Big Data – a New Challenge for Data Manipulation and Analysis

PETR BERKA
University of Economics, Prague, Czech Republic
and
University of Finance and Administration, Prague, Czech Republic

Abstract: This paper introduces the field of big data as a new concept related to data manipulation and analysis and reviews its main problems and challenges.

Keywords: big data, distributed computation, data mining, machine learning.

1 Introduction

It can be seen that the extent of data collected by various applications and systems doubles every three years. But not only the amount of data changes, also the types of collected data rapidly changes. While in the past most data collected was in the tabular (relational) form, now-days we can encounter data in the form of time series, data streams, text, images or videos. The source of such data can be various sensors, surveillance systems, mobile phones, GPS devices, RFID readers, social networks, computer networks, web logs, scientific data etc. This fact is behind a new concept that suddenly emerged in the data base and data analysis community, the concept of Big Data.

2 What is Big Data

According to Gartner, Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [1]. Big Data can be characterized by 4 V's - Volume, Velocity, Variety and Veracity:

- Volume: the size of Big Data goes beyond standard data storage and manipulation techniques,
- Velocity: Big Data is often available in real time,
- Variety: Big Data contains not only structured data (e.g. in tabular or relational form) but also texts, images, audio or video,
- Veracity: the quality and reliability of Big Data can vary.

What is Big Data may change over time, but generally speaking, Big Data are data that cannot be handled using standard data base systems and standard data analysis tools. Cf. the Mckinsey's report [9], where Big Data is defined as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." Big Data thus differs from very large databases (VLDB) where only the size of data was the problem.

3 Challenges for collecting, storing and manipulating Big Data

Traditional data management and analysis systems, mainly based on relational database management system (RDBMS), are inadequate in handling data that are characterized by 4 V's. RDBMS can efficiently handle structured data but offer little support for semi-structured or unstructured data. The scalability of RDBMS relies on expensive hardware and cannot be always guarantied.

The solution for the first problem can be NoSQL databases. NoSQL databases provide means for storing and retrieving data that cannot be expressed using tabular relations of relational databases. Beside this, these databases can be fast and highly scalable. The second problem can be handled (on the level of data storage) by the file systems or (on the level of data manipulation) by parallel computing, distributed computing, grid computing or cloud computing. Google designed and implemented Google File System (GFS) as a scalable distributed file system [6] for large distributed data intensive applications. GFS is designed to provide efficient, reliable access to data using large clusters of commodity hardware. The terms parallel computing, distributed computing, grid computing and cloud computing all refer to a processing where the computation is spread over more (or many) computing units. The computing units can be located within single computer ("classical" parallel computing), the units (computers in this case) can be interconnected within a network (distributed and grid computing) or can be put together on-demand via Internet (cloud computing).

Beside these particular approaches to Big Data, some complex solutions how to handle Big Data have been proposed as well. Let us introduce Apache Hadoop as such an example.

3.1 Apache Hadoop

Apache Hadoop is an open-source software framework that supports massive data storage and processing. Instead of relying on expensive, proprietary hardware to store and process data, Hadoop enables distributed processing of large amounts of data on large clusters of commodity servers. Hadoop has many advantages, and the following features make Hadoop particularly suitable for big data management and analysis. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. Hadoop can be deployed in a traditional onsite datacenter as well as in the cloud. Hadoop [11] thus integrates data storage, data processing, system management, and other modules to form a powerful system-level solution to handle Big Data.

4 Challenges for analyzing Big Data

The necessity to analyze huge amount of data that is collected on-line constitutes a paradigm shift for the field of data analysis (in particular for the areas of data mining and machine learning) since the idea of a standalone (desktop or workstation) analysis tool is abandoned in favor of process integrated, distributed and autonomous analysis systems. Instead of off-line learning in a batch setting, sequential learning, anytime learning, real-time learning, online learning etc. under real-time constraints from ubiquitous and distributed data is needed. Instead of learning from stationary distributions, concept drift is the rule rather than the exception. Instead of large stand-alone workstations, learning takes place inside small, unreliable devices. This paradigm shift is also expressed in new concepts for the whole knowledge discovery process, some examples are introduced bellow.

