Big Data: Challenges, Popular Tools Of Big Data - Benefits And Applications

Sadia Zafar, Haroon ur Rashid Kayani, Hafiz Burhan ul Haq, Imran Khalid, Ayesha Nasir

Abstract: Big Data is generated everywhere in the world in various digital formats. In 2020 the Big Data revolution estimated billion-billion devices connected to the fast internet, and massive data will be predicted at high speed and drawn researchers' attention in academia, governments, and industries. Big Data is valuable to enhance productivity in businesses and evolutionary breakthroughs in the many fields of sciences. However, there is no doubt that Big Data's handling produces many challenges, such as data analysis, data visualization, data storage, and new technology to deal with Big Data problems. This paper aims to demonstrate the challenges, the new tools of Big Data exploration, their benefits, and applications that can draw researcher's and users' attention to decide better tools for their businesses and need.

Index Terms: Big Data, Data Science, Big Data Tools, Open Source Tools, Artificial Intelligence, Machine learning, Data Analysis

1 INTRODUCTION

The world population is currently 7.2 billion [1], and out of these, approximately 2 billion people are linked with the internet. Besides this, 5 billion persons use mobile phones, as per McKinsey (2013). Due to this technological uprising, millions of individuals produce significant data volumes with these devices' increased usage. This continuous production data is called Big Data [2]. Big Data is a term that defines the enormous amount of data that can be unstructured and structured, which affects the business. In 2012, Gartner retrieved and gave a more detailed definition: ‘‘Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization’’. Big data require generalized tools for the treatment of data for generating significant results. Thus, the primary focus should not on high quantity data, but on the opportunity that data gives creative knowledge and information that make the public entities and company much competitive, which will help them offer improved services for citizens and customers [3]. Big Data has transformed the method adopted in doing research, Management, and businesses [4]. In 2010, Google assessed that every two days around, the world produced as much information the total it created up to 2003. Despite the exceptionally later "Huge Data Executive Survey 2013" by New Vantage Partners that states, "It is about assortment, not volume," numerous individuals (counting the creators) would still trust the premier issue with huge information is scale or volume. Beyond any doubt, huge information includes an incredible assortment of information shapes: content, pictures, recordings, sounds, whatever that may come into the play, and their subjective blends [2]. The selection of tools becomes difficult nowadays that are used for Big Data. Most of the tools have their pros and cons. Big Data is continuously increasing rapidly, and the selection of machine learning tools becomes insufficient for real-time processing and distributed processing. Sara Landset et al. discussed different tools for big data in their paper, such as Spark, Flink, Map Reduce and Storm etc. Their paper also provided much information for decision-makers for big data tools and gave future tool-based learning direction [5]. Big Data is divided into three parts, i.e., structured data that can be stored, retrieved, and can be processed in a fixed layout, unstructured data stored in any structure, and semi-structured data is a mixture of unstructured and structured data [6]. Furthermore, Big Data has different characteristics in the shape of seven V's, i.e., volume refers to the size of data generated from all sources together with audio, text, research studies, video, social networking, crime reports, space images, medical data, natural disasters, and weather forecasting, etc. Velocity is the speed or state of how fast the data can be accessed and processed. Variety means different data sources that are very difficult to arrange because data may be structured, unstructured, and semi-structured. Variability is all about the correct interpretation and understanding of the raw data with accurate meaning. Veracity refers to a term sure that collected data is accurate, and it also keeps worse data away from the system. Visualization means the presentation of data to Management for decision-making. Figure 1 shows the 7V's of big data. Values are an essential characteristic of big data as it adds the data's value, which may help exceed the cost [7].

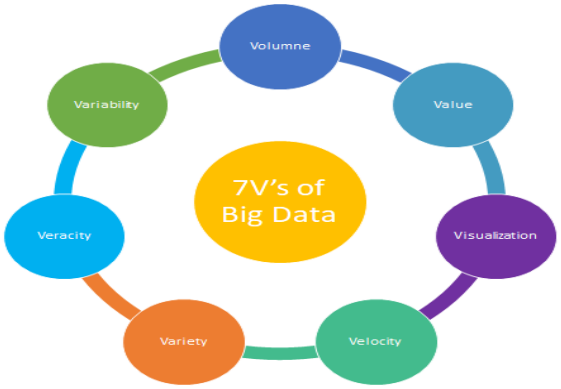

Fig. *1*. *7V's of Big Data*

- *Sadia Zafar, Department of Computer Sciences, Lahore Garrison University, Lahore, 75500, Pakistan. E-mail: sadiazafar@lgu.edu.pk*
- *Haroon ur Rashid Kayani, Data Scientist and Consultant, Lahore Pakistan. E-mail: hurkayani@gmail.com*
- *Hafiz Burhan ul Haq, Department of Computer Sciences, Lahore Garrison University, Lahore, 75500, Pakistan. E-mail: burhanhashmi64@lgu.edu.pk*
- *Imran Khalid, Department of Computer Sciences, Lahore Garrison University, Lahore, 75500, Pakistan. E-mail: imrankhalid@lgu.edu.pk*
- *Ayesha Nasir, Department of Computer Sciences, Lahore Garrison University, Lahore, 75500, Pakistan. E-mail: ayeshanasir@lgu.edu.pk*

2 CHALLENGES:

There are four significant challenges of big data that need to be addressed appropriately; processing issues, storage issues, and management issues. There are many sorts of analytics challenges, such as estimative, predictive, prescriptive, and descriptive, leading to numerous optimization and decision models. The design for components and systems that work with big data required a proper understanding of both the equipment and user requirements [8]. The output processing is also the major challenge of big data. Jacobs [9] summarized the issue very succinctly – "it is easier to get the data in than out." The main challenges of big data are briefly described as [8]:

2.1 Data Storage

The amount of data has blown up each time through social media, different IoT devices, etc.; for this, a storage medium has required an invention. Furthermore, every day, a large amount of data is created by everybody and everything (e.g., different, IoT devices, professionals, journalists, scientists, researchers, etc.), which is problematic to store.

2.2 Data Management

The Management of a large amount of storage data will, possibly, be the most significant complicated issue to address with big data. The United Kingdom eScience will first face this issue where data managed and owned by different entities. For resolving access matters, the metadata has been demonstrated as the main stumbling block.

