International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | Jun 2022 www.irjet.net p-ISSN: 2395-0072

CASE STUDY ON METHODS AND TOOLS FOR THE BIG DATA ANALYSIS

Chand Babu[1], Umesh Goyal[2]

[1]*Research Scholar, Dept. of Computer Science & Engineering, FCEM, Faridabad, Haryana, India*
[2]*Assistant Professor, Dept. of Computer Science & Engineering, FCEM, Faridabad, Haryana, India*

Abstract - *As the name of indicate, Big Data means huge collection of data that can't be proceed without traditional computing approach. To Compute the data it's need tool and technique. Having data bigger consequently requires different approaches, techniques, tools & architectures to manage the data in a better way. Big data technologies provide more accurate analysis which help in decision making. To manage and process huge volume of structured semi-structured and unstructured data you would require an infrastructure that can secure, privacy and protect the data. There are various tools and technologies in the market from different vendors IBM, Amazon, Microsoft, etc., to handle big data. The major challenges with big data are Capturing data, Storing data, Searching, Curation, Sharing, Transfer, Analysis, Presentation, To fulfill the above challenges, organizations normally take the help of enterprise servers.*

diverse methodologies (technologies/methods) based on certain parameters to chalk out which of them is best possible (optimal) or what else needs to be done, to have an optimal solution. As the data is being generated and accumulated at a very high velocity with diversity in addition to it, processing has become a tedious task. Whereas the fact remains that, by processing of this data we will uncover gold from these huge mountains but if it's left untreated it will become Everest's of garbage. Since the foundation of term Big Data many methodologies or technologies are present which are being used to process these data mountains. These technologies have their own area of interest due to which there are obvious drawbacks and mode of operations between them. After analysis we found that common to them is low agility and the Data management. Once the data Management issue of Cloud technology is resolved it will prove a boon and cure to the other drawbacks of cloud implementation for processing of Big Data. It makes as sure that Cloud technology with better Data Management will be implemented in every phase of life from governments to business.

Key Words: Big Data, Big data tools and methods.

1. INTRODUCTION

Every second sees huge amounts of exponentially growing data, being generated, acquired, stored, and analysed. The revolution in the generation of massive data amounts comes along with the Internet usage which allows data exchange between various electronic devices and humans. In this regard, the following fields are mentioned: mobile phones, social media, imaging technologies to determine a medical diagnosis, etc. The volume of available data continues to grow and it grows in different formats. Conversely, the price of data storage continues to fall which results in data storing being more reachable. Although creating data storage is getting cheaper and more available, the increasing volume of data in different formats and from diverse sources creates new problems with regards to data processing, including in its analysis and in integrating Big Data into business decision making processes.

In order to store and process Big Data, new technologies are evolving to address these problems. To deal with these challenges, there is a need for a new approach such as building a scalable and elastic architecture. The purpose of this study is to explore the domain of the Big Data problem; particularly, to create an overview among free available repositories of biomedical Big Data and to discover appropriate technologies and methods along with their limitations and use cases to be applied over the chosen data. Since the data, technologies and methods are chosen, a testing scenario is created and deployed over this data.

1.1 Big Data overview

There are many definitions for Big Data. Somebody defines Big Data as data that is quite complicated rather than "easy going" and it is hard to acquire, store, manipulate and process it, due to the fact it is "big". Another one, which is frequently mentioned is called "3V" described by three words (Volume, Variety and Velocity). These two definitions above and others are cited and mentioned below:

"Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions." [4]

Argues that there are three attributes standing out as defining Big Data characteristics:

"**Huge Volume of data:** Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns."

"**Complexity of data types and structures:** Big Data reflects the variety of new data sources, formats and structures, including digital traces being left on the web and other digital repositories for subsequent analysis."

"**Speed of new data creation and growth:** Big Data can describe high velocity data, with rapid data ingestion and near real time analysis."

The most known and used definition is originally described as "3V" (see Figure). Since then various authors have come up with other definitions and descriptions. The combination of these thoughts is described below:

1.1.1 Volume – the volume is increasing exponentially. There are many sources generating a huge amount of data which takes a long time to process.

1.1.2 Variety – Big Data is not always structured data. Actually, most parts of it are unstructured. In addition, it comes in different formats due to the variety of data sources; dealing with the data formats variety, increases the complexity of storing and analysing.