4.1 Ubiquitous Knowledge Discovery

Ubiquitous Knowledge Discovery can be defined as "Knowledge discovery process in mobile, distributed, dynamic environments, in presence of massive amounts of data". Knowledge discovery in ubiquitous environments is an emerging area of research at the intersection of the two major challenges of highly distributed and mobile systems and advanced knowledge discovery systems. Research areas as defined in the EU funded project KDUbiq (2005-2008 FP6 FET IST) are [5]:

- data mining in mobile systems, wireless communication networks, calm technologies,
- distributed architectures: distributed data mining, grid, P2P, autonomic computing,
- agents,
- learning components: statistical learning (incl. online learning), evolutionary computing,
- anytime algorithms data types: spatio-temporal, stream, multimedia,
- security and privacy: privacy preserving data mining, intrusion detection,
- HCI and cognitive modelling: user interfaces of ubiquitous discovery systems.

The idea of a standalone (desktop or workstation) analysis tool is abandoned in favor of process integrated, distributed and autonomous analysis systems of ubiquitous discovery systems.

4.2 Reality Mining

The term "reality mining" was coined by Nathan Eagle and Alex Pentland from Media Laboratory, Massachusetts Institute of Technology (MIT) about 10 years ago [4]. According to them, reality mining is the collection and analysis of machine-sensed environmental data pertaining to human social behavior, with the goal of identifying predictable patterns of behavior. Reality mining studies human interactions based on the usage of wireless devices such as mobile phones and GPS systems providing a more accurate picture of what people do, where they go, and with whom they communicate with rather than from more subjective sources such as a person's own account. Example applications of reality mining can be analysis of complex social systems, traffic monitoring and control, environmental monitoring or smart homes and ambient assisted living.

From the data mining point-of-view, reality mining deals with the most challenging data mining problems as defined in [12]. In particular, it tackles the issues of “scaling up for high dimensional data/high speed streams”, “mining sequence data and time series data”, and “data mining in a network setting”.

5 Conclusions

This paper tries to present an initial insight into the Big Data. Rather than an in-depth analysis of the field it just lists and briefly comments basic problems and basic notions of this interesting recently emerging area that is on the intersection of data storing and data analysis. Interested readers should refer to additional readings, some of them given in the list of literature.

Literature

1. Beyer, M., 2011. Gartner Says Solving ' Big Data' Challenge Involves More Than Just Managing V Volumes of Data. Gartner.
2. Chen, M., Mao, S. and Liu, Y., 2014. Big Data: A Survey, *Mobile Netw Appl*, 19 pp. 171–209.
3. Chen, C.L.P., and Zhang, C-Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275 pp. 314-347.
4. Eagle, N. and Pentland, A.S., 2006. Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, 10(4), 255–268.
5. Gama, J. and May, M., 2011. Ubiquitous Knowledge Discovery, *Intelligent Data Analysis,* 15(2011) pp. 1–2.
6. Ghemawat, S., Gobioff, H. and Leung, S.T., 2003. The Google file system. In Proc. *19th ACM Symposion Operating Systems Principles*, pp. 29-43.
7. Hu, H., Wen, Y., Chua, T.S. and Li, X., 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, doi 10.1109/ACCESS.2014.2332453
8. Jagadish, H.V., Gehrke, J., Labrindis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan,R. and Shahabi, C., 2014. Big Data and Its Technical Challenges. *Communications of the ACM*, Vol. 57, No. 7, pp. 86-94.
9. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung-Byers, A., 2011. *Big data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute.
10. Russom, P., 2011. *Big Data Analytics*. TDWI Research
11. White, T., 2012. *Hadoop: The Definitive Guide*. O'Reilly Media.
12. Yang, Q. and Wu, X., 2006. 10 Challenging Problems in Data Mining. *International Journal of Information Technology & Decision Making,* Vol. 5, No. 4, pp. 597–604.

Contact data:

Prof. Ing. Petr Berka, CSc.
University of Economics
W. Churchill Sq. 4
130 67 Praha 3, Czech Republic
and
University of Finance and Administration
Estonská 500
10100 Praha 10, Czech Republic
berka@vse.cz