2.3 Data Processing

The processing of a large amount of data is complicated. Suppose the Hexa-byte data is needed to be processed entirely. Much time is required to process the large volume of data. For better processing, a new algorithm has been required to give actionable and timely information.

2.4 Big Data analysis and visualization

Big Data analysis and visualization are significant problems for industries, organizations, and financial and research institutions. The new tools are developed to overcome Big Data analysis, visualization, and knowledge extraction, which are discussed in section.

3 METHODOLOGY

For the in-depth study of challenges and new tools of Big Data, a review approach was adopted. This study helped us to integrate the existing polished work and broadly identify our strategy. We made an in-depth study of the challenges and tools of Big Data, revealing their benefits, challenges benefits, and applications. Our study's prime focus is to provide comprehensive knowledge to all users to decide better tools to meet their businesses' needs.

4 LITERATURE SURVEY

The Big Data (structured and unstructured) generated by various sources such as social media, sensors, streaming data, or scientific data in various formats grows very fast. The biggest challenge is how to extract meaningful information from the processing of Big Data. The market is saturated with the commercial and open-source tools of Big Data that tools are used to extract the data and convert it into meaningful information. These big commercial tools or their licenses are costly. Big Data has been a developing area of research and study during the last decade. The numerous researcher papers have been published, and new tools of Big Data devolved. Big data also includes other fields like Mathematics, Machine Learning, Deep Learning, and Data Mining to extract valuable knowledge from unstructured and structured data [10, 11]. Parkin 2006 presented that big data are turning into another innovation center both in science and in industry. This paper discusses Big Data's difficulties on future Scientific Data Infrastructure (SDI) and the cutting edge. The paper talks about the nature and meaning of Big Data that incorporate such highlights as Velocity, Variety, Volume, Velocity, and Veracity. This paper alludes to various established researchers to characterize prerequisites on information the board, for getting on control and security. Barlas, Panagiotisl at el. 2015 showed a significant information examination framework that enables the government to quantify its security views from the Twitter interpersonal organization. The proposed framework was manufactured for utilizing a lot of enormous informational instruments to gather, pre-process, order, list, and picture information. The framework can identify whether a tweet is identified with security or not and utilized to introduce a city's security. A machine learning calculation was prepared to figure out how to perceive security qualities in tweets. Results demonstrate that this framework is a useful asset to dissect how natives see security outwardly [12]. Camargo, Jorge E., et al. 2016 presented that knows days the information is a significant factor that can lead toward the decision-making performance, so if the information is correct, then give a better decision of any problem. Nowadays, a large amount of data is Accessible in every organization. The knowledge extracted from the datasets is used for the decision-making process. The work's extreme goal and commitment are utilizing colossal information for the investigation to upgrade and bolster essential leadership in an organization [13]. Elgendy et al. discussed in 2016, a specific measurement of gathering data volume depends on the past in Exabyte format. Such volumes are considered to be exceeding the point of a limit of taking care of structures and storing frameworks. Learning data and taking care of systems are being made and accumulated rapidly toward the Exabyte/year run. This is the quickest path for putting away colossal measures of information for a long time. Volume is only a solitary piece of tremendous data; diverse characteristics are collection, regard, multifaceted nature, and regard. Exchanging the data requires a new perfect model to determine the delivery issue while information exchanging [14]. Iqbal, Muhammad Hussain at el. 2015 presented that huge information has turned into a featured popular expression since a year ago, colossal information mining, i.e., extracting from huge information, has quickly pursued up as a developing, interconnected region of research. This paper discussed the enormous mining of information and discusses related issues and new chances regarding information mining. The discussion includes a survey of best-in-class structures and stages apart for handling and supervising the vast information the same as the endeavors expected on enormous information mining. Similarly, broad issues are also addressed that were identified with colossal information. This will help reshape the knowledge regarding present information mining with advanced innovation toward handling future difficulties building in agreement with huge information [15]. Kaiser, Stephen, et al. presented that the data sets are developed more suddenly in

various structures due to digitalization. When the information or data sets are extensive and intricate, where conventional information-handling strategies are not ready to handle this complex information that is considered big information. Analysts, researchers, business associations, government offices, promoting offices, and medicinal analysts regularly experience more trouble managing information for any essential leadership. The information accessible for research should be controlled by utilizing various examining information methods, which is Big Data Analytics. These methods are beneficial for managing the vast volume in various forms like unstructured, organized, or semi-organized information content, suddenly evolving nature, which is unrealistic to process by accessing traditional database systems. This paper also presents the real use of enormous information examination after looking at various apparatuses accessible for massive information approval. This paper also talks about the contextual investigation that led to the defeat of the vast information issues and necessities [8]. Data sets were rapidly growing nowadays in different ways because of digitalization. Now, the information and data that are too complex and large and the traditional techniques processing cannot deal with that data are known as the Big data. Scientists, Government agencies, Advertising companies, medical researchers, etc., faced many difficulties in dealing with the data for giving out any decision. The data used by researchers has been passed out through different methods of data analysis. These methods give helpful information from the massive volume of structured, unstructured, and semi-structured data. The paper deliberates about the utilization of big data tools by their comparison [16].

5 Popular Tools of Big Data

Big Data is rapidly increasing nowadays, and very difficult to analyze and extract knowledge from Big Data. We have studied the various popular tools of Big Data that are briefly discussed subsection. Our study reveals the benefits, challenges, and applications of the Tools of Big Data.

5.1 Apache Drill:

Apache Drill is the tool of big Data that provides a Schema-free engine. It supports several file systems and databases i-e mongo dB, HDFS, amazon, and Hbase, etc. A single query will help to join data from different databases [17]. Its architecture includes distributed executed environment, which helps in processing a large volume of data. It is based on Drillbit's services, which is responsible for receiving requests from the client process them, and giving results. The drill is popular because of its Standard syntax because there is no need to learn any other SQL-like language [18].

5.2 Apache Hadoop

Apache Hadoop is a Big Data tool that provides us a framework for processing large data sets. It is designed to handle and detect failure in the application layer. It provides highly accessible services to the cluster of the computer. Hadoop is also designed for scaling up several computers or machines in a single server. Various organizations and companies now use Hadoop for research purposes for bringing efficiency and flexibility in information processing. It is also useful for data processing [19].

5.3 Apache HBase

HBase is a non-relational database management system that runs on top of the Hadoop Distributed File System (HDFS). Apache HBase is used when the user needs random, real-time write/read access to Big Data. This helps in hosting the large tables, billions of rows, and millions of columns. The applications of HBase are written in Java, like MapReduce( ) application [20].

