GUEST EDITORS' INTRODUCTION

Big Data: Promises and Problems

Venkat N. Gudivada, Marshall University

Ricardo Baeza-Yates, Yahoo Labs

Vijay V. Raghavan, University of Louisiana

Despite some key problems, big data could fundamentally change scientific research methodology and how businesses develop products and provide services.

0018-9162/15/$31.00 © 2015 IEEE

Smartphones with megapixel cameras, handheld computers, wireless sensor networks, ubiquitous social media, earth-orbiting satellites, space-bound telescopes—all generating more data than ever before: it is no exaggeration to say that over 90 percent of the world's data was produced in just the past two years. And this growth is on track to continue accelerating as we shift toward gigabit networks, gigapixel cameras, and a data-intensive Internet of Things (IoT).

NASA's Solar Dynamics Observatory uses four telescopes that gather eight images of the Sun every 12 seconds. In January 2015, the SDO captured its 100 millionth image of the Sun—just one example of the ways in which astronomers are collecting more and more data. Currently one petabyte of this data is publicly accessible online, and this volume grows at a rate of 0.5 petabytes per year. In CERN's Large Hadron Collider, 150 million sensors are capturing data about nearly 600 million collisions per second. On a similar scale, the work that won the 2013 Nobel Prize in chemistry involved measuring and visualizing the behavior of 50,000 or more atoms in a reaction over the course of a fraction of a millisecond.

In the social media domain, Facebook users add 300 million new photos a day; over 300 million Instagram users share 60 million photos every day; and more than 100 hours of video are uploaded to YouTube every minute.

Although business and social media data are rarely looked at more than once, much less analyzed in detail, we expect this to change. In 2013, only 22 percent of data was considered useful, and less than 5 percent of that amount was actually analyzed. By 2020, more than 35 percent of all data could be considered useful due to increased production from sensors and IoT devices, and because it is increasingly engineered to meet specific goals, such as scientific discovery or process optimization. For example, IoT data from giant gas turbines that generate electricity has tremendous value since this can optimize power generation and assist with maintenance and repair. Likewise, the Square Kilometre Array (SKA) radio telescope project, expected to be operational in 2020, will produce 2.8 Gbytes of astronomy data per second that will help create the biggest map of the universe ever made.

> **CURRENTLY ONE PETABYTE OF THIS DATA IS PUBLICLY ACCESSIBLE ONLINE, AND THIS VOLUME GROWS AT A RATE OF 0.5 PETABYTES PER YEAR.**

Big data, defined as data too large and complex to capture, process, and analyze using current computing infrastructure, is now popularly characterized by five V's (initially it was described as having three, but two have since been added to emphasize the need for data authenticity and business value):

- volume—data measurement is in terabytes (2^{40}) or even petabytes (2^{50}), and is rapidly heading toward exabytes (2^{60});
- velocity—data production occurs at very high rates, and, because of this sheer volume, some applications require real-time data processing to determine whether to store a piece of data;
- variety—data is heterogeneous and can be highly structured, semi-structured, or totally unstructured;
- veracity—due to intermediary processing, diversity among data sources and in data evolution raises concerns about security, privacy, trust, and accountability, creating a need to verify secure data provenance; and
- value—through predictive models that answer what-if queries, analysis of this data can yield counterintuitive insights and actionable intelligence.

Big data enables new directions for scientific research once limited by the volume of available data. For example, many natural language– and speech-related problems are ill suited for mathematically precise algorithmic solutions. To better address this problem, statistical machine-learning models—which require training data to build and evaluate—are often used.[1] For example, for the part-of-speech tagging problem, training data consists of several sentences and part-of-speech annotations for each word in the sentences. By harnessing the power of big

data, we have shifted the paradigm for solving these problems; now the accurate selection of a mathematical model loses its importance because there is *big enough data* to compensate.[2]

BIG DATA ENTRY

The ability to effectively process massive datasets has become integral to a broad range of scientific and other academic disciplines. However, this does not obviate the need for a deep understanding of a domain's theoretical foundations. Big data enables scientists to overcome problems associated with small data samples in ways that include relaxing theoretical model assumptions, avoiding overfitting models to training data, better handling noisy training data, and providing ample test data to validate models.

But big data ushers in several challenges, including how to capture, transfer, store, clean, analyze, filter, search, share, secure, and visualize data. Consider the problem of storing and retrieving big data. An array of new systems has emerged in recent years to address these kinds of big data challenges. Currently over 220 such systems fall under the NoSQL umbrella, with new ones emerging regularly (http://db-engines.com/en/). Hence, we face limited or missing theoretical bases for data models and query languages, and we lack clear standards that would help to avoid vendor lock-in.

Many NoSQL systems are designed for deployment on distributed-cluster computers, offer choices for data consistency level specification, provide built-in support for parallel processing using the MapReduce framework, and feature an assortment of application programming interfaces.[3] However, not all big data problems are amenable to MapReduce solutions. For example, time-varying graphs and dynamic networks, real-time processing requirements, and scalable stream data processing pose additional challenges.

Big data problems require making several tradeoffs among desired scalability, availability, performance, and security. For some problems, precise solutions are intractable, and may require faster and approximated algorithms that run the risk of decreasing the quality of the solution.

