Introduction to the special theme

Big Data

by Costantino Thanos, Stefan Manegold and Martin Kersten

Big data' refers to data sets whose size is beyond the capabilities of the current database technology.

The current data deluge is revolutionizing the way research is carried out and resulting in the emergence of a new fourth paradigm of science based on data-intensive computing. This new data-dominated science will lead to a new data-centric way of conceptualizing, organizing and carrying out research activities which could lead to an introduction of new approaches to solve problems that were previously considered extremely hard or, in some cases, impossible to solve and also lead to serendipitous discoveries.

The recent availability of huge amounts of data, along with advanced tools of exploratory data analysis, data mining/machine learning and data visualization, offers a whole new way of understanding the world.

In order to exploit these huge volumes of data, new techniques and technologies are needed. A new type of e-infrastructure, the Research Data Infrastructure, must be developed to harness the accumulation of data and knowledge produced by research communities, optimize the data movement across scientific disciplines, enable large increases in multi- and interdisciplinary science while reducing duplication of effort and resources, and integrating research data with published literature.

Science is a global undertaking and research data are both national and global assets. A seamless infrastructure is needed to facilitate collaborative arrangements necessary for the intellectual and practical challenges the world faces.

Therefore, there is a need for Global Research Data Infrastructures to overcome language, policy, methodology, and social barriers and to reduce geographic, temporal, and national barriers in order to facilitate discovery, access, and use of data.

The next generation of global research data infrastructures is facing two main challenges:
- to effectively and efficiently support data-intensive science
- to effectively and efficiently support multidisciplinary/interdisciplinary science.

In order to build the next generation of Global Research Data Infrastructures several breakthroughs must be achieved. They include:

Data modelling challenges

There is a need for radically new approaches to research data modelling. Current data models (relational model) and management systems (relational database management systems) were developed by the database research community for business/commercial data applications. Research

data has completely different characteristics and thus the current database technology is unable to handle it effectively.
There is a need for data models and query languages that:

- more closely match the data representation needs of the several scientific disciplines;
- describe discipline-specific aspects (metadata models);
- represent and query data provenance information;
- represent and query data contextual information;
- represent and manage data uncertainty;
- represent and query data quality information.

Data management challenges

There is a clear need for extremely large data processing. This is especially true in the area of scientific data management where, in the near future, we expect data inputs in the order of multiple Petabytes. However, current data management technology is not suitable for such data sizes.
In the light of such new database applications, we need to rethink some of the strict requirements adopted by database systems in the past. For instance, database management systems (DBMS) see database queries as contracts carved in stone that require the DBMS to produce a complete and correct answer, regardless of the time and resources required. While this behaviour is crucial in business data management, it is counterproductive in scientific data management. With the explorative nature of scientific discovery, scientists cannot be expected to instantly phrase a crisp query that yields the desired (but a priori unknown) result, or to wait days to get a multi-megabyte answer that does not reveal what they were looking for. Instead, the DBMS could provide a fast and cheap approximation that is neither complete nor correct, but indicative of the complete answer. In this way, the user gets a ‘feel’ for the data that helps him/her to advance his/her exploration using a refined query.

The challenges faced include catching the user’s intention and providing the users with suggestions and guidelines to refine their queries in order to quickly converge to the desired results, but also call for novel database architectures and algorithms that are designed with the intent to produce fast and cheap indicative answers rather than complete and correct answers.

Data Tools challenges

Currently, the available data tools for most scientific disciplines are inadequate. It is essential to build better tools in order to improve the productivity of scientists. There is a need for better computational tools to visualize, analyze, and catalog the available enormous research datasets in order to enable data-driven research.
Scientists need advanced tools that enable them to follow new paths, try new techniques, build new models and test them in new ways that facilitate innovative multidisciplinary/interdisciplinary activities and support the whole research cycle.

Please contact:

Costantino Thanos
ISTI-CNR Italy
E-mail: thanos@isti.cnr.it

Stefan Manegold, Martin Kersten
CWI, The Netherlands
E-mail: Stefan.Manegold@cwi.nl, Martin.Kersten@cwi.nl