5.4 Apache Hive

Apache Hive is a big data tool that is open-source to help write, read, and maintain large files that are directly stored in the HDFS and another storage system like Hbase. The data summarization can be done quickly with the help of the Apache Hive. It gives an easy/simple way of structuring a large volume of data. It is also integrative with the data centers. It makes the MapReduce( ) language easier as there is no need to write and heavy Java codes. Hive helps SQL developers to write queries in Hive Query Language as it is similar [21].

5.5 Apache Flink

Apache Flink is a scalable platform and distributing process engine. It helps in processing data on a large scale. It is designed to run the cluster environment and also to perform the computations. Apache Flink is a tool that can solve real-world problems more efficiently. Nowadays, a company needs a platform which single can solve all the major problems. Additionally, the Flink also gives fault tolerance, communication, and distribution for the data stream [22].

5.6 Apache Mahout

Apache Mahout is a big data tool that is open-source and helps to develop machine learning libraries. It is a distributed linear algebra framework that helps statisticians, mathematicians, and data scientists implement their algorithms quickly and efficiently [23]. Developers also used mahout for mining a massive amount of data. It is also used to create an application with machine learning techniques like collaborative filtering, clustering, and categorization [24].

5.7 Apache Spark

Apache Spark is a platform that helps in the processing of data on a massive amount of data sets. It can be organized in many ways that give native binding for Python, Scala, Java, and R programming language. It also gives support to the machine learning algorithm, processing of graphs, and streaming of data. The architecture of apache spark includes two major components; the driver, which helps in converting user code into numerous tasks distributed in different worker nodes, and the other part is the executor that runs on these particular nodes and helps in the execution of tasks assigned to them [25].

5.8 Apache SAMOA

Apache SAMOA (Scalable Advanced Massive Online Analysis) is an Open-Source platform for processing a massive volume of data. It belongs to the Apache group of tools utilized for Data handling and uses distributed stream algorithms for sharing common machine learning and data mining tasks, such as clustering, regression, and classification, to develop new algorithms. It also provides an application programming interface (API) that the programmer uses to implements new distributed algorithms [26].

5.9 Apache Storm

Apache Storm is a free distributed computational tool. It helps in the processing of an unbounded stream of data. A Storm is

a simple tool that can easily use with any programming language. It provides fault tolerance, scalability, and a guarantee to data that will be processed and operate efficiently [27]. Many software industries use Storm to process data as it efficiently processes 100 bytes of data in a single node. It is exceedingly flexible, easy to utilize, and provides low latency with guaranteed information handling [28].

5.10 BigMLer

BigMLer is an Open-Source tool that helps perform machine learning (ML) work. This tool provides a command prompt-based interface for implementing ML tasks. BigMLer is very useful and performs any task by typing a single command prompt. It is beneficial in optimizing the model by automatically selecting the feature. However, this tool is also speedy and used for model tuning [29].

5.11 Cassandra

Cassandra is a No SQL database that is highly scalable and used for fault tolerance by replicating data into multiple nodes [30]. It is highly elastic with no downtime or any interrupts while running the applications. Cassandra is durable and suitable for online transactions and applications because data can be safe even when servers go down. Cassandra is designed to handle a massive amount of data and available as an open-source [31].

5.12 Cloudera

Cloudera is an Open Source and extensible platform that is used for promoting business and information technology. However, this platform is also having high popularity among data scientists and data engineers. Cloudera provides multi-cloud with several functions. Furthermore, this platform is easily manageable, scalable, easy to use, and contains highly secure designs to handle various data types [32].

5.13 Data Applied

Data Applied is a non-programming tool intended to construct, share structure information investigation reports, and represent substantial informational indexes. However, this tool helps visualize data with the help tree maps. Furthermore, this tool is also used for multiple purposes, such as anomaly detection, data transformation, and meaningful data extraction from crude information [33].

5.14 Data Wrapper

Data wrapper is a data visualization tool that is available online and used for creating collaborative charts. Once the data inserted into the Data wrapper, it will quickly convert the information into interactive charts, bars, and other visualization forms. This graphical representation of data can be easily embedded into websites and articles for making interactive [34].

5.15 Dryad

Dryad is a big data tool that helps in parallel and appropriated programs for dealing with setting bases on dataflow outlines. It involves a gathering of enrolling center points, and a customer takes the benefits of a PC gathering to run their desired programs disseminated. A dryad customer utilizes countless, all of them with various processors or focuses. The critical favored point of view is that customers never need to know anything about concurrent programming. A dryad application runs a computational composed outline that is made of correspondence channels and computational vertices. In this way, Dryad provides a vast number of useful counting creating of employment diagram, planning the machines for the accessible procedures, taking care of the group for changing disappointment, gathering execution measurements, and picturing the activity [35].

5.16 H2O

H2O is an Open-Source platform and has distributed in-memory machine learning with linear scalability. H2O is highly applicable for statistical and machine-learning algorithms such as generalized linear models, gradient boosted machines, deep learning, etc. H2O is also used AutoML functionality that automatically runs through all the algorithms and their hyper-parameters to produce the best models. Globally, H2O is approximately used by 18,000 organizations and also achieved the same popularity as R & Python communities [36].

5.17 High-Performance Cluster Computing (HPCC)

HPCC stands for High-Performance Cluster Computing, which describes the computational environment that utilizes computer clusters to support the different applications within a significant amount of time and process a significant amount of the data. HPCC is a tool of big data that Lexis Nexis Risk Solution develops. It also provides scalability and gives us better performance [37].

5.18 Jaspersoft

The Open-Source platform produces reports from the database sections. One imperative property of Jaspersoft is that it can rapidly investigate huge information rather than extraction, changing, and stacking. It is an intelligent tool that allows the users to blend with different databases without moving the data to other databases. It is scalable, cost-effective, and easy to use [38].

5.19 Lumify

Lumify is one of the powerful platform for big data visualization. It helps the user to invent a very complex connection and explore the relationship in the data. It works with amazon and provides support to the environment based on cloud. It has also built-in technologies like Accumulo and Hadoop. Lumify is efficient and scalable. For Lumify, there is no need to install a workstation individually as it works on the server. It also allows integration with the other tools that are working in the background [39].

