Lecture Notes
in Business Information Processing   337

Series Editors

Wil van der Aalst
RWTH Aachen University, Aachen, Germany
John Mylopoulos
University of Trento, Trento, Italy
Michael Rosemann
Queensland University of Technology, Brisbane, QLD, Australia
Michael J. Shaw
University of Illinois, Urbana-Champaign, IL, USA
Clemens Szyperski
Microsoft Research, Redmond, WA, USA

More information about this series at http://www.springer.com/series/7911

Malu Castellanos · Panos K. Chrysanthis ·
Konstantinos Pelechrinis (Eds.)

Real-Time Business Intelligence and Analytics

International Workshops
BIRTE 2015, Kohala Coast, HI, USA, August 31, 2015
BIRTE 2016, New Delhi, India, September 5, 2016
BIRTE 2017, Munich, Germany, August 28, 2017
Revised Selected Papers

Editors
Malu Castellanos
Teradata
Santa Clara, CA, USA

Panos K. Chrysanthis
University of Pittsburgh
Pittsburgh, PA, USA

Konstantinos Pelechrinis
University of Pittsburgh
Pittsburgh, PA, USA

Real-Time Business Intelligence and Analytics (BIRTE)

Preface

This LNBIP special volume on the BIRTE workshops serves as the proceedings of the last two editions of its first decade. Since its inception in 2006, BIRTE evolved its focus and its name to more accurately capture its scope but kept its acronym. This volume reflects its new name "Real-Time Business Intelligence & Analytics" and includes extended versions of selected papers presented in the 2015 and 2016 editions: The 9th International Workshop on Business Intelligence for the Real-Time Enterprise (BIRTE 2015) was held on August 31, 2015, in conjunction with the VLDB 2015 Conference that took place during September 1–4, 2015 in Kohala Coast, Hawaii. The 10th International Workshop on Enabling Real-Time Business Intelligence (BIRTE 2016) was held on September 4, 2016, in conjunction with the VLDB 2016 Conference that took place during September 5–9, 2016 in New Delhi, India.

Following the previous workshop editions, these two editions of the BIRTE workshop aimed at providing a forum for presentation of the latest research results, new technology developments, and new applications in the areas of business intelligence and analytics in the real-time enterprise. To ensure the quality of the program, each paper was reviewed by three or four reviewers who provided their feedback to authors. Authors of a conditionally accepted paper had to address the reviewers' comments and provide a summary of each improvement done to their paper. All authors had the opportunity to further improve their papers with the feedback received at the workshop.

The two BIRTE editions presented in this volume were as successful as their predecessors, with a high-quality program, evidenced by having full room in most sessions. Both had the traditional structure of a blend of research paper presentations, keynotes, and invited presentations from industry and academia as well as panel discussions.

During the first decade, the BIRTE workshops had post-proceedings published after the event by Springer as LNBIP volumes. Starting with the 11th International Workshop on Real-Time Business Intelligence and Analytics (BIRTE 2017), which took place on August 28, 2017, again in conjunction with the VLDB 2017 Conference, held on August 28–September 1, in Munich, Germany, the BIRTE workshop switched to regular proceedings, published at the time of the workshops (ACM ICPS 2017). This switch to regular proceedings reflected the feedback from the authors and participants during BIRTE 2016. However, keeping with the tradition, the presenters of BIRTE 2017 were invited to submit extended versions of their papers for this LNBIP special issue. The three selected papers, one from the keynote talk and two research papers, from BIRTE 2017 are included as an extension to the post-proceedings of BIRTE 2015 and 2016.

In the following we present a summary of the BIRTE 2015, BIRTE 2016, and BIRTE 2017 programs with emphasis on the papers included in this volume.

BIRTE 2015

BIRTE 2015 was organized in four sessions, of which the first session was devoted to the four accepted research papers. The other three sessions were the keynote talk on "Real-Time Analytics: The Third Time's a Charm?" by Michael Franklin, and the invited talks "Twitter Heron: Stream Processing at Scale" from Tweeter, "High-Availability at Massive Scale: Building Google's Data Infrastructure for Ads" from Google, and "Apache Flink: Scalable Stream and Batch Data Processing" from TU Berlin. The closing session included the panel on the topic "Can Big Data Platforms EVER Deliver Real-Time Analytics?" moderated by Guy Lohman. We briefly describe next the four research papers and the invited paper from Google, which appear in this volume.

