2013 IEEE 37th Annual Computer Software and Applications Conference

Big Data - Opportunities and Challenges

Panel Position Paper

Elisa Bertino
Cyber Center, CERIAS and CS Department
PurdueUniversity
West Lafayette, Indiana (USA)
bertino@cs.purdue.edu

***Abstract*— This paper summarizes opportunities and challenges of big data. It identifies important research directions and includes a number of questions that have been debated by the panel.**

Keywords—data management; data analytics; data security; data privacy

I. INTRODUCTION

Recent technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics, and social networks, are making possible to capture, process, and share huge amounts of data – referred to as *big data* - and to extract useful knowledge, such as patterns, from this data and predict trends and events. Big data is making possible tasks that before were impossible, like preventing disease spreading and crime, personalizing healthcare, quickly identifying business opportunities, managing emergencies, protecting the homeland, and so on [1]. As discussed by The Economist [2] "*Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to accounts*". Unlocking the potential of big data requires however addressing several major challenges. The goal of this panel is to identify and discuss research directions to address these challenges.

In what follows, we first discuss the notion of big data and application domains where big data is relevant. We then outline relevant challenges and summarize questions addressed by the panel.

II. WHAT IS BIG DATA

In order to discuss about research issues for big data management, it is crucial to better understand all dimensions related to big data. In this respect four characteristics define big data:

- Volume – data sizes will range from terabytes to zettabytes (that is, 10^{21} bytes).
- Variety – data comes in many different formats from structured data, organized according to some structures like the data record, to unstructured data, like image, sounds, and videos which are much more difficult to search and analyze.
- Velocity – in many novel applications, like smart cities and smart planet, data continuously arrives at possible very high frequencies, resulting in continuous high-speed data streams. It is critical that the time required to act on this data be very small.
- Huge number of data sources – the real value of data sets is when these data sets are integrated and cross-correlated. Integration and cross-correlation among data sets from different sources allow one to uncover information and trends that often cannot be uncovered by looking at a data set in isolation. It is critical that effective automated approaches to large scale data integration be devised.

The above short discussion emphasizes that volume alone is perhaps the least difficult problem to address then dealing with big data. The real challenge arises when we have big volumes of unstructured and structured data continuously arriving from a large number of sources. Addressing such challenge require a new generation of "*technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery and/or analysis"* [3].

III. FOR WHOM IS BIG DATA RELEVANT

Big data is relevant for all components of our society. Industry is using big data for shifting business intelligence from reporting and decision support to prediction and next-move decisions. This use of big data emphasizes that big data is critical for obtaining actionable knowledge. Governments are also interested in using big data and predictive analytics to improve decision making and transparency, to engage citizens in public affairs, to improve national security. Healthcare represents another major area to which big data may offer novel opportunities [4]. Learning health systems are currently focusing on turning health care data into knowledge, translating that knowledge into practice, and creating new data by means of advanced information technology. As pointed out in [5], the use of big data technologies can reduce the cost of healthcare while improving its quality by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring.

Big data is also crucial for research. Many areas of science and engineering are currently facing from a hundred to a

0730-3157/13 $26.00 © 2013 IEEE
DOI 10.1109/COMPSAC.2013.143

thousand-fold increase in the volume of data generated compared to only one decade ago. This data is produced by many sources including simulations, high-throughput scientific instruments, satellites, and telescopes. While the availability of big data is revolutionizing how research is conducted and is leading to the emergence of a new paradigm of science based on data-intensive computing, at the same time it poses a significant challenge for scientists. In order to be able to leverage these huge volumes of data, new techniques and technologies are needed. A new type of e-infrastructure, the Research Data Infrastructure, must be designed, implemented and optimized to support the full life cycle of scientific data, its movement across scientific disciplines, and its integration with published literature.

IV. TEACHNICAL RESEARCH CHALLENGES

There are many technical challenges that must be addressed to realize the full potential of big data. Jagadish et al. [5] provide a comprehensive discussion of such challenges based on the notion of data analysis pipeline:

- Data Acquisition and Recording: it is critical to capture the context into which data has been generated, to be able to filter out non relevant data and to compress data, to automatically generate metadata supporting rich data description and to track and record provenance.
- Information Extraction and Cleaning: data may have to be transformed in order to extract information from it and express this information in a form that is suitable for analysis. Data may also be of poor quality and/or uncertain. Data cleaning and data quality verification are thus critical.
- Data Integration, Aggregation and Representation: data can be very heterogeneous and may have different metadata. Data integration, even in more conventional cases, requires huge human efforts. Novel approaches that can improve the automation of data integration are critical as manual approaches will not scale to what is required for big data. Also different data aggregation and representation strategies may be needed for different data analysis tasks.
- Query Processing, and Analysis: methods suitable for big data need to be able to deal with noisy, dynamic, heterogeneous, untrustworthy data and data characterized by complex relations. However despite these difficulties, big data even if noisy and uncertain can be more valuable for identifying more reliable hidden patterns and knowledge compared to tiny samples of good data. Also the (often redundant) relationships existing among data can represent an opportunity for cross-checking data and thus improve data trustworthiness. Supporting query processing and data analysis requires scalable mining algorithms and powerful computing infrastructures.
- Interpretation: analysis results extracted from big data needs to be interpreted by decision makers and this may require the users to be able to analyze the assumptions at each stage of data processing and possibly re-tracing the analysis. Rich provenance is critical in this respect [6].

In addition to the above challenges, privacy and security are critical issues. Privacy in particular raises many concern as big data could be used to re-identify privacy-sensitive data even when this data has been anonymized. Also big data can be used to create profiles of user groups that may be used for discriminating specific groups of individuals or even single individuals. In this respect, population privacy is as crucial as personal privacy. Security also raises challenging issues including scalable security administration, management and integration of heterogeneous data security policies, and the security of data when hosted clouds.

V. PANEL QUESTIONS

The panel debated several aspects of big data management and applications that are relevant from a researcher perspective. Questions asked to the panelists include:

- Are there additional big-data applications that we should consider?
- Which are policy issues related to big data that we should address, in addition the well-known privacy policies?
- How can academia, industry and governments engage in projects and initiatives focusing on big data and data intensive applications?
- Are there national and international initiatives that we should engage with?
- What would be technology transfer and commercialization opportunities for research focusing on big data and Research Data Infrastructures?
- Can we quantify the economic value of data?
- Which are the incentives and impediments to data sharing?

ACKNOWLEDGMENT

The work reported here has been partially supported by the Purdue Cyber Center (Discovery Park).

REFERENCES

[1] E. Bertino, Data Protection from Insider Threats. Morgan&Claypool, 2012.

[2] "Data, data everywhere", The Economist, 25 February 2010, available at http://www.economist.com/node/15557443 (Downladed on April 30, 2012).

[3] O'Reilly Radar Team, Big Data Now: Current Perspeectives from O'Reilly Radar. O'Reilly, 2011.

[4] T. Murdoch, A. Detsky, "The Inevitable Application of Big Data to Health Care", JAMA, 2013, 309(13):1351-1352.

[5] H.V. Jagadish et al., "Challenges and Opportunities with Big Data", 2012, available at http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf (Downladed on April 30, 2012).

[6] S. Salmin, E. Bertino, "A Comprehensive Model for Provenance", Invited Paper, Proceedings of the First International Workshop on Modeling Data-Intensive Computing (MoDIC 2012), Florence, Italy, october 15-18, 2012, LNCS 7518, Springer.