5.20 Map Reduce

Map Reduce is a programming tool that is related to execution for handling and creating huge informational collections. The MAP and REDUCE are the two elements of this model. The MAP( ) function work produces sorting and filtering, and REDUCE( ) function works on summary operations. By improving the execution, the map-reduce programming model includes being specific, adaptable, and adaptation to internal failure. Map Reduce libraries have been written in various programming, with essential enhancement. The name Map Reduce was the technology that Google proposed. By 2014, Google was never again using Map Reduce as primary information for preparing the model and proceeded onward improvement on Apache Mahout to progressively able and

less plate arranged systems that consolidated full guide [40].

5.21 MongoDB

Mongo DB is a popular database in the modern era, especially in big data. It is a decent asset to oversee information that reflects the information in a semi-organized or unstructured form. It is used to store information in versatile applications, item lists, content administration, and applications that give a single view over different frameworks. This Mongo DB connects through data visualization tools, including Tableau, IBM Cognos Business Intelligence, Objects, Qlik, and SAP Business [41].

5.22 No-SQL Database

This type of database is not bound by customary diagram models permitting them to collect unstructured datasets. The adaptability of No-SQL databases like Mongo DB, Cassandra, and HBase makes them a superior choice for colossal information investigation. No-SQL DB is used to store increased volume or variety of the data and its performance compared to the traditional database is much effective and scalable. No-SQL is now maintaining and mange by the foundation of the Apache [42].

5.23 Open Refine

It is the most dominant Big Data tool that helps clean data and changes it into different layouts. It can also help to explore a large amount of data sets. Open Refine has some essential features like importing the data in different formats. It helps apply the advanced and fundamental transformation of cells, using refined language to perform advanced data operations [43].

5.24 Open Text

Open Text is an efficient platform used by business users to explore, analyze, and blend the data fast without relying on IT or data exporters. Users can explore the 360 degrees view for their business, and also, in a second, the user can explore the billions of records. It is effortless to use does not require complex coding. It is a complete solution that combines the advanced level of the software and supports the AI and the machine learning algorithms; most organizations are used to improve decision-making [44].

5.25 Pentaho

Pentaho, the tool of big data analysis, was developed in 2004. Its offered solutions help to maintain the Business Intelligence (BI) projects. Pentaho supports multiple components. The data integration of the Pentaho helps in the Extract, Transform and Load (ETL) solution, and it is much valued in the market. It works is very simple and efficient. The BI server of the Pentaho allows managing the BI resources. It also helps manage the reporting solutions and fulfills all the reporting environment requirements [45].

5.26 PolyBase

PolyBase is an updated tool that helps enable the queries across data stored in Parallel Data Warehouse (PDW). PDW is a data warehousing contraption that worked to set up any social data volume and coordinate Hadoop to get non-social data. The big data tool's Ploybase facilitates joining relational and the relational database. It can easily import the data from Hadoop. To make the thing easy, it does not require any other software integration [46].

5.27 Python

Guido van Rossum developed Python in the 1980s. Python is a high-level language and Open Source in nature. Python is very popular in data science in the past decades. It also supports multiple libraries to perform different tasks. Its use is increasing day by day due to its flexible nature. The significant benefit of using Python is easily understandable and also supports multiple platforms. Python is much productive as compared to the other languages uses so far [47].

5.28 R Programming

R is the combination of lexical scoping and S programming language. R language is utmost essential for performing visualization and calculation. R is a factual programming language that is a part of insights, investigation, and representation. In the present information, researchers and business pioneers utilize R to settle on power business choices. The R tool is adaptable and open source. R incorporates special bundles that are helpful in the investigation of information. It contains Deploy R API's, Deploy R server, and Deploy R store, which are utilized to transfer and confirm information. R language is also extensible that can be assimilated with other languages and accessible as an open-source [48].

5.29 Splunk

Splunk is a platform helpful in searching, monitoring, visualizing, and analyzing machine-generated data in real-time environments. A tremendous amount of data is created through the machine from business endeavors. It joins the regularly updated cloud propels and enormous data. Consequently, it makes customers chase, screen, and separate their machine-created data through the web interface. The results are appeared in a characteristic course, for instance, diagrams, reports, and alerts. Its qualities fuse requesting composed, unstructured machine-made data and uncovering indicative results. Splunk's most basic focus is to offer metrics to various applications, investigate issues for structure and information development establishments, and help for business tasks [49].

5.30 Sqoop

Sqoop is a big data tool for transferring a large amount of data. The primary purpose of Sqoop is to connect the Hadoop with different databases to transfer the data. It is also useful to transfer the structured data to Hive and Hadoop. The Map Reduce is used by the Sqoop tool order to import and export the data. It also helps to manipulate the data, easier to use, cost-effective, gives security to the data, and many of the processes in the Sqoop are automated [50].

5.31 Statistical Package for the Social Sciences (SPSS) Modeler

SPSS Modeler is efficient software for machine learning and data science. It helps the organization in achieving the desired goal. It is used to build predictive models and conducting analytical tasks. It works efficiently in a hybrid environment. The model deployment becomes more comfortable with the help of SPSS; the machine learning models are created efficiently [51].

5.32 Tableau

A tableau is a platform of big data that visualizes the data. Tableau is the most powerful tool from the business point of view. It is a quick representation programming that allows investigating information, each perception utilizing different conceivable diagrams. Its wise calculations make sense of independence from anyone else about the kind of information, best technique accessible. The intelligent algorithm of Tableau will figure out by itself the type of data and tell which methodology is best for this type of data [52].

5.33 Talend

Talend is an Open-Source platform that helps in data integration. Supervisors and experts never again settle on a gut-based choice. They require an instrument that can help them rapidly. Talend can assist them with exploring information. Moreover, it also helps to produce clean information. It also offers exciting automation features that highlight where users can spare and re-try past assignments on another informational index. This element is remarkable and has not been found in numerous instruments [53].

5.34Tanagra

Tanagra is an Open-Source tool that helps in academia and research purposes. Tanagra project began as free programming for research purposes. Being an open-source venture and, it gives enough space to devise calculation and contribution. Alongside directed learning calculations, it is empowered with standards, for example, grouping, factorial investigation, parametric and nonparametric measurements, affiliation rule, highlight determination, and development calculations. A portion of its restrictions incorporates inaccessibility of comprehensive information sources, direct access to data warehouses and databases, interactive usage, and so forth [54].

