# COMPARISION OF TOOLS AND TECHNIQUES OF BIG DATA

PREETI[1], Dr. CHHAVI RANA[2]

STUDENT UIET MDU ROHTAK[1]

ASSISTANT PROFESSOR UIET MDU ROHTAK[2]

**ABSTRACT:-**
Big data is used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. There are various challenges in big data. In this, we use a framework of map reducing using hadoop. MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Hadoop is an open-source software framework for distributed storage and distributed processing of big data on clusters of commodity hardware.

**KEYWORDS:-** Big Data, Hadoop, Matlab, cluster, distributed computing, EEG/ERP.

## 1. INTRODUCTION

Big Data refers to the efficient handling of large amount of data that is impossible by using traditional or conventional methods such as relational databases or it is a technique that is required to handle the large amount of data that is generated with advancements in technology and increase in population. Big data helps to store, retrieve and modify these large data sets. For example with the advent of smart technology there is rapid increase in use of mobile phones due to which large amount of data is generated every second, so it is impossible to handle by using traditional methods hence to overcome this problem big data concepts were introduced most analysts and practitioners currently refer to data sets from 30-50 tera bytes (10 12 or 1000 gigabytes per terabyte) to multiple peta-bytes (1015 or 1000 terabytes per peta-byte) as big data. Figure No. 1.1 gives Layered Architecture of Big Data System.

Big data is a new data challenge that requires leveraging existing systems differently. It is classified in terms of 4Vs – Volume, Variety, Velocity and Veracity.

4 Vs of Big Data:-

A. Data Volume (Data in rest)

Data volume refers to the amount of data. At present the volume of data stored has grown from megabytes and gigabytes to peta-bytes and is supposed to increase to zeta-bytes in nearby future.

B. Data Variety (Data in many forms)

Variety refers to the different types of data-– text, images video, audio, etc and sources of data. Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail,

documents. In other words, we all know that data is being generated at a very fast pace. Now, this data is generated from different types of sources, such as internal, external, social, and behavioral. Even a single source can generate data in varied formats.

C. Data Velocity (Data in motion)

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows and aggregated. In other words, Velocity describes the rate at which data is generated, captured and shared. Enterprises can capitalize on data only if it is captured and shared in real time. The sources of high velocity data include the following: IT devices, routers, switches, firewalls etc, constantly generate valuable data.

D. Data Veracity (Data in Doubt)

Veracity generally refers to the uncertainty of data, i.e. whether the obtained data is correct or consistent. Out of huge amount of data that is generated in almost every process, only the data is correct and consisted can be used for further analysis. Data when processed becomes information; however, a lot of effort goes in processing the data. Big Data, especially in the unstructured and semi-structured forms, is messy in nature.

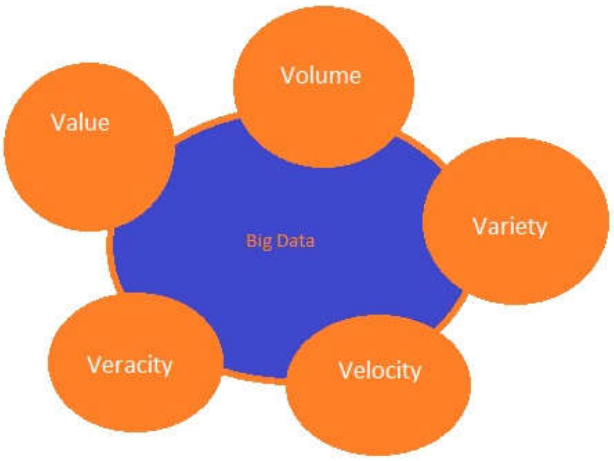


**Fig. :- Characteristics of Big Data**

## 2. Characteristics of Big Data

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data [9]. Data volume is the primary attribute of big data.

## 3. LITERATURE REVIEW

**1. Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey:-** Their survey presents the concept of *Big Data*. Firstly, definition of Big Data and then its features are describe. Secondly, how Big Data is processing with step by step and the main problems encountered in big data management are described. After that a basic overview of an architecture for handling is illustrated. Then, a problem is discussed which is already exist in information system about merging Big Data architecture . Finally their survey tackles semantics in the *Big Data* context. **[1]**

**2. Nada Elgendy, Ahmed Elragal, Big Data Analytics: A Literature Review Paper:-**In today's statistical era, huge amounts of data have become available on hand to decision makers. Big data refers to those datasets which are not only big but also have high variety and velocity that's why they face difficult to handle with traditional tools and techniques. Because of the rapid growth of such data, there is need to study about solutions and provide in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gather valuable awareness about such varied and rapidly changing data which is happen because of daily transactions of customer interactions and social network data. This can be provided by using big data analytics which is the application of advanced analytics techniques on big data. Aim of their paper is to analyze the tools and some of the different analytics methods which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains. **[2]**