1.1.3 Velocity – describes the rate of data change. Data volume changes dynamically. There is a need to manipulate with data in real-time. However, for some tasks in fields such as business and health-care, it can be somehow demanding.

1.1.4 Value – it is often argued that the value is the most critical part of Big Data. Most of projects have to produce appropriate results, unless the company/institution can waste financial sources. It costs a lot of money to implement IT infrastructure systems to store data. If the subject which stores data is not able to extract data value, it is desirable to consider if the intention of the investment to an implementation of IT infrastructure systems was suitable.

1.1.5 Veracity – not all data has to be perfectly good, but it does need to be almost good to give relevant sight. Upon considering the errors rate and the data incompleteness, ambiguity in the dataset is desirable and even necessary for further data analysis. Credibility is another aspect of data veracity which is worth mentioning.

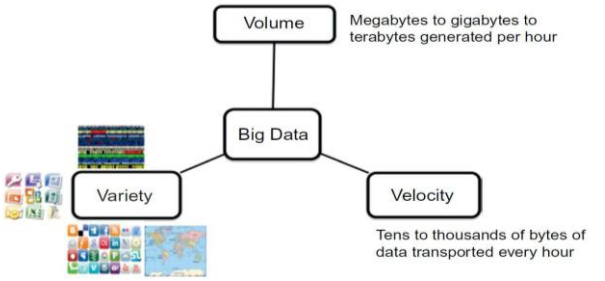

Figure 1: Big Data characteristic

1.2 Big Data evolution

To better understand what Big Data is and where it comes from, it is crucial to first understand some past history of data storage, repositories and tools to manage them. As shown in Figure , there has been a huge increase of data volume during the last three decades.

As we can see in the decade of 1990s the data volume was measured in terabytes. Relation databases and data warehouses representing structured data in rows and columns were the typical technologies to store and manage enterprise information.

Subsequent decade data started dealing with different kinds of data sources driven by productivity and publishing tools such as content managed repositories and networked attached storage systems. Consequently, the data volume was started being measured in petabytes.

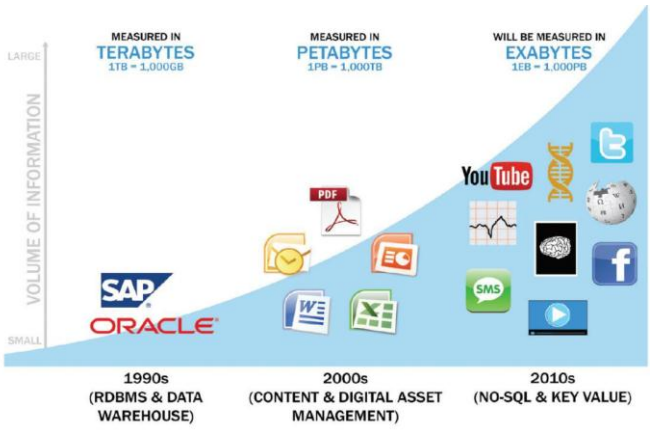

Figure 2: Data evolution and the rise of Big Data sources

1.3 Source of Data

Typically, the first type is bound up with the spread of machines digitalisation which is related to sensors integration, the connectivity increase, and devices recording sounds, images or videos and to machines communication between each other. Particularly, devices such as cameras recording videos, cell phones collecting

geospatial data, machines in production lines of industrial systems, are exchanging important information while processing their activities. More examples are mentioned below:

Medical information – machines recording EEG (Electroencephalography), heartbeats, genomic sequencing.

Multimedia – photos and videos uploaded on the Internet,

Mobile devices – provide geospatial data (location, gyroscope), as well as metadata about phone calls, messages, internet usage and data gathered by mobile applications,

Other devices – Warehouse Management Systems (WMS) providing inside location realised by Wi-Fi, identification of stored products or material by BAR, QR (Quick Response) codes or RFID (Radio-Frequency Identification) chips. There is a presence of many other technologies such as navigation systems, seismic processing, etc.

It is appropriate to consider other sources generated by activities on the Internet in general. Let us imagine having a set of servers which run web pages that can be determined for retail business. The servers can collect records of all activities of the websites' customers, users, transactions, applications and servers own activity and behaviour. For instance, there are logs which can be collected of [2].

1.4 Characteristics of Big Data

In relation to the definitions, the reason why there is intense complexity in processing Big Data is shown. Along with the Big Data there also exists ambiguity, viscosity and virality (see Figure).