5.35 Yahoo S4

Yahoo S4 is a platform that helps in the processing of data streams. Yahoo S4 engages designers to develop applications that process in real-time for getting information. It is impressed by the Map-Reduce model and procedure for the information in the circulated design. For example, designers can create fitting and play modules in Java. Modules are created in Yahoo S4 can be combined to plan progressively advance ongoing handling applications [55].

COMPARATIVE STUDY OF BIG DATA TOOLS

We studied the various popular tools of big data. The most demanding and open source tools are used in academic institutions, governments, banking, industries, non-public groups, monetary establishments and researchers, and students. Comparative research of Big Data tools in conjunction with their advantages, challenges, and applications are discussed within the table [1].

TABLE 1
Names, Challenges, Benefits, and Applications/Projects of Big data tools.

S. No	Names	Advantages	Challenges	Applications/Projects
1	Apache Drill	Easily integrate with the SQL tools. Scalable in nature, having good /high performance [56].	Not effective for executing long quarries require more space [56].	Apache drill give support to user defined functions, its simple model help in operating and deploying huge clusters. Drill has extensible architecture and malleable data model [57].
2	Apache Hadoop	Apache Hadoop is scalable, fast, and cost-effective. It provides security to users and also helps in detecting frauds [58].	It will not efficiently work for small data and have stability issues. It is also vulnerable [58].	Techniques of early detection of events like Bus Beat that utilized GPS trajectories of the periodic-cars, which can be done by detection of events using networks [59]. Hadoop will ensure the user security [60].
3	Apache HBase	Provide scalability, fast processing, and offers consistent write and read [61].	Apache Hbase has Memory issues and does not support database structure. Handling queries is tough in Hbase [61].	It is helpful in medical field to gather people's disease history that belongs to a specific area [62]. It is also used on the Web for storing the history or preferences of users [62].
4	Apache Hive	It is a Declarative language like SQL and able to store files [61].	It does not support the updates functions and Subqueries [61].	Apache Hive is used to perform analysis in the airline dataset. It is also used for building a data warehouse for an E-commerce Environment [63].
5	Apache Flink	Apache Flink is applicable for streaming and processing real-time data [57].	It does not have APIs in matured form and has less maturity in the industry [64].	Apache Flink is used by Alibaba, King, websites etc. Bouygues also used this tool for real-time processing and analysis of messages per day [57].
6	Apache Mahout	It can process data more efficiently—support different algorithms [65].	Visualization is not good and it does not support scientific libraries [66].	Modified version of Mahout algorithms utilized by Cull.tv for content recommendations [67].
7	Apache Spark	Apache spark is the best tool for big data due to its high speed, easily useable, and supports multiple languages [68].	It supports a few algorithms, the issue in handling small files. Unable to support multiple users [68].	Building a data warehouse for an E-commerce Environment [69].
8	Apache Samoa	Simple to use, scalable, and fast in use [70].	N/A	Helpful in normalization, Clustering, Regression, Classification of the dataset [70].
10	BigMLer	Highly scalable and valuable	BigMLer on use decision tree model	N/A

		for real-time prediction [74].	which has their own restriction, and only accepts up to 64 GB of dataset [74].	
11	Cassandra	Massive data is easily handled through Cassandra and and its architecture is simple [70].	The maintenance of Cassandra is complicated and required much effort [70].	Cassandra has abroad coverage of storage that helps users to store data of any kind. Users can also use Cassandra for backend development of Applications [75].
12	Data Wrapper	Friendly in nature, compatible with IoT devices, fast, and fully responsive [70].	Limited for certain cases [70].	Creating an interactive chart, maps, and table by just uploading a dataset/content in CSV format. [70].
13	Dryad	Dryad is better in performance-wise as compared to other software. It also allows run-time optimization [76].	N/A	Provide a better way of Financing for a viable community [77].
14	H2O	The advanced machine learning model is created easily within no time [78].	The software is not scalable and easily manageable [78].	Helpful in exploring data and creation of the model. It will integrate with both R and Python and other big data tools [78].
15	HPCC	HPCC is robust, scalable, helpful in parallel processing, and Cost-Effective [70].	N/A	HPCC allows medical professionals to digitalize the more complex processes such as genome sequencing and testing the drugs [70].
16	Jasper soft	It is Cost-Effective, easy to use, and supportive [79].	Optimization of memory is not good; performance is not much efficient for extensive reports [79].	Creating the documents in several complex formats like OpenOffice, PowerPoint, RTF, Word, and spreadsheet documents can generate raw CSV, JSON, or XML by using Jaspersoft-Studio [80].
17	Lumify	Highly secure and scalable, also supporting a cloud-based environment [70].	N/A	Lumify's infrastructure works in the background to observe changes also provides an API that allows us to map analytic inputs and outputs to object types [70].
18	Map Reduce	Performance is good, easily handles the data-intensive App [81].	More memory is needed for Map Reduce [81].	Used by Social Media like Facebook, and Twitter, for finding familiar followers/friends on a different social network [81].
19	Mongo DB	Low-Cost, Reliable, easily installed, it gives support to multiple platforms [70].	Slow in terms of speed [70].	Used by Weather Channel's to deliver weather alerts to millions of users in real-time by using Android apps and iOS [82].
20	No-SQL Database	It is scalable that can easily store massive types of data. [83].	Need more technical skills, less Supportive, less mature [83].	No-SQL database as compared to traditional database have simple and flexible and simple structure. It is Open-Source and does not required expensive licensing to run [84].
21	Open Text	The tool is more comfortable and faster in up-gradation, efficient in manipulating the data [85].	Response time is slow; the learning curve needs to be improved [86].	Open Text have built-in statistic algorithms that give opportunity to business analysts to do profiling, mapping, fore-casting and clustering without writing number of lines of code [87].
22	PolyBase	Easily to import data from different software, cost-effective, scalable [88].	To use it, the user must have the permission of Sysadmin [89].	N/A
23	Python	Flexible in nature [90].	Python has limited speed, and does not support threading. It is not helpful for mobile apps [91].	The fantastic projects are supported by Python like Pytorch, Home-Assistance, Grimpy [91].
24	Splunk	Splunk helps in generating reports with interactive graphs, tables and charts. It is easy to use and automatically embed useful information in data. [92].	It required many resources for maintenance. Expensive when Handel large datasets [93].	Splunk helps in designing information's rich dashboard and views that fulfil the need of users. It is trusted by wide number of uses [92].
25	SPSS Modeler	It provides data-classification, data-clustering, and Automated modeling of data [94].	It is not easily integrated with some of the most valuable tools like Tableau and Qilk, and also formula writing is tricky for excel users [94].	Very useful for data preparation and model management, and deployment in an organization [51].
26	Tableau	It is the best tool for Data visualization and exploration. It is easily connected with most of the databases [70].	The formatting needs to be improved [70]. N/A	This tool is much suitable for data visualization and exploration and can be integrated with databases [70].
27	Talend	Handle multiple platforms—Speedy in nature [70].	Interface is difficult [70].	Provide market precision in the banking sector by business intelligence and improve stock analysis and sales [70].