**3. Bo Li, Prof. Raj Jain, Survey of Recent Research Progress and Issues in Big Data:-** The term "Big data" is used for large and complicated data sets which make difficult to process using traditional data management tools or processing applications. Their paper tells about recent progress on big data networking and big data. They have categorized reported efforts into four general categories. First, efforts related to classic big data technology such as storage, Software-Defined Network, data transportation and analytics are reported. Second, important aspects of big data in cloud computing such as recourse management and performances optimization are introduced. Lastly, they introduce interesting benchmarks and progress in both search engines and mobile networking. With the help of detailed summary and analysis, limitations of the proposed works and possible future research directions have been proposed. **[3]**

**4. Samiddha Mukherjeet, Ravi Shaw, Big Data-Concepts, Applications, Challenges and Future Scope:-**The term, "Big Data" has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. Big Data is still a novel concept, and in the following literature we intend to elaborate it in a palpable fashion. It commences with the concept of the subject in itself along with its properties and the two general approaches of dealing with it. The comprehensive study further goes on to elucidate the applications of Big Data in all diverse aspects of economy and being. The utilization of Big Data Analytics after integrating it with digital capabilities to secure business growth and its visualization to make it comprehensible to the technically apprenticed business analyzers has been discussed in depth. Aside this, the incorporation of Big Data in order to improve population health, for the betterment of finance, telecom industry, food industry and for fraud detection and sentiment analysis have been delineated. The challenges that are hindering the growth of Big Data Analytics are accounted for in depth in the paper. This topic has been segregated into two arenas- one being the practical challenges faces whilst the other being the theoretical challenges. **[4]**

**5.M.Dhavapriya, N.Yasadha, Big Data Analytics: Challenges and Solutions Using Hadoop, MapReduce and Big Table:-** We live in on-demand, on-command Digital universe with data prolife ring by Institutions, Individuals and Machines at a very high rate. This data is categories as "Big Data" due to its sheer Volume, Variety, Velocity and Veracity. Most of this data is unstructured, quasi structured or semi structured and it is heterogeneous in nature. The volume and the heterogeneity of data with the speed it is generated, makes it difficult for the present computing infrastructure to manage Big Data. Traditional data management, warehousing and analysis systems fall short of tools to analyze this data. Big Data has specific nature that's why it is stored in distributed file system architectures.

Hadoop and HDFS by Apache is widely used for storing and managing Big Data. To analyze the Big Data is a challenging task with it's large distributed file systems which should be fault tolerant, flexible and scalable. MapReduce has been used for the efficient analysis of Big Data. For classification and clustering of Big Data, traditional DBMS techniques like Joins and Indexing and graph search is used . These techniques are being a dopted to be used in MapReduce. In this research paper the authors suggest various methods for catering to the problems in hand through MapReduce framework over Hadoop Distributed File System (HDFS). MapReduce technique is used for file indexing with mapping, sorting, shuffling and finally reducing. MapReduce techniques have been studied at in this paper which is implemented for Big Data analysis using HDFS. [5]

## 4. Implementation

Keeping in mind the end goal to coordinate these five advances (Matlab, EEGLAB, BiosigToolbox and FileToolbox among each other in the earth of Metacentrum, it is important to get comfortable with its design.

### 4.1 Metacentrum Architecture

The engineering of Metacentrum comprises of three principle parts: frontends, PBS (Portable Batch System)/Torque servers, figuring hubs and circle stockpiles. A portion of the hubs can be virtual ones, as such that there is one physical machine where are some virtual machines keep running on. The frontends are machines used to clients to sign into the framework straightforwardly without a reservation. Especially, the frontends hubs give the passage purpose of the Metacentrum framework for clients to set up their assignments (employments) to run, check the activity genuine state and to get created comes about. PBS/Torque servers care for employments planning, for example, work need calculation, booking occupations (placing them into lines), doling out assets and giving data about genuine conditions of occupations. Figuring hubs are the end gadgets where the employments are at last run. They are resolved to use non-intuitive errands which are conveyed to them by a scheduler framework. They are put together with plate stockpiles in a few urban areas all through Czech Republic, for example, (Pilsen, Prague, Liberec, České Budějovice, Jihlava, Brno, Olomouc and Ostrava). Circle stockpiles are put away inside specific bunches in the urban areas said above to give elite without having bothersome deferrals. Also, all stockpiles are accessible on all frontend hubs.

