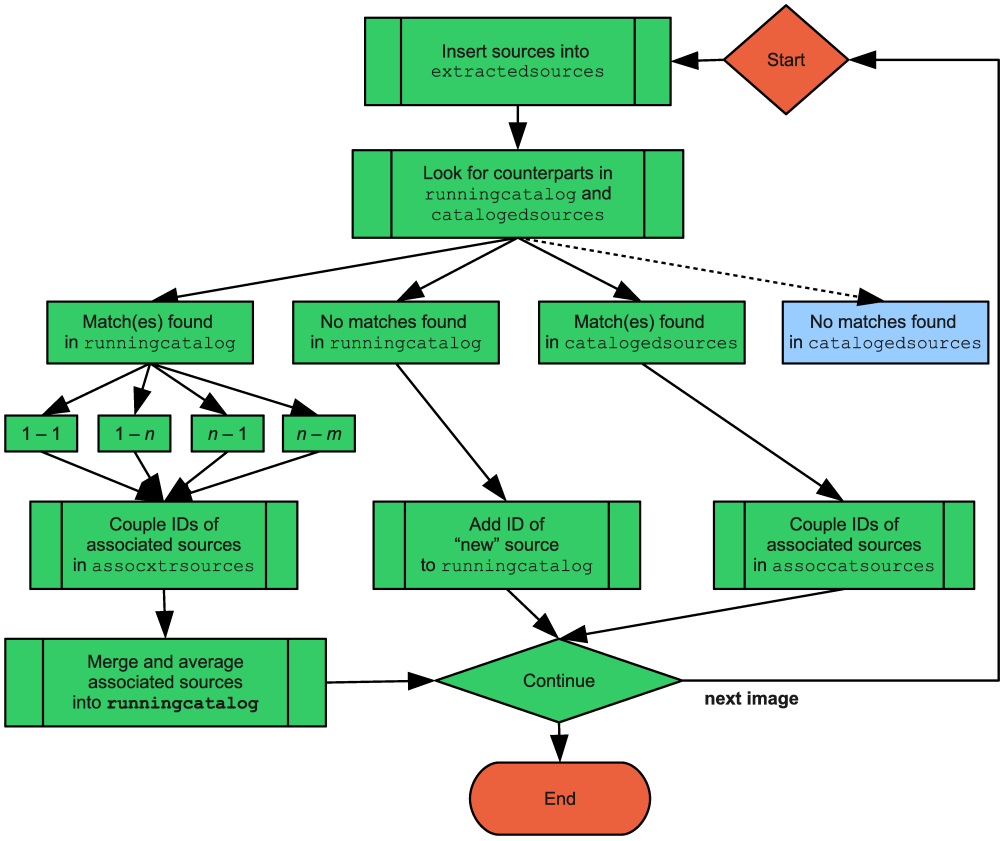

Figure 1. Flow diagram of the association of extracted sources against LOFAR sources (in `runningcatalog`) and the external catalogues (in `catalogedsources`).

consequences are that queries only touch the relevant columns, and when in contiguous memory it allows compression and good cache-hit ratios. Furthermore, MonetDB's kernel is a programmable relational algebra machine operating on "array"-like structures, exactly what CPUs are good at. The experimental SciLens platform is a 330 node, 4-tier locally distributed infrastructure based on MonetDB technology and focusses on massive I/O, instead of raw computing power. The SciLens infrastructure (`http://www.scilens.org`) is envisaged to be the prime choice for a scalable LOFAR light-curve database.

The Transients Pipeline Database forms the heart of the Transients Pipeline, the automated software framework that aims for detecting transient and variable sources in the high-cadence calibrated LOFAR images (Swinbank 2010; Scheers 2011). During LOFAR observations, the Transients Pipeline stores all sources extracted from the images in the Transients Database. While observing, sources are cross-matched with counterpart candidates in the spatial, spectral, polarisation and temporal domains of LOFAR and non-LOFAR catalogues as well, as depicted in Fig. 1. It is crucial to keep the query processing relatively constant over time, therefore, cross-matching is done against a statistical representation of the LOFAR catalogue (the so-called *running catalogue*), instead of the whole data volume (available in *extracted sources*). The running

catalogue, a statistical summarisation of the millions of sources and their millions of measurements, serves as a global all-sky model as well, that is being used in the image calibration steps. The catalogue and the sky model will evolve and improve over time.

3. Maintaining a Statistical Representation of the LOFAR Catalogue

Association parameters and variability indices determine whether sources may be genuinely associated, and if so, whether their light curves show variability, respectively. The values of the parameters and indices are related to the corresponding probabilities of their null hypotheses.

The magnitude of the flux variability of a source can be expressed as the ratio of the sample flux standard deviation, and the sample arithmetic mean flux. A second indicator, which expresses the significance of the flux variability, is based on reduced χ^2 statistics. We assume that the weighted average flux value is a fitted parameter, so that the number of degrees of freedom is $N - 1$. It is given by the sum of the squared deviations from the weighted average, weighted by the errors, and divided by the number of degrees of freedom. Both indices, respectively, are defined as

$$V_\nu = \frac{1}{\overline{I_\nu}} \sqrt{\frac{N}{N-1}(\overline{I_\nu{}^2} - \overline{I_\nu}^2)} \tag{1}$$

$$\eta_\nu = \frac{N}{N-1}\left(\overline{w_\nu I_\nu{}^2} - \frac{\overline{w_\nu I_\nu}^2}{\overline{w_\nu}}\right), \tag{2}$$

where N is the number of source measurements and I_ν is the flux measured at frequency ν, and $w_\nu \equiv 1/\sigma_\nu^2$, where σ_ν is the corresponding flux error and the bar represents the average. Maintaining only these statistical components in the respective columns of the running catalogue, the indices can be created on the fly, taking advantage of being expressed in aggregate form.