28	Tanagra	Helpful in creating machine learning models [95].	Old fashioned that why most of the user does not like it [70].	It is an open-source tool that is helpful in the medical field. It is the best statistical tool that helps to analyze the data set [95].
29	Yahoo S4	Scalable in nature, detect the frauds, helpful for intensive data App [96].	Yahoo S4 use zookeper for configuration of node and clustering setup, but zookeeper have some limitations like lack of replication, and also only 3 or 5 zookeeper nodes, are inherited [96].	It focuses on data locality and detecting the fault also used by Twitter [97].
30	R Language	R language is accessible for a variety of hardware and software. The R gives variety of functions i.e. data manipulation, statistic modelling, and also useful in graphics [98].	R language packages are slower as compared to MATLAB and python. It is difficult to understand because of it steep curve [99].	R is very effective in drugs discovery and also in risk modelling. Its help manufacturing companies for the evaluation of better opinions [99].

CONCLUSION

In this paper, we have briefly discussed the challenges and the popular Tools of Big Data. Big Data analysis and visualization's fundamental problem was overcome with the development of new tools of Big Data, and we reviewed numerous research papers for the study of Big Data tools. However, there no doubt, the Open-Source Tools (Python and R-Programming) were used by IBM, MS, Oracle, and SAP for Big Data analysis and visualization. Our study provides researchers or users guidelines to know their benefits, challenges, and tools of Big Data applications. Different Big Data tools are used in industries, organizations, and research institutions depending upon their needs and requirements. The researchers or users may benefit from our study and decide the right tool of Big Data for their research or their organization depending on their needs and requirements.

REFERENCES

[1] World Population.C19-World News. Retrieved January 19, 2021. https://live-c19-worldnews.com/worldPopulation.php

[2] D. Che, M. Safran, Z. Peng, "From big data to big data mining: challenges, issues, and opportunities." International conference on database systems for advanced applications. Springer, Berlin, Heidelberg, 2013.

[3] F. L. Almeida, "Benefits, challenges, and tools of big data management." Journal of Systems Integration 8.4 (2017): 12-20.

[4] C.L.P. Chen, C.Y. Zhang, "Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data." Information sciences 275 (2014): 314-347.

[5] S. Landset, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." Journal of Big Data 2.1 (2015): 24

[6] B. Curtis, (2020, October 21). What are the 7 V's of Big Data? YourTechDiet.
   a. https://www.yourtechdiet.com/blogs/7vs-big-data/

[7] M.F. Uddin, N. Gupta, "Seven V's of Big Data understanding Big Data to extract value." Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education. IEEE, 2014.

[8] S. Kaisler, F. Armour, J.A. Espinosa, W. Money, "Big data: Issues and challenges moving forward." 2013 46th Hawaii International Conference on System Sciences. IEEE, 2013.

[9] A. Jacobs, "The pathologies of big data." Communications of the ACM 52.8 (2009): 36-44.

[10] M. Zaharia, "Introduction to MapReduce and Hadoop." UC Berkeley RAD Lab.

[11] K. Adnan, R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data." Journal of Big Data 6.1 (2019): 91.

[12] K. Dahal, Y. Ouzrout, P. Barlas, I. Lanning and C. Heavey "A survey of open source data science tools." International Journal of Intelligent Computing and Cybernetics (2015).

[13] J.E. Camargo, C.A. Torres, I.H. Martínez, AND F.A. Gómez, "A big data analytics system to analyze citizens' perception of security." 2016 IEEE International Smart Cities Conference (ISC2). IEEE, 2016.

[14] N. Elgendy, A. Elragal, "Big data analytics in support of the decision-making process." Procedia Computer Science 100 (2016): 1071-1084.

[15] M.H. Iqbal, T.R. Soomro, "Big data analysis: Apache storm perspective." International journal of computer trends and technology 19.1 (2015): 9-14.

[16] S. K. Sahu, Jacintha M. M., & A. P. Singh, (2017, May). "Comparative study of tools for big data analytics: An analytical study." In 2017 International Conference on Computing, Communication, and Automation (ICCCA) (pp. 37-41). IEEE.

[17] AMIS Conclusion. (2019, April 9). What is Apache Drill, and how to set up our Proof-of-Concept? AMIS, Data-Driven Blog - Oracle & Microsoft Azure. https://technology.amis.nl/big-data-database/what-is-apache-drill-and-how-to-setup-your-proof-of-concept/

[18] G. (2020a, April 8). The world of Big Data: Apache Drill and why I need it. Galaktikasoft.
   a. https://galaktika-soft.com/blog/apache-drill.html

[19] Rungta, K. (2021, February 7). Top 15 Big Data Tools | Open Source Software for Data Analytics. BigData Tool. https://www.guru99.com/big-data-tools.html

[20] What is HBase? IBM. https://www.ibm.com/analytics/hadoop/hbase

[21] What is Apache Hive? IBM. https://www.ibm.com/analytics/hadoop/hive

[22] What is Apache Flink? Apache Flink. https://docs.cloudera.com/csa/1.2.0/flink-overview/topics/csa-flink-overview.html

[23] Apache Mahout. Apache Mahout. https://mahout.apache.org/

[24] Techopedia. (2014, August 14). Apache Mahout. Techopedia.Com. https://www.techopedia.com/definition/30301/apache-mahout