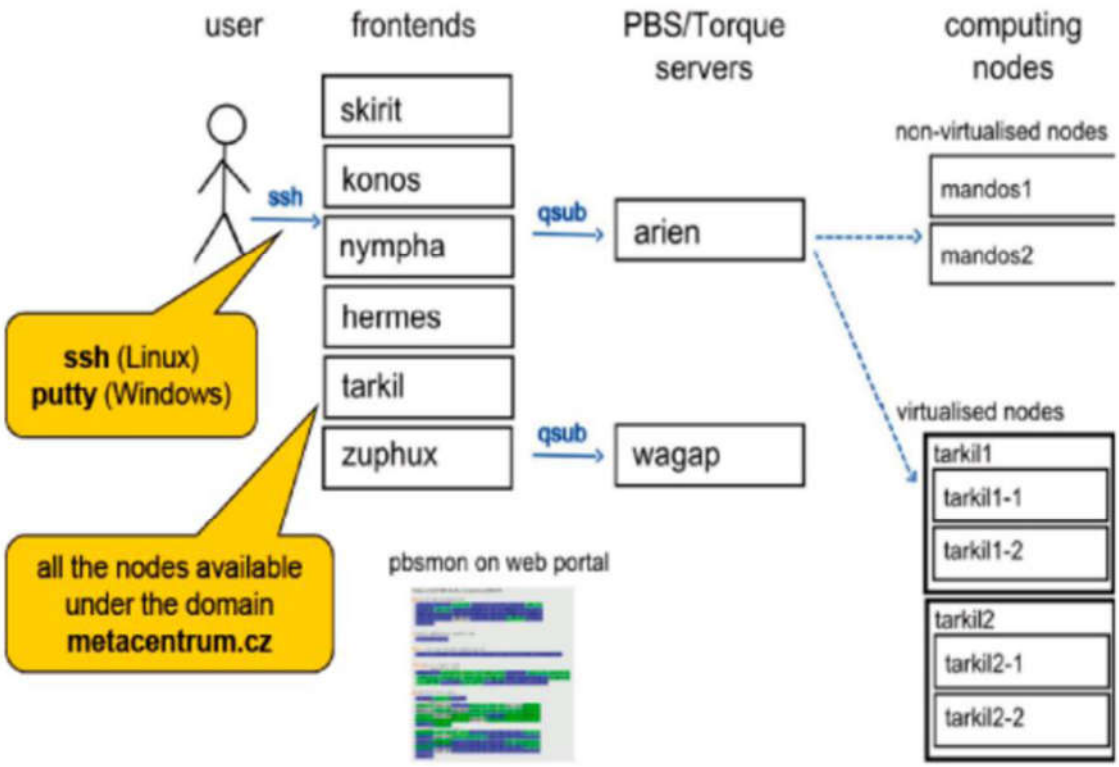


**Fig. - 1 Metacentrum Architecture**

**4.1.1 Scenario**

As it is portrayed in Figure 1 , the run of the mill movement work process is made out of a couple of exercises:

1. Sign in – clients sign into the framework by means of a ssh (secure shell) customer to one of accessible frontend hubs.

2. An occupation readiness – in this stage, the client readies his business to run, he appoints different parameters to the activity and submits it to the scheduler through an order called qsub.

3. Arranging – the scheduler designs the activity as per its need and asset necessities.

4. Running – after occupation's holding up cycle, it is kept running on specific hubs picked by the scheduler as indicated by the client's prerequisites.

5. Completed occupation – when the activity is done the client can be educated through an email or he can check its real state by means of qstat summon from his frontend hub where he is signed in or from Metacentrum sites.

**4.1.2 Connection and Access to Sources of Metacentrum**

Bunches, frameworks, servers and hubs are associated by a system which most extreme exchange speed is up to 10 GB/s. To get to machines of Metacentrum ssh convention is utilized to associate with a frontend hub terminal which keeps running on GNU/Linux. Ordinarily, it is prescribed to interface by means of an application called PuTTy on the off chance that one is utilizing MS Windows stage or by means of ssh order on GNU/Linux stage. Validation and security is acknowledged by a verification framework called Kerberos.

To exchange documents between a client and Metacentrum's stockpiling a convention called scp is commonly utilized. For MS Windows stage, programming called Winscp is suggested while for GNU/Linux is FilleZilla.

**4.1.3 Torque Scheduler**

Because of the reality, that Metacentrum is utilized by numerous clients, employments are run by means of a scheduler framework called Torque which guarantees a reasonable way to deal with the sources assignments to Metacentrum's clients. It is a PBS scheduler which empowers you to run assignments in two unique modes and with various undertakings prerequisites. These modes are characterized by the client's point of view where the first permits a cooperation (intuitive mode) while the later one is kept running on the foundation (non-intelligent mode). Moreover, the scheduler gives a meaning of undertakings' prerequisites, for example, execution, area, evaluated time to process the assignment, and so forth. Keeping in mind the end goal to serve numerous clients, Torque needs to manage the administration of sources portion to give a reasonable sources task among the group clients. Torque is the scheduler that is utilized all through all groups of Metacentrum. An essential term identified with Torque is known as an occupation. Because of the reality, that there are two methods of the association, when a client needs to make work, he needs to decide whether he needs to make an intelligent or non-intuitive (bash) work. On the off chance that he picks a bash work, after he presents the activity by a specific summon, the bash work is allocated to one of client's accessible lines, which was indicated by the client previously. The activity is holding up in the server line on a reasonable time to be run. Then, PBS is settling on choices relying upon the genuine condition of accessible assets and clients' needs. At last, the activity is kept running when there are important occupation assets accessible and furthermore when the activity has achieved the significant need. As it was specified above, occupations are holding up in lines. The lines are sorted by the most extreme time term of employments to be done. Especially, there are eight gatherings: 2h, 4h, multi day, 2 days, multi week, 2 weeks and over 2 weeks.

