International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
العدد 28
Volume 28

Using Technology and components for Big data Hadoop Framework

Khaled Elabbani[1], Ahmed Jafar[2] , WaleedKhalafullah[3] and Salem Alfrjani[4]

Collage of Computer Technology Benghazi

elabani1980@yahoo.com[1] , ahmed.s.jaffar@gmail.com[2]

waleed.younus.kh@gmail.com[3], Salemteacher15@gmail.com[4]

الملخص

بعد ظهور الحوسبة السحابية ومنصات التواصل الاجتماعي وزيادة عدد المستخدمين على شبكة الانترنت، ازداد حجم البيانات بكافة انواعها (هيكلية و غير هيكلية) بشكل متسارع وكبير ، مما ادى الى ظهور مصطلح البيانات الضخمة ، حيث اصبحت الشركات والحكومات تحتاج الى تحليل تلك البيانات الضخمة للحصول على معلومات مفيدة تساعد في أتخاذ القرارات باستخدام تقنية تساعد في تحليل البيانات الضخمة التي من الصعب التعامل معها بالأدوات التقليدية ، الهدف من هذه الورقة هو التطرق الى نظرة عامة على مفهوم البيانات الضخمة والادوات التي تساعد في تحليل البيانات.

.

Abstract:

Data has grown rapidly after the advent of social networks and cloud computing, with the difficulty of analyzing that data. The term Big Data appeared to work with all types of data, whether it is structured or unstructured. Companies and governments need to analyze data to get useful information from the huge amount of data. Big data is difficult to use with traditional database tools. This paper aims to give an overview of the concept of big data and the technology that help in data analysis.

Keywords: Big data, Hadoop, HDFS , Map Reduce

1. Introduction

The modern world is interested in data to obtain knowledge, Data is the main factor for collecting information, which leads to knowledge to make decisions and solve problems. Over the years, the increased flow of data has made it hard to deal with information, most companies are facing challenges from data growth leading to the presence of huge information in servers, and the rapid growth of data has obtained a large amount of data from variety sources and variety of formats called Big Data. Big data is a word that describes itself; Big Data is a type of data that is massive. Big Data is a term that refers to a large volume of data that is rising rapidly over time. In other words, such data is so large and complicated that no standard data management solutions can effectively store or process it [1]. This paper is organized as follows: Section 2 defines Big data; Section 3 describes Big data factors; Section 4 Challenges; Section 5 Tools and Technology; Section 6 Hadoop architecture; Section 7 Hadoop Ecosystem and Section 8 Hadoop components comparison.
.

2. Big Data:

In 2005, 130 Exabyte of data grew and exploded to 1,227 Exabyte in 2010 and data inflation increased by a high rate of 7,910 Exabyte in 2015 according to IDC Digital Universe Study [2].According to projections from Statista 2021, the amount of data generated globally is expected to increase to 64.2 Zettabytes in 2020. In 2025, the amount of data is expected to increase to more than 180 Zettabytes as shown in figure 1. In 2020 due to the COVID-19 pandemic, the volume of data generated is higher than previously expected due to increased usage, as more people are working from home[3].

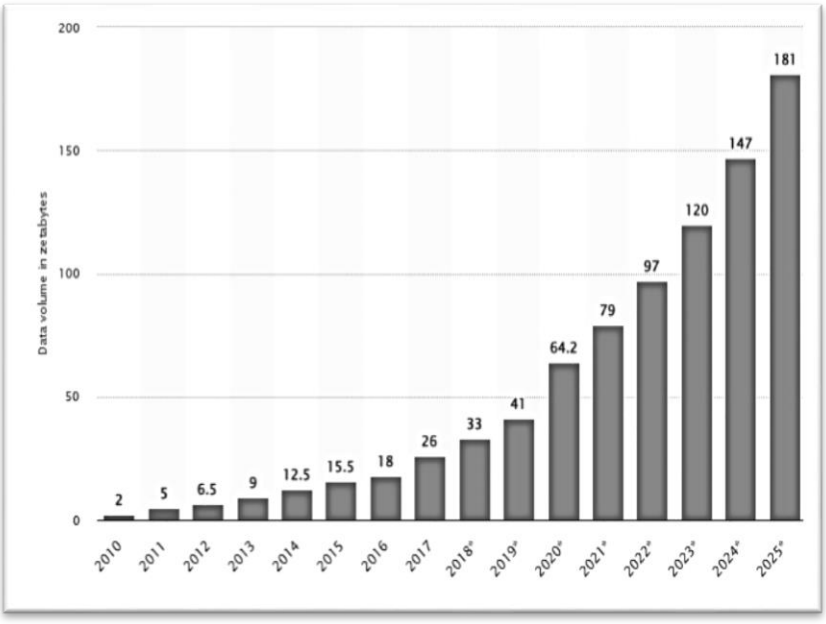

Figure 1: Showing the Increase data volume from 2010 to 2025.

