Big Data Applications and Challenges

Milan Simaković
School of electrical engineering, University of Belgrade
Belgrade, Serbia
milanrus@hotmail.com

Zoran Čiča
School of electrical engineering, University of Belgrade
Belgrade, Serbia
cicasyl@etf.rs

Abstract—**Big Data is an uprising technology that will have a huge impact on future business strategies. Big Data is a platform that is used for efficient data storing and processing whose main characteristics are theoretically unlimited storage capacity and computing power, high availability and reliability, scalability and large tool portfolio. This technology is unavoidable if there is a need to store and process large amount of data. This paper presents an overview of the current state of the Big Data technology. Special attention will be given to Big Data architecture and its main components. We will also discuss the solutions currently available on the market. The paper proposes the Big Data applications grouped by different industries and emphasizes the current challenges of the Big Data platform that need to be resolved. The aim of this paper is to present the current state and remaining issues of the Big Data platform.**

Keywords-Big Data; Hadoop; Cloud; Data mining

I. INTRODUCTION

The amount of the generated data constantly increases due to the expansion of the Internet and continuous development of the computer technologies. Huge amount of data and data generation speed lead to difficulties in processing. With the exponential data growth, current methods for data storing and processing become practically useless because of limited resources and lack of scalability. Due to hardware and computing limitations, most of the data is discarded and analysis is performed on the small data subset. Analyzing the entire dataset, it is possible to gain a better insight into the data, and therefore draw a better conclusion. To achieve storing and processing of the large amount of data at high speeds, IT world entered into a new era of the Big Data. Business and market requirements are the main reason for opening a new chapter in data processing and for development of Big Data technologies. The main challenges of today's data processing are represented by "5V" concept [1]. "5V" term stands for Volume, Variety, Velocity, Veracity and Value that will be briefly discussed in the remainder of this section.

The growing number of data sources lead to constant increase of data amount (Volume). Devices such as mobile phones, cameras, automobiles, televisions and machines in industry and health care contribute to the exploding data volumes that we see today [2]. For example, airline companies collect data from airplanes in order to predict dysfunctionality and malfunctions of its components. In one aircraft there is over 6000 sensors. One Boeing jet generates 10 TB of information per engine every 30 minutes of flight, according to Stephen Brobst, the CTO of Teradata [3]. Hence, for one hour flight on a Boeing 737 with two engines, total amount of generated data is about 40 TB. In other words, a single flight requires processing of 40TB of data per hour. According to [4], there are about 100000 commercial flights worldwide every day, so it is easy to see that the Volume represents a very important issue in Big Data.

Velocity stands for the speed at which data is being generated. This high speed generated data must be dealt with in timely manner. Velocity opens a new branch in Big Data technology that deals with real time data stream processing and in-memory processing. Within this problem, two challenges that are frequently encountered in practice, can be identified.

In many cases, there is no time for the data storage and its subsequent processing, but the analysis needs to be done on the fly. In already mentioned example of collecting data from the aircraft, it is easy to conclude that the data processing must be performed during a flight in order to identify potential problems and prevent catastrophic event. If the data is collected and processed afterwards, the generated results would have no value because disaster had already occurred. In this case, processing power is a limited factor considering the fact that it is necessary to process a huge amount of data per time unit.

Another problem that falls within velocity is related to bursts that can occur in streams. Bursts represent a problem not only in terms of data storage, but also in real-time processing. For example, one mobile base station is collecting the data near the city hall. In normal occasions this base station covers this area with some daily average capacity. On the other hand, during a concert, the audience makes phone calls, texting, tweeting, sharing photos, updating statuses on social networks, etc. Consequently, the traffic and the amount of data suddenly increase hundreds, even thousands times. Traditional systems cannot efficiently deal with this type of problems and this is the reason why the solution needs to be searched within Big Data technology.

Nowadays there are many data sources (Variety). All of them can be classified into 3 groups: structured (databases), semi-structured (LOG files) and unstructured (documents, emails, web pages, video…) data. Also, keeping in mind the usage of many different file formats, data processing becomes more challenging. Usually, there is no such raw data that can be directly used, but it is necessary to do some preprocessing. Moreover, correlation with data received from other sources represents unavoidable step. The variety of data that needs to

be processed and stored additionally increases complexity of the Big Data solution.