To empower client to deal with an occupation, PBS gives an arrangement of orders where the fundamental ones are associated with:

- submitting employment to the line,
- cancellation of a pausing or a running employment,
- receiving data about the present hub state and its properties,
- appearing graphical diagram of lines and employments, and so on.

Another arrangement of summon is considered to the necessities particular for computational assets. Necessities for the properties and assets are determined by exceptional imprints which can be a piece of the summon of work submit. These settings enable the client to set up the accompanying errands:

- the time line where the activity will be kept running in,
- how numerous hubs the activity needs and their compose,
- how numerous CPUs must be apportioned,
- which specific hubs or their physical area,
- required measure of the required physical memory,
- temporary capacity with short access span for computational purposes,
- required programming licenses,
- other choices, for example, the activity framework, area in the group, organizing cards and the record framework.

There is a work process of running a vocation:

- A client sign in the framework by means of ssh convention to one of frontend hubs,

- He runs a vocation by presenting an order called qsub which empowers to set up hotspots for a specific time and passes it to a scheduler,

- After some season of sitting tight for the sources, the scheduler appoints the required sources (machines, processors, memory, licenses) to the activity.

There are four primary sorts of occupation states:

- Q – The activity is lined,

- R – The activity is running,

- E – The activity is leaving in the wake of running,

- C – The activity is finished in the wake of running.

The need of a vocation is related with a few principles and there is a method of the need foundation. Initially, employments are arranged in by the accompanying criteria:

- The need of the line – occupations in lines with the higher need have favorable position,

- Fair share – occupations from the client who has invested less energy by handling employments Has leeway,

- Job time – the activity which has the minimum time prerequisite has preference.

Since the employments are arranged by the past advances, the scheduler experiences the arranged occupations and chooses if the activity is allowed to run in regards to the sources requested and which assets are accessible with thought of the ideal area..

**4.1.4 Data**

Metacentrum utilizes a common record framework to ensure a strong approach to deal with information all through the bunch which guarantees that a client can get to his information notwithstanding to the area where it is put away and furthermore gives a speedier stockpiling considered to meet information requests of being put away on a gadget that isolates negligible deferrals while employments are being processed.

There are to two principle sorts of the capacity:

- Storage – it comprises of shared and went down NFSv4 volumes accessible from all the frontend and in addition all the registering hubs. Plate exhibits are identified with the area of the specific bunches where they are physically set close-by to guarantee the least dormancy for the most ideal execution.

- Home – it is a mutual file system devoted for clients' home organizers.

- Scratch – the scratch is a sort of capacity giving the quickest gadgets where information can be put away.Primarily, it is devoted for applications' worldly/working information. This neighborhood volume is accessible on all registering hubs all through the entire bunch. Despite the fact that this stockpiling does not give back-ups, it has fundamental assurance of RAID 10 to keep away from equipment disappointments.

**4.2 Matlab with Distributed Computing Server**

Matlab permits run programs on a bunch. It depends on the ace slave design where one hub/occasion is the ace and different hubs are slaves. A program resolved to keep running on a Matlab Distributed Computing Server is deteriorated on undertakings which are parts of an occupation. The activity's part is to speak to an arrangement of errands as one question be taken care of by a scheduler. Undertakings are disseminated to accessible hubs of the bunch. The ace runs just a single example of Matlab to deal with the correspondence among slaves' hubs. Assignments are conveyed to accessible hubs of the group. There are two employment classifications:

- Distributed work – an occupation comprised of errands which don't speak with themselves. There is no requirement for the correspondence and synchronization. The undertakings are autonomous on each other.

- Parallel work – it is the inverse of the circulated work. For this situation, undertakings should be synchronized and impart among the activity setting to get an outcome with is come to by their collaboration.

**4.2.1 Architecture**

The Matlab session in which the activity and its assignments are characterized is known as the customer session. All customers getting to a Matlab Cluster have their own session/setting in which their projects run. The session contains meaning of ways, occupations and errand. The inward Matlab Job Scheduler (MJS) conveys the undertakings for assessment to the server's individual sessions called laborers. The engineering is portrayed in Figure 3. To be capable run the customer session, the client needs the permit called Parallel Computing Toolbox. So also, the client needs to have different licenses for running errands on works which is overseen by Distributed Computing Server. The licenses' composition is appear in Figure 2.