There are thousands of scientific and article papers and millions of web pages that provide information and concepts about big data. Big data comes big thing after Cloud [4]. In many fields such as health, education and earth sciences, it contains a large amount of data that is difficult to deal with and analyze using traditional models and software that need to develop and design efficient computing models for Big data analysis [5]. In healthcare, big data helps make critical decisions and make predictions. For example, Google Flu Trends (GFT) provided a service for influenza trends and provided estimates of influenza activity for more than 25 countries. It also made accurate predictions about influenza activity [6].Big data analysis leads to improving business processes, developing products, providing new services and creating new businesses, as companies face difficulty in exploiting data that affects decision-making and cost savings[7].

Data storage, modification and retrieving data are the essential operations in data management. In contrast, Big data is diverse and irregular with a lack of clarity and needs data processing and speed.

Access, Assembly, Analyse and Action are known as the stages of Big data analysis. Big data requires modern and advanced analytical techniques because Big data analysis is not an easy method. Big data analysis needs tools and storage capabilities in a way that enables it to deal with huge data, so it is easy to provide large statistical samples and results of experiments. Companies and governments have knowledge benefits of the digital economy of big data.
There are several techniques used for big data analytics:

1. Association rule learning is used to find relationships between entities.
2. Machine learning is used to make computers learn complex patterns to make decisions.
3. Data mining can be thought of as a combination of statistics and machine learning.
4. Cluster analysis aims to break data into smaller clusters that have the same set of previously unknown characteristics [8].

3. Big data factors

Big data is called in the existence of the characteristics that are known as big data volumes (V’5) , as shown in figure 2.

Figure 2: Showing the Big data factors.

3.1 Volume :

Every second, a proportion of data, or a huge of data, is created. An example of these components is forecasting weather and sensors

data. Nowadays, the amount of data is also increasing dramatically up to Zettabyte which is 300 times from 2005 [9].

3.2 Variety :

Data variety is one of the most important characteristics of big data. It is different formats such as text, video and images data. Also, data variety refers to categories of structured, semi-structured and unstructured [10].

3.3 Velocity :

Velocity in big data refers to the speed of data flowing continuously and in a short time from different sources. Thus, traditional systems are unable to deal with data and perform analyzes [11].

3.4 Value :

It is important to note that not all big data has value. Valuable data need to be extracted from a huge amount of data, and for this reason, a data analyst has to learn when retrieving big data is to quickly identify the important and valuable data [12].

3.5 Veracity :

The Veracity of big data indicates biases, noise, and anomalies in the data. When scope out big data, keep the data clean and prevent the accumulation of useless data [13].

4. Big Data Challenges

There are many challenges related to big data, including the complexity of the data, how it is processed, incomplete, scalable, and security, it is important to build the data appropriately before analyzing the data. To enhance the results of the analysis and the quality of the data, it is necessary to consider an understanding of the appropriate method of data processing [14].Data privacy is a serious concern. Some recent disputes have highlighted how some security organizations are improperly utilizing data created by individuals for their gain. As a result, policies should be designed that address all user privacy concerns. Users' data should not be exploited or disclosed, and rule-breakers should be discovered [15]. Big data handling is complicated by system arrangements. Data transmission for big data services necessitates a lot of bandwidth. The internet is used to communicate with big data services for both

data collecting and service delivery. Data integrity is difficult to maintain, and data loss during transmission is always a possibility. Furthermore, there is always the issue of data security. The cloud environment has now taken up the challenge of storing large amounts of data. With cloud technologies, a slew of big data solutions is emerging. The fundamental issue confronting the new field is a severe lack of human resources. To exploit the value of big data, people with strong mathematical ability and relevant professional expertise are required for big data application development [8].