Veracity claims confidence factor of the generated results. This characteristic is related to the uncertainty of the data, no confidence in data source (typographical mistakes, missing points because of latency or timeouts…) and mistrust of the process that is generating results. The Veracity problem appears when it is not possible to rely on the obtained results. Usually, the analysis is performed on certain data samples because of the huge amount of data. This kind of analysis can lead to insufficiently precise conclusions. Hence, there is a need to observe the complete set of data rather than just some specific samples. On the other hand, the processing of the complete dataset increases the load of the processor and time needed for data processing.

Value refers to the ability to turn data into value that is easy to understand and comprehend. It is easy to fall into a buzz trap and embark on Big Data initiatives without a clear understanding of the business value it will bring [1]. From a huge amount of the unstructured data it is important to extract reasonable results that could be used further, for example in the development of the business strategy.

In this paper, Big Data technology is presented. In the following section, Big Data technology is defined and Big Data architecture is described together with its most important components. Also, multiple vendors are discussed and special attention is given to cloud solutions. In the third section, the practical use of the Big Data technology is analyzed and applications for various industries are proposed. In the fourth section, overview of the current problems and challenges in the Big Data technology is given, including security issues, data mining challenges and the lack of skilled engineers. The fifth section concludes the paper.

II. BIG DATA OVERVIEW

Big Data is growing very fast as new problems and solutions are being introduced. On its own, Big Data is a term that is too abstract. In this section, Big Data term is discussed and an architecture overview is given together with the description of its main components. Also, a review of the current Big Data "on-premises" solutions will be presented, as well as the ones on-cloud.

A. Definition

Big Data is a term that is hot topic in recent years in the world. Since this term is new, there is no unique definition of it. Someone connects Big Data term to huge amount of data, but it is not quite correct. According to Gartner, Big Data are information assets with volumes, velocities and/or variety requiring innovative forms of information processing for enhanced insight discovery, decision-making and process automation [5]. IBM considers Big Data as a form that is uniquely defined by "5V" – Volume, Variety, Velocity, Veracity and Value that are discussed in the previous section. Big Data term should stand for the data that is not manageable using traditional processing systems and/or cannot be stored in the traditional data warehouses due to their size. Big Data term is relative and cannot be attached to exact numerical values. What is considered Big Data varies depending on the capabilities of the organization managing the dataset. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration [6].

B. Architecture

Big Data in practical terms is initiated in the form of Hadoop ecosystem. Hadoop is an open source project hosted by Apache Software Foundation. It consists of many small subprojects which belong to the category of infrastructure for distributed computing [7]. Big Data is a distributed, scalable platform for storing and processing huge amounts of data generated with great speed that cannot be treated in the traditional manner. It is a combination of different technologies and tools that operate amongst themselves. Together they form Hadoop ecosystem. Considering the fact that the platform is in constant development, the number of these tools steadily increases. All other solutions available on the market are based on Apache Hadoop. Therefore, in the following, the most important components of Hadoop ecosystem will be described.

One of the two main components of the Hadoop ecosystem is Hadoop Distributed File System (HDFS). HDFS is a framework that provides unique file system across multiple servers that form one cluster for data storage. It is based on master/slave architecture where Namenode is a master (with secondary Namenode used for prevention of single point of failure (SPOF)) and Datanodes are slaves. Namenode splits data into blocks and manages data storing, while Datanodes are used for the storage of data. Incoming file is divided into blocks that are stored on Datanodes. The main advantage of this file system, in comparison to others, is possibility to store big files that cannot be saved on single storage due to capacity limitation, as well as high availability and scalability. Total HDFS capacity is equal to the aggregated capacity defined on Datanodes. In this way, unlimited file system capacity can be achieved allowing storage of large files. High reliability is enabled by data replication. Each block is replicated on HDFS specified number of times (usually three). If necessary, new Datanodes can be added without downtime which increases capacity and computing power that will be explained later in this section.