4. Sharded Database for LOFAR Light curves

A sharded database cluster, keeps the entire volume quickly accessible without replication as opposed to a distributed database environment. Queries are fired at all machines in the cluster through the multiplex funnel. In the set-up of Fig. 2, data are sharded by declination zones. Joins with steering tables, on all nodes, determine at which node a query is actually executed. The tables in red boxes cover the whole sky, whereas the tables in the blue boxes cover only parts of the sky which are bound by the declination zones. `ITLLite` is a *stripped* LOFAR Catalogue view and contains all the unique sources ($\sim 10^8$), i.e. the positions where at least once a LOFAR detection was made, but nothing more than that. This table is fully replicated over all the nodes, and it has only the essential columns needed for source association, in order to fit into memory. The collection of `ITL_XL` tables over all nodes (i.e. zones) represents the full LOFAR Catalogue.

SciQL (Kersten et al. 2011; Zhang et al. 2011, 2012) eases the scientifically very relevant light-curve analyses by query window processing, where moving averages, Fourier transformation, correlation and convolution are directly applied inside the database engine.

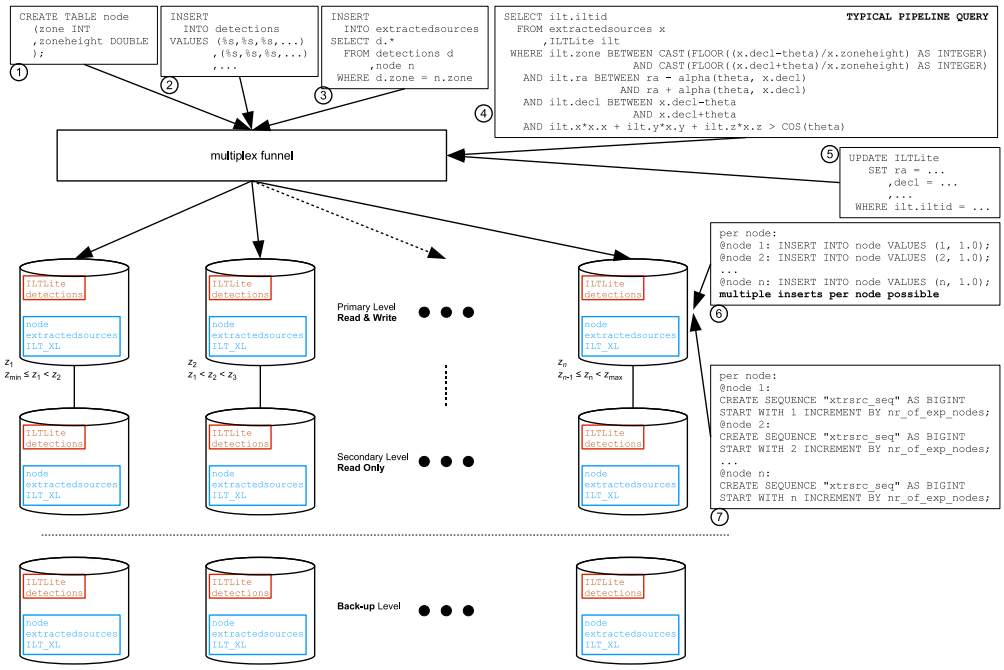

Figure 2. Schematic set-up of the sharded database cluster for the LOFAR light-curve archive. The numbers 1 – 7 are some of the relevant queries, whereas the arrows show the points of execution.

5. Conclusion

A statistical representation of the full catalogue relaxes the source association and transient detection query processing times.

A sharded database keeps the entire light-curve archive quickly accessible without replication. SQL and the SciQL extension in combination with a column-store opens up a lot of data mining advantages, where next to exploring the traditional domains, we may now include source-model domains as well.

References

Fender, R. P. 2007, in Bursts, Pulses and Flickering: Wide-Field Monitoring of the Dynamic Radio Sky, vol. Dynamic 2007 of Proceedings of Science, 30

Kersten, M. L., Zhang, Y., Ivanova, M., & Nes, N. J. 2011, in Proceedings of EDBT/ICDT - Workshop on Array Databases 2011, 12

Scheers, L. H. A. 2011, Ph.D. thesis, University of Amsterdam

Swinbank, J. D. 2010, in ISKAF2010 Science Meeting, vol. ISKAF 2010 of Proceedings of Science, 82

Zhang, Y., Kersten, M. L., Ivanova, M., & Nes, N. J. 2011, in Proceedings of IDEAS 2011, 10

Zhang, Y., Scheers, B., Kersten, M. L., Ivanova, M., & Nes, N. 2012, in ADASS XXI, edited by P. Ballester, & D. Egret (San Francisco: ASP), vol. TBD of ASP Conf. Ser., TBD