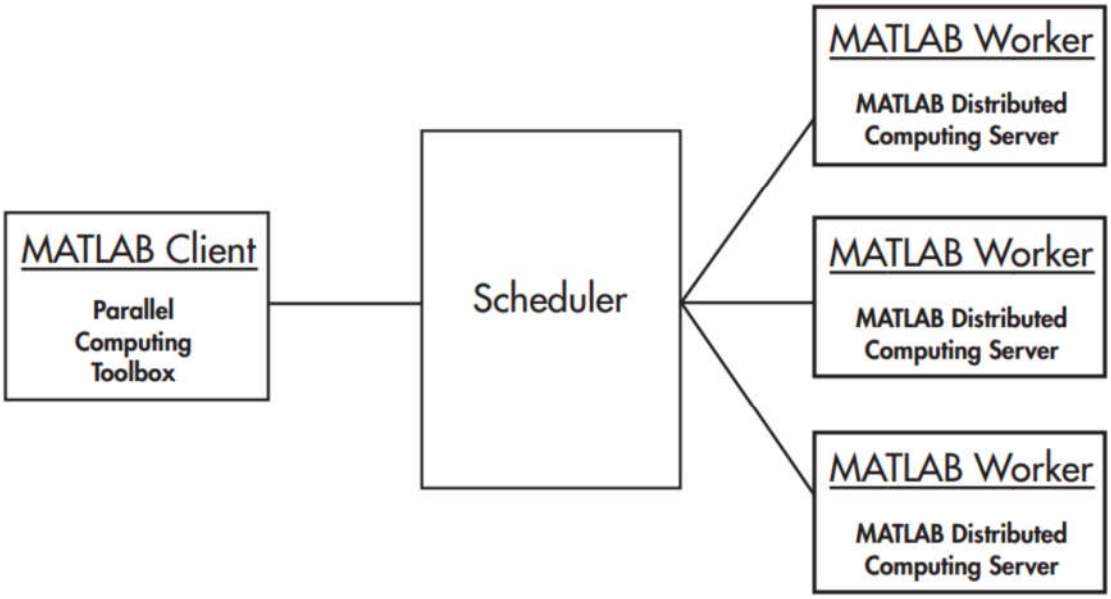


**Fig. 2 - Licences' Schema**

**4.2.2 Matlab Job Scheduler**

MJS is the piece of the Matlab Distributed Computing Server that arranges the execution of occupations and the assessment of their assignments. MJS scheduler runs occupations in the request in which they are submitted, except if any employments in its line are advanced, downgraded, dropped, or obliterated. Also, if there is a need, there is a plausibility to utilize MJS in a collaboration with another scheduler of an outsider. The part of the scheduler and a vocation work process is delineated in Figure 3.

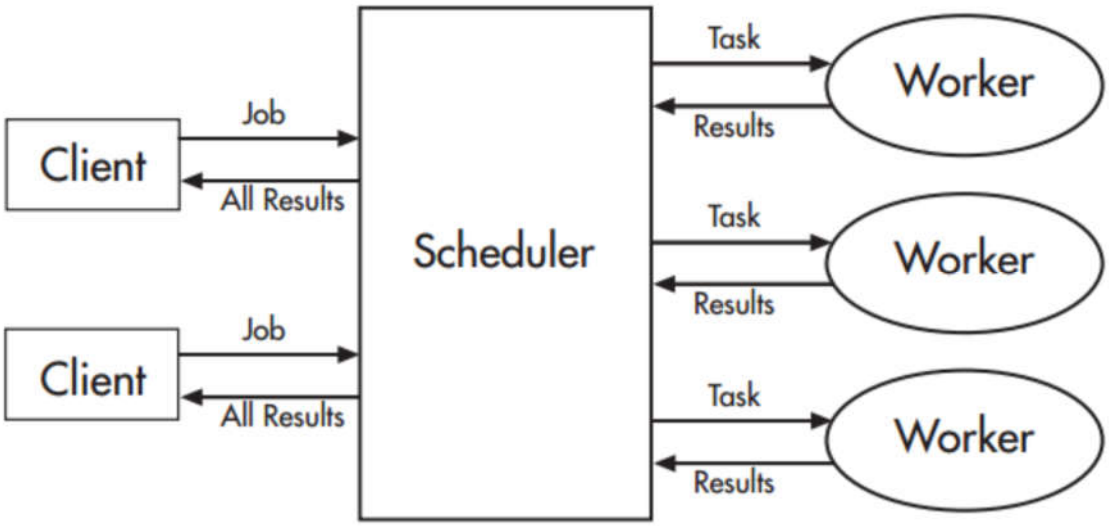


**Fig. 3 - The Role of The Scheduler and The Job Workflow**

**4.2.3 Preparation**

Right off the bat, a client needs to make a Matlab work which he might want to keep running over different information (UserFunction.m). To run it over a Matlab bunch, it is important to make another Matlab record which speaks to the customer session (ClientSession.m). This document contains settings of the Matlab bunch and the deterioration of the application onto Matlab employments and undertakings (see 7.3.2 Matlab with Distributed Computing Server). For this situation, we consider to have the application decayed of only one occupation which comprises of various undertakings.

To run an application over a Metacentrum group, there is a need to make an occupation where the Matlab employment and assignment will run. The activity is made by presenting a bash content (Script.sh) to the Metacentrum's assignment scheduler Torque. The bash content contains Torque mandates to characterize the activity's parameters, settings, and the product which should be kept running on the bunch (Matlab), and to allot equipment assets which are required for the running of the program, for example, the memory sum, number of processors and hubs. Besides, the content needs to likewise characterize required programming licenses, for example, one permit of Parallel Computing Toolbox for the Matlab session and Distributed Computing Server licenses for Matlab laborers, whose check relies upon the quantity of the CPUs where the disseminated application should be kept running on.