Second main component of the Hadoop Big Data cluster is MapReduce. It is a programming model for processing large datasets with a parallel, distributed algorithm on a cluster [8]. Nowadays, MapReduce is replaced with YARN (Yet Another Resource Negotiator, MRv2) which is an updated version of MapReduce (MRv1). Big data processing is realized through map/reduce algorithm. During mapping process, each Datanode performs operations on data blocks. In this way, it is possible to perform distributed programming and achieve theoretically unlimited computing power. Map results are being collected in the Namenode where their aggregation is performed, i.e. reduce part of program. MapReduce is Java-based framework, but there are a lot of interfaces that enable

the usage of other programming languages for MapReduce coding. These engines are installed on top of the MapReduce framework. The most used ones are Hive (SQL like), Pig (high level programming), JAQL (Java QL), Pydoop (python based), etc.

Spark is a new computing framework in Hadoop family. Like MapReduce and YARN, it is built on top of HDFS. Spark provides an easier to use alternative to Hadoop MapReduce and offers performance 10 to 100 times faster than previous generation systems (like Hadoop MapReduce for certain applications) [8]. Unlike MapReduce, Spark is focused on in-memory computation, which gives him superior computing performance. In-memory computing allows Spark to efficiently deal with stream processing challenges and query big sets of data. To make programming faster, Spark provides clean, concise APIs in Scala, Java and Python. In addition, it is also possible to use Spark interactively from the Scala and Python shells to rapidly query big datasets [8].

Since Hadoop was initially developed like data storage and post processing engine, it could not respond to challenges of data processing and storing in a real time. Also, there are structured data sources that could not be saved in traditional relational databases because of their large amount. Therefore, HBase has been developed. HBase is a non-relational distributed database. It presents Big Data component which is used to store time sensitive structured data. It enables very high read/write performance on tables with millions of columns and billions of rows.

Big Data ecosystem grew over time and it was very difficult to monitor work of its components. Therefore, Zookeeper has been developed. Zookeeper is used to coordinate the cluster and provide highly-available distributed services. Zookeeper simplifies the development process, making it more agile and enabling more robust implementations [8].

C. Vendors

The main credit for rapid development of Big Data technologies takes Apache Hadoop. Hadoop is an open-source project of Apache. Main goal of this project is Big Data tools and components development. The fact that Big Data market is in trend encouraged many other vendors and companies to develop and offer their Big Data solutions. All those solutions actually represent customization of the core Apache Hadoop platform giving them different flavor. Some of the most important vendors on the market are Cloudera, Hortonworks, IBM, Microsoft, Google, Yahoo, Oracle, etc.

D. Big Data on Cloud

Cloud computing is a powerful technology for performing massive-scale and complex computing. It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Addressing Big Data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis [9]. Today's big cloud providers (IBM Bluemix, Microsoft Azzure, Google Cloud, etc.) offer Big Data solutions on cloud. There are many advantages of using this method in comparison to in-house Big Data solutions like reducing costs, easy scalable infrastructure, disaster recovery, security, licensing minimization, improving accessibility… Big Data on cloud is one of the newest discoveries in the computing world. At this moment, there are a lot of challenges in front of the cloud platforms that need to be solved. Some of them are guaranteeing upload speed and delays in transmission for time sensitive data. Power and utilization of this technology will be expressed in following years.

III. BIG DATA PROJECTS AND APPLICATIONS

The possibility of processing huge amounts of different data types creates new opportunities and plays a large role in generating new insights [10]. Big Data technology is of great importance for many companies. Using this technology, it is possible to improve business management and gain higher income. A large number of companies have initiated a research in this area and development of new solutions. Big Data additionally earns on its importance, due to such high interest.

TABLE I. BIG DATA APPLICATIONS GROUPED BY INDUSTRIES

Industry	Big Data applications
IT	Log analysis Network optimization Failure prediction
Telecommunications	Performance monitoring Billing Prediction of contract termination 360 degrees view of the customer
Banking	Fraud detection Operational analysis Customer analysis Prediction of contract termination 360 degrees view of the customer
Government	National security Crime Prediction and Prevention Cybersecurity National Security Scientific Research Tax Compliance [11]
Retail (traditional and online)	Billing Customer understanding Monitoring and insight into history of every product Prediction in terms of recommendation offer to the customer 360 degrees view of the customer
Internet of Things	Smart cities (intelligent street lighting, traffic jam prediction, traffic light control, etc.) Prediction of catastrophic events (earthquake, flood, fire, demolition of building, etc.)
Social media analysis	Customer Sentiment analytics Society opinion Customer experience
Industries	Detection of unexpected events in production chain Quality control Machine failure prediction
Health	Telemedicine analysis Clinical pathway extraction Prediction of stroke Disease detection

Based on Big Data product portfolio, this platform can be used for solving various problem types that occur in practice. The most primitive way to use Big Data platform is as a data storage. Different types of data from different sources interflow to one place. With the help of MapReduce, aggregated data can be correlated among themselves and processed in order to extract hidden information and generate better insight into data. Using Spark it is possible to monitor and process data streams and generate results in real time. Moreover, the usage of Spark gives possibility for designing models regarding machine learning and predictive analytics.