Ambiguity – emerges when there is less or no metadata in Big Data. An example can be a graph or something that usually needs a description. Letters M and F in a graph can depict genders or they can even represent Monday and Friday.

Viscosity – this term is often used to describe the latency time in the data relative to the event of being described.

Virality – describes how quickly data is shared throughout a network among people who are connected. The measurement result is the rate of spread of data in time. For instance, Twitter can be a relevant example when the tweets are spreading from the first (root one) original tweet among people throughout the network.

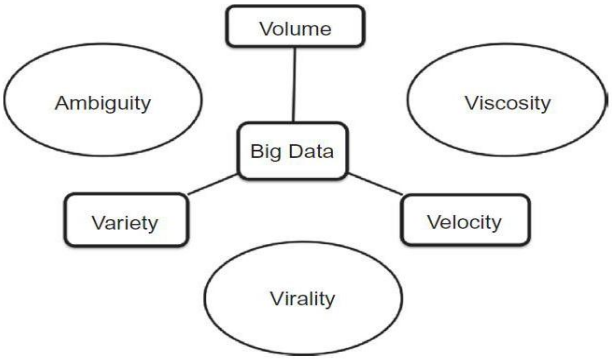

Figure 3: Big Data characteristics derived from '3 V' definition

1.5 Big Data vs Small Data

Big Data is not simply small data that has grown. There are more aspects that define differences between these two categories. A subset of the aspects can be derived from the "3V" definition described above and the others are argued by [6] (see Table 1).

SMALL DATA	BIG DATA
GOALS	
Usually designed to answer a specific question or to achieve a particular goal	Nobody knows what the exact output of the project is. Usually it is designed with a goal in mind, but the goal is flexible
LOCATION	
Typically, small data is contained within one institution, often on one computer or even in one file	Big data are located throughout the company network or throughout the Internet. Typically, it is kept onto multiple servers, which can be everywhere
DATA STRUCTURE	
Ordinarily contains highly structured data. Commonly, the data domain is restricted to a single discipline or its sub-sequence. The typical forms of its storage are uniform records or spreadsheets	Has to be capable of absorbing unstructured data such as text documents, images, sounds and physical objects. The subject of disciplines can vary throughout the data
DATA PREPARATION	

In many cases, the data is prepared by its own user for his own purposes	The data is collected from many different sources. People who the data comes from are rarely the same people who use the data
LONGEVITY	
The data is kept for a limited time (academic life). After few years when the data project is finished, the data is usually discarded	A Big Data project typically contains data which has to be stored in perpetuity
MEASUREMENTS	
Typically, the data is measured using one experimental protocol	Many various types of data are measured in many different electronic formats
REPRODUCTIBILITY	
The projects are typically repeatable. If there is a problem with the data quality, the entire project can be repeated	Replication of the project is seldom feasible. There is nothing more than optimism that the bad quality data is found and flagged, rather than be replaced by a better one
STAKES	
Project costs are limited. The institution can usually recover from small data failure	Big data projects can be really expensive. A failed Big Data project can lead to bankruptcy
INTROSPECTION	
Data points are identified by rows and columns within a spreadsheet or a database. It enables to address a particular data point unambiguously	It is more difficult to access the data. The organisation and the context can be inscrutable. Access to the data is achieved by a special technique referred to as introspection
ANALYSIS	
In most cases, all the data in the project can be analysed together	Big data is typically analysed in incremental steps. It is extracted, reviewed, reduced, normalised, transformed, visualised, interpreted and reanalysed with different methods

Additionally, in relation to the volume of Big Data and its incompleteness, corruption and ambiguity that are caused by its large volume, the Small Data is generally more organised, structured and it is also possible and desirable to avoid the incompleteness together with the other attributes of Big Data (see Figure).

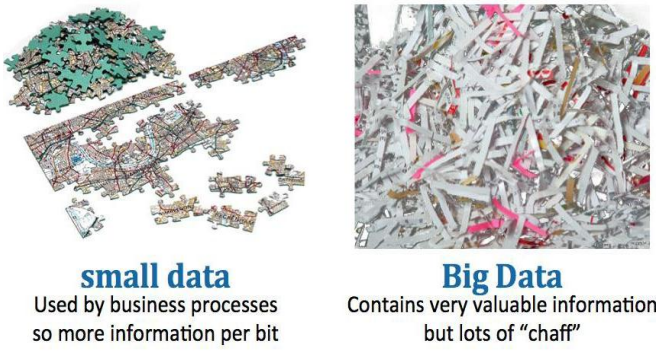

Figure 4: Small and Big Data illustration1

2. BIG DATA TOOLS

Although Big Data principles and approaches are frequently discussed, there are not many technologies which are convenient to deal with such data. Due to the definitions of the volume and the velocity, the tools which are supposed to deal with Big Data have to offer a distributed computing approach. There are the following approaches: multiple data and single program, and single data and multiple program.