**4.3 Testing**

In order to test the model's appropriate functionality, two different datasets were chosen. The first dataset is determined to test the model functionality over a dataset which is characterised by a huge amount of small files. The second dataset is used to test the appropriate functionality over the few largest files of the EEG/ERP database.

- P300 Components Dataset – a dataset determined for finding P300 components. This dataset consists of 247 measurements recorded in BrainVision Format with an overall size of 1 GB.

- Driver's Attention Dataset – a subset of measurement which is characterised by the large size of the files. There are 14 EEG records with a total size of 3 GB.

The following Table 1 and Table 2 represent the hardware configurations and the elapsed times of the performed use cases which were run over the datasets.

| Attribute | Setting 1 | Setting 2 | Setting 3 | Setting 4 | Setting 5 | Setting 6 | Setting 7 | Setting 8 | Setting 9 |
|---|---|---|---|---|---|---|---|---|---|
| Number of nodes | 1 | 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Number of CPUs | 4 | 8 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| Total available memory [GB] | 10 | 10 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| Location | X | X | X | Brno | Brno | Brno | Brno | Brno | Brno |
| Number of workers | 3 | 7 | 15 | 31 | 47 | 63 | 79 | 80 | 83 |
| Elapsed time – P300 Dataset [Min] | 34 | 17 | 14 | 9 | 9 | 8 | 8 | 8 | 8 |
| Elapsed time – Driver's attention Dataset [Min] | 9 | 6 | 6 | 7 | 6 | 8 | 8 | 8 | 8 |

**Table 1: Dividing on Epochs - Hardware Configuration**

| Attribute | Setting 1 | Setting 2 | Setting 3 | Setting 4 | Setting 5 | Setting 6 | Setting 7 | Setting 8 | Setting 9 |
|---|---|---|---|---|---|---|---|---|---|
| Number of nodes | 1 | 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Number of CPUs | 4 | 8 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| Total available memory [GB] | 10 | 10 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| Location | X | X | X | Brno | Brno | Brno | Brno | Brno | Brno |
| Number of workers | 3 | 7 | 15 | 31 | 47 | 63 | 79 | 80 | 83 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Elapsed time – P300 Dataset [Min] | 84 | 31 | 18 | 12 | 11 | 10 | 10 | 10 | 10 |
| Elapsed time – Driver's attention Dataset [Min] | 187 | 46 | 31 | 46 | 42 | 35 | 41 | 42 | 47 |

**Table 2 : Signal Filtering - Hardware Configuration**

As depicted in Figure 3 and Figure 4, the distribute computation of SPMD seems to be more efficient in the case of the signal filtering than of the cutting on the epochs, due to the higher complexity of the signal filtering method which can be seen in the considerable differences between curves of these two methods in Figure 5 .

