Report on the Second International Workshop on Cloud Intelligence (Cloud-I 2013)

Jérôme Darmont
Université de Lyon (Laboratoire ERIC)
5 avenue Pierre Mendès-France
F-69676 Bron Cedex – France
jerome.darmont@univ-lyon2.fr

Torben Bach Pedersen
Aalborg University (Daisy)
Selma Lagerløfs Vej 300
DK-9220 Aalborg Ø – Denmark
tbp@cs.aau.dk

http://eric.univ-lyon2.fr/cloud-i/

1. INTRODUCTION

Business intelligence (BI) is a broad field related to integrating, storing and analyzing data to help decision-makers in many domains (from "real" business to administration, health, and environment) make better decisions. Front-end analytics methods include reporting, on-line analytical processing (OLAP), and data mining.

With the increasing success of cloud computing, cloud BI "as a service" offerings have started appearing, both from cloud start-ups and major BI industry vendors. Beyond porting BI features into the cloud, which already implies numerous issues (e.g., BigData/NoSQL database modeling and storage, data localization, security and privacy, performance, cost and usage models...), this trend also poses new, broader challenges for making data analytics available to small and middle-size enterprises (SMEs), non-governmental organizations, Web communities (e.g., supported by social networks), and even the average citizen; this vision presumably requiring a mixture of both private and open data.

Thus, Cloud Intelligence is not only a current technological and research challenge, but also an important economic and societal stake, since people increasingly demand open data, which must be easily accessible on the Web, possibly mixed with private data, and analyzed with intelligible on-line tools with advanced collaborative features enabling users to share and reuse BI concepts and analyses in large scale fashion. Analysis results can then be be shared world-wide.

The Second International Workshop on Cloud Intelligence (Cloud-I 2013) [3] was held in conjunction with VLDB 2013 in Riva del Garda, Italy on August 26, 2013. In continuation of the first edition, it brought together researchers and engineers from academia and industry to discuss and exchange ideas related to Cloud Intelligence. The workshop featured a joint keynote with the BIRTE workshop and two research sessions, the latter including a panel discussion. The topics of this year's accepted papers mainly focused on MapReduce-based computations and indexing.

2. KEYNOTE

The keynote entitled "AsterixDB: A New Platform for Real-Time Big Data BI" was given by Prof. Michael J. Carey, from the University of California, Irvine. AsterixDB is a so-called "BDMS (Big Data Management System)" grounded in both parallel database systems and data-intensive computing frameworks (i.e., MapReduce and NoSQL-like frameworks) [2]. AsterixDB notably features a flexible, semi-structured data model (Asterix Data Model - ADM) based on JSON, a declarative query language (Asterix Query Language - AQL) for expressing a wide range of BI queries, and a parallel runtime engine, Hyracks, that has been scale-tested to thousands of cores. This feature set makes AsterixDB ideally suited to modern needs such as Web data warehousing, social data storage and analysis, and other real-time BI applications.

3. RESEARCH PAPERS

3.1 Session 1: MapReduce

Guoliang Zhou, Yongli Zhu and Guilan Wang, in "Cache Conscious Star-Join in MapReduce Environments", propose two join strategies for star schemas in MapReduce environments. Two algorithms, namely Multi-Fragment-Replication Join (MFRJ) and MapReduce-Invisible Join (MRIJ), avoid fact table data movement and are cache conscious in each MapReduce node. In addition, in MFRJ, the fact table is partitioned into several column groups for cache optimization; one group contains all the foreign key columns and each measure column becomes a separate group. In MRIJ, each column is processed separately one by one, giving a higher cache utilization and avoiding frequent cache misses from one column to the other column. Experimental results in cluster environments show that MFRJ and MRIJ outperform existing approaches in the Hive system.

In "Toward Intersection Filter-Based Optimization for Joins in MapReduce", Thuong-Cang Phan, Laurent d'Orazio and Philippe Rigaux criticize MapReduce for not directly supporting join operations with multiple inputs. To address this problem, they propose a Bloom filter-based intersection filter that exploits probabilistic models to remove most disjoint elements between two datasets. The cost of two-way and cascade joins is analyzed to minimize disk I/O and communication costs. Comparison experiments show that this approach is more effective than existing state-of-the-art solutions.

Yanfeng Zhang and Shimin Chen, in the position paper entitled "i^{2}MapReduce: Incremental Iterative MapReduce", aim to to support incremental iterative computation (e.g., PageRank) on constantly changing datasets (e.g., the Web graph). Since in many cases, data changes impact only a very small fraction of the datasets, and the new iteratively converged state is quite close to the previously converged state, i^{2}MapReduce saves recomputation by starting from the previously converged state, and by performing incremental updates on changing data. Based on this, i^{2}MapReduce achieves a significant performance improvement over recomputing iterative jobs in MapReduce.

3.2 Session 2: Emerging Topics

In the paper entitled "Bloofi: A Hierarchical Bloom Filter Index with Applications to Distributed Data Provenance", Adina Crainiceanu considers the use of Bloom filters in federated cloud environments. With hundreds of geographically distributed clouds participating in a federation, information needs to be shared by the semi-autonomous cloud providers. This can be done by encoding the information using Bloom filters and sharing the Bloom filters with a central coordinator. Bloofi, an efficiently constructed and maintained hierarchical index structure for Bloom filters, is thus proposed to speed-up the search process. Theoretical and experimental results show that Bloofi provides a scalable and efficient solution for searching through a large number of Bloom filters.

