Report on the Fourth International Workshop on Cloud Data Management (CloudDB 2012)

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

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

Adam Silberstein
Trifacta Inc.
San Francisco, USA
aesilberstein@yahoo.com

1. INTRODUCTION

The workshop series on Cloud Data Management (CloudDB) has been held successfully in the past four years [4, 2, 3, 5]. CloudDB serves as a premier forum for researchers and practitioners to present research results and share ideas and progress in the area of data management within cloud computing infrastructure.

Technology advances in communications, computation, and storage result in huge collections of data, capturing information of value to business, science, government, and society. Data volumes are currently growing faster than Moore's law. Looking forward, the exponential growth is not likely to stop. The huge volumes of data pose major infrastructure challenges, including data storage at Petabyte scale, massively parallel query execution, facilities for analytical processing, and online query processing. Meanwhile, the rise of large data centers and cluster computers has created a new business model, cloud-based computing, where businesses and individuals can rent storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale computer installations. Cloud-based data storage and management is a rapidly expanding business. Whilst these emerging services have reduced the cost of data storage and delivery by several orders of magnitude, there is significant complexity involved in ensuring that large data services can scale when needed to ensure consistent and reliable operations under peak loads. Cloud-based environment has the technical requirements to manage data center virtualization, lower cost and boost reliability by consolidating systems on the cloud.

CloudDB brings together researchers and practitioners in cloud computing and data-intensive system design, programming, parallel algorithms, data management, scientific applications, and information-based applications to maximize performance, minimize cost and improve the scale of their endeavors.

The fourth ACM international workshop on cloud data management was held in Hawaii, USA on October 29, 2012, co-located with the ACM 21st Conference on Information and Knowledge Management (CIKM) [1]. The call for papers attracted a wide range of submissions on query optimization, data security and privacy, big data analytics, and system development. The program committee accepted seven papers from fourteen submissions by authors from Asia, Europe, North America and South America. In addition, the program included three keynote speakers from leading cloud computing researchers.

2. KEYNOTE TALKS

The keynote talks covered topics on OLTP benchmarking in the cloud, data analytics in the cloud, and challenges in enabling social applications at scale.

The first keynote talk was delivered by Carlo Curino from Microsoft on *Benchmarking OLTP/Web Databases in the Cloud: the OLTP-Bench*. The speaker shared the experience in building several ad-hoc benchmarking infrastructures for various research projects targeting several OLTP DBMSs, ranging from traditional relational databases, main-memory distributed systems, and cloud-based scalable architectures. OLTP-Bench is capable of controlling transaction rate, mixture, and workload skew dynamically during the execution of an experiment, thus allowing the user to simulate a multitude of practical scenarios that are typically hard to test (e.g., time-evolving access skew). OLTP-Bench also provides ten workloads derived from synthetic micro benchmarks, popular benchmarks and real world applications.

The second keynote talk *Large Scale Data Analytics on Clouds* was delivered by Geoffrey Fox from Indiana University. The speaker summarized major issues affecting the use of clouds to support data science, and discussed the major characteristics between cloud and traditional high performance computing systems. While cloud is on-demand driven and provides scalable elastic services, traditional supercomputers can achieve high performance through large scale highly parallelized jobs. Thus, when analyzing large scale data, the characteristics of different categories of applications should be considered. These include map-only applications, MapReduce applications, classic MPI applications, and iterative MapReduce based applications. To achieve high performance, the speaker discussed the mapping of different applications to HPC and Cloud systems.

Ashwin Machanavajjhala from Duke University presented in his keynote talk *Challenges in Enabling Social Application at Scale* on major challenges associated with social network data and new applications for social discovery and engagement. He summarized three applications that should be considered properly to solve the data management and privacy issues: i) Feed Following, or the problem of delivering highly personalized feeds based on content generated by one's friends; ii) Social Coordination, or the problem of jointly planning and coordinating on a task, and iii) Social Recommendations, or recommending objects and people based on one's social connections. He also shared his experience on working with these challenges in data management and privacy research.

3. RESEARCH PAPER PRESENTATIONS

3.1 Workload-aware Processing

Computation throughput maximization, efficient resource scheduling, and query optimization are critical components for cloud computing. These are covered by three papers from different perspectives: load balancing through automatically imbalance detecting and mitigating, elastic query processing to maximize the efficiency of providers' environment, and statistical based estimation of the data in the cloud to support query optimization in the cloud.

Cloud data stores achieve high scalability and elasticity by partitioning data across a large number of servers. These stores must detect and cope with load imbalance. Markus Klems et al. develop a cloud datastore load balancer in the paper *The Yahoo! Cloud Datastore Load Balancer.* The load balancer is called *Yak*, which now provides load balancing for Yahoo!'s cloud storage system *Sherpa*. The authors describe the key design principles for Yak: understanding the goal, measurable, simple, extensible and configurable, conservative and knowing the limit. Based on the design principles, Yak defines and monitors load metrics to detect imbalance, and provides a set of rules that decide when to invoke load balancing actions. When hotspots are detected, Yak automatically balances the load by migrating tablets from the overloaded servers, and also by splitting data into new tables.

The paper *Towards Non-Intrusive Elastic Query Processing in the Cloud* by Ticiana Coelho Da Silva et al. focuses on taking full advantages of the potential flexibility of cloud computing systems. One major benefit of cloud computing is the elasticity which enables the systems to provide and remove resources according to the applications needs in real-time. They develop a non-intrusive approach that monitors the performance of relational DBMSs in a cloud infrastructure, and automatically makes decisions to maximize the efficiency of providers' environment while still satisfying "service level agreements". The workflow of their approach contains four modules: the *Partition Engine* partitions the input query Q to achieve the query's service level objective (SLO); the *Monitor Engine* is executed within each VM allocated to process Q and aims at guaranteeing that each VM meets the expected SLO; the *Capacity Planner* provides a number of VMs initially to process Q within the agreed SLO, minimizing the computational cost and penalty; and the *Orchestration Engine* is responsible for the communication between the modules.