The first paper titled "A Federated In-Memory Database System for Life Sciences" was from the Enterprise Platform and Integration Concepts group at Hasso Plattner Institute and Max Delrück Center. The authors present their hybrid cloud computing approach that integrates decentralized computing resources to form a federated in-memory database system that combines cloud computing with local storage and processing of highly sensitive data and huge amounts of data that would be too costly to transfer. The authors illustrate their approach on a real-world life sciences use case.

The second paper "An Integrated Architecture for Real-Time and Historical Analytics in Financial Services" was a short paper from Datometry Inc. The authors propose adaptive data virtualization (ADV) as an approach to satisfy the contradicting requirements of high-speed real-time analytics by some applications like analysis of ticker data, and scale by other applications like daily reporting. The authors illustrate how ADV breaks data silos by letting applications use different data management technologies without the need for database migrations or re-configuration of applications.

In the third paper, "Processing of Aggregate Continuous Queries in a Distributed Environment," the authors from the University of Pittsburgh explore how the sharing of partial aggregations can be implemented in the challenging environment of distributed data stream management systems (DSMSs) performing on-line analytics where the efficient execution of large numbers of aggregate continuous queries (ACQs) is crucial. They formulate the problem as an optimization that combines sharing of partial aggregations and assignment to servers to produce high-quality plans that keep the total cost of execution of ACQs low and balance the load among the computing nodes. The authors classify, implement, and compare optimizers and present their conclusions.

The fourth paper titled "Collecting and Analyzing User-Behavior Information in Instant Search Using Relational DBMS" is a short one from the University of California at Irvine where the authors study how to systematically collect information about user behaviors when they interact with an instant search engine, especially in a real-time environment. They present a solution, called RILCA, which uses front-end techniques, instead of methods based on traditional Web servers, to keep track of rich

information about user activities, a relational DBMS to store and analyze the log records efficiently, and a dashboard to monitor and analyze log records in real time.

In the invited industrial talk "High-Availability at Massive Scale: Building Google's Data Infrastructure for Ads" from Google, the authors described the evolution of the high availability strategies of the ads infrastructure from the first generation to today's approach based on multi-homed systems. These systems run hot in multiple datacenters all the time, and adaptively move load between datacenters, with the ability to handle outages of any scale transparently. The authors described their approaches for, and experiences with, availability and consistency in multi-homed data storage and processing systems. They shared details of the large-scale streaming systems Mesa and Photon and a new system, Ubiq, a highly-available multi-homed system, that scales to extremely high throughput by continuously processing events in small batches in near real-time.

BIRTE 2016

The 10th BIRTE 2016 was a special workshop edition, celebrating ten years of bridging academic and industrial innovation, as such it featured three keynote talks in addition to the traditionally invited talks from academia and industry. The overall program was organized into five sessions that included the keynotes "A Tale of Quest for Real-Time Business Intelligence" by Rakesh Agrawal (Data Insight Labs), "Data Exploration Challenges in the Age of Big Data" by Surajit Chaudhuri (Microsoft) and "Hybrid Transaction and Analytics Processing (HTAP): State of the Art" by C. Mohan (IBM); three invited talks and the closing panel which debated the key question, "The Singularity of Real-Time Analytics, Are We There Yet?" moderated by Panos K. Chrysanthis; and presentation of two research papers, and one research demo.

This volume includes extended versions of the two research papers and the demo paper, which covered the broad spectrum from the use of real-time analytics for recommendations (the PolyRecs system), to enabling real-time analytics over raw XML data, to scheduling complex analytical workflows executing over multi-engine environments (the IReS system). It also includes two papers of the invited talks from Harvard University on "Data Systems That Are Easy to Design, Tune and Use in Real-Time", and from Google on "Ubiq: A Scalable and Fault-tolerant Log Processing Infrastructure".

In their presentations, the authors of "PolyRecs: Improving Page-View Rates Using Real-Time Data Analysis" from the University of Athens, Greece, outlined their effort to enhance the pageview rates of e-content in order to achieve higher market penetration and gains in advertisements. They presented the design decisions underlying PolyRecs, which enable PolyRecs to be both effective on-the-fly calculations of what might be interesting to the browsing individuals at specific points in time as well as produce accurate results capable of improving the user-experience. They finally shared the experimental results of the real deployment of PolyRecs, pointing out that PolyRecs is a domain-agnostic hybrid-recommendation system that successfully functions regardless of the underlying data and/or content infrastructure.