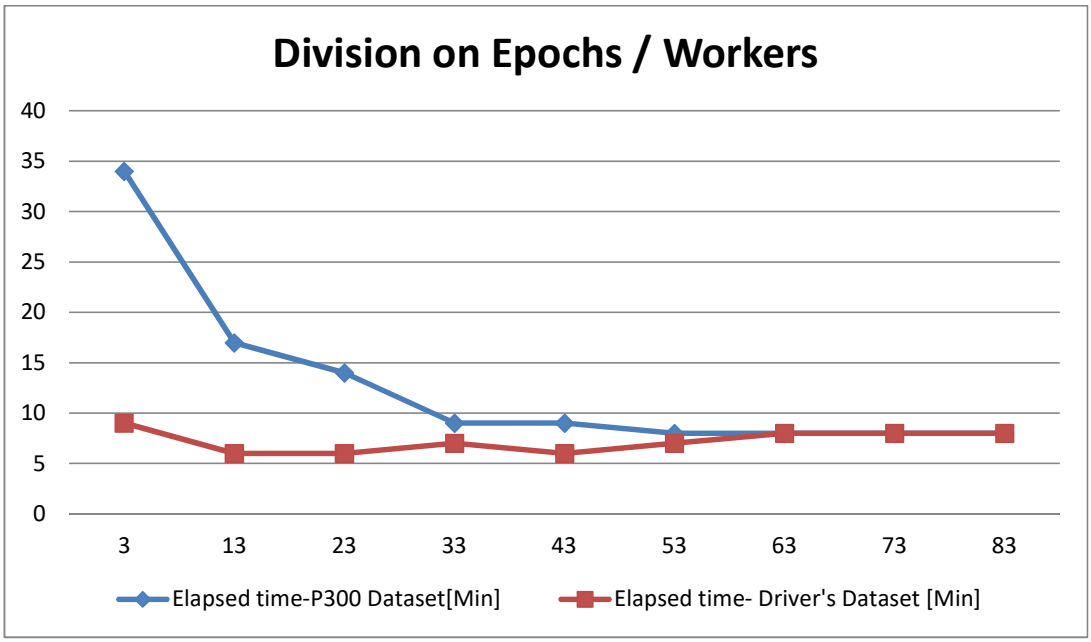


**Figure 3 : Division on Epochs / Workers**

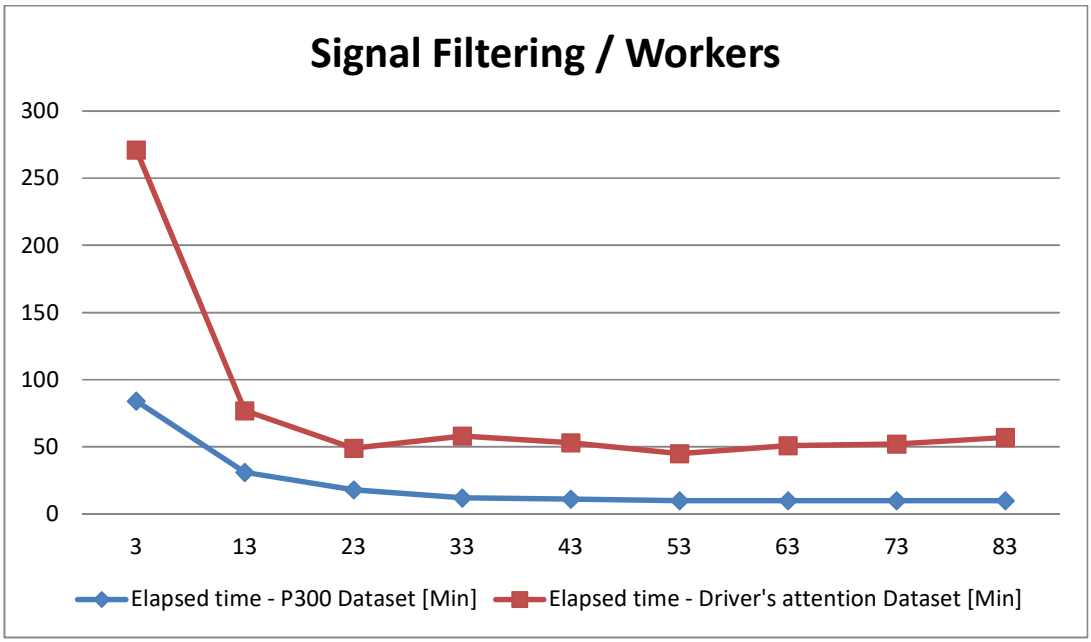


**Figure 4 : Signal Filtering / Workers**

Additionally, Figure 4 and Figure 5 depict the processing dependency on the particular datasets. A huge amount of small files tends to have more deterministic characteristic than the small amount of large files of Driver's Attention Dataset. As seen in Figure 5, the ideal setting for the best efficiency of processing of the Driver's Attention dataset is to use the same amount of workers as the count of the dataset's files which is 13.

