

## Preface

In today's competitive and highly dynamic environment, analyzing data to understand how the business is performing, and to predict outcomes and trends have become critical. The traditional approach to reporting is no longer adequate. Instead users now demand easy-to-use intelligent platforms and applications capable of analyzing real-time data to provide insight and actionable information at the right time. The end goal is to support better and timelier decision making, enabled by the availability of up-to-date, high-quality information. Although there has been progress in this direction and many companies are introducing products toward meeting this goal, there is still a long way to go. In particular, the whole lifecycle of business intelligence requires innovative techniques and methodologies capable of dealing with the requirements imposed by these new generation BI applications. From the capture of real-time business data to the transformation and delivery of actionable information, all the stages of the Business Intelligence (BI) cycle call for new algorithms and paradigms to support value-added functionalities. These functionalities include dynamic integration of real-time data feeds from operational sources, optimization and evolution of ETL transformations and analytical models, and dynamic generation of adaptive real-time dashboards, just to name a few. In addition, the need to handle the 3 V's of Big Data, in particular "Velocity" of fast dynamic data streams, has boosted research efforts both in academia and in industry, leading to the emergence of new technologies and platforms for big data; and the expectation is that this trend will continue.

The BIRTE (Business Intelligence for the Real-Time Enterprise, which later became simply Real-Time Business Intelligence) 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 enterprise. Building on the success of its previous six editions, BIRTE continued the tradition of being colocated with the VLDB Conference. BIRTE 2013 was held in Riva del Garda, Italy on August 26, 2013 and BIRTE 2014 was held in Hangzhou, China on September 1, 2014.

Both workshops featured exciting technical programs including a total of three keynote speeches, three invited industrial talks, a panel, plus a number of peer-reviewed papers from different countries in USA, 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, a total of six full research papers and one short position paper, along with two demo papers, were selected for presentation. Based on the feedback of the reviewers and the feedback received during the workshops, the authors prepared revised versions of their papers. We are happy to present these contributions in this joint post-proceedings volume.

Both BIRTE 2013 and BIRTE 2014 were extremely well attended: the former with a peak audience of over 70 people, making it by far the most attended of the VLDB workshops held on August 26, 2013; and the latter breaking an attendance record in the

history of the BIRTE workshop series with about 80 participants during the keynote session. In what follows, we provide an overview of each workshop.

## **BIRTE 2013**

After the welcome by the chairs, the program started with a highly interesting 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, it 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, which 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, it has a range of secondary indexing options, including B+ tree, R tree, and inverted files. Eighth, it has support for fuzzy, spatial, and temporal queries as well as for parametric queries. Ninth, the notion of “data feeds” 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 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 model 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.

After lunch, the program continued with the second very interesting keynote, by Prof. Johann-Christoph Freytag from Humboldt-Universität zu Berlin. This keynote 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 that will be built into 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.

After the keynote, the 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 +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.

Next, 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).

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. Each panelist gave a short presentation on his perspective on the general topic and his responses to 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 we see real-time analytics relate? After the presentations a lively (and somewhat controversial) debate ensued between the panelists and the highly active audience.

## **BIRTE 2014**

The workshop opened with a session of accepted papers after a short introduction welcoming the participants. This session consisted of a position paper of co-authors from TU Dresden in Germany and the SAP Labs in USA, and a research paper of co-authors from University at Buffalo, SUNY, and the Oracle Corporation in USA. First, Michael Rudolf presented a flexible approach for multi-dimensional graph data analysis in their paper entitled “SynopSys: Foundations for Multidimensional Graph Analytics.” The key feature that distinguishes SynopSys from existing technologies, which require upfront modeling of analytical scenarios and are difficult to adapt to changes, is the ability to express ad-hoc analytical queries over graph data. The second paper, entitled “Detecting the Temporal Context of Queries” and presented by Ying Yang, focuses on the concept of contextual dependency – a term used by the authors to explain and attribute mistaken assumptions made by end users of BI applications. A formal definition for contextual dependence is given, followed by several strategies to efficiently detect and quantify the effects of contextual dependence on query outputs.