5. Big Data Tools and Technology

There are several ways to deal with Big Data. Different approaches and technologies have been developed for manipulating, analyzing, and displaying large amounts of data. Big data demands advanced technology [16]. One of the most extensively utilized technologies is Hadoop.

6. Hadoop

Doug and Mike created Hadoop, an open-source framework for processing enormous volumes of data, in 2005. It is the most essential Apache large data distributed tool. Its components include simple languages, graphical user interfaces, and administrative tools for processing petabytes of data across thousands of machines [17]. Hadoop is used by the majority of social networking sites, including Google, Yahoo, and Facebook. It can bind difficult-to-understand data and serve as a tool or data organizer. Hadoop is a framework for processing massive amounts of data with varied or no structure [18].The HDFS and Map Reduce are two main components in Hadoop architecture which is illustrated in Figure 3.

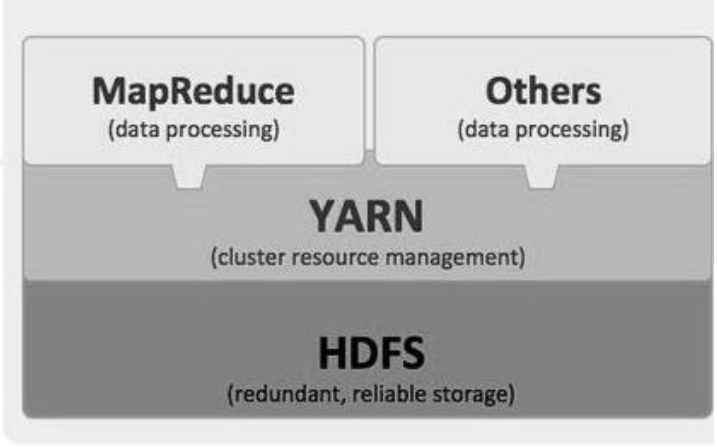

Figure 3: Hadoop Architecture.

6.1 Hadoop Distributed File System (HDFS) :

Hadoop's storage component is known as Hadoop Distributed File System (HDFS) as shown in figure 4. it keeps track of file system metadata on clusters. To ensure dependability, availability, and performance, it stores three copies of each data block by default [4]. Because the data is written once and read multiple times, it is also a strong choice for facilitating large data analysis. When data quantities and velocity are large, this type of data service provides a new set of capabilities. HDFS divides huge files into little chunks called blocks.

The blocks are saved on data nodes, and an HDFS cluster is made up of only one NameNode[5]. The NameNode is responsible for noticing which blocks on which data nodes make up the whole file. Furthermore, a number of DataNodes, generally one per node in the cluster, handle storage attached to the nodes on which they operate [19]. The GNU/Linux operating system is included in the term node, which is software that can operate on commodity hardware, the master server is the system with the name node, and it performs the following tasks: It controls client access to files and performs file system operations, as well as managing the file system namespace [10]. Another element is data node performs operations such as block creation and deletion according to the instructions of the name node. It performs read and write operations on the file systems, as

per client request. Block size is 64MB that is the smallest quantity of data that HDFS can read or write [5].

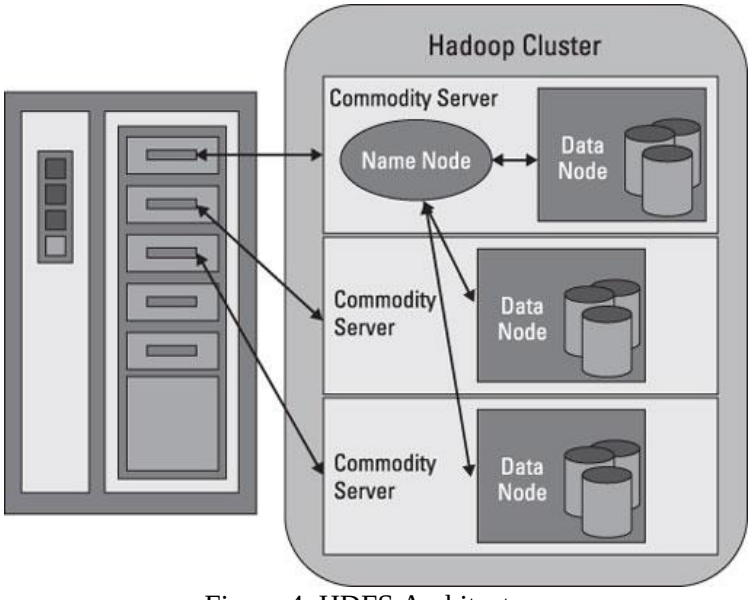

Figure 4: HDFS Architecture

6.2 Map Reduce:

Hadoop provides Map Reduce for distributed computing applications that are appealing owing to its scalability. In other words, its software for analyzing massive datasets.it has two primary functions Map and Reduce. Figure 5 showsMap Reduce Architecture. The Map function is mostly used first to filter, manipulate, or parse data. Reduce receives the output of Map as an input [9].