In the first case, there is a single program, which is run on more nodes, where all nodes process different data. On the contrary, the second case is considered to have only one dataset, which is processed by a program divided on small tasks that are run on different nodes in parallel.

Due to it, there are tools that try to abstract from the physical distribution as much as possible. Since the Apache company released its new implementation of Map Reduce paradigm, a whole ecosystem called Hadoop has started evolving. The MapReduce paradigm offers the means to break a large task into smaller tasks, run in parallel, and consolidate the outputs of the individual tasks into the final output. The significant ecosystem expansion was caused by using simple programming models to process large datasets across clusters as well as was amplified by the fact that the whole solution has started as open source software.

Hadoop as the first publicly known and discussed technology of Big Data processing has been used as the base of open source and commercial extensions. In other words, most of the set of Big Data tools are based on the Hadoop solution.

These solutions offer methods and approaches to load, pre-process, store, query and analyse data.

In the following subchapters the Hadoop ecosystem will be described along with other technologies which have been evolved from it or others which are using its technologies.

2.1 MapReduce

As mentioned earlier, the MapReduce paradigm provides the means to break a large task into smaller tasks, run the tasks in parallel and consolidate the outputs of the individual tasks into the final output. MapReduce consists of two basic parts: a map step and a reduce step. [1]

- **Map –** performs an operation to a piece of data which generates some intermediate output.

- **Reduce –** gathers the intermediate outputs from the map steps, processes it and provides the collected final output.

The main advantage of MapReduce is the workload distribution over a cluster of computers (to run tasks in parallel). Particularly, MapReduce provides a technique, which allows the processing of one portion of the input which can be run independently of the other input parts. In other words, the workload can be easily distributed over the cluster.

2.2 Distributed File System – HDFS

The Hadoop Distributed File System (HDFS) is a file system which provides the capability to distribute data across a cluster to take advantage of the parallel processing of MapReduce.

HDFS is designed to run on common low-cost hardware. Consequently, it means there is no need to deploy it only on super computers. Although, it is implemented in Java, HDFS can be deployed on a wide range of machines apart from a node, which is dedicated to manage namespace services (see Architecture below).

2.2.1 Architecture

HDFS has a master/slave architecture as shown in Figure . It consists of a single master server which manages the filesystem namespace and manages access to files by clients and a single NameNode. In addition, there are DataNodes which are usually bound up with a node in the cluster. These DataNodes manage storage within their nodes that they run on. A file in HDFS is split into one or more blocks that are stored by a set of DataNodes. Moreover, they are responsible for serving read and write requests from the file system clients and performing blocks' creations, deletions and replications which are requested by NameNode. Whenever there is a request for an operation as opening, closing, renaming of a file or a folder, it is handled by the NameNode. Additionally, it is also in charge of the blocks mapping to DataNodes. [11]

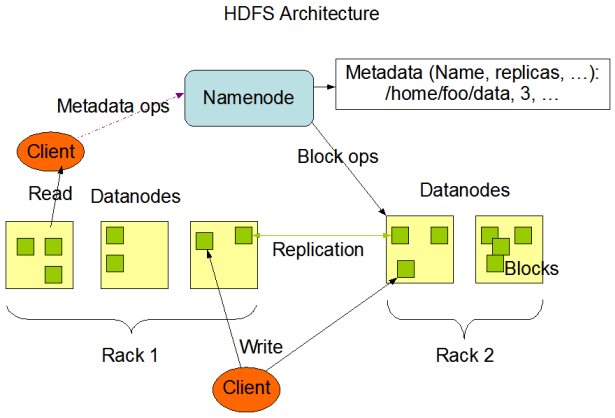

Figure 5: HDFS Architecture

2.2.2 File Storage

HDFS breaks files typically into 64 MB blocks and stores the blocks throughout the cluster. Whenever possible, HDFS attempts to store file blocks on different machines to allow the map step to operate on each block of a file in parallel. If a file size is 200 MB, the file is stored in four blocks: three 64 MB blocks and one 8 MB block. If the file is smaller than 64 MB, the block represents the whole file. HDFS is determined to deal with large files such as files of up to 1 GB. Files with a size smaller than 64 MB are considered as small sizes which are not efficiently maintained by HDFS.