4. PANEL

Finally, Jérôme Darmont and Torben Bach Pedersen launched a panel discussion themed "Cloud Intelligence – Challenges for Research and Industry", by presenting Fusion Cubes [1], a collaborative framework aiming to support *self-service business intelligence*. Self-service BI enables non-expert users to make well-informed decisions by enriching the decision process with situational data, i.e., data that have a narrow use for a specific business problem, typically a short lifespan for a small group of users, and are usually not owned and controlled by the decision maker. This way of working can even be pushed towards *personal intelligence*, i.e., cloud decision-support services (in SaaS mode) that would be accessible to individuals and very small companies, from "NoETL" data integration to easy OLAP, including the automatic multidimensional modeling of the underlying datamart [4].

Morten Middelfart, from the BI vendor TARGIT, also stressed the need to allow easy integration of existing BI data with external data using a "NoETL" paradigm.

Adina Crainiceanu questioned how to increase the impact of academic research, which fueled a stimulating discussion between academics and industrials about building systems *vs.* writing scientific publications; and the value and quality of (seldom well-documented) code produced by researchers *vs.* academia-industry collaborations.

Finally, the challenges ahead for cloud intelligence systems were discussed, such as extracting data and/or knowledge from low-availability sites, making self-service BI user-friendly, mobile cloud issues, hybrid SQL/NoSQL systems and interoperability, and finally privacy issues that integrating aggregate data can solve while complying to various international regulations.

5. DISCUSSION AND OUTLOOK

If we first look at the topics of the presented papers, we see that 3 out of 4 accepted papers concerns processing various types of queries in using a MapReduce platform. This witnesses the fact that one of the key challenges in cloud intelligence is the ability to process really massive datasets within an acceptable time frame. This is even more the case than in cloud computing in general, since cloud intelligence use cases tend to involve very expensive "deep analytics" computations.

The keynote addresses the same challenge, but in a broader and more advanced sense, since it presents a considerably more advanced platform than standard MapReduce, including a more advanced data model and query language, and much more advanced optimization techniques. This shows that the cloud intelligence community is, in some sense, returning to the "old virtues" of semantic data models and high-level query languages, only now with the key difference that massive scalability must be supported at all times.

The fourth paper considered scalability in a somewhat different sense, namely in large scale and wide area distributed cloud environments. This is a reminder of the fact that cloud intelligence is not only about scaling up within a single data center, but also to enable efficient integration of geographically dispersed information resources.

The panel considered wider issues such as personal and collaborative cloud intelligence, including ETL and data integration for cloud intelligence, human computer interaction and user friendliness, and privacy issues. In comparison, these topics are more open and the exact problems and solutions perhaps less obvious, than for the issue of scaling up and out specific computations. Thus, we expect that solutions will eventually emerge, but that it will take some time. The same is true for some topics from the Call for Papers that were not presented or discussed at all. These include developing novel payment models for amortizing the cost of cloud intelligence solutions over time and different users and how to provide cloud intelligence as a service. We attribute the lack of papers on such issues to the fact that the cloud intelligence field is still young and that other issues are more pressing for the majority of users.

Summing up, we conclude that there is a lot of interesting work going on in the area of cloud intelligence, but that many challenges are still remaining. Thus, there is a continued need for venues that focus on this issue. We hope to continue the Cloud-I workshop series in connection with future VLDB conferences. For the third edition of the workshop, it is again the intention to organize a special issue of a journal for extended versions of the best papers.

6. ACKNOWLEDGEMENTS

The Cloud-I Chairs would like to thank all the authors of submitted papers for their interest in the workshop and the high quality of the submitted papers. We would also like to thank all the referees (both PC members and external reviewers) for their careful and dedicated work, both during the reviewing and the discussion phases. Working in cooperation with this Program Committee has been both an honor and a pleasure. Finally, we would like to express our gratitude to the members of the Organizing Committee of VLDB 2013, especially the Workshop Chairs Tiziana Catarci, AnHai Doan and Tova Milo, for their support in organizing this workshop; and to Craig Rodkin, Publications Operations Manager at the ACM, for his assistance in editing the proceedings.

7. REFERENCES

[1] A. Abello, J. Darmont, L. Etcheverry, M. Golfarelli, J.-N. Mazon, F. Naumann, T.-B. Pedersen, S. Rizzi, J. Trujillo, P. Vassiliadis, and G. Vossen. Fusion Cubes: Towards Self-Service Business Intelligence. *International Journal of Data Warehousing and Mining*, 9(2):66–88, April-June 2013.

[2] S. Alsubaiee, Y. Altowim, H. Altwaijry, A. Behm, V. R. Borkar, Y. Bu, M. J. Carey, R. Grover, Z. Heilbron, Y.-S. Kim, C. Li, N. Onose, P. Pirzadeh, R. Vernica, and J. Wen. ASTERIX: An Open Source System for &quot;Big Data&quot; Management and Analysis. *PVLDB*, 5(12):1898–1901, 2012.

[3] J. Darmont and T. B. Pedersen, editors. *2nd International Workshop on Cloud Intelligence (colocated with VLDB 2013), Cloud-I '13, Riva del Garda, Italy, August 26, 2013*. ACM, 2013.

[4] C. Phipps and K. C. Davis. Automating data warehouse conceptual schema design and evaluation. In *4th International Workshop on Design and Management of Data Warehouses (DMDW 2002), Toronto, Canada*, volume 58 of *CEUR Workshop Proceedings*, pages 23–32, 2002.