Large-Scale Data Analytics

Aris Gkoulalas-Divanis • Abderrahim Labbi
Editors

Large-Scale Data Analytics

Springer

Editors
Aris Gkoulalas-Divanis
IBM Research – Ireland
Damastown Industrial Estate
Mulhuddart, Ireland

Abderrahim Labbi
IBM Research – Zurich
Rüschlikon, Switzerland

ISBN 978-1-4614-9241-2 ISBN 978-1-4614-9242-9 (eBook)
DOI 10.1007/978-1-4614-9242-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013954541

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In recent years, we are witnessing a data explosion: almost 90 % of today's data have been produced only in the last 2 years, with data being nowadays produced in the order of Zettabytes! This data comes from various sources, including sensors, social networking sites, mobile phone applications, electornic medical record systems and e-commerce sites, just to name a few. Apart from its massive volume, this data is also characterized by variety (heterogeneity) and velocity (streams of data).

Traditional approaches and algorithms are not able to process and analyze such massive and complex datasets. This has signified the need for a paradigm shift, where new hardware and software technology is emerging to efficiently and reliably manage, store, process, analyze and synthesize very large amounts of complex data generated by massively distributed data sources. Beside their massively distributed nature, which requires new distributed architectures for data analysis, the heterogeneity of such sources imposes significant challenges for the efficient analysis of the data under numerous constraints, such as consistent data integration, data homogenization and scaling, privacy and security preservation. Moreover, the emerging real-world applications in domains such as healthcare, weather forecasting, financial engineering, urban planning, traffic management and environmental monitoring impose extra requirements for large-scale data analysis.

This edited book contains contributions on cutting edge research related to large-scale data analytics in the following core areas: databases, data mining, supercomputing, data visualization and privacy. Our goal is to present to students, researchers, professionals and practitioners the state-of-the-art research, which will help shape up the future of large-scale analytics, leading the way to the design of new approaches and technologies that can analyze and synthesize very large amounts of heterogeneous data, generated by massively distributed data sources.

Each chapter of the book presents a survey of an area in large-scale data analytics, or individual results of the emerging research in the field. Chapters 1 and 2 are devoted to the MapReduce framework. In particular, the first chapter provides a comprehensive survey for a family of approaches and mechanisms of large scale data analysis that have been implemented based on the MapReduce framework. Chapter 2 focuses on optimization approaches for plain MapReduce

jobs, as well as for parallel data flow systems. Chapters 3 and 4 present two important application areas of the MapReduce framework: mining tera-scale graphs for patterns and anomalies (Chap. 3), and analyzing customer behavioral data for the Telecom industry (Chap. 4). In Chap. 5, the authors describe a unified heterogeneous architecture that integrates massively threaded shared-memory multiprocessors into MapReduce-based clusters to enable executing Map and Reduce operators on thousands of threads, across multiple GPU devices and nodes. The proposed hybrid system can be used to accelerate machine learning algorithms, such as support vector machines, achieving significant speedup. Chapter 6 is devoted to large-scale social network analysis, offering a comprehensive survey of the state-of-the-art in this area, with focus on parallel algorithms and libraries for the computation of network centrality metrics. An overview of data visualization methods that help users to gain insight into large, heterogeneous, dynamic textual datasets is provided in Chap. 7. The last chapter of the book is devoted to technologies for offering security and privacy at large scale. The authors of this chapter present a novel framework for privacy-preserving, distributed data analysis that is practical for many real-world applications.

We, as editors, are genuinely grateful to all contributors of this book for the time and effort they put into this project, despite the heavy burden that we put on them. We also owe special thanks to the effort of the external reviewers for their help in this effort. Last but not least, we are indepted to Susan Lagerstrom-Fife and Courtney Clark from Springer, for their great support towards the preparation and completion of this work. Their editing suggestions were valuable to improving the organization, readability and appearance of the manuscript.