The Reduce is an optional function that is often used to analyse the data collected from the Map function. Hadoop distributes Map and Reduce jobs to the cluster's relevant nodes [14]. The majority of computation is done on nodes, with data stored on local storage, which decreases network traffic. The cluster gathers and reduces data to provide a suitable result before returning it to the Hadoop

server [6]. In Figure 6 illustrates how the MapReduce nodes and the HDFS work together.

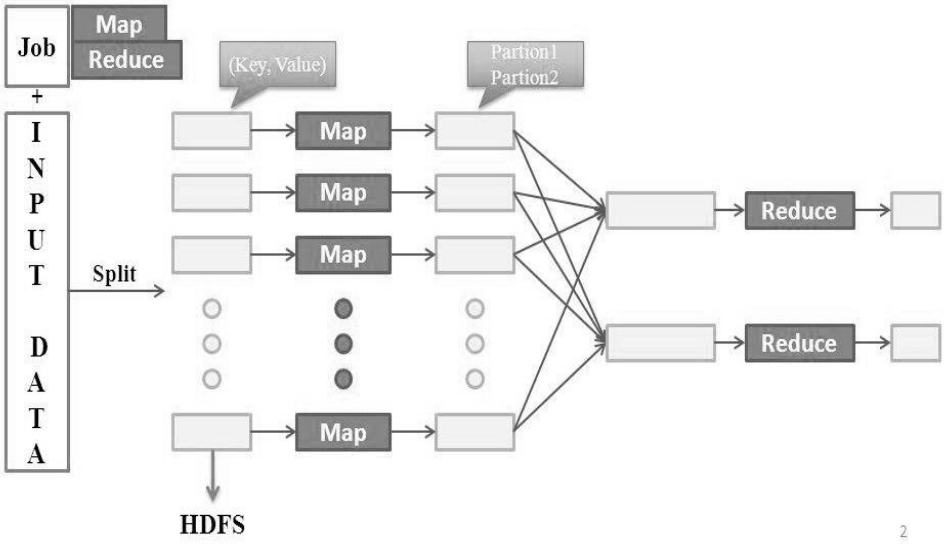

Figure 5: MapReduce Architecture

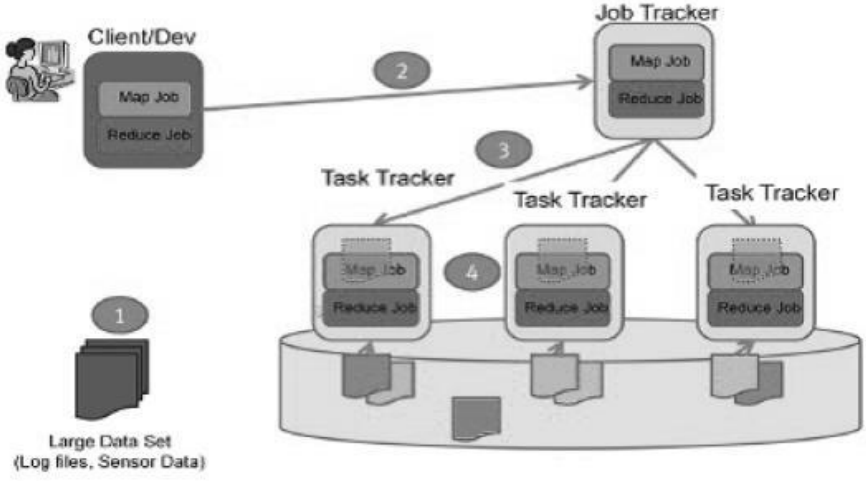

Figure 6: MapReduce and HDFS

7. Hadoop Ecosystem:

Hadoop comes with several Apache-built tools for dealing with Big data. Hadoop Ecosystem as shown in figure 7 is a term for these technologies, some of the tools as following :