In practice, there are many fields in which Big Data technologies can be applied or have already been applied. Table I proposes Big Data technology applications grouped by industries in which they can be used. From Table I it can be seen that the usage and potential of this technology is enormous. Table I shows only a part from huge portfolio of possibilities. This technology can be used to solve many other problems in different fields where Big Data problem is present.

IV. BIG DATA CHALLENGES

Considering the fact that Big Data technology is still young, there are still problems and challenges to be solved. In this section, an overview of main problems and challenges related to this technology is given.

A. Privacy and security

Personal customer information can be combined with huge amount of data collected by service provider (i.e. Telco operator) and generate information about customer that violates his privacy. This information is used to improve the business of the company, but still without user permission and knowledge [7]. If the generated information is used to harm the customer, it directly leads to security violation.

B. Data mining

Let's assume that company is using Big Data technology and collecting all kind of data from different sources and different types. When data is collected, the question is what to do with the data and how to extract some valuable information that will be used to improve business. Analytic process that deals with data analysis in terms of extraction of useful information is called data mining. Company establishes a data mining team whose only task is to get familiar with the data and to try to generate useful output. Each company generates its own set of data that is typical only for that company. Because of that, developed models from one data mining team cannot apply to other datasets of other companies which limits the application of the developed data mining solutions to one company or in the best case small number of companies.

C. Lack of skilled personnel

Intensive development of Big Data technologies leads to opening new job positions. Huge demand and fact that this technology is quite new leads to the lack of skilled people. Scope of work for one software engineer is not only limited to technical area, but should be extend to research and analytical fields [7]. Big companies and open source communities create plenty of online courses in order to arm engineers with new skills. Moreover, the universities should introduce students with Big Data in order to produce skilled employees in this area of expertise [7].

V. CONCLUSION

The usage of Big Data technologies can give deep insight into the data providing conclusions based on the whole set of data, not just on a small subset. This technology has a great potential and represents unavoidable tool for every serious company because it responds to many challenges and maximizes business value. Giving the fact that this technology is on the rise, there are plenty questions to be answered. Large number of vendors on the market shows that the potential of this technology is recognized and that its widespread usage is expected in the following years. This paper presents overview of the current state for Big Data technologies and as such can be used as an introduction to this area of expertise.

REFERENCES

[1] B. Marr, "Why only one of the 5 Vs of big data really matters", IBM Big Data and Analytics hub, March 2015, http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters.

[2] D. deRoss, P. C. Zikopoulos, B. Brown, R.Coss, R. B. Melnyk "*Hadoop For Dummies*", John Wiley & Sons, Inc., 2014.

[3] S. Higginbotham, "Sensor Networks Top Social Networks for Big Data", GigaOM, September 2010, http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/.

[4] G. Garfors, „100,000 Flights a Day", June 2014, http://www.garfors.com/2014/06/100000-flights-day.html.

[5] D. Laney, "3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety", *META Group*, February 2001.

[6] J. Guterman, "Big Data, Release 2.0: Issue 11", Radar, O'Reilly, June 2009.

[7] A. Katal, M. Wazid, R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices", *Contemporary Computing (IC3)*, August 2013.

[8] J. Roman, and contributors, "The Hadoop Ecosystem Table", GitHub, https://hadoopecosystemtable.github.io/.

[9] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, "The rise of "Big Data" on cloud computing: Review and open research issues", *Information Systems*, Vol. 47(1), pp. 98–115, January 2015.

[10] Jonathan Shaw, "Why "Big Data" Is a Big Deal", Harvard Magazine, March – April 2014. http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal.

[11] "Big Data and Apache Hadoop for Government", MapR Technologies, 2016, https://www.mapr.com/solutions/industry/big-data-and-apache-hadoop-government.