Report on the Seventh International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE 2013)

Torben Bach Pedersen
Aalborg University
Denmark
tbp@cs.aau.dk

Malu Castellanos
HP Vertica
USA
malu.castellanos@hp.com

Umesh Dayal,
Hitachi Labs,
USA
umeshwar.dayal@hal.hitachi.com

1. INTRODUCTION

This paper reports on the 7th International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE 2013), co-located with the VLDB 2013 conference. The BIRTE workshop series aims at providing a forum for presentation of the latest research results, new technology developments, and new applications in the areas of business intelligence and real time enterprises. Building on the success of the previous BIRTE workshops, co-located with the VLDB conferences in Seoul, Auckland, Lyon, Singapore, Seattle, and Istanbul, the seventh workshop in the series was held in Riva del Garda, Italy, on August 26, 2013.

Today, business analytics have to use new data sources and technologies in order for the business to be completely up-to-date. Traditional “in-house” data sources about transactions, sales, and finances still form the cornerstone of business analytics applications, but this is no longer enough. Instead, “Big Data” with high *velocity* such as tweets and other social network updates and sensor data from RFID, GPS, Bluetooth, etc. must be captured and analyzed instantly to understand the latest customer and market trends. Further, analyzing the past and even the present is no longer enough, so predictive analytics solutions are used to make decisions based on the expected future. These new applications and data sources mean that existing business intelligence methods and techniques must be revisited to provide better efficiency, scalability, expressiveness, and ease-of-use.

BIRTE 2013 featured an exciting technical program including two keynotes, an invited industrial talk, a panel, and a number of peer-reviewed papers from different countries in Europe, Africa, and Asia. Each submission received three reviews from the members of the distinguished program committee consisting of leading researchers in the field from academia and industry. From these submissions, two full research papers and one short position paper, along with two demo papers, were selected for presentation at the conference. Based on the feedback of the reviewers and the feedback at the workshop, the authors have made revised versions of their papers which will be published in a joint post-proceedings volume of BIRTE 2013 and 2014 in the Springer LNBIB series [1]. BIRTE 2013 was extremely well attended, with a peak audience of over 70 persons.

2. KEYNOTES

After the welcome by the BIRTE 2013 chairs, the program started with a keynote by Michael J. Carey from UC Irvine, entitled “AsterixDB: A New Platform for Real-Time Big Data BI”. In this keynote, Prof. Carey explained the key ideas and principles behind the AsterixDB BDMS (Big Data Management System). AsterixDB has a number of features that sets it apart from other systems for managing Big Data. First, it has a unique flexible, semi-structured data model (Asterix Data Model) based on JSON. Second, is has a high-level declarative query language (AQL - Asterix Query Language) that can express a wide range of BI-like queries. Third, it has a highly scalable parallel runtime engine, Hyracks, that has been tested up to thousands of cores. Fourth, it supports new data intake very efficiently through its partitioned LSM-based data storage and indexing. Fifth, it has support for externally stored data (e.g., in HDFS) as well as natively managed data. Sixth, it features a rich set of primitive types, including spatial, temporal, and textual data types. Seventh, is has a range of secondary indexing options, including B+ tree, R tree, and inverted files. Eighth, is has support for fuzzy, spatial, and temporal queries as well as for parametric queries. Ninth, the notion of “datafeeds” supports continuous ingestion from relevant data sources. Finally, it has basic transactional capabilities like those of a NoSQL data store. Asterix is a system where 'one size fits a bunch'.

The second keynote, by Prof. Johann-Christoph Freytag from Humboldt Universität zu Berlin was entitled "Query Adaptation and Privacy for Real-Time Business Intelligence" and aimed at taking a holistic view of the challenges and issues that relate to real-time business intelligence systems, by discussing both technical and non-technical aspects. First, the keynote introduced a number of real-world applications and used these to derive technical and non-technical requirements for real-time business intelligence. Based on these requirements and the experience of Prof. Freytag in co-developing the Stratosphere database management system with other Berlin research groups, the talk described techniques for query adaptation and histogram building in Stratosphere to support real-time business intelligence. The second part of the keynote discussed important aspects of privacy when dealing with personal data. It then outlined the necessary requirements for implementing real-time business intelligence systems to protect privacy, and discussed the trade-off between the level of privacy and the utility expected by those who perform real-time business analytics.

