XLDB Asia 2012: the First Extremely Large Databases Conference at Asia

Xiaofeng Meng
Renmin University
Beijing, China
xfmeng@ruc.edu.cn

Fusheng Wang
Emory University
Atlanta, Georgia, USA
fuseng.wang@emory.edu

ABSTRACT

The Extremely Large Databases (XLDB) series of conferences/workshops have been held successfully six times in recent years. The First XLDB Conference at Asia (XLDB Asia) was held at Beijing, China on June 22-23, 2012. The conference attracted nearly 200 participants. XLDB takes a fresh format on the organization through invited talks, lightning talks and open discussions. Most invited speakers are also owners of real extremely large data from industries and scientific research, practitioners who are handling the real data, or DBMS researchers who are researching new solutions. Based on the enthusiastic embrace and positive feedbacks from participants, we believe the conference series will continue as a venue for the discussion on the management and analysis of extremely large data sets with increasing popularity.

1. INTRODUCTION

The Extremely Large Database Conferences (XLDB) [1] were established by people with highly demanding data challenges, and researchers and solution providers who are developing systems to address such challenges. Encouraged by the success holding of the conferences in the past five years – four times in the USA and twice at Europe, this year the XLDB conference was extended to the Asian community. The First Extremely Large Database Conference at Asia [2] was held at Beijing, China on June 22-23, 2012, which brought together premium speakers around the world and attracted nearly 200 participants.

There is no strict definition of "XLDB" [5], and in many cases it represents a major trend of increasing scales of data and the associated complexity and challenges on managing and analyzing the amount of data. The goals of the conferences are to provide a meeting place for database researchers, for businesses with advanced solutions, and for people from many research disciplines, industries and organizations who need to urgently address real data challenges. Topics include: the state of the art data handling technologies on extremely large datasets; practical use cases of current and anticipated data challenges; lessons and innovations on building extremely large databases; and trends and strategies for surmounting current hurdles. Different from traditional technique conferences, XLDB conferences are based on invited premium talks by pioneers and leaders in the field, especially those who are owners of real extremely large data from industries and scientific research, practitioners who are handling the real data, or DBMS researchers who are researching new solutions.

The program of XLDB Asia 2012 consisted of four sessions: reference cases from scientific communities, reference cases from industries, research topics on big data management, and lightning talks from a wide spectrum of topics. The conference also provided stimulating discussions with three intensive panel discussions entitled "the Challenges and Requirements for Handling Extremely Large Scientific Data", "NoSQL: the Cure for Big Data?", and "Evolution or Revolution: Database Research for Big Data".

2. INVITED SPEAKERS

To provide broad discussions, invited speakers came from scientific research communities, industries, and the database research community. Speakers include Alexander Szalay from John Hopkins University, a pioneer in astronomic data management, who builds one of the largest scientific databases together with Jim Gray from Microsoft; Joel Saltz from Emory University, a pioneer in biomedical informatics, who works on extreme scale data analytics and queries of biomedical data; Kian-Tat Lim from SLAC National Accelerator Laboratory, Stanford University, who works on designing and building the petabyte-scale data management system for the Large Synoptic Survey Telescope project [3], one of the coming largest scientific databases; Chenzhou Cui from National Astronomical Observatories, Chinese Academy of Sciences; and Lizhe Wang from the Center of Earth Observation and Digital Earth, Chinese Academy of Sciences.

Industry speakers include Milind Bhandarka, Chief Architect of Greenplum Labs in EMC; Tomasz Nykiel from Facebook; Zhengkun Yang, senior scientist from

Taobao, the largest online bidding company in China; Masaya Mori, the founding director of Rakuten Institute of Technology, Japan; Shohei Hido, co-leader of Jubatus, Preferred Infrastructure, Inc., Japan; and Eddy Cai, Manager of Data Platform Engineering, eBay.

Academic speakers include Laura Haas, director of IBM massive data, analytics and modeling research at IBM Almaden Research Center; Martin Kersten, one of the founders of MonetDB, and a pioneer of column store and array database; Xiaodong Zhang from the Ohio State University; and Haixun Wang from Microsoft Research Asia.

Besides invited talks, the conference also accepted a set of high quality lightning talks and poster presentations about XLDB related research and systems.

3. SCIENTIFIC COMMUNITIES

This session featured five invited talks from scientific communities. Alexander Szalay in the talk entitled "Extreme Data-Intensive Scientific Computing" presented use cases on data intensive scientific computing. He introduced the Sloan Digital Sky Survey (SDSS) project, in which large amount of star data from sky images were collected, analyzed and managed. He also discussed a cheap, yet high performance multi-petabyte system currently under construction at John Hopkins University. Joel Saltz introduced their work on high performance pathology image processing pipeline with hybrid CPU/GPU architecture to support feature extraction, machine learning, and querying and comparing results on "big image data". Kian-Tat Lim from SLAC introduced an open-source database management system "qserv", which aims to manage massive amount of astronomical data. Preliminary experiments have demonstrated the feasibility on managing and processing 32TB data with a cluster of 150 nodes. Lizhe Wang presented their work on data intensive computing for earth observation, and outlined their major focuses and challenges. Chenzhou Cui introduced projects on virtual observatories, and increasing data challenges for coming projects in the area at extreme scale. The session was concluded with the panel discussion "the Challenge and Requirements for Handling Extremely Large Scientific Data", in which the panelists discussed the challenges, the experiences and prospectives on analyzing and managing big scientific data.