[25] Pointer, I. (2020, March 16). What is Apache Spark? The big data platform that crushed Hadoop. InfoWorld. https://www.infoworld.com/article/3236869/what-is-apache-

spark-the-big-data-platform-that-crushed-hadoop.html
[26] N. Kourtellis, G. D. F. Morales, & A. Bifet, (2019). "Large-scale learning from data streams with apache Samoa". In Learning from Data Streams in Evolving Environments (pp. 177-207). Springer, Cham.
[27] Apache Storm. Apache Storm. https://storm.apache.org/
[28] Doddamani, S. (2020, October 8). What is Apache Storm? Intellipaat Blog. https://intellipaat.com/blog/what-is-apache-storm/
[29] BigMLer-The command-line tool for Machine Learning | BigML.com. BigML.Com - Machine Learning Made Easy. https://bigml.com/tools/bigmler
[30] DataStax. (2020b, March 6). Hadoop Vs. Apache CassandraTM | Comparison. https://www.datastax.com/products/compare/hadoop-vs-cassandra
[31] Eliazat, A. (2018, May 16). 18 Big Data tools you need to know - Towards Data Science. Medium. https://towardsdatascience.com/18-big-data-tools-you-need-to-know-ebdb82f2c608
[32] Data Platform (CDP) Big Data Platform. (2021, February 9). Cloudera. https://www.cloudera.com/products/cloudera-data-platform.html
[33] Verma, A. (2020b, January 3). Top 10 Open Source Big Data Tools in 2020 [Updated]. Whizlabs Blog. https://www.whizlabs.com/blog/big-data-tools/
[34] Choi, N. (2018, November 16). Top 30 big data tools for data analysis. Big Data Made Simple. https://bigdata-madesimple.com/top-30-big-data-tools-data-analysis/
[35] D.P Acharjya, K Ahmed, "A survey on big data analytics: challenges, open research issues, and tools." International Journal of Advanced Computer Science and Applications 7.2 (2016): 511-518.
[36] E.LeDell, & S. Poirier, (2020, July). H2o auto ml: Scalable automatic machine learning. In 7th ICML workshop on automated machine learning.
[37] Home Page | HPCC Systems. HPCC. https://hpccsystems.com/
[38] Solutions, E. Jaspersoft Big Data services. Jaspersoft. https://www.e-zest.com/jaspersoft-big-data-services
[39] Team, T. (2020, September 20). Top 10 Big Data Tools for Analysis. TechVidvan. https://techvidvan.com/tutorials/big-data-analytics-tools/
[40] Rungta, K. (2021b, February 3). What is MapReduce in Hadoop? Architecture | Example. MapReduce. https://www.guru99.com/introduction-to-mapreduce.html
[41] G. (2020b, December 23). What Is MongoDB? G Teknoloji. https://www.gtech.com.tr/en/what-is-mongodb/
[42] MongoDB. What is NoSQL? NoSQL Databases Explained. https://www.mongodb.com/nosql-explained
[43] Sharma, R. (2021, January 11). Top 5 Big Data Tools [Most Used in 2021]. UpGrad Blog. https://www.upgrad.com/blog/big-data-tools/
[44] Big Data Discovery | OpenText Magellan. OpenText. https://www.opentext.com/products-and-solutions/products/ai-and-analytics/opentext-magellan-data-discovery
[45] Vargas, V., Syed, A., Mohammad, A., & Halgamuge, M. N. (2016). Pentaho and Jaspersoft: a comparative study of business intelligence open-source tools processing big data to evaluate performances. International Journal of Advanced Computer Science and Applications, 7(10), 20-29.
[46] Top big data tools used to store and analyze data text-magellan-data-discovery. Designing Buildings Wiki. https://www.designingbuildings.co.uk/wiki/Top_big_data_tools_used_to_store_and_analyze_dataentext-magellan-data-discovery
[47] H. B. U. Haq, H.U. R. Kiyani, S. K. Toor, S. Zafar, I. Khalid."The Popular Tools of DataSciences: Benefits, Challenges, and Applications." IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.5, 2020.http://search.ijcsns.org/07_book/html/202005/202005008
[48] Choi, N. (2018b, November 16). Top 30 big data tools for data analysis. Big Data Made Simple. https://bigdata-madesimple.com/top-30-big-data-tools-data-analysis/
[49] Splunk for big data analytics. Splunk. https://www.splunk.com/en_us/big-data/splunk-for-big-data-analytics.html
[50] N. (2020e, October 15). What Is Apache Sqoop? Intellipaat Blog. https://intellipaat.com/blog/what-is-apache-sqoop/
[51] SPSS Modeler - Overview. SPSS. https://www.ibm.com/products/spss-modeler
[52] Vidhya, A. (2020c, July 5). 18 Free Exploratory Data Analysis Tools For People who do not code so well. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/09/18-free-exploratory-data-analysis-tools-for-people-who-dont-code-so-well/
[53] Rungta, K. (2021b, February 1). Talend Tutorial for Beginners: What is Talend ETL Tool [Example]. Talend. https://www.guru99.com/talend-tutorial.html
[54] ReviewDesk, P. (2020b, November 21). TANAGRA. PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices. https://www.predictiveanalyticstoday.com/tanagra/
[55] Xhafa, F., Naranjo, V., & Caballé, S. (2015, March). Processing and analytics of significant data streams with yahoo! s4. In 2015 IEEE 29th International Conference on Advanced Information Networking and Applications (pp. 263-270). IEEE.
[56] Narasimman, L. (2020b, April 17). Apache Drill vs Apache Hive - A comparative analysis. Indium Software. https://www.indiumsoftware.com/blog/apache-drill-vs-apache-hive/
[57] TechCrunch is now a part of Verizon Media. (2012b, August 17). Big-Data-TooL. https://techcrunch.com/2012/08/17/googles-real-time-big-data-tool-cloned-by-apache-drill/
[58] You are being redirected... HADOOP. https://www.mindsmapped.com/hadoop-advantages-and-disadvantages/
[59] S-Logix. BusBeat: Early Event Detection with Real-Time Bus GPS Trajectories –. https://slogix.in/busbeat-early-event-detection-with-real-time-bus-gps-trajectories
[60] Pedamkar, P. (2021, March 3). Uses of Hadoop. EDUCBA. https://www.educba.com/uses-of-hadoop/
[61] Team, D. (2018b, September 14). HBase Pros and Cons | Problems with HBase. DataFlair. https://data-flair.training/blogs/hbase-pros-and-cons/
[62] Big Data. https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783985944/1/ch01lvl1sec15/applications-of-hbase