In the second research paper, titled "Enabling Real Time Analytics over Raw XML Data," its authors from Microsoft Bing, IIT Bombay, and Flipkart pointed out that transforming data from a semi-structured format, such as XML, to a structured one is often time consuming and inappropriate for real-time analytics. To avoid losing, or rendering unactionable crucial, time-sensitive insights, they posed the question of how to expose analytical insights in the raw XML data. They proposed the idea of extracting complementary information (CI) from the raw semi-structured data repository for a given user query to enhance users' ability to better comprehend the original query response. CI examples include the context of the user query and the hidden patterns in the underlying XML data. Their solution is based on the node categorization model and a new ranking function for candidate entity nodes based on the underlying XML data structure. Crowd-sourced feedback on the CI discovered by their solution showed its ability to reveal useful CI in real-world XML data repositories with high precision and recall.

The demo presentation was on the Intelligent Resource Scheduler (IReS) from the National Technical University of Athens, Greece. The details of the resource allocation, optimizations and decision making of IReS are described in the developers' demo paper titled "Robust and Adaptive Multi-Engine Analytics Using IReS." During the demo, the attendees observed how to create, optimize, and execute workflows that match real use cases over multiple compute and data engines, imposing their preferred optimization objectives. Furtermore, the attendees had a chance to confirm the resilience and adaptability of IReS in the presence of failed nodes, unavailable engines, and load surges.

The paper "Data Systems That Are Easy to Design, Tune and Use in Real-Time" captures the invited talk by Stratos Idreos, from Harvard University. In his thought-provoking talk, the speaker discussed the vision of a future where businesses are driven by explosive amounts of data, where data becomes readily available and its power can be harnessed by everyone in real-time. He made the point that for this to happen, it is crucial that data systems be easy to design, tune, and use in real-time, and then described his group's efforts towards this vision. These efforts include the development of (1) adaptive data systems that can adjust to data and access patterns on-the-fly, (2) self-designing data systems that make it easy to spin-off and test new data system architectures in near real-time, and (3) curious data systems that make it easy to explore data in real-time, even if we do not know what queries to ask.

The second invited paper from Google presents "Ubiq: A Scalable and Fault-Tolerant Log Processing Infrastructure." This infrastructure has been in production for Google's advertising system for many years, and has served as a critical log processing framework for several dozen pipelines. During his talk, Manpreet Singh, Ubiq's lead developer, described Ubiq's architecture ability to continuously process log files in real time, while fully tolerating infrastructure degradation and data center-level outages without any manual intervention. He pointed out that Ubiq guarantees exactly-once semantics for application pipelines and provides an end-to-end latency of under a minute.

BIRTE 2017

The overall program of the 11th edition of BIRTE, BIRTE 2017, was organized into four sessions that included a keynote, an invited talk, five research paper presentations, an industrial paper presentation, and a closing panel. Extended versions of two of the research papers and the keynote paper were selected to be included in this volume. We briefly describe these three papers below.

The paper titled "Towards Interactive Data Exploration" corresponds to the keynote talk in the workshop under the same title. In this talk, Carsten Binning from TU Darmstadt and Brown University made the point that even though technology has been a key enabler of the ongoing big data trend toward datafication of almost every research field and industry with open-source tools like R and Hadoop and the advent of cheap, abundant computing and storage in the cloud, the current big data tool set is ill-suited for interactive data exploration. Consequently, the knowledge discovery process is a major bottleneck in our data-driven society. He gave an overview of challenges for interactive data exploration on large data sets and then presented current research results that revisit the design of existing data management systems, from the query interface to the underlying hardware, to enable interactive data exploration.

The paper titled "DCS: A Policy Framework for the Detection of Correlated Data Streams" is an extended version of the paper "Detection of Highly Correlated Live Data Streams" presented in the workshop. This is a product of a research collaboration. The authors from the Universities of Pittsburgh, USA, and of Queensland, Australia, proposed a solution, called DCS (Detection of Correlated Data Streams), which quickly identifies windows of highly correlated data streams and provides results in real-time. DCS achieves this by combining priority scheduling, pruning, and early termination to maximize the detection of correlated pairs within a micro-batch. DCS uses the Pearson correlation coefficient as a metric of correlation of two sliding windows of data streams and supports two modes of operation, "cold start" and "warm start." In the former mode, the analysis of a micro-batch starts with no prior knowledge of correlated pairs of streams and initializes the parameters of the priority scheduler *PriCe*'s utility function to its default values, whereas in the latter mode, the utility function is initialized based on the results of the latest micro-batch analysis for deeper exploration.

