Facts about XLDB-2011

Jacek Becla Kian-Tat Lim Daniel L. Wang

February 21, 2012
SLAC-TN-12-001

This note provides details of the 5th Extremely Large Databases Conference and Invitational Workshop that were held in 2011 on 18-19 October and 20 October, respectively, at the SLAC National Accelerator Laboratory in Menlo Park, California, USA. The website is located at: `http://www-conf.slac.stanford.edu/xldb2011/`.

1 Conference

The main goals of the conference were:

- Encourage and accelerate the exchange of ideas between users trying to build extremely large databases worldwide and database solution providers
- Share lessons, trends, innovations, and challenges related to building extremely large databases
- Facilitate the development and growth of practical technologies for extremely large databases
- Strengthen, expand, and engage the XLDB community

Attendance was limited to 280 people due to the auditorium's capacity. All talks were held in SLAC's Panofsky Auditorium The conference included a reception and dinner at the Stanford Faculty Club.

1.1 Invited talks

Reference cases from industry

- "Real-time Analytics at Facebook," Zheng Shao (Facebook)
- "Data Infrastructure at LinkedIn," Shirshanka Das (LinkedIn)
- "Extreme Analytics at eBay," Thomas Fastner (eBay)
- "Youtube Data Warehouse," Biswapesh Chattopadhyay (Google)

A panel discussion with the speakers followed this session.

Statistics at scale

- "Big Data System Metrics - Managing Systems of Extreme Scale," Nachum Shacham (eBay)
- "Extremely Large Data Challenges - What R Can and Can't Do," Susan Holmes (Stanford Statistics Dept)

Visualization

- "Visualizing Large, Complex Data," Kwan-Liu Ma (UC Davis)
- "Integrated Analysis and Visualization for Data Intensive Science: Challenges and Opportunities," Attila Gyulassy (UC Davis)
- "Database Requirements for Visualizing Large Multiscale Simulation Data," Ralf Kaehler (SLAC/Kavli)

Work supported in part by US Department of Energy contract DE-AC02-76SF00515.

Reference cases from science

- "Sequence Read Archive: Validation, Archival, and Distribution of Raw Sequencing Data," Eugene Yaschenko (NCBI/NIH)
- "Managing the Data Bonanza: Generating, Analysing and Sharing Data for Megasequencing Projects," Narayan Desai (ANL)
- "Functional Annotation of the Protein Sequence Universe," Eugene Kolker (Seattle Children's Hospital)

Growing to large scale stories

- "Drug Discovery in the Era of Big Data," Gregory McAllister (Novartis)
- "Growing to Large Scale at Netflix," Eric Colson (Netflix)

Short surveys

- "Value of Train Scheduling," Daniel Wang (SLAC/LSST)
- "Shared-nothing vs Shared-disk?" Michael Stonebraker (MIT)

Data intensive simulation

- "In-situ Scientific Data Processing for Extreme Scale Computing," Scott Klasky (ORNL)

Platinum sponsor talk

- "Building Blocks for Large Analytic Systems," Andrew Lamb (Vertica Systems)

Cloud coumputing at scale

- "Scaling Up Quickly on the Cloud," Edmond Lau (Quora)
- "One Billion Rows a Second: Fast, Scalable OLAP in the Cloud," Michael Driscoll (Metamarkets)
- "Cloud Computing at Scale," Roger Barga (Microsoft)

A short discussion followed this session.

1.2 Lightning talks

The conference included a two sessions of 5-minute lightning talks. In order of presentation, these were:

Session 1

- "Techniques for Discovering Relationships in Massive-Scale Data," Peter J. Haas / IBM
- "Lightning Queries," Miguel Branco (EPFL)
- "The 1000 Genomes Project, User Accessibility," Laura Clarke / EBI
- "Scalable Analytics on SciDB, a scientific data management & analytics platform," Paul Brown (SciDB)
- "bigBed/bigWig file format and usage case for efficient remote access to large data sets," Hiram Clawson (UCSC)
- "InfiniteGraph - A Scalable, Distributed Graph Database," Leon Guzenda (Objectivity)
- "MCDB–The Monte Carlo Database System," Chris Jermaine (Rice University)
- "Hadoop jobs require one-disk-per-core, Myth or Fact?," Min Xu (Seamicro)

Session 2

- "Serving Analytics in Real-time to 10000s Servers," Rushan Chen (Zynga)
- "A Science Benchmark on SciDB," Michael Stonebraker (MIT)
- "Minerva: A Compute Capable SSD Architecture for Next-Generation Non-Volatile Memories," Maya Gokhale (LLNL)
- "The Next EPICS Data Storage," Nikolay Malitsky (BNL)
- "Resource Management in the Greenplum Parallel Database," Sivaramakrishnan Narayanan (EMC)
- "Lustre, Scalable Storage for Big Data," Robert Read (Whamcloud)
- "Data Infrastructure for Massive Scientific Visualization and Analysis," Christopher Mitchell (LANL)
- "Make it work: Stewardship of Digital Information," Jane Mandelbaum (Library of Congress)

Lightning talk abstracts are available online: `http://www-conf.slac.stanford.edu/xldb2011/LT_Abstracts.asp`

1.3 Posters

Posters were accepted for presentation during the conference.

Session 1

- "Fail-Proofing Hadoop Clusters with Automatic Service Failover," Michael Dalton (Zettaset)
- "Data is Dead – Without What-If Models," Pat Selinger (IBM)
- "Backup and Restore strategies for a Giant," Ruben Gaspar (CERN)
- "Implementing the SSDB Benchmark on SciDB," DIFALLAH Djellel Eddine / UNIFR
- "Big Data Challenges in Application Performance Management," Tilmann Rabl (University of Toronto)
- "Large-Scale Databases in Gaia," Pilar de Teodoro (ESA)

Session 2

- "From High-throughput Single Molecule Measurement Data to Actionable Bioinformatics Knowledge," Jason Chin (Pacific Biosciences)
- "Federal Market Information Technology in the Post Flash Crash Era: Roles for Supercomputing," John Wu (LBL)
- "Genome Annotation with Full-text Biomedical Research Articles," Maximilian Haeussler (CBSE, UC Santa Cruz)
- "Modeling Dengue Fever: An extra large data challenge," Kun Hu (IBM)
- "GLADE: A Scalable Framework for Efficient Analytics,," Florin Rusu (UC Merced)
- "The CGHub Cancer Genomics Data Repository," Brian Craft (UC Santa Cruz)

Poster abstracts are available online: `http://www-conf.slac.stanford.edu/xldb2011/Poster_Abstracts.asp`

2 Workshop

The invitational workshop fostered close discussion on large data management in health care, genomics, spreadsheet-based analysis, and statistics and machine learning at scale. It is described in a forthcoming comprehensive report.

3 Organization

XLDB-2011 was organized by a committee consisting of: Anastasia Ailamaki(École Polytechnique Fédérale de Lausanne), Jacek Becla [chair] (SLAC), Peter Breunig (Chevron), Bill Howe (University of Washington), David Konerding (Google), Samuel Madden (MIT), Jeff Rothchild, (Facebook), and Daniel L. Wang (SLAC). It was sponsored by eBay, Vertica, Chevron, SciDB&Paradigm4, MonetDB, IBM, and Exxon Mobil.

4 Acknowledgements

This technical note was supported in part by the U.S. Department of Energy under contract number DE-AC02-76SF00515.