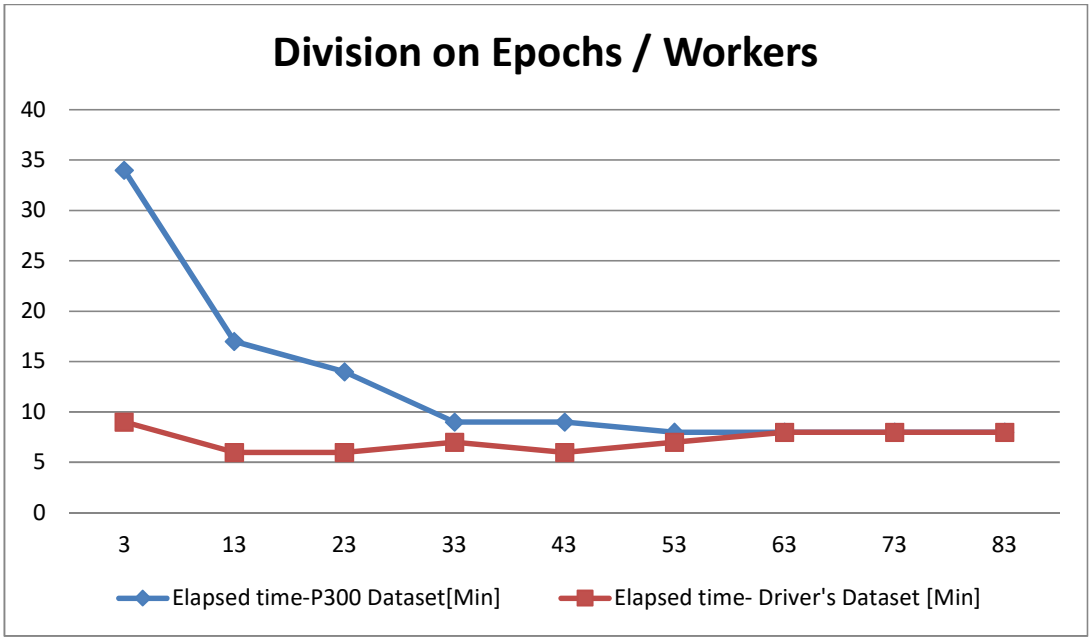


**Figure 5 : Signal Processing / Workers - P300 Dataset**

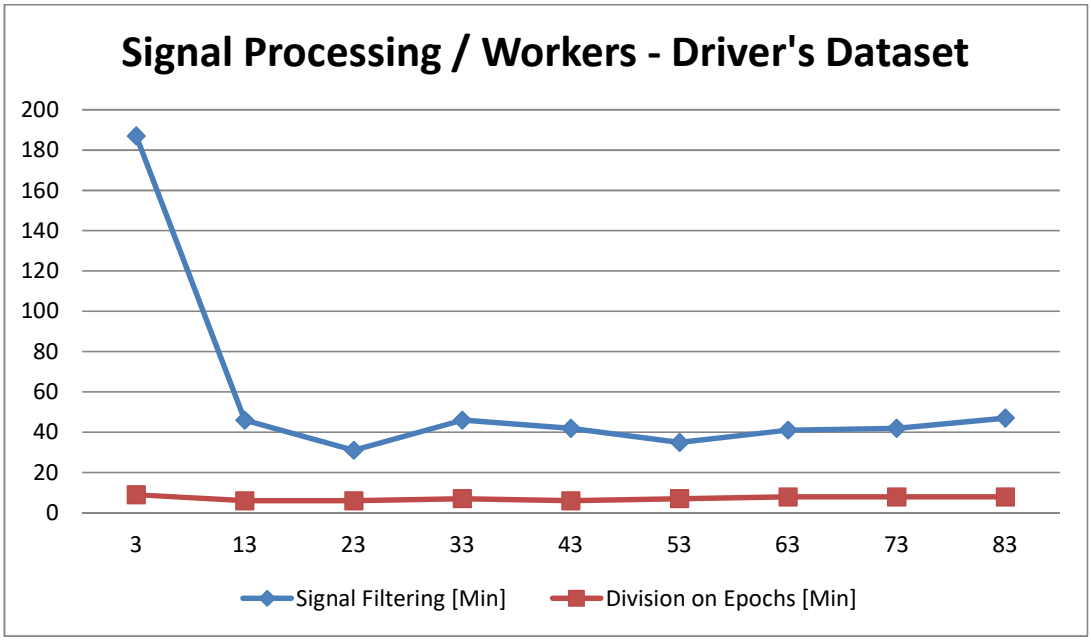


**Figure 6 : Signal Processing / Workers - Driver's Dataset**

Although the files of epochs, filtered signals and their configurations were created from all measurements without any occurred errors, when the computation is distributed over more nodes, the location of each node has to be equal, otherwise the computation ends up with an error. Similarly, the same problem occurs when non-existent nodes are required. These problems have been discussed with the support of Metacentrum, likely the problems are caused by machine virtualisations along with the access to the file shared system where many technologies have to be integrated into each other.

## 5. Results

As the consequence of the testing, it can be led that the created demonstrate is completely utilitarian. Subsequently, the incorporation of these four advancements Matlab, EEGLAB, Biosig tool stash and FileIO tool stash can be conveyed together finished a Matlab group. This model can be a major resource for flag pre-handling and examination inside the space of the EEG/ERP venture. The principle resource is considered to having an adaptable answer for performing different strategies over huge datasets of information. With everything taken into account, a technique for investigation of Big Data has been connected in the EEG/ERP area.

## 6.CONCLUSION

This study has shown you Big Data definition and its usage and its future challenges . We are living in the era of data deluge. The term *Big Data* had been coined to describe this age. This paper defines and characterizes the concept of *Big Data*. It gives a definition of this new concept and its characteristics. In addition, a supply chain and technologies for *Big Data* management are presented. During that management, many problems can be encountered, especially during semantic gathering. Thus it tackles semantics (reasoning, coreference resolution, entity linking, information extraction, consolidation, paraphrase resolution, ontology alignment) with a zoom on “V’s”

## 7. REFERENCE:-

**[1].** Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey, Univ. Bourgogne Franche-Comte, France, 2015.

[2]. Nada Elgendy, Ahmed Elragal, Big Data Analytics: A Literature Review Paper, Article in Lecture Notes in computer science, 2014.

[3]. Bo Li, Prof. Raj Jain, Survey of Recent Research Progress and Issues in Big Data, 2013.

[4].Samiddha Mukherjeet, Ravi Shaw, Big Data-Concepts, Applications, Challenges and Future Scope, IJARCCE,2016.

[5]. M.Dhavapriya, N. Yasadha, Big Data Analytics: Challenges and Solutions Using Hadoop, MapReduce and Big Table, IJCST, 2016.

[6]. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, A Review Paper on Big Data and Hadoop, IJSRP, 2014.

[7]. Varsha B.Bobade, Survey Paper on Big Data and Hadoop, IRJET, 2016.

[8].Bijesh Dhyani, Anurag Barthwal, Big Data Analytics Using Hadoop, International Journal of Computer Applications, 2014.

[9]. Sulochana Panigrahi, S Mohan Kumar, A Survey on Social Data Processing Using Apache Hadoop, MapReduce, IJSTA, 2016.

[10]. Ashwini A. Pandagale, Anil R. Surve, Big Data Analysis Using Hadoop Framework, IJRAR, 2016.

[11]. Ms. Tarunpreet Chawla, Mr. Lalit, A Review Paper on MapReducing Using Hadoop, IJRTER, 2016.