The paper entitled *HEDC: A Histogram Estimator for Data in the Cloud* by Yingjie Shi et al. introduces an approach for histogram estimate for data in the cloud. With increasing popularity of cloud based data management, improving the performance of queries in the cloud is an urgent issue to solve. Summary of data distribution and statistical information has been commonly used in traditional database to support query optimization and histograms are of particular interest. Since it could be much expensive to construct the exact histogram on massive data, building the approximate histogram is a more feasible solution. They propose a histogram estimator called *HEDC*. The workflow of HEDC is built on an extended MapReduce framework, and takes a novel block based sampling mechanism to leverage the sampling efficiency and estimate accuracy.

3.2 Energy Efficient Data Centers

In the paper entitled *Cloud Computing for Environment-Friendly Data Centers*, Michael Pawlish et al. consider the carbon footprint and utilization rates in data centers. Previous literature shows that low utilization rates in data centers are due to the forecasting of demand to meet spikes in data center use. This management policy has led to many servers running idle the majority of the time which is a waste of resources. The authors argue that a majority of the data centers should be downsized through decommissioning of phantom servers, virtualization, and shifting spikes in demand to a cloud provider. They adopt data mining techniques of decision trees and case-based reasoning to conduct analysis for decision support in cloud computing at data centers.

3.3 Computing Models

Zhuhua Cai et al. propose a system called *GraphInc* in the paper *Facilitating Real Time Graph Mining.* While incremental processing is critical for real-time graph mining, designing incremental graph algorithms is challenging. GraphInc is built on top of the Pregel model and provides efficient incremental processing of large graphs. Users are allowed to write programs as if on batch workloads and the algorithms are automatically converted to an incremental one by memorizing and reusing subcomputations. Programmers thus can develop graph analytics in the Pregel model without worrying about the continuous nature of the data. GraphInc integrates new data in real-time in a transparent manner, by automatically identifying opportunities for incremental processing.

3.4 Privacy and Security

Privacy and security are critical issues to solve for processing sensitive data in the cloud. These are covered from two different aspects in following two papers respectively: privacy preserving query processing in the cloud, and efficient encryption based approach to achieve data security and query security for data processing in the cloud.

Xu Han et al focus on the privacy protection problem in the cloud in their paper entitled *Differentially Private Top-k Query over MapReduce.* They propose an efficient privacy protection algorithm called *DiffMR*, which aims to process top-k query as well as satisfying differential privacy. They adopt an exponential mechanism to select top-k records from big data sets based on specified score function to avoid the privacy leak. In order to get more accurate results, they reduce the reject rate and perform exponential selection multiple times during the MapReduce progress. Laplace noise will be added at last and then post-processing will be performed to improve the quantity of results.

The paper *A Security Aware Stream Data Processing Scheme on the Cloud and its Efficient Execution Methods* by Katsuhiro Tomiyama et al. evaluates queries over encrypted data streams in the cloud. A public cloud may be managed by a third party and outside the firewall of the organization, which brings the problem of data security and query security. The authors propose a scheme based on CryptDB to evaluate queries over encrypted data streams. They describe performance issues incurred by the proposed scheme, and present an approach to reduce the encryption cost and amounts of data to be transmitted. In addition, they propose an approach to reduce memory usage by analyzing a plan tree in a stream processing engine (SPE). The experiments demonstrate that their approaches can improve the memory utilization.

4. CONCLUSIONS

CloudDB has been held successfully four times associated with CIKM since 2009. During the four years, cloud computing has undergone significant development and attracted major interest from both industry and academia. CloudDB workshop series aims to address the challenges of large scale database services based on the cloud computing infrastructure in a timely fashion, with increasing number of participants. Topics of CloudDB 2012 covered query optimization, data security and privacy, big data analytics, and large scale social applications in the cloud. The participants agreed that many open challenges remain open, such as cloud data security and privacy, and the efficiency of big data management in the cloud.

5. ACKNOWLEDGMENTS

The workshop was partially supported by the Research Fund of Renmin University of China (No. 11XNL010) and by grant R01LM009239 from the National Library of Medicine. We would like to thank the program committee members, keynote speakers, authors and attendees, for making CloudDB 2012 a very successful workshop. We also express our great appreciation for the support from Renmin University of China and Emory University.

6. REFERENCES

[1] X. Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors. *CIKM '12: Proceedings of the 21st ACM international conference on Information and knowledge management*, New York, NY, USA, 2012. ACM. 605120.

[2] X. Meng, Y. Chen, J. Lu, and J. Xu. Report on the second international workshop on cloud data management (clouddb 2010). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1969–1970, New York, NY, USA, 2010. ACM.

[3] X. Meng, Z. Ding, and H. Hu. Report on the third international workshop on cloud datamanagement (clouddb 2011). In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2637–2638, New York, NY, USA, 2011. ACM.

[4] X. Meng, J. Lu, J. Qiu, Y. Chen, and H. Wang. Report on the first international workshop on cloud data management (clouddb 2009). *SIGMOD Rec.*, 39(1):58–60, Sept. 2010.

[5] X. Meng, A. Silberstein, and F. Wang. Clouddb 2012: fourth international workshop on cloud data management. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2754–2755, New York, NY, USA, 2012. ACM.