2.3 MapReduce Job in Hadoop

A typical MapReduce program in Java is composed of three classes: the driver, mapper and reducer.

The driver – contains the job details and its configurations such as input file locations, details for assigning the input file to the map task, the names of the mapper and reducer Java classes and it also contains the location of the reducer task output.

The mapper – represents the logic to be processed on each data block related to the defined input files in the driver code.

The reducer – represents the logic of the gathering of intermediate results from the mappers.

2.4 The Hadoop Ecosystem

As mentioned above, Hadoop evolution comes all along with open source and commercial extensions to make Apache Hadoop easier to use and provide additional functionality and features. This subchapter examines the

following Hadoop-related Apache projects which all together form the Hadoop Ecosystem.

2.5 Spark

It is an open source Big Data processing framework for performing data analytics on a distributed computing cluster such as Hadoop. Spark supports in-memory processing to increase speed and data process over MapReduce. It is discussed as a more powerful analysis replacement of Hadoop. Spark is deployed on the top of existing Hadoop cluster which enables Spark to access data via HDFS. Additionally, it can also process structured data via Hive and stream data from HDFS. [13]

3. BIG DATA TOOL SELECTION

3.1 Storage

Nowadays, there are two databases where the data is being stored within the virtual machine which has 100 GB allocated. As result of the data analysis, without external stimuli the data volume will probably reach the level of 1 TB in approximately the year 2032. The current data storage system is sufficient for next following years. Most of PostgreSQL database size is occupied by binary files, which cannot be queries. With this in mind, there is no need to exchange this RDBMS for other more scalable technology such as a distributed file system. Moreover, the metadata of experiments are currently being stored in a NoSQL database where it can be queried in a more efficient way.

3.2 Pre-processing and Analysis

As said before, the whole process of analysis is performed either within Matlab along with EEGLAB or by applications written in Java.

3.3 Hadoop Based Solution

Most of technologies providing a scalable solution for dealing with Big Data are based on Hadoop. Due to the fact that Hadoop is implemented in Java, it provides suitable API which enables writing distributed Java programs based on MapReduce paradigm.

Hadoop is built on its distributed file system HDFS. HDFS solution is not efficient when the file system is supposed to deal with a huge amount of small files where a small file is the file which is smaller or almost equal to the size of block (default 64 MB). With this in mind, we can state that this solution could not be efficient in the domain of the EEG/ERP because the size of the most of the data files is less than 27 MB in 63 % of cases. Moreover, the remaining rest 37% of other files have different sizes, it means, which means, that if it they were stored in HDFS, most of the files would not fill HDFS blocks appropriately due to its inner fragmentation. Although Hadoop provides some approaches how to deal with small files, neither of them is convenient to be applied in this context. The process of the data generation cannot be changed because the generated files are associated with particular experiments. Similarly, the same problem would occur if we used the method of batch file consolidation. Sequenced files technique, seems to be better, but if we consider that we would like to run a method over the dataset of one experiment, it would not be efficient either because one map operation is run over one data block, where the files of one dataset could be stored anywhere. In general, the solutions based on Hadoop are not appropriate Big Data technology for the domain of the department's EEG/ERP project. Moreover, there is no need to store data in a distributed file system.

3. 4 MATLAB

Matlab provides a number of techniques and approaches to handle Big Data. Although there are tools such as memory mapped variables, disc variables and datastore which enable to deal with problems that occur when a large data has to be loaded into a computer's memory at once, they are considered as a method which provides a scalable solution. This functionality is enabled by Parallel Computing Toolbox along with Distributed Computing Server that provide a scalable distributed solution. This solution enables to run methods for data pre- processing and analysing over a scalable Matlab cluster.

3.5 Evaluation

A Matlab cluster seems to be the best solution for the application of a Big Data approach within the EEG/ERP project. Moreover, the researchers are familiar with this environment, which contains many available methods for data pre- processing and analysis either _rogrammed as Matlab scripts by the researchers or as functions provided by EEGLAB plugin which is widely used among this community.