3. RESEARCH PAPERS

The next session featured two full research papers and a position paper. The paper "LinkViews: An Integration Framework for Relational and Stream Systems"' by Yannis Sotiropoulos and Damianos Chatziantoniou from Athens University of Economics and Business, addresses the current lack of a unified framework for querying (persistent) relational and stream data. Concretely, the authors proposed a view layer defined over standard relational systems to handle the mismatch between relational and stream systems. Here, database administrators define a special type of views (called LinkViews) which combine relational data and stream aggregates. The authors showed how this could achieve transparent integration of relations and streams and how queries could be optimized. Next, the paper "OLAP for Multidimensional Semantic Web Databases" by Adriana Matei, Kuo-Ming Chao, and Nick Godwin from Coventry University, proposed a new framework for doing OLAP over Semantic Web data. The framework has multiple layers including additional vocabulary, extended OLAP operators, and the SPARSQL query language, allowing the modeling of heterogeneous semantic web data, the unification of multidimensional structures, and enabling interoperability between different semantic web multidimensional databases. Finally, the paper "A Multiple Query Optimization Scheme for Change Point Detection on Stream Processing System" by Masahiro Oke and Hideyuki Kawashima from University of Tsukuba, showed how to apply multiple query optimization, well-known from relational database technology, to change point detection (CPD) queries. The authors propose a two-stage learning approach based on autoregressive models and divide CPD into four operators. To accelerate multiple CPD executions -needed for parameter tuning- they use multi-query optimization (MQO). The authors showed how MQO enables sharing a large part of the CPD processing, leading to significantly improved performance.

4. DEMOS

As a novel addition to the BIRTE program, two demo papers were presented. First, the demo paper "Big Scale Text Analytics and Smart Content Navigation" by Karsten Schmidt, Philipp Scholl, and Sebastian Bächle from SAP AG, and Georg Nold from Springer Science and Business Media, showed how to use the SAP Hana platform for flexible text analysis, ad-hoc calculations and data linkage. The goal is to enhance the experience of users navigating and exploring publications, and thus to support intelligent guided research in big text collections. Case data from the major scientific publisher Springer SBM was used. Second, the demo paper "Dynamic Generation of Adaptive Real-time Dashboards for Continuous Data Stream Processing" by Timo Michelsen, Marco Grawunder, Dennis Geesen, and H.-Jürgen Appelrath from University of Oldenburg presented a novel dashboard concept for visualizing the results from continuous stream queries, based on several individually configurable dashboard parts, each connected to a (user defined) continuous query, the results of which are received and visualized in real-time.

5. INDUSTRIAL INVITED TALK

Dr. Morten Middelfart from TARGIT gave an inspiring invited industrial talk on "The Inverted Data Warehouse based on TARGIT Xbone - How the biggest of data can be mined by the 'little guy'." The talk presented TARGIT's Xbone memory-based analytics server and defined the concept of an Inverted Data Warehouse (IDW), a DW storing query results rather than raw data. The concept and system were exemplified with a large-scale solution in which TARGIT Xbone and IDW were applied on Google search data with the aim of Search Engine Optimization (SEO).

6. PANEL

The workshop ended with a panel on "Real Time Analytics on Big Data" moderated by Meichun Hsu from HP Labs. The panel featured six distinguished panelists: Alejandro Buchmann from TU Darmstadt, Shel Finkelstein from SAP, Johann-Christoph Freytag from Humboldt University of Berlin, C. Mohan from IBM, Ippokratis Pandis from IBM, and Torben Bach Pedersen from Aalborg University. The panelists gave short presentations on their perspectives on the general topic and their responses to the four questions posed by the moderator: *what does real time analytics on big data really mean?, what are the compelling applications that motivated such capabilities? what is the status of the technology stack that delivers this capability and what are the gaps and challenges? Relative to the technology attributes often used to characterize big data such as extreme scale-out, NoSQL, and open source, and the emerging technologies such as SQL-on-Hadoop and in-memory stores, how do real time analytics relate?* After the presentations a lively (and somewhat controversial) debate ensued between the panelists and the highly active audience.

7. DISCUSSION AND OUTLOOK

We now summarize the discussions and contributions in the panel and the workshop overall, structured according to the four questions from the panel.