The next session was dedicated to the invited industrial talks. Inviting industrial speakers to present their perspective on real-world BI problems, solutions, and applications has been a tradition of BIRTE since its inception in 2006. This year’s workshop featured two industrial talks. First, in his talk entitled “Building Analytics Engines for the Big Data Age,” Dr. Badrish Chandramouli of Microsoft Research presented the challenges of a temporal streaming engine called Trill. Trill has been architected as a library to support embedded execution within cloud applications and distributed fabrics. Second, Dr. Qiming Chen of HP Labs gave a talk about “Optimistic Failure Recovery in Distributed Stream Processing.” More specifically, he presented the backtrack-based and the window-oriented recovery mechanisms in the Fontainebleau distributed stream analytics system built on top of the Storm platform. Both of these talks covered industry-scale stream processing applications and solutions, and demonstrated the importance of stream processing technology for real-time business intelligence and other big-velocity applications.

After lunch, the program continued with the keynote speech. This year’s keynote speaker was Dr. C. Mohan from the IBM Almaden Research Center. Dr. Mohan has been a well-known pioneer in database systems and has made numerous contributions to relational database research and technology in various different roles at IBM for more than 30 years. In his talk “Big Data: Hype and Reality,” he presented a concrete and detailed picture of the current landscape of big data systems. According to Mohan,

as users and developers gain a deeper understanding of the needs of real use cases (including real-time BI applications), the initial hype around big data systems (including noSQL, newSQL, and others) has been fading away. It is now becoming clearer that most of the so-called distinctive features of big data systems have in fact been well-known principles of relational database systems for decades. Mohan's comprehensive and critical survey of this popular field attracted much attention from a big audience and was very well received.

Finally, the last session of the workshop consisted of two paper presentations: an application paper jointly written by co-authors from Aalborg University in Denmark and Universite Libre de Bruxelles in Belgium, and a research paper from University of Southern Denmark. First, Dilshod Ibragimov explained during his talk, entitled "Towards Exploratory OLAP over Linked Open Data - A Case Study", how to integrate real-time data from web sources described in RDF into the analysis process in BI environments. To achieve this, a system that uses a multi-dimensional schema of the OLAP cube expressed in RDF vocabularies is proposed. The second presentation of this session was on "Efficient Pattern Detection over a Distributed Framework" by Ahmed Khan Leghari. In this talk, Leghari described an event stream partitioning scheme that partitions streams over time windows without considering any key attributes.

Overall, BIRTE 2014 was a great success. We have once again witnessed that real-time business intelligence continues to be a critical topic for both database researchers and practitioners. Talks covered a diverse set of real-world BI use cases, and indicated strong collaborations between academic and industrial community in this field. This year, we have observed that big data analytics has been a common theme for all BIRTE presentations, with a striking emphasis on the analysis of streaming and graph-structured data.

The BIRTE 2013 and BIRTE 2014 chairs would like to thank all the authors of submitted papers for their interest in the workshop and the high quality of their papers, and the distinguished PC members for their conscientious work, both during the reviewing and the discussion phases. Our workshop would not have been as successful as it was without the talks given by our invited speakers Prof. Michael J. Carey, Prof. Johann-Christoph Freytag, Dr. Morten Middelbart, Dr. C. Mohan, Dr. Badrish Chandramouli, and Dr. Qiming Chen, to whom we are deeply grateful. We would like to also express our gratitude to the Organizing Committees of VLDB 2013 and VLDB 2014, especially the General Chairs and the Workshop Chairs, for their support to BIRTE. Thanks also go to our Proceedings Chairs, Katja Hose (2013) and Jennie Duggan (2014), for their excellent job in putting this combined proceedings together, as well as to our webmaster, Emmanouil Valsomatzis (2013), for efficiently managing the workshop web site. Last but not least, we are grateful to the BIRTE 2014 session chairs, Prof. Damianos Chatziantoniou and Dr. Qiming Chen for their great support during the workshop in Hangzhou.

February 2015

Malu Castellanos  
Umeshwar Dayal  
Torben Bach Pedersen  
Nesime Tatbul

Enabling Real-Time Business Intelligence

International Workshops, BIRTE 2013, Riva del Garda, Italy, August 26, 2013, and BIRTE 2014, Hangzhou, China, September 1, 2014, Revised Selected Papers  
Castellanos, M.; Dayal, U.; Pedersen, T.B.; Tatbul, N. (Eds.)

2015, XX, 175 p. 80 illus., Softcover

ISBN: 978-3-662-46838-8