SpringerBriefs in Computer Science

More information about this series at http://www.springer.com/series/10028

Sherif Sakr

Big Data 2.0 Processing Systems

A Survey

Springer

Sherif Sakr
University of New South Wales
Sydney, NSW
Australia

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-319-38775-8 ISBN 978-3-319-38776-5 (eBook)
DOI 10.1007/978-3-319-38776-5

Library of Congress Control Number: 2016941097

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To my wife, Radwa,
my daughter, Jana,
and my son, Shehab
for their love, encouragement,
and support.

Sherif Sakr

Foreword

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 change fundamentally 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, for example, as illustrated by current industry trends such as "Industry 4.0" and "Internet-of-Things." Data-driven research approaches utilizing Big Data technology and analysis have become increasingly commonplace, for example, in the life sciences, geosciences, 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 and using Big Data are successfully addressed. Some of the technical challenges have been associated with different "V" characteristics, in particular Volume, Velocity, Variety, and Veracity that are also discussed in this book. Other challenges relate to the protection of personal and sensitive data to ensure a high 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 is the increasingly powerful and relatively inexpensive computing platforms allowing 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 giants 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 a 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, for example, to support the processing of fast changing data such as datastreams or to process highly iterative algorithms, for example, for machine learning or graph processing. Furthermore, the Hadoop world has been largely decoupled from the widespread data management and analysis approaches based on relational databases and SQL. 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, such as 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 net effect of all these developments is that the current technological landscape for Big Data is not yet consolidated but there are many possible approaches within the Hadoop ecosystem and also within the product portfolio of different database vendors and other IT companies (Google, IBM, Microsoft, Oracle, etc.). The book *Big Data 2.0 Processing Systems* by Sherif Sakr is a valuable and up-to-date guide through this technological "jungle" and provides the reader with a comprehensible and concise overview of the main developments after the initial MapReduce-focused version of Hadoop. I am confident that this information is useful for many practitioners, scientists, and students interested in Big Data technology.

University of Leipzig, Germany Erhard Rahm

Preface

We live in an age of so-called Big Data. The radical expansion and integration of computation, networking, digital devices, and data storage have provided a robust platform for the explosion in Big Data as well as being the means by which Big Data are generated, processed, shared, and analyzed. In the field of computer science, data are considered as the main raw material which is produced by abstracting the world into categories, measures, and other representational forms (e.g., characters, numbers, relations, sounds, images, electronic waves) that constitute the building blocks from which information and knowledge are created. Big Data has commonly been characterized by the defining 3V properties which refer to huge in volume, consisting of terabytes or petabytes of data; high in velocity, being created in or near realtime; and diversity in variety of type, being both structured and unstructured in nature . According to IBM, we are currently creating 2.5 quintillion bytes of data every day. IDC predicts that the worldwide volume of data will reach 40 zettabytes by 2020 where 85 % of all of these data will be of new datatypes and formats including server logs and other machine-generated data, data from sensors, social media data, and many other data sources. This new scale of Big Data has been attracting a lot of interest from both the research and industrial communities with the aim of creating the best means to process and analyze these data in order to make the best use of them. For about a decade, the Hadoop framework has dominated the world of Big Data processing, however, in recent years, academia and industry have started to recognize the limitations of the Hadoop framework in several application domains and Big Data processing scenarios such as large-scale processing of structured data, graph data, and streaming data. Thus, the Hadoop framework has been slowly replaced by a collection of engines dedicated to specific verticals (e.g., structured data, graph data, streaming data). In this book, we cover this new wave of systems referring to them as Big Data 2.0 processing systems.

This book provides the big picture and a comprehensive survey for the domain of Big Data processing systems. The book is not focused only on one research area or one type of data. However, it discusses various aspects of research and development of Big Data systems. It also has a balanced descriptive and analytical content. It has information on advanced Big Data research and also which parts

of the research can benefit from further investigation. The book starts by introducing the general background of the Big Data phenomenon. We then provide an overview of various general-purpose Big Data processing systems that empower the user to develop various Big Data processing jobs for different application domains. We next examine the several vertical domains of Big Data processing systems: structured data, graph data, and stream data. The book concludes with a discussion of some of the open problems and future research directions.

We hope this monograph will be a useful reference for students, researchers, and professionals in the domain of Big Data processing systems. We also wish that the comprehensive reading materials of the book may influence readers to think further and investigate the areas that are novel to them.

To Students: We hope that the book provides you with an enjoyable introduction to the field of Big Data processing systems. We have attempted to classify properly the state of the art and describe technical problems and techniques/methods in depth. The book provides you with a comprehensive list of potential research topics. You can use this book as a fundamental starting point for your literature survey.

To Researchers: The material of this book provides you with thorough coverage for the emerging and ongoing advancements of Big Data processing systems that are being designed to deal with specific verticals in addition to the general-purpose ones. You can use the chapters that are related to certain research interests as a solid literature survey. You also can use this book as a starting point for other research topics.

To Professionals and Practitioners: You will find this book useful as it provides a review of the state of the art for Big Data processing systems. The wide range of systems and techniques covered in this book makes it an excellent handbook on Big Data analytics systems. Most of the problems and systems that we discuss in each chapter have great practical utility in various application domains. The reader can immediately put the gained knowledge from this book into practice due to the open-source availability of the majority of the Big Data processing systems.

Sydney, Australia Sherif Sakr

Acknowledgements

I am grateful to many of my collaborators for their contribution to this book. In particular, I would like to mention Fuad Bajaber, Ahmed Barnawi, Omar Batarfi, Seyed-Reza Beheshti, Radwa Elshawi, Ayman Fayoumi, Anna Liu, and Reza Nouri. Thank you all!

Thanks to Springer-Verlag for publishing this book. Ralf Gerstner encouraged and supported me to write this book. Thanks, Ralf!

My acknowledgments end with thanking the people most precious to me. Thanks for my parents for their encouragement and support. Many thanks for my daughter, Jana, and my son, Shehab, for the happiness and enjoyable moments they are always bringing to my life. My most special appreciation goes to my wife, Radwa Elshawi, for her everlasting support and deep love.

Sherif Sakr

Contents

About the Author

Sherif Sakr is a professor of computer and information science in the Health Informatics department at King Saud bin Abdulaziz University for Health Sciences. He is also affiliated with the University of New South Wales and DATA61/CSIRO (formerly NICTA). He received his Ph.D. degree in Computer and Information Science from Konstanz University, Germany in 2007. He received his BSc and M. Sc. degrees in Computer Science from the Information Systems department at the Faculty of Computers and Information in Cairo University, Egypt, in 2000 and 2003, respectively. In 2008 and 2009, Sherif held an Adjunct Lecturer position at the Department of Computing of Macquarie University. In 2011, he held a Visiting Researcher position at the eXtreme Computing Group, Microsoft Research Laboratories, Redmond, WA, USA. In 2012, he held a Research MTS position in Alcatel-Lucent Bell Labs. In 2013, Sherif was awarded the Stanford Innovation and Entrepreneurship Certificate. Sherif has published more than 90 refereed research publications in international journals and conferences, (co-) authored three books and co-edited three other books. He is an IEEE Senior Member.