

Editor's Notes

Welcome to the December 2011 issue of the ACM SIGMOD Record! This issue appearing in December 2011 puts us back on the regular publication schedule.

I am glad to start this issue's notes by congratulating the members of our scientific community which have been recently named ACM Fellows (<http://www.acm.org/press-room/news-releases/2011/fellows-2011>): Serge Abiteboul, Divyakant Agrawal, Christian S. Jensen, Beng Chin Ooi, Margo Seltzer, Divesh Srivastava, Dan Suciu and Meral Özsoyoglu. Thanks to them all for representing our community within the larger computer science family, and many congratulations!

The first article by Lee, Lee and Park studies index scan operations on Flash Memory solid state drives (or SSDs, in short). The authors investigate the relation between the selectivity of an index scan and the performance that the operation can attain on an SSD. Index scans turn out to be inefficient for very selective look-ups, therefore the authors consider as an alternative a sorted index scan, which first sorts the index entries by record ID. However, the sorted index scan loses the index key order. To mitigate this, the authors propose a new index-based sort algorithm, which can sort the data in one pass regardless of the available sort memory size.

The survey by Lee, Lee, Choi, Chung and Moon is a very timely one: it provides an overview of parallel data processing techniques based on the principles of MapReduce. This is a field in which the state of the art moves faster than most of us manage to follow! The authors start by describing a generic MapReduce-based architecture, discuss advantages and pitfalls, before moving on to describe variants and improvements. Important extensions such as higher-level languages, flexible data flows, schema support, scheduling, I/O optimizations and others are discussed. An interesting list of applications is also presented, making the survey a valuable collection of information and pointers.

The system paper by Marcus, Bernstein, Badar, Karger, Madden and Miller is a nice example of how database technologies and a database-oriented mind set can help make the most out of data sources which exist today, but would have seemed outlandish in the days when the first bricks of our discipline were laid out! The authors devise a model (akin to data streams), a query language and elements of interface with a system that queries and integrates data from microblogs such as Twitter streams. TweepQL, the language they propose, mixes stream query features with built-in UDFs e.g. for text processing, geo-codes, classification and more.

Jiawei Han is this issue's guest in the Distinguished Profiles in Databases. The processing of Jiawei's interview transcript, through the meanders of the Record's editorial process (Marianne's assistant, Donna Coleman, does a great job at transcribing the audios gathered by Marianne, before Vanessa makes a second beautification pass), can be measured in time by the evolution of Jiawei's H-index: 76 at the time of the interview, it is 101 by the time of the publication! We are glad to send this issue to print before this number becomes obsolete. Read Jiawei's interview for many unique stories and insights: how he was admitted to an US university long before Chinese students could take GREs, where the border lies between data mining and databases and how to organize a very active research and supervisor career.

In the Research Centers column, Kersten, Manegold and Mullender summarize the twenty-six years of history of the CWI database research group from Amsterdam. Longtime a stalwart of the database community, CWI has recently enjoyed significant visibility in our community, such as, for instance, the 2011 SIGMOD Jim Gray Doctoral Dissertation Award to Stratos Idreos on database cracking. An important part of the CWI recent successes can be attributed to their longstanding work around MonetDB.

The paper surveys other areas of the group's work, including hardware-conscious data management, scientific databases, graphs and stream processing.

The Industry Perspective column is present with a report on HANA, SAP's new database, by Färber, Cha, Primsch, Bornhövd, Sigg and Lehner. HANA is a memory-centric database, built to leverage the capability of modern hardware (such as multi-core processors and SSDs) for the benefit of analytical and transactional applications. Interestingly, HANA supports both structured and semi-structured data, as well as entity extraction, or graph processing – clearly not the seventies' database! HANA is made to fit within the complex SAP Business Intelligence product suite. The authors end by discussing various levels of evolution and integration of HANA within the suite, perspectives for their work.

Three workshop reports are part of this issue.

First, Maurino, Capiello, Vassiliadis and Sattler outline the proceedings of the 8th International Workshop on Quality in Databases (QDB 2010). The workshop focused on issues such as data quality assessment frameworks, data privacy and visualization, as well as the applications of data quality analysis to new domains such as the LOD (Linked Open Data) movement.

The report by Bizer, Boncz, Brodie and Erling results from discussions held at 2011 STI Semantic Summit within a group of twenty-five researchers, on the meaningful use of Big Data. The four authors each put forward a challenge: multi-disciplinary Big Data integration, the Billion Triple Challenge from the Semantic Web community, getting value out of government linked open data, and integrating linked data in regular DBMSs.

Last but not least, the report on the 4th Workshop on Very Large Digital Libraries, by Candela, Manghi and Ioannidis, highlights recent works at the confluence of very large digital libraries, and very large data archives. The workshop proceedings consider topics such as scalable support for digital libraries, archiving scientific data and data archive federations.

The call for papers of the 5th International Workshop on Testing Database System closes the issue.

On behalf of the Record's editorial board, let me wish you pleasant holidays and a Happy New Year!

Your contributions to the Record are welcome via the RECESS submission site (<http://db.cs.pitt.edu/recess>). Prior to submitting, be sure to peruse the Editorial Policy on the SIGMOD Record's Web site (<http://www.sigmod.org/publications/sigmod-record/sigmod-record-editorial-policy>).

Ioana Manolescu

December 2011

Past SIGMOD Record Editors:

Harrison R. Morse (1969)
Daniel O'Connell (1971 – 1973)
Randall Rustin (1975)
Thomas J. Cook (1981 – 1983)
Jon D. Clark (1984 – 1985)
Margaret H. Dunham (1986 – 1988)
Arie Segev (1989 – 1995)
Jennifer Widom (1995 – 1996)
Michael Franklin (1996 – 2000)
Ling Liu (2000 – 2004)
Mario Nascimento (2005 – 2007)
Alexandros Labrinidis (2007 – 2009)