Report on the Second International Workshop on Cloud Data Management (CloudDB 2010)

Xiaofeng Meng[1] Ying Chen[2] Jiaheng Lu[1] Jianliang Xu[3]
{xfmeng,jiahenglu}@ruc.edu.cn, yingch@cn.ibm.com, xujl@comp.hkbu.edu.hk
[1]School of Information and DEKE, MOE, Renmin University of China, Beijing China
[2]IBM Research - China
[3]Hong Kong Baptist University, Hong Kong

Categories and Subject Descriptors
A. General Literature; A.1 INTRODUCTORY AND SURVEY

General Terms
Documentation

Keywords
Cloud Data Management

1. INTRODUCTION

The second ACM international workshop on cloud data management was held in Toronto, Canada on October 30, 2010 and co-located with the ACM 19th Conference on Information and Knowledge Management (CIKM). The main objective of the workshop was to address the challenge of large data management based on cloud computing infrastructure. The workshop brings together researchers and practitioners in cloud computing and data-intensive system design, programming, parallel algorithms, data management, scientific applications and information-based applications interested in maximizing performance, reducing cost and enlarging the scale of their endeavors.

The workshop attracted 11 submissions from Asia, Canada, Europe and the United States, out of which the program committee finally accepted 8 full papers. The accepted papers focused on cloud-based indexing and query processing, cloud security, cloud replication and system development.

1. RESEARCH PAPERS

The technical paper session consisted of eight presentations, whose main points are summarized next. Together, they give a glimpse to the exciting new developments spurred by data management in the cloud. These papers cover a variety of topics. We believe that these papers will provide researchers and developers with a brief glimpse into this exciting new technology, specifically from the perspective of cloud data management.

The paper entitled *ESQP: An Efficient SQL Query Processing for Cloud Data Management* focuses on How to improve query efficiency in cloud data management system, especially query on structured data. J. Zhao, X. Hu and X. Meng first analyzed the main shortcomings of query processing in the existing cloud computing products. Then they presented an efficient query processing algorithm to support search on structured data. Their approach is inspired by the idea of MapReduce, in which a job is divided into several tasks. Based on the distributed storage of one table, this algorithm divides a user query into different subqueries, at the same time, with replicas in cloud, a subquery is mapped to k+1 subqueries. Every subquery has to wait in the queue of the slave where the query data store. In order to balance the load, their algorithm also takes two scheduling strategies to dispatch the subquery.

C. Ordonez and S. Pitchaimalai proposed some novel techniques to handle data mining computations inside the DBMS in the paper *"Comparing SQL and MapReduce to compute Naive Bayes in a Single Table Scan"*. The Naive Bayes classification algorithm is used to demonstrate the usefulness of these techniques in the paper. The techniques work completely inside the DBMS exploit the DBMS programmability wherein the user has complete access to the data but transparent to the DBMS internals. SQL and User Defined Functions (UDFs) are used to program the Naive Bayes algorithm in the DBMS. Also, they compare these techniques with MapReduce. Finally, they consider two phases of the classifier: building the model and scoring a new data set. Both building and scoring phases involves a single table scan on the input data set for discussed techniques.

The adaptive scheme considered in *Adaptive Query Execution for Data Management in the Cloud, A. D. Popescu, D. Dash, V. Kantere and A. Ailamaki* uses a cost model to switch between MapReduce, and a DBMS. They showed that "light" queries, with small response times, can also be processed in the cloud, while existing cloud-based data analysis systems using MapReduce, are geared towards long running "heavy" analytical queries. In their mind the cloud DBMS should satisfy both the heavy queries and the light queries efficiently. The heavy queries are required to build the knowledge and run large scale analysis, while the small queries are required to enable interactive analysis of the data and real-time feedback. In order to make an informed decision on where to execute a query, they propose a cost model that takes as input the query and the failure rate of the hardware nodes on which the query will execute, and determines the best option for running the query.

In the Paper entitled *Towards a Data-centric View of Cloud Security*, W. Zhou, M. Sherr, W. Marczak, Z. Zhang, T. Tao, B. T. Loo and I. Lee took an alternative perspective and proposed a data-centric view of cloud security while others focused primarily concentrated on securing the operating systems and virtual machines on which the services are deployed. They first discussed data management challenges that face multiuser cloud environments based on some analysis, they identify three security

Copyright is held by the author/owner(s).
CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0099-5/10/10.

challenges associated with cloud data management: (i) Secure Query Processing and Data Sharing, (ii) System Analysis and Forensics, and (iii) Query Correctness Assurance. To meet these challenges, they proposed the Declarative Secure Distributed Systems (DS2) platform. DS2 allows cloud users to both seamlessly integrate their services without exposing their confidential information as well as verify the authenticity of received data.

In the paper *Contract-based Cloud Architecture*, M. Schnjakin, R. Alnemr and C. Meinel proposed an architecture to facilitate the integration of these security requirements in the cloud environment and to address the legal issues attached. Their approach customized the selection of a service provider based on the companies' preference. They also defined a trusted third party to handle the monitoring and auditing processes over different service providers.

K. Daudjee and S. Savinov investigated data replication in a virtualized environment, focusing on provisioning when the master database server is heavily loaded or when it fails in the paper titled *Dynamic Data Replication through Virtualization*. They showed that using virtualization, provisioning of the backup replica can be done more quickly than in traditional environments, thereby reducing service interruption time. Fianlly, they also showed that load balancing using replication can be done efficiently in a virtualized environment.

In the Paper entitled *Benchmarking Cloud-based Data Management Systems*, Y. Shi, X. Meng, J. Zhao, X. Hu, B. Liu and H. Wang investigated that several benchmarks have been proposed to evaluate the performance. However, there were no reported studies about these benchmark results which provide users with insights on the impacts of different implementation approaches on the performance. So they conducted comprehensive experiments on several representative cloud-based data management systems to explore relative performance of different implementation approaches.

The topic in *Towards Bipartite Graph Data Management* by B. Zhao, W. Qian and A. Zhou is to propose a logic graph structure for indexing bipartite graph to improve common operations efficiently after raising the issues of BGDM and present architecture of BGDM. Their structure can avoid loading the whole block for vertex queries by using bloom filter methods.

2. CONCLUSION

CloudDB 2010 was the second CIKM-associated workshop addressing the challenges of large database services based on the cloud computing infrastructure. 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 large data service can scale when one needs to ensure consistent and reliable operation under peak loads. Cloud-based environment has the technical requirement to manage data center virtualization, lower cost and boost reliability by consolidating systems on the cloud. The existing research works in the area of cloud-based data management are still somehow immature and significant room for progress exists. The works presented in the workshop mainly focused on adapting existing Grid and Map/Reduce techniques to the cloud environment. The participants agreed that many open challenges still remain such as cloud data security and the efficiency of query processing in the cloud.

3. Acknowledgement

We would like to thank the program committee members, authors and attendees, for making CloudDB 2010 a successful workshop. Finally, we also express our great appreciation for the support from Renmin University of China, Hong Kong Baptist University and IBM research-China.