What does it mean? The first observation is that "real-time" is used with two different meanings: real-time as in streaming versus real-time as in agile business real-time, i.e., minutes/hours versus days. From a user perspetive, what really matters is to get current info/knowledge from data, i.e., get changes in the real world reflected in data asap. This means increasing data freshness demands and low query response times, but does not necessarily mean continuous queries/streams. It also means automatic notifications and responses to business events. For business real-time, it should be easy to ask new questions on new data, e.g., as supported by the paper on OLAP on Semantic Web data, enabling agile OLAP on new data sources. Another paper addressed stream query optimization. One paper combined the two meanings, business real-time and streaming, by auto-generating dashboards for streaming data. As for the Big Data buzz, the database industry has always worked on "bigger", the new value lies instead in using un- and semi-structured data. Finally, it was argued that "real-time" should not only mean the past and current, but also the future, i.e., tightly integrating forecasting and prediction with database and stream queries.

Compelling applications? The compelling applications discussed included CRM, brand sentiment, predictive maintenance, network optimization, security, fraud detection, text analytics and smart content navigation, the last two in an SAP paper. Major issues are discovering trends early and finding outliers. The analytics applications should be optimized for people, not machines, e.g., possibilities for user feedback are missing. A new type of applications concern cyber-physical systems producing huge amounts of data and events. One type of cyber-physical system is the emerging smart grid. Here, demand/supply flexibilities and forecasts must be tightly integrated and managed, as data is "born" in long-term forecasts, later re-forecasted, and finally measured and captured, before they are used as a basis for further planning and optimization.

Status of technology stack and relation to technology attributes? The last two questions are treated together. On the one hand, we see that several "new" data management technologies are emerging in this field. Examples include the AsterixDB system with its semi-structured data model, SAP Hana, based on main memory and compressed storage and offering integrated analysis and transactions in real-time in a single system, the Bubblestorm "data rendezvous" system which is self-organizing -correcting, -optimzing, and the TimeTravel system based on hierarchical models allowing efficient integrated querying of past, present, and future data. On the other hand, we see a reverse trend that "new" technologies such as column and compressed storage, vector processing, multi-core, and cache-awareness are now integrated into classical systems providing order of magnitude performance leaps, e.g., as exemplified by the DB2 Blu system. The technology stack is also aiming at integrating historical and streaming data as exemplified by the LinkView paper.

Finally, if we look at the topics listed in the Call for Papers that were not (in one way or another) discussed in the workshop, they were analytics as a service, cloud intelligence, collaborative real-time BI, crowdsourcing and crowd intelligence, and data quality and cleansing. The first two relate to running analytics as cloud-based services, which will be very relevant in the future, and the preferred option for many enterprises. However, this option is not wide-spread yet, which explains the lack of contributions for these topics. The next two topics are related to the role of people in real-time analytics, either a small-scale collaboration by analysts or a large-scale "collaboration" of many people in a crowd. Again, this is not yet done by most enterprises, but will surely become more used in the coming years. Data quality and cleansing in the context of real-time analytics is a non-resolved topic, so papers on this topic will emerge.

In summary, we can conclude that business intelligence for the real-time enterprise is as relevant as ever, addressing in particular, the velocity aspect of Big Data and the new challenges imposed by this new trend. Thus, the outlook for BIRTE seems to be on the forward path with more editions planned for the future.

8. ACKNOWLEDGEMENTS

The BIRTE 2013 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 the distinguished PC members for their careful and dedicated work, both during the reviewing and the discussion phases. To the Organizing Committee of VLDB 2013, especially the General Chairs Themis Palpanas and Yannis Velegrakis, and the Workshop Chairs Tiziana Catarci, AnHai Doan, and Tova Milo, we would like to express our gratitude for their support to the BIRTE 2013 workshop. Finally, we would like to thank the BIRTE 2013 Proceedings Chair, Katja Hose and the web master Emmanouil Valsomatzis for their excellent job.

9. REFERENCES

[1] M. Castellanos, U. Dayal, K. Hose, T. B. Pedersen, and N. Tatbul (Eds.). Joint Proceedings of the 7th and 8th International Workshops on Business Intelligence for the Real Time Enterprise. Springer Lecture Notes in Business Information Processing, forthcoming.