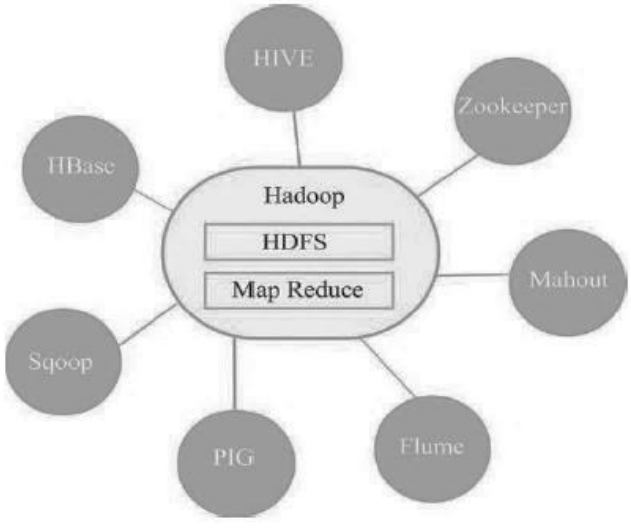

Figure 7: Hadoop Eco System

7.1 HBase :

Apache HBase allows you to access your data in Hadoop in a random and real-time fashion. It was designed to accommodate extremely big tables, making it an excellent choice for storing multi structured or sparse data. This system is column-based rather than row-based, which speeds up operations on data sets with comparable values. APIs are used to access HBase data [17].

7.2 Hive :

Hive is a type of data warehousing software that is used to handle structured data As a SQL language, it's known as HiveQL. it was created by Facebook . Many businesses utilize it for data analysis.Many businesses utilize it for data analysis as well. Integral data types, literal data types, and string data types are the three types of data types supported by Hive [18]. Data from HDFS may be queried, and these queries are then turned into Mapreduce tasks.

7.3 Zookeeper :
Zookeeper is an open-source system powered by Apache that provides a distributed service with master and child nodes that store configuration data. Moreover, it provides a service for maintaining configuration information and distributed synchronizers that give a centralized infrastructure [19].

7.4 Spark:
Apache Spark is an open-source software used for data analysis. The largest part of the components, experts describe it as a computing tool for the data analytics suite. It can be used with Distributed File System (HDFS), a specific Hadoop component that facilitates complex file processing. It has an architectural basis in Resilient Distributed Data Set (RDD), a read-only multitest of data elements distributed over a set of machines, which are maintained in a fault-tolerant way. Spark's RDDs act as a working set of distributed software that offers a restricted form of distributed shared memory. In 2012 Spark and its RDDs were developed in response to limitations in the MapReduce cluster computing model [20].

7.5 Apache Pig:
Apache Pig is an Apache Foundation-developed high-level scripting language. Pig is known for its extensibility and ease of programming. Pig Latin is the name given to the language spoken by pigs. Pig Latin is made up of numerous operations that, when combined, allow programmers to create their reading and writing functions [21]. The pig can accept programs written in any language, including Java and Python, and it supports Hadoop streaming. To handle data, Pig uses the MapReduce framework.

8 Hadoop MapReduce and tools comparison:
The previous section presented Hadoop components, which are developed on the software foundation of Apache. There are several other tools found on the support of Hadoop such as Apache Sqoop, Apache Flume, Oozie, and Cassandra [22]. The difference between Hadoop tools is in terms of data processing, data management, data access, and scripting. In addition to the common factor in most of

the tools of Hadoop is that it is open-source, and its support and help for big data issues. Table 1 shows the most differences between some tools of Hadoop.

TABLE 1 . MapRedues Vs Hive Vs Pig

Hadoop MapReduce	Hive	Apache Pig
Apache	Originally developed by Facebook	Originally developed by Yahoo
Compiled language	SQL Query Language	Scripting language
Implementation language: Java	Implementation language: Java	Implementation language: Pig Latin
Supported programming language: java , C++, python ,Ruby	Supported programming language: C++ , PHP , Python	Supported programming language: Java, Jython, JavaScript, Python, Ruby
Code efficiency : high	Code efficiency : low	Code efficiency : low
Supported (UDFs)	Supported (UDFs)	Supported (UDFs)
Write several lines to basic code	Not real time to access data	Pig is still in the development
Used for programming	Used for reports and data analysts	Used to process data flow

Through the information presented in this research, the problem of increasing the volume of data significantly and rapidly, the most common and used solution is the use of the Hadoop framework, where Yahoo has overcome the analysis of its big data using Hadoop. After that, major companies used it, and its effectiveness and success appeared, despite the difficulty of dealing with it. Also needs special expertise.