4. LITERATURE REVIEW

The characteristics of Big Data were first described in 2001, when Laney [4] identified three key attributes of large data amounts: high variety, volume, and velocity. To date, these attributes have become the defining characteristics of Big Data. However, contemporary authors and business specialists enlarged these defining characteristics with further aspects such as dedicated storage, management, and analysis techniques [8], [9], [10]. Further amendments to the definition include the addition of a fourth V, veracity, by IBM [11], emphasizing the aspect of data quality. Taking these different extensions of the original definition into account, we define Big Data as a phenomenon characterized by an ongoing increase in volume, variety, velocity, and veracity of

data that requires advanced techniques and technologies to capture, store, distribute, manage, and analyze these data.

Mark Beyer, Douglas Laney Published on 21 June 2012 "Big data" warrants innovative processing solutions for a variety of new and existing data to provide real business benefits. But processing large volumes or wide varieties of data remains merely a technological solution unless it is tied to business goals and objectives.

Ming Ke, Yuxin Shi published on September 17, 2014 "Big Data, Big Change: In the Financial Management" In recent years, "Big Data" has attracted increasing attention. It has already proved its importance and value in several areas, such as aerospace research, biomedicine, and so on. In "Big Data" era, financial work which is dominated by transaction, business record, business accounting and predictions may spring to life. This paper makes an analysis about what change that "Big Data" brings to Accounting Data Processing, Comprehensive Budget Management, and Management Accounting through affecting the idea, function, mode, and method of financial management. Then the paper states the challenges that "Big Data" brings to enterprise aiming to illustrate that only through fostering strengths and circumventing weaknesses can an enterprise remain invincible in "Big Data" era.

Aicha Ben Salem, Faouzi Boufares, Sebastiao Correia published on April 2014 "Semantic Recognition of a Data Structure in Big-Data" In fact, good governance data allows improved interactions between employees of one or more organizations. Data quality represents a great challenge because the cost of non-quality can be very high. Therefore the use of data quality becomes an absolute necessity within an organization. To improve the data quality in a Big-Data source, our purpose is to add semantics to data and help user to recognize the Big-Data schema. The originality of this approach lies in the semantic aspect it offers. It detects issues in data and proposes a data schema by applying a semantic data profiling

5. CONCLUSIONS

Each second sees a huge amount of data being generated either by human collaboration or by machines which are all around us. The Age of Big Data has come and there is a need to address the challenges which come along with it.

Consequently, the problem of Big Data is widely discussed; many books and journals have been published to address its challenges, definitions and recommendations on how to deal with it. Moreover, the terms such as privacy, security and ethical problems are also considered.

Although Big Data is a frequently discussed topic in theoretical manners, there is deficiency in publications and sources dedicated to its practical usage. Nevertheless, there are some evolving technologies such as Apache Hadoop along with its ecosystem. This technology is considered as the first open-source and widely used Big Data technology, building upon a distributed filesystem and an implementation of MapReduce paradigm. Most of the other technologies dedicated to deal with Big Data are based on the Hadoop solution. Although Hadoop is often discussed as a universal Big Data platform, it cannot address all Big Data problems. There are still some available solutions which are not based on Hadoop such as Matlab, a system which provides a different approach of a cluster computation run over a shared filesystem.

Furthermore, it is convenient to mention where Big Data is stored and how much it is available. Although there are many possibilities on how to access Big Data of various types, biomedical data is mostly not much available. There is only one publicly opened biomedical Big Data database which is intended to preserve genomes. On the other hand, due to its big volume, there is considerable quality variety among publications where the database is described. Other available databases cannot compete with the volume of the genome database. However, they are still characterised by some qualities which should be considered. (consider revising sentence to make it clearer) Nevertheless, due to its domain, availability, many publications and information within the project, the database of the EEG/ERP project was evaluated as the most suitable one.

Depending on the EEG data characteristics, which were obtained by the deep database analysis, Matlab solution was evaluated as suitable technology for application on EEG data. Additionally, Matlab is widely used within the EEG project for the data processing and its programs can be deployed over a Matlab cluster. Although the Computer Science and Engineering Department does not possess all required licences, there is the possibility to use the project Metacentrum where all licences are available along with countless hardware resources.

To confirm that a Matlab cluster is a solution which can be an asset for the EEG/ERP project, a model which allows the running of a Matlab program over multiple data on a Matlab cluster was created. The model was tested by performing two use case, where EEG signals were either divided on epochs or filtered over two experiment datasets. This testing has shown that the model is functional and can be considered as an asset for the EEG/ERP project. Additionally, a few recommendations are proposed on how the project can be improved further.

