Editorial

Erhard Rahm*

Big Data Analytics

DOI 10.1515/itit-2016-0024

Big Data has become a core topic in different industries and research disciplines as well as for society as a whole. This is because the ability to generate, collect, distribute, process and analyze unprecedented amounts of diverse data has almost universal utility and helps to fundamentally change the way industries operate, how research can be done and how people live and use modern technology. Different industries such as automotive, finance, healthcare or manufacturing, can dramatically benefit from improved and faster data analysis, e.g., as illustrated by current industry trends like "Industry 4.0" and "Internet of Things". Data-driven research approaches utilizing Big Data technology and analysis become increasingly commonplace, e.g., in the life sciences, geo sciences or in astronomy. Users utilizing smartphones, social media, and web resources spend increasing amounts of time online, generate and consume enormous amounts of data and are the target for personalized services, recommendations and advertisements.

Most of the possible developments related to Big Data are still in an early stage but there is great promise if the diverse technological and application-specific challenges in managing und using Big Data are successfully addressed. Some of the technical challenges have been associated to different "V" characteristics, in particular Volume (support of very high data volumes), Velocity (fast analysis of data streams), Variety (support for diverse kinds of data) and Veracity (support for high data quality). Other challenges relate to the protection of personal and sensitive data to ensure a high a degree of privacy and the ability to turn the huge amount of data into useful insights or improved operation.

A key enabler for the Big Data movement are increasingly powerful and relatively inexpensive computing platforms allowing the fault-tolerant storage and processing of petabytes of data within large computing clusters typically equipped with thousands of processors and terabytes of main memory. The utilization of such infrastructures was pioneered by internet enterprises such as Google and Amazon but has become generally possible by open source system software such as the Hadoop ecosystem. Initially there have been only few core Hadoop components, in particular its distributed file system HDFS and the MapReduce framework for the relatively easy development and execution of highly parallel applications to process massive amounts of data on cluster infrastructures. The initial Hadoop has been highly successful but also reached its limits in different areas, e.g., to support the processing of fast changing data such as data streams or to process highly iterative algorithms, e.g. for machine learning or graph processing. These aspects have led to a large number of additional components within the Hadoop ecosystem, both general-purpose processing frameworks such as Apache Spark and Flink as well as specific components, e.g., for data streams, graph data or machine learning. Furthermore, there are now numerous approaches to combine Hadoop-like data processing with relational database processing ("SQL on Hadoop").

The wide spectrum of Big Data related challenges is being addressed worldwide in research, development and diverse applications. This is also the case for Germany with numerous Big Data projects and initiatives. Since 2014, the Federal Ministry of Education and Research (BMBF) is funding two competence centers for Big Data: the Berlin Big Data Center (BBDC) and the Competence Center on Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig. For this special issue we have invited contributions from these two centers as well as from other institutes with Big Data projects on current topics. After a careful reviewing by several experts and revision of the papers, we have finally accepted six contributions for this special issue on "Big Data Analytics".

The first three papers come from the two Big Data competence centers BBDC and ScaDS Dresden/Leipzig. The paper "Apache Flink in Current Research Projects" by T. Rabl, J. Traub and V. Markl from TU Berlin outlines the functionality of the general-purpose Big Data framework Apache Flink and its use in current research projects and applications. The broad spectrum of use cases shows the ver-

***Corresponding author: Erhard Rahm,** Universität Leipzig, Institut für Informatik, Augustusplatz 10, 04109 Leipzig, Germany, e-mail: rahm@informatik.uni-leipzig.de

satility of this platform and its specific strengths such as support for data stream processing.

A. Petermann and M. Junghanns from the University of Leipzig focus in their article "Scalable Business Intelligence with Graph Collections" on the use of graph analytics for business intelligence. They compare different graph processing approaches and propose the use of a new Hadoop-based graph processing system called Gradoop for enhanced and scalable business intelligence. The approach supports the analysis and mining of collections of graphs, e.g., representing business transactions.

M. Hahmann and colleagues from TU Dresden advocate in their paper "Big by Blocks: Modular Analytics" for a modular specification, implementation and optimization of analysis and data mining techniques. They observe that analysis techniques consist of several phases with different implementation alternatives. These alternatives can be provided as building blocks to compose different variations within a modular and flexible analysis framework. The approach is described for clustering and forecasting algorithms.

The last three papers focus on the efficient processing of data streams which poses many new challenges and has wide applicability, e.g., to analyze messages in social networks, click streams or stock changes. W. Wingerath and colleagues from the University of Hamburg provide an overview about current stream processing systems in their article "Real-Time Stream Processing for Big Data". They focus on Apache Storm/Trident, Spark Streaming and LinkedIn's Samza but also discuss further systems.

The paper "Stream Processing Platforms for Analyzing Big Dynamic Data" by S. Hagedorn and colleagues from TU Ilmenau also surveys popular systems for analyzing data streams (Storm, Spark and Flink Streaming) with a focus on their APIs and user perspective. The authors further present an extension of the PigLatin dataflow language (called Piglet) for stream processing. The Piglet scripts can be executed on different streaming platforms.

The last paper "Clustering Big Data Streams: Recent Challenges and Contributions" by M. Hassani and T. Seidl from RWTH Aachen and LMU Munich provides an overview about the challenges and recent techniques for clustering high-dimensional data streams. The new approaches address-ing the introduced challenges fall into different categories, in particular density-based, anytime and subspace stream clustering and combined techniques. Subspace stream clustering is supported within an open-source framework developed by the authors.

Bionotes

Prof. Dr. Erhard Rahm
Universität Leipzig, Institut für Informatik,
Augustusplatz 10, 04109 Leipzig, Germany
rahm@informatik.uni-leipzig.de

Prof. Dr. Erhard Rahm studied Computer Science at the University of Kaiserslautern (Diplom 1984, Promotion 1988, Habilitation 1993). From 1988 to 1989 he was a post-doctoral fellow at the IBM Research Center in Hawthorne, NY. Since 1994 he is a Full Professor for Databases at the University of Leipzig. He spent sabbaticals at Microsoft Research (Redmond, WA) and the Australian National University. His current research focusses on Big Data and data integration. He has authored several books and more than 200 peer-reviewed journal and conference publications. His research on data integration has been awarded several times, in particular with the renowned 10-year best-paper award of the conference series VLDB (Very Large Databases) and the Influential Paper Award of the conference series ICDE (Int. Conf. on Data Engineering). Prof. Rahm is one of the two scientific coordinators of the German Big Data competence center ScaDS Dresden/Leipzig.