The second extended research paper reported the results of a collaboration among University of Toronto, Brown University, Intel Labs, and MIT. In their paper, titled "Towards Dynamic Data Placement for Polystore Ingestion," the authors proposed a streaming ETL architecture to support heterogeneous workloads with data ingestion and analytical queries, which are executed with strict performance guarantees. Their prototype system consists of a transactional streaming engine (S-Store), an OLAP back-end engine (Postgres), and a middleware (Big-DAWG) that controls the data migration and execution of the queries. Using this prototype, the authors studied the ingestion performance in terms of latency of various data placement and migration (Copy and Move) strategies between S-Store and OLAP engines under different mixed (read and write) ETL workloads.

February 2018   Malu Castellanos
                                                    Panos K. Chrysanthis

Organization

BIRTE 2015 and 2016

General Chair

Meichun Hsu   Hewlett-Packard, USA

Program Chairs

Malu Castellanos   Hewlett-Packard, USA
Panos K. Chrysanthis   University of Pittsburgh, USA

Program Committee (BIRTE 2015)

Christof Bornhoevd   RMS, USA
Alejandro Buchmann   Technische Universität Darmstadt, Germany
Badrish Chandramouli   Microsoft Research, USA
Shimin Chen   Chinese Academy of Sciences, China
Ben Chin Ooi   National University of Singapore, Singapore
Howard Ho   IBM, USA
Wolfgang Lehner   Dresden University of Technology, Germany
Fatma Ozcan   IBM Almaden, USA
Torben B. Pedersen   Aalborg University, Denmark
Kostantinos Pelechrinis   University of Pittsburgh, USA
Evaggelia Pitoura   University of Ioannina, Greece
Karthik Ramasamy   Twitter, USA
Elke Rundensteiner   Worcester Polytechnic Institute, USA
Mohamed Sharaf   Queensland University, Australia
Eric Simon   SAP-BO, France
Nesime Tatbul   Intel Labs and MIT, USA

Program Committee (BIRTE 2016)

Roger Barga   Amazon, USA
Christof Bornhoevd   RMS, USA
Alejandro Buchmann   Technische Universität Darmstadt, Germany
Shimin Chen   Chinese Academy of Sciences, China
Howard Ho   IBM, USA
Wolfgang Lehner   Dresden University of Technology, Germany
Olga Papaemmanouli   Brandeis University, USA
Evaggelia Pitoura   University of Ioannina, Greece
Krithi Ramamrithan   IIT Bombay, India
Karthik Ramasamy   Twitter, USA
Mohamed Sharaf   Queensland University, Australia

| Eric Simon | SAP-BO, France |
| Nesime Tatbul | Intel Labs and MIT, USA |

Sponsor and Student Travel Scholarships Chair

| Panos K. Chrysanthis | University of Pittsburgh, USA |

Proceedings Chair

| Kostantinos Pelechrinis | University of Pittsburgh, USA |

Webmaster

| Anatoli Shein | University of Pittsburgh, USA |

Sponsors

HP VERTICA
Google Inc.

BIRTE 2017

General and Program Chairs

| Malu Castellanos | Hewlett-Packard, USA |
| Panos K. Chrysanthis | University of Pittsburgh, USA |

Program Committee (BIRTE 2017)

Stefan Appel	Siemens, Germany
Alejandro Buchmann	Technische Universität Darmstadt, Germany
Shimin Chen	Chinese Academy of Sciences, China
Avrilia Floratou	Microsoft, USA
Ashish Gupta	Google, USA
Howard Ho	IBM Almaden, USA
Wolfgang Lehner	Dresden University of Technology, Germany
Marco Mellia	Politecnico di Torino, Italy
Themis Palpanas	Paris Descartes University, France
Panickos Neophytou	NetBeez, USA
Olga Papaemmanouil	Brandeis University, USA
Evaggelia Pitoura	University of Ioannina, Greece
Mohamed Sharaf	Queensland University, Australia
Nesime Tatbul	Intel Labs and MIT, USA

Sponsor and Student Travel Scholarships Chair

| Panos K. Chrysanthis | University of Pittsburgh, USA |

Proceedings Chair

| Damianos Chatziantoniou | Athens University of Economics and Business, Greece |

Webmaster

Anatoli Shein   University of Pittsburgh, USA

Sponsors

National Science Foundation
Google Inc.

Contents

BIRTE 2017