in (https://www.linkedin.com/groups/ERCIM-81390) ✖ (https://twitter.com/ercim_news) f
(http://www.facebook.com/ercimnews)
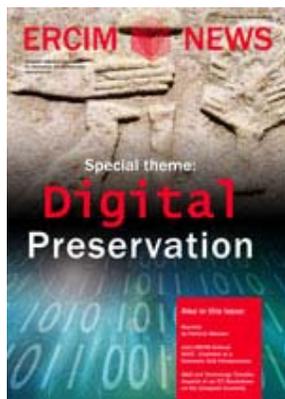
(/images/stories/EN80/EN80-web.pdf)

January 2010
Special theme: Digital Preservation
This issue in pdf (/images/stories/EN80/EN80-web.pdf)
(64 pages; 15 Mb)
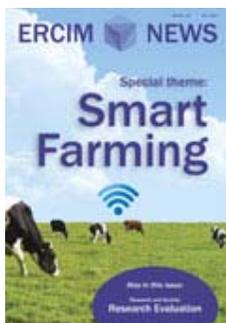
(/ERCIM News 113

(/ERCIM News 112

(/ERCIM News 111

(/ERCIM News 110

(/ERCIM News 109

1 / 10

CONTENTS

Keynote (/en80/keynote)

Joint ERCIM Actions (/en80/jea)

European Scene (/en80/es)

Next issue: **July 2018**
Special theme:
Human-Robot Interaction
Call for the next issue (/call)

Ranking the Stars with MonetDB (/en80 /rd/ranking-the-stars- with-monetdb)

R&D And Technology Transfer (/En80/rd)

📅 14 January 2010   👁 Hits: 6334

by Annette Kik and Milena Ivanova
The international Sloan Sky Server project has put together a huge scientific database of information on celestial bodies. Querying these vast amounts of data is a great challenge. Using the MonetDB database system, Milena Ivanova and her fellow researchers at CWI in the Netherlands have implemented the first open-source solution. The MonetDB/SkyServer project now provides a valuable experimentation platform for developing new techniques for scientific data management.

The Sloan Digital Sky Survey (SDSS) started in 2000. Its aim is to map a quarter of the sky and to obtain observations of 100 million objects, including galaxies, nebulas and quasars. The SkyServer application gives public access to these data through a Web site. Both researchers and school children can now easily learn more about temperature, mass or the chemical composition of objects like the Whirlpool Galaxy and Owl Nebula. The survey is immense. In 2008, the sky object catalogue alone contained four terabytes (4000 GB) of information. The data are organized in a relational database, containing tables with millions of rows and hundreds of columns.

These vast amounts of data stress the capabilities of most database management systems (DBMSs), with efficient querying being a particular problem. The architectures of most modern DBMSs are based on an original design that is now three decades old, and was originally intended for business applications (eg bank transactions) having many small and frequently updated records. Scientific databases, however, have large records, with data that stay unchanged once they have been put in the

Get the latest issue to your desktop

database. They require a different type of database management.

Figure 1: The famous Whirlpool Galaxy is one of the many objects in the SDSS database. SDSS acts as a well-documented benchmark for scientific database management. (Picture: The Sloan Digital Sky Survey.)

The original SDSS Skyserver, based on Microsoft SQL Server, was the first to bridge the gap between databases and astronomy. It became a successful showcase of scalable database support for scientific applications. There were several other attempts to port the complete SkyServer application to other commercial and open-source systems, but they did not succeed. That is, until the MonetDB solution was implemented.

MonetDB is an innovative open-source database system that has been under development at CWI for over a decade. MonetDB has several advantages, such as efficient data access patterns, flexibility with changing workloads, reduced storage needs, and run-time query optimization. It is a column-store database system. Where other systems organize data in rows, MonetDB reads and stores columns. This approach minimizes the data flow from disk through memory into the CPU caches since only the columns relevant for processing have to be fetched from disk. This can be especially favourable in data analysis applications that need to efficiently retrieve and process large portions of stored data, as in this real-life astronomy application.

Figure 2: The Sloan Telescope, a 2.5-meter telescope at Apache Point Observatory, did all SDSS imaging and spectroscopy. (Picture: The Sloan Digital Sky Survey.)

The team of CWI researchers - Milena Ivanova, Martin Kersten, Niels Nes, Arjen de Rijke and Romulo Goncalves - intended to test the maturity of this column-store technology by working on a new version of the SkyServer database that is both scalable to growing amounts of data and more efficient to query. The researchers considered MonetDB the best candidate to act as an experimentation platform, since it enables experimentation at all levels of a DBMS architecture.

To make the new SkyServer version, the team members optimized MonetDB for scientific data. They improved its scalability through partitioning and distribution, and made it more efficient. The first functional prototype of MonetDB/SkyServer went live in 2006. It was a 1% subset of the archive, called 'Personal SkyServer', having a size of 1.5GB. The large vendor-specific database schema and its extensive use of a specific SQL functionality required a significant engineering effort. The initial performance was competitive with the reference platform, MS SQL Server 2005. This demonstrated the benefit of column-stored database techniques for scientific database management. The full-size version went live at the end of 2008 - a major achievement.
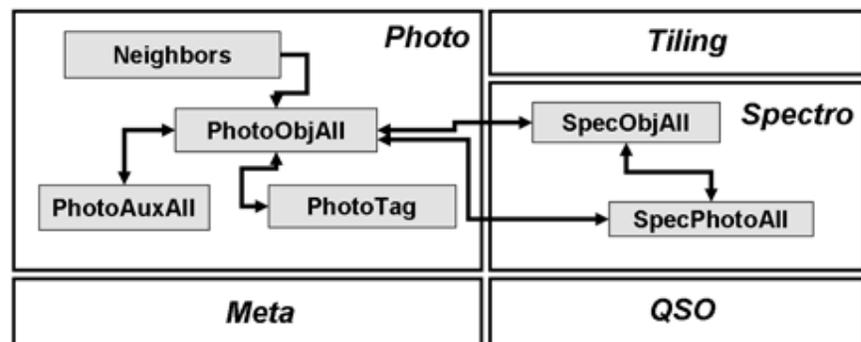

Figure 3: The database schema of SkyServer. The photometric data are stored in the photo-section. The PhotoObjAll table has 454 columns and over 585 million rows. (Picture: The Sloan Digital Sky Survey.)

The team is currently investigating a number of techniques to increase the system's efficiency, such as parallel load, interleaving of column I/O with query processing, exploitation of commonalities in query batches, and self-

organizing indexing schemes like 'crackers'.

Crackers, developed by other members of the MonetDB team, Stratos Idreos, Stefan Manegold, and Martin Kersten, are methods that 'crack' the database into smaller pieces based on querying. A cracker structure converges quickly towards a partial index for fast access. Instead of ordering the data in a fixed manner, this is performed dynamically during query processing.

This project is a good example of one of the CWI key research themes: the data explosion. The current explosion in the amount of digital data confronts science and society with new questions. How can concise and relevant information be extracted from this flood of data? There is great need for models, methods and techniques to control it. The MonetDB/SkyServer project contributes to this objective.

Links:
SDSS (Sloan Digital Sky Survey/SkyServer): http://cas.sdss.org (http://cas.sdss.org)
MonetDB: http://monetdb.cwi.nl (http://monetdb.cwi.nl)

Please contact:
Milena Ivanova
CWI, The Netherlands
Tel: +31 205924317
E-mail: Milena.Ivanova@cwi.nl (mailto:Milena.Ivanova@cwi.nl)