Second International Workshop on Data Management in the Cloud

1 Introduction

The Second International Workshop on Data Management in the Cloud took place on April 8, 2013 in Brisbane, Australia, on the day before ICDE. The DMC workshop aims to bring researchers and practitioners in cloud computing and data management systems together to discuss the research issues at the intersection of these two areas, and also to draw more attention from the larger data management and systems research communities to this new and highly promising field. The DMC Workshops are sponsored by the IEEE TCDE Workgroup on Cloud Computing.

The DMC 2013 program began with a keynote presentation by Amr El Abbadi, Professor at the University of California, Santa Barbara. This was followed by six technical papers presented in two sessions. The workshop concluded with a panel discussion on research challenges in data management for the cloud.

2 Keynote

Amr El Abbadi is a Professor of Computer Science at the University of California, Santa Barbara. He is an established researcher in the areas of databases and distributed systems. He is an ACM Fellow and an AAAS Fellow, and has been Program Chair for multiple database and distributed systems conferences, including the ACM Symposium on Cloud Computing (SoCC) 2011. Amr's keynote talk was titled "MIND THE GAP: Managing Multi-Data Center Data." It focused on transaction management issues that arise when the data is distributed across a "wide gap", that is, in multiple data centers.

Amr pointed out that storing data in a geographically distributed, multi-data center setting is required by many applications for performance and fault tolerance. In such an environment, large round trip times for network messages increase the cost of coordinating transactions among replicas of a data item. One can address this problem by weakening the transactional guarantees offered to applications, but that makes application development quite difficult for programmers. Instead, it would be highly desirable to give programmers full ACID transactions in a multi-data center setting. Amr provided an overview of the problems that need to be solved to achieve this goal, and presented work from his group that addresses several of these problems. Amr concluded by pointing out the relationship between the topic of consensus from distributed systems and the topic of commitment from database transactions, and noted that the solutions to many of the problems in multi-data center data management may be found at the confluence of these two areas.

3 Paper Sessions

The six papers presented at the workshop dealt with a wide range of research issues related to data management in the cloud. The first three papers, presented in the morning session, dealt mostly with issues related to cloud data management infrastructure. The three papers in the afternoon session dealt mostly with issues related to cloud data management applications.

Rao et al. [1] presented a protocol and system for replicating data on-demand between different clusters running analytics systems, in particular Hadoop. The protocol is motivated by the observation that data on different clusters is shared by different applications, and replicating this data on-demand saves storage cost and bandwidth as compared to bulk copying the data.

Jin et al. [2] presented a mechanism for implementing materialized views in the Cassandra distributed storage system. The mechanism enables fast access based on secondary keys in Cassandra, and keeps the materialized views updated in the face of insertions and deletions. The paper compares different alternatives for secondary key access and explores the tradeoffs in these alternatives.

Duggan et al. [3] presented a model for predicting database performance under changing hardware configurations. Unlike previous work in this area, the proposed model can be trained on one hardware configuration and be used to predict performance on other hardware configurations. The model relies on workload fingerprinting to characterize workloads and collaborative filtering models to predict performance.

Baralis et al. [4] presented an approach for supporting frequent itemset data mining on multi-core servers. The approach relies on a disk-based data structure called the VLDBMine data structure. The paper presents techniques to create this data structure in parallel and to exploit it for data mining.

Li et al. [5] proposed that data centers can be used to reduce the variance in electrical load on a power grid. The basic idea is that data centers can migrate jobs between them to adjust their energy consumption in response to fluctuations in the load on the power grid. The paper presents a dynamic pricing approach that can be used to provide incentives for data center operators to perform this migration.

Finally, Künsemöller et al. [6] presented a study of business models for caching in internet service providers (ISPs). Content providers on the internet can use content delivery networks (CDNs) to cache their data in order to improve performance. It should be possible for ISPs to provide similar caching, and the paper uses a game-theoretic approach to investigate models under which it would be beneficial to the ISP and/or the content provider to adopt a particular style of caching.

4 Closing Panel

The workshop concluded with a panel entitled "Data Management in the Cloud - Where are We? And Where to Next?" The panel consisted of five distinguished panelists: Hakan Hacıgümüş from NEC Labs America, Sriram Rao from Microsoft, Boon Thau Loo from the University of Pennsylvania, David Maier from Portland State University, and Markus Weimer from Microsoft.

Hacıgümüş presented an overview of database management issues in the cloud, noting that there is a strong desire to support SQL on the cloud. This requires advances in several areas, such as service level agreements (SLAs) and workload prediction.

Rao noted that for analytical workloads, a significant fraction of MapReduce jobs in production workloads is written in high-level languages such as Pig Latin or Hive, so the focus of research in this area should expand beyond supporting individual MapReduce jobs to supporting Pig Latin and Hive on the cloud. He also noted the importance of investigating multi-tenancy for MapReduce.

Loo suggested that using the cloud for mission critical applications gives rise to many interesting research problems. He also pointed out opportunities for leveraging Openflow software defined networks in cloud data management. In addition, he suggested that formal methods can hold promise for solving some cloud data management problems.

Maier presented work on serving scientific data from a cloud environment. He pointed out research problems that are unique to this application area, especially for scientific domains that produce massive data sets. Issues such as latency and monetary cost become important in this context.

Weimer pointed out limitations in the support for machine learning provided by current cloud data management technologies. He argued that there is a need to better support the end-to-end machine learning workflow, from data capture through to learning and all the way to deployment. He also noted that better integration of machine learning concepts such as error tolerance for certain computations can result in higher performance and better models.

The panel concluded with a discussion session in which the audience presented their questions and views on the workshop topic.

References

[1] Sriram Rao, Benjamin Reed, and Adam Silberstein. HotROD: Managing grid storage with on-demand replication. In *Proc. IEEE Int. Conf. on Data Engineering, Workshop on Data Management in the Cloud (DMC '13)*, pages 243–249, 2013.

[2] Changjiu Jin, Rui Liu, and Kenneth Salem. Materialized views for eventually consistent record stores. In *Proc. IEEE Int. Conf. on Data Engineering, Workshop on Data Management in the Cloud (DMC '13)*, pages 250–257, 2013.

[3] Jennie Duggan, Yun Chi, Hakan Hacıgümüş, Shenghuo Zhu, and Ugur Çetintemel. Packing light: Portable workload performance prediction for the cloud. In *Proc. IEEE Int. Conf. on Data Engineering, Workshop on Data Management in the Cloud (DMC '13)*, pages 258–265, 2013.

[4] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, and Alberto Grand. P-Mine: Parallel itemset mining on large datasets. In *Proc. IEEE Int. Conf. on Data Engineering, Workshop on Data Management in the Cloud (DMC '13)*, pages 266–271, 2013.

[5] Yang Li, David Chiu, Changbin Liu, Linh T.X. Phan, Tanveer Gill, Sanchit Aggarwal, Zhuoyao Zhang, Boon Thau Loo, David Maier, and Bart McManus. Towards dynamic pricing-based collaborative optimizations for green data centers. In *Proc. IEEE Int. Conf. on Data Engineering, Workshop on Data Management in the Cloud (DMC '13)*, pages 272–278, 2013.

[6] Jörn Künsemöller, Nan Zhang, and João Soares. ISP business models in caching. In *Proc. IEEE Int. Conf. on Data Engineering, Workshop on Data Management in the Cloud (DMC '13)*, pages 279–285, 2013.

Ashraf Aboulnaga and Carlo Curino
Program Committee Chairs
University of Waterloo and Microsoft