Mulhuddart, Ireland — Aris Gkoulalas-Divanis
Rüschlikon, Switzerland — Abderrahim Labbi

Contents

Contributors

John Canny Computer Science Division, University of California, Berkeley, CA, USA

Yitao Duan NetEase Youdao, Beijing, China

Christos Faloutsos School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Michael Granitzer University of Passau, Passau, Germany

Sergio Herrero-Lopez Technologies, Equities and Currency (TEC) Division, SwissQuant Group AG, Zurich, Switzerland

Fabian Hueske Technische Universität Berlin, Berlin, Germany

U Kang Department of Computer Science, KAIST University, Republic of Korea

Wolfgang Kienreich Know-Center Graz, Graz, Austria

David Konopnicki IBM Haifa Research Lab, Haifa, Israel

Mattia Lambertini Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Elisabeth Lex Know-Center Graz, Graz, Austria

Anna Liu NICTA and University of New South Wales, Sydney, NSW, Australia

Matteo Magnani Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden

Volker Markl Technische Universität Berlin, Berlin, Germany

Moreno Marzolla Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Danilo Montesi Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Carmine Paolino Department of Computer Science, Vrije Universiteit, Amsterdam, The Netherlands

Vedran Sabol Know-Center Graz, Graz, Austria

Sherif Sakr NICTA and University of New South Wales, Sydney, NSW, Australia

Christin Seifert University of Passau, Passau, Germany

Michal Shmueli-Scheuer IBM Haifa Research Lab, Haifa, Israel

John R. Williams Massachusetts Institute of Technology, Cambridge, MA, USA

List of Figures

List of Tables

Acronyms

AHP	Analytical hierarchical process
APA	Austrian Press Agency
API	Application Programmer Interface
APP	Mobile application
AQL	Asterix query language
BFS	Breadth-First Search
BGL	Boost Graph Library
BOW	Bag of Words
C2S	Client to server
CC	Connected components
CDR	Call data record
CMV	Coordinated multiple views
DAG	Directed acyclic Graph
DBMS	Database Management System
DC	Disconnected component
DDL	Data definition language
DFS	Depth-First Search/Distributed File System
DML	Data manipulation language
EC2	Amazon Elastic Compute Cloud
ECC	Elliptic curve cryptography
EDR	Event data/detail record
EFF	Electronic Frontier Foundation
EM	Expectation maximization
FIFO	First In First Out
GCC	Giant connected component
GFS	Google File System
GIM-V	Generalized iterative matrix-vector multiplication
GPU	Graphics processing unit
HDFS	Hadoop Distributed File System
IAAS	Infrastructure As A Service
IE	Information extraction

IRAM	Implicitly Restarted Arnoldi Method
JSON	JavaScript object notation
KD	Knowledge discovery
LAN	Local area network
LINQ	Language INtegrated Query
LSI	Latent semantic indexing
MDS	Multidimensional scaling
MPC	Secure multi-party computation
MPI	Message Passing Interface
MPP	Massively parallel processing/processor
MPPN	Massively parallel processing node
MRF	MapReduce framework
MSA	Massive scale analytics
MST	Minimum spanning tree
MTGL	Multi-Threaded Graph Library
NAS	Network attached storage
NUMA	Non-uniform memory access
ODP	Open Directory Project
PACT	Parallelization contract
PBGL	Parallel Boost Graph Library
PCA	Principal component analysis
POS	Part of speech
PPDM	Privacy preserving data mining
PRG	Pseudo-random number generator
QP	Quadratic programming
RAID	Redundant array of independent/inexpensive disks
RAIN	Redundant array of independent/inexpensive nodes
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RRS	Recursive random search
RWR	Random Walk with Restart
S2S	Server to Server
SAN	Storage area network
SCC	Strongly connected components
SCOPE	Structured Computations Optimized for Parallel Execution
SMO	Sequential minimal optimization
SMP	Symmetric multi-processing machine
SNA	Social network analysis
SNAP	Small-world Network Analysis and Partitioning
SNS	Social network site
SQL	Structured query language
SSSP	Single source shortest path
SVD	Singular value decomposition
SVM	Support vector machine
TCP	Transmission Control Protocol

UDF	User-defined function
UMA	Uniform memory access
URL	Uniform resource locator
VSS	Verifiable secret sharing
ZKP	Zero-knowledge proof