Of course, a bit of caution is imperative when dealing with a trendy topic. Most problems do not need big data: they need the right data.[4] Often data can be conflicting, incomplete, imprecise, subjective, redundant, biased, and noisy. Such data has the potential to create confusion and misinformation rather than provide actionable insights. Indeed, we need to avoid the temptation of following a data-driven approach instead of a problem-driven one. As described in "Big Data or Right Data?,"[4] we need to ask the right kind of questions:

- How do we process, filter, and sample the source data to obtain the right data?
- How do we determine the trustworthiness of such data?
- How much noise is there?
- How do we distinguish between valid data and spam? Filtering spam is a nontrivial problem and a possible bias source for any data.
- Is the data distribution valid, or is there a hidden bias that needs to be corrected? How do we correct bias?
- How do we determine and eliminate duplicates?

Privacy is also relevant, as it deals with legal and ethical restrictions. Which privacy issues must be taken care of? Do we need to anonymize the data? This is a concern because data provenance tracking is a major requirement in many big data applications: provenance information is used for data transformations, enabling auditing, modeling authenticity, implementing access control for derived data, and evaluating the quality of and trust in data.

IN THIS ISSUE

With all of its promise, big data clearly presents problems for researchers. In this special issue, five feature articles provide solutions to some of these challenges.

In the first, "In-Memory Graph Databases for Web-Scale Data," Vito Giovanni Castellana, Alessandro Morari, Jesse Weaver, Antonino Tumeo, David Haglin, Oreste Villa, and John Feo present a software framework named Graph database Engine for Multithreaded Systems (GEMS). Because resource description framework (RDF) databases have emerged as a preferred solution for organizing, integrating, and managing very large, heterogeneous, and loosely structured data that is prevalent in a variety of scientific and commercial domains, the authors describe how the GEMS framework successfully implements RDF databases on commodity, distributed-memory, high-performance clusters. GEMS is designed from the ground up to natively implement graph-based methods.

In “Taming Replication Latency of Big Data Events with Capacity Planning,” Zhenyun Zhuang, Haricharan Ramachandra, and Chaoyue Xiong discuss how they have addressed the challenge of minimizing latency in replicating database events. Based on their observations of LinkedIn’s production traffic and various systems’ moving parts, they developed a model to predict incoming traffic rates, reduce replication latency, and answer a set of business-critical questions related to capacity planning.

In “Integrating Big Data: A Semantic Extract-Transform-Load Framework,” Srividiya Bansal and Sebastian Kagemann present a semantic extract-transform-load (ETL) framework, which uses semantic technologies to integrate and publish data from multiple sources. The authors present two case studies, one that integrates household travel and fuel economy data and one that integrates datasets from three massive open online course providers.

In “Managing Data in Healthcare Information Systems: Many Models, One Solution,” Karamjit Kaur and Rinkle Rani describe a multimodel-based healthcare information system. The system uses a relational database system, a document database, and a graph database to manage data.

Finally, in “Optique: Zooming in on Big Data,” Martin Giese, Ahmet Soylu, Guillermo Vega-Gorgojo, Arild Waaler, Peter Haase, Ernesto Jiménez-Ruiz, Davide Lanti, Martin Rezk, Guohui Xiao, Özgür Özçep, and Riccardo Rosati present the Optique platform. The platform eliminates the data access bottleneck and enables end users to directly specify their information needs through an intuitive and visual query interface. The user query is transformed into highly optimized queries for the underlying heterogeneous data sources.

About 400 years ago, Galileo observed that “the book of nature is written in the language of mathematics.’’ This is even more relevant today, given the potential opportunities big data presents for groundbreaking discoveries through data-driven science and analytics. Big data may well be the next frontier for innovation, competition, and productivity.

We welcome readers to explore the articles in this special issue, to use them to solve business problems, and to contribute their own new research findings to the growing big data area.

ABOUT THE AUTHORS

VENKAT N. GUDIVADA is a professor of computer science and interim division chair at Marshall University, Huntington, West Virginia. His research interests include database management, information retrieval, high-performance computing, and personalized e-learning. Gudivada received a PhD in computer science from the University of Louisiana, Lafayette, Louisiana. He is a member of the IEEE Computer Society. Contact him at gudivada@marshall.edu.

RICARDO BAEZA-YATES is vice president of research at Yahoo Labs, Sunnyvale, California. His research interests include Web search, data mining, and scalability. Baeza-Yates received a PhD in computer science from the University of Waterloo, Ontario, Canada. He is a Fellow of both ACM and IEEE. Contact him at rbaeza@acm.org.

VIJAY V. RAGHAVAN is the Alfred and Helen Lamson/BoRSF Endowed Professor in Computer Science at the Center for Advanced Computer Studies and the director of the National Science Foundation–sponsored Industry/University Cooperative Research Center for Visual and Decision Informatics at the University of Louisiana, Lafayette. His research interests include data mining, information retrieval, machine learning, and Internet computing. Raghavan is a senior member of the IEEE Computer Society. Contact him at vijay@cacs.louisiana.edu.

REFERENCES

1. V. Gudivada, D. Rao, and V. Raghavan, “Big Data–Driven Natural Language–Processing Research and Applications,” *Big Data Analytics*, V. Govindaraju, V. Raghavan, and C.R. Rao, eds., Elsevier, 2015 (in press).
2. A. Halevy, P. Norvig, and F. Pereira, “The Unreasonable Effectiveness of Data,” *IEEE Intelligent Systems*, vol. 24, no. 2, 2009, pp. 8–12.
3. V. Gudivada, D. Rao, and V. Raghavan, “Renaissance in Data Management Systems: SQL, NoSQL, and NewSQL,” *Computer* (in press).
4. R. Baeza-Yates. “Big Data or Right Data?” *Proc. 7th Alberto Mendelzon Int’l Workshop on Foundations of Data Management* (AMW 13), 2013, vol. 1087, paper 14; http://ceur-ws.org/Vol-1087/paper14.pdf.