Report on the First International Workshop on Cloud Intelligence (Cloud-I 2012)

Jérôme Darmont
Université de Lyon (ERIC Lyon 2)
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

1. INTRODUCTION

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

With the increasing success of cloud computing, cloud business intelligence "as a service" offerings have arisen, 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, data marketplaces, security and privacy, performance, cost and usage models...), this trend also poses new, broader challenges for making data analytics available to small and medium-size enterprises (SMEs), non-governmental organizations, web communities (e.g., supported by social networks), and even the average citizen. This vision requires new integration and deployment models. For example, some deployments would benefit from an integrated database of private and open data.

Thus, Cloud Intelligence is not only a current technological and research challenge, but also an important societal stake, since people increasingly demand open data (e.g., the Spanish *indignados*), which they possibly mix with private data, and analyze with tools with advanced collaborative features, enabling users to share and re-use business intelligence concepts and analysis results world-wide.

The First International Workshop on Cloud Intelligence (Cloud-I 2012) [1] was held in conjunction with VLDB 2012 in Istanbul, Turkey on August 31, with the aim of becoming an interdisciplinary, regular exchange forum for researchers, industry and practitioners, as well as all potential users of Cloud Intelligence. This full-day event brought together researchers and engineers from academia and industry to discuss and exchange ideas related to BI and the cloud. The workshop featured one industrial keynote, three research sessions, and a panel.

The topics of the accepted papers spanned a number of exciting topics within Cloud Intelligence, including (in no particular order) RDF triple stores for the cloud, secure and private data sharing and analytics outsourcing in the cloud, MapReduce-based computations, and domain-specific cloud-based BI solutions.

2. INDUSTRIAL KEYNOTE

The keynote entitled "Analytic Lessons: in the Cloud, about the Cloud" was given by Dr. Morten Middelfart, CTO of TARGIT, Europe's largest pure-play developer of business intelligence products and a top 15 international vendor. With a single quote in mind: "The journey to courageous leadership, where organizations compete at a new level is: eliminate fear and trust computing as partner in a high-performance team.", Dr. Middelfart shared his experience about designing two different approaches to cloud-based deployment of analytics and business intelligence, namely an analyst specialist platform and a social platform. The analyst specialist platform helps model and share data, and has proven particularly useful in the analysis of large amounts of streaming unstructured data; aka Big Data. In the social networking approach, users can friend, share, analyze and discuss datasets. So far, the analyst platform has been the most popular. However, Dr. Middelfart finally discussed the trending behavior of the social platform and its current and potentially game-changing impact on the industry, as analytics shifts from being inside-out to embracing entire industries from the outside and in.

3. RESEARCH PAPERS

3.1 Session 1: Data and Knowledge Management

The position paper entitled "Towards a Hybrid Row-Column Database for a Cloud-based Medical Data Management System", by Baraa Mohamad, Laurent d'Orazio and Le Gruenwald, pinpoints the challenges in integrating high-volume, heterogeneous medical data in the form of DICOM files in the cloud. A novel hybrid "row-column", two-level database architecture is pro-

posed, where mandatory/frequently used attributes and attributes frequently accessed together are stored in a row-oriented database, and optional/private attributes are stored in a column-oriented database. This architecture is easy to use, extensible, efficient and allows ad-hoc queries over DICOM files, while benefiting from the elasticity, billing by use, and scalability of the cloud.

Yasin Silva, Jason Reed and Lisa Tsosie, in "MapReduce-based Similarity Join for Metric Spaces", study cloud-based similarity joins (a sparsely studied issue up till now). They propose a MapReduce-based algorithm called MR-SimJoin that efficiently partitions and distributes the data until the subsets are small enough to be processed in a single node. MR-SimJoin is general enough to be used with data that lies in any metric space, thus it can be used with multiple data types and distance functions. It is implemented in Hadoop and has good execution time and scalability properties.

Roshan Punnoose, Adina Crainiceanu and David Rapp propose "Rya: A Scalable RDF Triple Store For The Clouds". This scalable RDF data management system uses Accumulo, a Google Bigtable variant. Storage methods, indexing schemes and query processing techniques allow to scale to billions of triples across multiple nodes, while providing fast and easy access to the data through conventional query mechanisms such as SPARQL. Performance evaluations show that Rya outperforms existing distributed RDF solutions in most cases.

3.2 Session 2: Data Analytics

The position paper entitled "Integrity Verification of Cloud-hosted Data Analytics Computations", by Wendy Wang, introduces efficient and practical integrity verification techniques that check whether an untrusted cloud returns correct results of outsourced data analytics computations including a large class of machine learning and data mining methods. Verication techniques work for both non-collusive and collusive malicious workers in MapReduce.

Thanh Binh Nguyen, Fabian Wagner and Wolfgang Schöpp, in "Cloud Business Intelligent Services of well-established modeling tools to explore the synergies and interactions among climate change, air quality objectives", design a Cloud-based Business Intelligent Application Framework that includes a set of services grouped into Data warehousing Services and Business Intelligent Services. The former are used to specify the GAINS (Greenhouse Gas – Air Pollution Interactions and Synergies) data warehouse, while the latter help publish key data of scientific analysis in a transparent manner.