4. INDUSTRIES

The industry session featured six invited talks. Masaya Mori presented a real case of utilizing Hadoop and coping with BigData in Rakuten, to support business data mining, product ranking, product search and online advertisement for e-commerce. He also introduced the new trends of "Online to Offline" (O2O) and potential requirements and challenges. Satoshi Oyama in his talk entitled "Distributed Online Machine Learning Framework for Big Data" presented Jubatus, the first open source platform for online distributed machine learning on the data streams of big data. Jubatus takes a loose model sharing architecture for efficient training and sharing of machine learning models, by defining three fundamental operations, which matches the Map and Reduce operations in Hadoop. Milind Bhandarkar reported the results of the "Workshop on Big Data Benchmarking" [4] held at San Jose on May 8-9 2012, in which a large number of companies and institutions agreed to work on establishing a benchmark framework for big data. Zhenkun Yang introduced Oceanbase, an open source distributed database system which supports extreme scale of transactions for Taobao. Oceabase provides a hybrid architecture which combines high throughput transactions on current data and large scale analytics on history data. Tom Nykiel reported the scalability challenges and solutions of the Hadoop Distributed Filesystem at Facebook, driven by the extreme scale of data, for example, the Hive based data warehouse stores tens of petabytes of data, in hundreds of millions of files. Eddy Cai presented how eBay handles big data with low cost way and how to resolve business question based on technical solutions.

The Panel discussion entitled "NoSQL: the Cure for Big Data?" had a broad discussion of a variety of topics on NoSQL, and the panelists shared their experiences on the difficulty to find a single perfect solution, and their visions on how NoSQL and RDBMS could interplay.

5. THE DATABASE COMMUNITY

The session of research on big data management featured four invited talks. Laura Haas analyzed four types of data integration problems, and introduced the challenges of integrating extremely large data. Xiaodong Zhang introduced a scale-out model for big data software development in distributed systems. The model generalizes critical computation and communication behavior and computation-communication interactions for big data analytics in a scalable and fault-tolerant manner. Haixun Wang in his talk entitled "Managing and Mining Billion-Node Graphs" presented the challenges posed by big graph data generated from Web and social network applications, the constraints of architectural design, the different types of application needs, and the power of different programming models that support such needs. Martin Kersten presented a new query language SciQL to support powerful queries on top of the array database MonetDB, and introduced use cases to support scientific applications.

This session was concluded with enthusiastic discussions in the panel entitled "Evolution or Revolution: Database Research for Big Data", with panelists con-

sisting of Laura Hass, Martin Kersten, Haixun Wang, Min Wang, and Xiaodong Zhang. The panel had an open discussion of a wide range of questions, including the ones from the audience. Example questions include:

- What are the essential needs of big data applications and/or users that are not being met by DBMS?
- What changes in dbms would better support i) analytics on big data or ii) management of big data?
- Can we define a general-purpose system infrastructure the DB community accepts, and that big data processing systems fit? Or do we need multiple interoperating systems?
- What do you consider the most promising trend in db research for big data? Why?

6. LIGHTNING TALKS

The lightning talks included 9 talks with a variety of topics, ranging from big data collection, movement, storage, queries, and online aggregation. Weisong Shi presented their work on streamlining processing for big data on metagenomics software, Jennie Zhang presented the work on how an array query language can be used to support a scientific use case – the LOw Frequency ARray radio telescope project. Fabian Groffen presented "Jacqueline: JSON/JAQL for MonetDB", and Abhishek Parolkar introduced a massive data collection tool Fluentd. Three cloud based data management systems of different flavors were presented by Jan-Jan Wu, Yunpeng Chai, and Jidong Chen, respectively. Yingjie Shi introduced COLA: a cloud-based on-line aggregation system.

7. CONCLUSIONS

The Extremely Large Database Conference series has been founded to provide a meeting place for people from different domains and background who need to urgently address real big data challenges. XLDB takes a fresh format and becomes a popular venue for discussions on extremely large databases. The First XLDB Asia conference attracted a good number of participants (nearly 200 people) and received very positive feedbacks from both speakers and participants. The conference came with a format with invited talks by those who are owning or handling real extremely large data, multiple panels for stimulus discussions, and lightning talks for broader participation. Future improvement includes a formal poster session to enable more users' participation and interactive discussions.

8. ACKNOWLEDGMENTS

Through the generous support from our sponsors EMC, MonetDB, Microsoft Research Asia, HP Lab China, DataTang and HZ Books, along with a contribution from the National Science Foundation of China, we were able to keep the conference fees to a minimum that allowed many students to participate. We are also grateful for the support from organizations such as China Computer Federation Technical Committee on Databases, Lab of Mobile and Web Data Management at Renmin University of China, the Department of Biomedical Informatics at Emory University, and National Engineering Lab for Video Technology at Peking University. We also thank Jacek Becla for his encouragement and many insightful suggestions for the organization of the conference. Thanks are also due to many student volunteers who work hard on the coordination of the conference.

9. REFERENCES

[1] The extremely large database conferences. http://xldb.org.
[2] The first extremely large database conference at asia. http://xldb-asia.org.
[3] The large synoptic survey telescope project. http://www.lsst.org.
[4] Workshop on big data benchmarking, san jose, may 8-9, 2012. http://clds.ucsd.edu/wbdb2012/.
[5] Xldb. http://en.wikipedia.org/wiki/XLDB.