[63] Hadoop Hive Projects | Hive Real Time Projects. Hadoop Hive. https://www.dezyre.com/projects/big-data-projects/apache-hive-projects
[64] John, T. Data Lake for Enterprises. O'Reilly Online Learning. https://www.oreilly.com/library/view/data-lake-for/9781787281349/3ced0f87-601d-4016-9285-359a45bcdf8b.xhtml
[65] Shah, P. (2018, September 14). Why the Apache Mahout Framework is So Popular. Open Source For You. https://www.opensourceforu.com/2018/09/why-the-apache-mahout-framework-is-so-popular/
[66] Apache Mahout and Spark Comparison – matthew's blog. (2015c, October 22). Apache Mahout. http://matthewbarga.com/blog/index.php/2015/10/22/apache-mahout-and-spark-comparison/
[67] Quora. Mahout. https://www.quora.com/What-are-the-real-world-use-cases-of-the-mahout-Which-all-companies-are-actively-using-mahout-for-machine-learning-purposess
[68] K. Apache Spark Pros and Cons. Apache Spark. https://www.knowledgehut.com/blog/big-data/apache-spark-advantages-disadvantages
[69] Singh, U. (2020, October 7). Top 3 Apache Spark Applications / Use Cases & Why It Matters. UpGrad Blog. https://www.upgrad.com/blog/apache-spark-applications-use-cases/
[70] See some Best-Known Big Data tools, their Advantages and Disadvantages to Analyze your Data. (2019c, May 22). Big Data Tools. https://www.houseofbots.com/news-detail/12023-1-see-some-best-known-big-data-tools,-there-advantages-and-disadvantages-to-analyze-your-data
[71] B. (2018a, December 25). Apache SAMOA – Scalable Advanced Massive Online Analysis. Big Data and Security. https://www.bigdata-security.net/samoa-scalable-advanced-massive-online-analysis/
[72] Companies Using Apache Storm. Apache Storm. https://storm.apache.org/Powered-By.html
[73] Wisdom Jobs. (2019, December 4). Apache Storm Applications - Apache Storm. https://www.wisdomjobs.com/e-university/apache-storm-tutorial-1298/apache-storm-applications-19117.html
[74] BigML Review: Pricing, Pros, Cons & Features. (2019, July 15). CompareCamp.Com. https://www.quora.com/What-are-the-limitations-of-BigML
[75] Team, D. (2018a, September 13). Cassandra Applications | Why Cassandra Is So Popular? DataFlair. https://data-flair.training/blogs/cassandra-applications/
[76] Dryad Data -- Social learning and the demise of costly cooperation in humans. Dryad. https://datadryad.org/stash/dataset/doi:10.5061/dryad.10g95
[77] DRYAD: Financing Sustainable community forest enterprises in Cameroon. World Agroforestry | Transforming Lives and Landscapes with Trees. https://www.worldagroforestry.org/project/dryad-financing-sustainable-community-forest-enterprises-cameroon
[78] H2o. https://www.quora.com/What-are-the-risks-of-using-H2O-ai-framework-When-would-my-company-need-to-pay-anything-to-H2O-ai-Is-the-framework-buggy-somehow-or-is-it-hard-to-install-configure-extend-Do-I-need-to-pay-for-consultancy-eventually.
[79] TrustRadius CAPTCHA. (Jaspersoft. https://www.trustradius.com/products/jaspersoft/reviews?qs=pros-and-cons
[80] Jaspersoft® Studio. Jaspersoft Community. https://community.jaspersoft.com/project/jaspersoft-studio
[81] https://www.researchgate.net/figure/The-advantages-and-disadvantage-of-MapReduce-applications_tbl1_303286828
[82] The Weather Channel Launches New Features in Hours, Not Weeks. MongoDB. https://www.mongodb.com/customers/weather-channel
[83] Chaudhri, A. (2015b, September 24). Advantages and Disadvantages of NoSQL databases – what you should know. Hadoop360. https://www.hadoop360.datasciencecentral.com/blog/advantages-and-disadvantages-of-nosql-databases-what-you-should-k
[84] UK Essays. Applications of Using NoSQL Databases. UKEssays.Com. https://www.ukessays.com/essays/information-technology/applications-of-using-nosql-databases.php
[85] Business Benefits. OpenText. https://www.opentext.com/products-and-solutions/partners-and-alliances/strategic-partners/accenture-and-opentext/business-benefits
[86] OpenText EnCase eDiscovery Pros and Cons | IT Central Station. Open Text. https://www.itcentralstation.com/products/guidance-software-encase-pros-and-cons
[87] Enterprise Content Management Software – ECM Software. OpenText. https://www.opentext.com/about/press-releases?id=7CC4DAE7BE9849D180922A5B3865F9E7
[88] M. (2019e, December 14). What is PolyBase? - SQL Server. Microsoft Docs. https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide?view=sql-server-ver15
[89] M. (2020e, November 13). PolyBase features and limitations - SQL Server. Microsoft Docs. https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-ver15
[90] Basel, K. (2020b, November 2). Python Pros and Cons. Netguru. https://www.netguru.com/blog/python-pros-and-cons
[91] M. (2019c, May 2). 30 Amazing Python Projects for the Past Year (v.2018). Medium. https://medium.mybridge.co/30-amazing-python-projects-for-the-past-year-v-2018-9c310b04cdb3
[92] Splunk Advantages. https://www.learnsplunk.com/splunk-advantages.html
[93] Apache. https://www.quora.com/What-are-the-disadvantages-of-splunk
[94] Excellent Review of IBM SPSS Modeler by a Real User.). SPSS. https://www.itcentralstation.com/product_reviews/ibm-spss-modeler-review-48144-by-altanatabarut
[95] Daniela, J. TANAGRA-A USEFUL TOOL FOR STATISTICS IN MEDICAL APPLICATIONS. In The International Conference Education and Creativity for a knowledge based Society–Computer Science, 2012 (p. 17). Brindusa Covaci.
[96] Chauhan, J., Chowdhury, S. A., & Makaroff, D. (2012, November). Performance evaluation of Yahoo! S4: A

first look. In 2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 58-65). IEEE.

[97] G. (2018c, October 4). Top big data tools used to store and analyse data. Big Data Made Simple. https://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data/

[98] The Benefits of Using R. (2016, March 26). Dummies. https://www.dummies.com/programming/r/the-benefits-of-using-r/

[99] D. Team, (2019, December 31). Pros and Cons of R Programming Language – Unveil the Essential Aspects! DataFlair. https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/