In their position paper "On Saying "Enough Already!" in MapReduce", Christos Doulkeridis and Kjetil Nørvåg criticize the brute force approach of MapReduce, which leads to performing redundant work, especially in the case of top-k queries. Different techniques that allow the efficient processing of top-k queries without exhaustively accessing input data are investigated. Various individual approaches and combinations of such approaches are proposed to provide the first steps towards integrating efficient top-k processing in MapReduce.

3.3 Session 3: Security and Privacy

Bharath Samanthula, Gerry Howser, Yousef Elmehdwi and Sanjay Madria, in the paper entitled "An Efficient and Secure Data Sharing Framework using Homomorphic Encryption in the Cloud", propose an efficient and Secure Data Sharing (SDS) framework using homomorphic encryption and proxy re-encryption schemes that prevents the leakage of unauthorized data when a revoked user rejoins the system. This framework is generic and secure under the security definition of Secure Multi-Party Computation (SMC). Any additive homomorphic encryption and proxy re-encryption scheme can be used. In addition, the underlying Secure Data Sharing (SDS) framework features a new solution based on data distribution to prevent information leakage in the case of collusion between users and Cloud Service Providers.

Mehdi Bentounsi, Salima Benbernou, Mikhail Atallah and Cheikh Deme present "Anonyfrag: An Anonymization-Based Approach For Privacy-Preserving BPaaS", which is an anonymization-based approach to preserve the client business activity while sharing process fragments between organizations on the cloud, i.e., when using on demand applications as Business Process as a Service through multi-tenant cloud platforms.

4. PANEL

Finally, Jérôme Darmont, Torben Bach Pedersen and Morten Middelfart launched a panel discussion themed "Cloud Intelligence: What is REALLY New?" to sort out what is new and not so new in cloud business intelligence as a service.

Torben Bach Pedersen defined three new things in cloud intelligence: elasticity, including the ability to bring in new data sources; a bottom-up, user-driven approach (in opposition to a top-down, enterprise-driven approach); and the fundamentally new economic model needed for cloud intelligence (pay-as-you-go instead of large prior investment).

Jérôme Darmont stressed out that security issues were even more critical in the cloud. Although some of these issues are inherited from classical distributed architecture, some directly relate to the new framework of the cloud, with privacy being of premium importance. Moreover, the social aspect of cloud intelligence involves sharing analysis results without necessarily disclosing source data.

Morten Middelfart eventually discussed the challenges about interpretation, bias, and completeness of external data gathered from the Web. Cloud intelligence presents an entirely new era of analytically founded strategic thinking, but on the other hand, it elevates the need for user understanding of the "truth behind the chart". A rich discussion ensued with the audience, the conclusions of which are included in the next section.

5. DISCUSSION AND OUTLOOK

The lessons that can be drawn from the workshop fall along several directions. First, when comparing the wide range of themes within cloud intelligence, e.g., as outlined in the call for papers and the panel discussions,

with the papers that actually appeared in the workshop, it is clear that the presented papers mainly focus on rather specific, mostly technical, issues. These are more precisely data management architectures and systems for cloud platforms, MapReduce-based algorithms for specific problems, and issues related to privacy, security, and integrity in the more technical sense. As a side note, the non-accepted papers also mainly fell in these areas. The only outlier to this pattern is the paper on cloud business intelligent services that mainly focus on the new user-oriented functionality enabled by cloud deployment. These issues were also covered in some of the panel statements.

Second, we can look at for which topics no papers appeared. One such issue is elasticity in the wider sense of the word. Another important "missing" set of topics relates to the social aspects of cloud intelligence, e.g., sharing results and new collaborative bottom-up approaches for building BI systems in the cloud. This leads to a demand for exploring new ways of using analytics enabled by the new opportunities in the cloud. However, these opportunities will only be used if the delivered results are backed up by work on truth and trust in the more intuitive sense of the word. Finally, new economic models for pay-as-you-go cloud intelligence will have to be developed. A long discussion on this topic concluded that typical web economic models like online ads and app stores were not well suited for this scenario. Micro-payment models had a better fit, but conflicted with the need for enterprises to above all have *predictable* costs.

We can thus conclude that there is a large demand for further research within cloud intelligence. As a first facilitating activity, the two best papers have been invited to submit extended versions to a special issue/section of *Information Systems* which also has an *open* call for papers[1]. We thus encourage the readers of SIGMOD Record to submit papers on cloud intelligence topics. Next, we hope that the workshop is just the first in a hopefully long series, and we certainly hope to hold the workshop again in 2013 and beyond.

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 2012, especially the Workshop Chairs James Joshi, Hakan Ferhatosmanoglu, and Andreas Wombacher, for their support in organizing this workshop.

[1] http://eric.univ-lyon2.fr/cloud-i/?p=269

7. REFERENCES

[1] J. Darmont and T. B. Pedersen, editors. *1st International Workshop on Cloud Intelligence (colocated with VLDB 2012), Cloud-I '12, Istanbul, Turkey, August 31, 2012*. ACM, 2012. http://dl.acm.org/citation.cfm?id=2347673.