Conclusion:

Big data provides a vision for solving data problems that are increasing in a huge size and very quickly over time. Big data has

been contributed to reducing difficulties and provided solutions in dealing with data in a short time instead of traditional solutions. Big data is one of the most important leading topics in computer science research. In this paper, an overview of big data was given and identified main characteristics (V'5)، as well as the challenges that faced big data. In addition, this paper has described big data techniques and some of the tools. In addition, the paper presents one of the most important open-source framework (Hadoop) to deal with big data.

References

[1] Tyagi, H., 2018. *Big Data-A Review Study with Comparitive Analysis of Hadoop*.

[2] Giri, K.J. and Lone, T.A., 2014. Big Data-Overview and Challenges. International Journal of Advanced Research\in Computer Science and Software Engineering, 4(6).

[3] Arne Holst, 2021. Amount of data created, consumed, and stored 2010-2025. Statista [viewed 12 December 2021] https://www.statista.com/statistics/871513/worldwide-data-created.

[4] Zanoon, N., Al-Haj, A. and Khwaldeh, S.M., 2017. Cloud computing and big data is there a relation between the two: a study. *International Journal of Applied Engineering Research*, *12*(17), pp.6970-6982.

[5] Agrahari, A. and Rao, D., 2017. A review paper on Big Data: technologies, tools and trends. *Int Res J EngTechnol*, *4*(10), p.10.

[6] Ravichandran, G., 2017. Big Data processing with Hadoop: a review. *Int. Res. J. Eng. Technol*, *4*, pp.448-451.

[7] Lee, I., 2017. Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*, *60*(3), pp.293-303.

[8] Malik, D. and Goel, P.K., 2020. A Brief about Big Data, It's Technology and Challenges. Volume 22, PP 01-05

[9] Kaur, M.G. and Kaur, M., 2015. Review Paper On Big Data Using Hadoop. *International Journal of Computer Engineering & Technology (IJCET)*, *6*(12), pp.65-71.

[10] Sethy, R. and Panda, M., 2015. Big data analysis using Hadoop: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(7).

[11] Beakta, R., 2015. Big data and Hadoop: a review paper. *Baddi University of EmergingSciences& Technology, Baddi, India Volume 2, Spl. Issue 2 ISSN: 1694-2329 2015*.

[12] Giri, K.J. and Lone, T.A., 2014. Big Data-Overview and Challenges. *International Journal of Advanced Research\in Computer Science and Software Engineering*, *4*(6).

[13] Reimer, A.P. and Madigan, E.A., 2019. Veracity in big data: How good is good enough. *Health informatics journal*, *25*(4), pp.1290-1298.

[14] Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), pp.1-16.

[15] Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M. and Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. *The scientific world journal*, *2014*.

[16] Agrahari, A. and Rao, D., 2017. A review paper on Big Data: technologies, tools and trends. *Int Res J EngTechnol*, *4*(10), p.10.

[17] Dagade, V., Lagali, M., Avadhani, S. and Kalekar, P., 2015. Big data weather analytics using hadoop. International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN, pp.0976-1353..

[18]]Vo, A.V., Konda, N., Chauhan, N., Aljumaily, H. and Laefer, D.F., 2018, June. Lessons learned with laser scanning

point cloud management in HadoopHBase. In *Workshop of the European Group for Intelligent Computing in Engineering* (pp. 231-253). Springer, Cham.

[19] Anuradha, J., 2015. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. Procedia computer science, 48, pp.319-324.

[20] Erraissi, A. and Belangour, A., 2018, December. Meta-modeling of Zookeeper and MapReduce processing. In *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)* (pp. 1-5). IEEE.

[21] Mavridis, I. and Karatza, H., 2017. Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark. *Journal of Systems and Software*, *125*, pp.133-151.

[22] Swarna, C. and Ansari, Z., 2017. Apache pig-a data flow framework based on hadoop map reduce. *International Journal of Engineering Trends and Technology (IJETT)*, *50*(5), pp.271-275.