REFERENCES

[1] **EMC Education Services.** *Data Science & Big Data Analytics.* Indianapolis : John Wiley & Sons, 2015. 978-1-118-87613-8.

[2] **Splunk Inc.** Machine Data. *Splunk.* [Online] Splunk Inc., 2005-2016. [Cited: March 8, 2016.] http://www.splunk.com/en_us/resources/machine-data.html.

[3] **Mayer-Schönberger, Viktor and Cukier, Kenneth.** *A Revolution That Will Transform How We Live, Work and Think.* New York : Houghton Mifflin Harcourt Publishing Company, 2013. 978-0-544-00269-2.

[4] **Krishnan, Krish.** *Data Warehousing in the Age of Big Data.* San Francisco : Morgan Kaufmann Publishers, 2013. 9780124059207.

[5] **Manyika, James and Chui, Michael.** *Big data: The next frontier for innovation, competition, and productivity.* s.l. : McKinsey Global Institute, 2011. 978-0983179696.

[6] **Berman, Jules J.** *Principles of Big Data.* Boston : Elsevier, 2013. 9780124045767.

[7] **Iafrate, Fernando and Front, Matter.** *From Big Data to Smart Data.* Chap : John Wiley & Sons, 2015.

[8] **Jagadish, H. V., Gehrke, Johannes and Labrinidis, Alexandros.** Big Data and Its Technical Challenges. *Communications of the ACM.* Month, 2014, Vol. 57, 7.

[9] **Hurvitz, Judith, Kaufman, Marcia and Bowles, Adrian.** *Cognitive Computing and Big Data Analytics.* Hoboken : John Wiley & Sons,, 2012. 978-1- 118-89662-4.

[10] **Minelli, Michael, Chambers, Michele and Dhiraj, Ambiga.** *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses.* s.l. : John Wiley & Sons, 2013. 9781118562260.

[11] **Borthakur, Dhruba.** HDFS Architecture Guide. *Hadoop.apache.org.* [Online] The Apache Software Foundation, 8 4 2013. [Cited: 19 April 2016.] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Data+Organization.

[12] **The Apache Software Foundation.** Hadoop. *Hadoop.* [Online] The Apache Software Foundation. [Cited: 15 May 2016.] http://hadoop.apache.org/.

[13] Spark. *Spark.* [Online] The Apache Software Foundation. [Cited: 22 May 2016.] http://spark.apache.org/.

[14] **Zaharia, Matei; Chowdhury, Mosharaf; Das, Tahagata; Dave, Ankur; Ma, Justin; McCauley, Murphy; Franklin, Michael; Shenker, Scott; Stoica, Ion;.** *Resilient Distributed Datasets: A Fault-Tolerant.* Berkeley: University of California at Berkeley, Electrical Engineering and Computer Sciences, 2011.

[15] **MathWorks.** Makers of MATLAB and Simulink. *Mathworks.com.* [Online] [Cited: 24 April 2016.] http://www.mathworks.com/.

[16] **Holubová, Irena, a další.** *Big Data a NoSQL databaze.* Praha : Grada, 2015. 978-80-247-5466-6.

[17] **Rigden, Daniel J. , et al.** The 2016 database issue of Nucleic Acids Research. *Nucleic Acids Research.* 2016,, Vol. 44, D1–D6.

[18] **Pennisi, Elizabeth.** 1000 Genomes Project Gives New Map of Genetic Diversity. *Science.* 2010, Vol. 330, 6004.

[19] *Citizen Science: The Law and Ethics of Public Access to Medical Big Data.* **Hoffman, Sharona.** 3, Cleveland : Case Western Reserve University School of Law, 2014, Vol. 30.

[20] **The GovLab.** The Open Data Era in Health and Social Care. *GOVLAB.* [Online] May 2014. [Cited: 5 June 2016.] http://images.thegovlab.org/wordpress/wp-content/uploads/2014/10/nhs-full- report-21.pdf.

[21] **Bydžovský, Martin.** *Relational and non-relational modeling for portal of electrophysiological experiments.* Pilsen : University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, 2014.

[22] **Řeřicha, Jan.** *Software tool for management of neuroinformatics data.* Pilsen : University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, 2013.