Comprehensive Analysis of Big Data Tools

ISSN (e) 2520-7393
ISSN (p) 2521-5027
Received on 2 Sept, 2018
Revised on 2, Oct, 2018
www.estirj.com

Amna Asif[1] and Irshad Ahmed Sumra[2]

Department of Computer Science, University of South Asia Lahore, Pakistan

Abstract: Big data visualization and classification becomes the main job for database management systems, dealing with big data is not an easy scenario specifically when you are getting different type of data from different places and you have to maintain data security and efficiency. 5 V's of big data introduced to overcome some data transforming issues but still there are many challenges that a big data system may face. There are many tools and techniques that have been designed and developed specifically to handle big data; in this paper we will discuss some focused challenges of big data along with some basic big data tools and techniques, at the end we will compare all the tools to demonstrate that which big data tool is better for big data analysis and visualization.

Keywords: *Big Data ,5 V's of big data, veracity, visualization, classification*.

1. Introduction

Nowadays working with data visualization and classification is one of the fore most tasks for database management systems, for this purpose there are many techniques and tools used for larger data analysis but big data is on top of all these techniques. In today's life dealing with a very large amount of data is not an easy task therefore big data became one of the most important aspect for every organization. Big data makes it reality for many large and small online websites and social media services like Google, yahoo and Facebook to upload and handle terabytes and gigabytes of unstructured data without any loss of data or difficulty in processing. Every organization either large or small have to use Big Data approach for monitoring and organization of large databases through cloud computing and internet. The data to be analyzed or manage by big data processing will be a structured data, semi structured data or unstructured data; for many larger companies the main challenge is not just the amount of data but its variety; as big data systems receive data in many forms and simple systems may not be able to distinguish between un-faulty and faulted data so big data processing may help them to analyze and manage these type of datasets. The main concerns for which big data was introduced are management of continuous large amount of unstructured incoming data, the various varieties of data coming from different places and the rate of growth at which the data is coming to the system. There are many pros of big data but still there are many concerns that should be in researcher or developer's mind when they are handling data with big data, these challenges are data security, cost of data source servers, quality of data, statistical methods and technology defies. There are many large companies that are using big data approach for their database systems such as a group of statistics is using big data for traffic and transport statistics. The same group is also using this approach for handling of their social media messages. Software that is using big data technique developed at Euro stat for price sparring on internet. Other one project is for collection of prices from worldwide retailers.[1]

In today's life big data processing not only made data handling comfortable but there are still some areas this technique facing some problems such as, Heterogeneity and Incompleteness, to avoid these the system should have very good techniques to analyze larger data into databases, this could be done by making tools for data analyzing and classification. Another problem to be faced by big data is the scaling of data, as there may be a lot of data contain by its database and also there will be more coming, normal CPU's with less storage and ram may not be able to handle this. So there should be fast and reliable technique that can handle that data. There is also a problem of handling the rate of speed at which this continuous data is coming in the system, big data processing may get very slow and less effective progress if it doesn't contains faster systems for data transferring. In the maintenance of all these data transferring and analyzing the problem of data security may occurs so all processing systems should be more reliable and secure enough to make data private. Another problem with big data is that it should have some mechanism in which human can interact as sometime systems may not be smart enough to give required information; at that time system should get some input from human to clarify and verify the resultant output.

The big data architecture contains three layers, the first one from bottom is called infrastructure layer that contains networking and connection between storage devices and computations. The central layer is called computing layer or data layer that contains database management and analysis of data through some programming model. The last layer is application layer that contains handling of required classified data.[2]

Corresponding author Email: amna_aasif@yahoo.com

The rest of the paper is organized as follow: in next section we will discuss 5 important V's of big data; in third section we will discuss some tools of big data along with their key features; fourth section shows some real time challenges for big data and at the end the fifth section will show the conclusion based on information we acquired from this paper.

2. 5 V's of Big Data

The handling of big data is not an easy task and many innovative tools and techniques are being developed to handle Big Data in efficient way; to make big data easy to understand the whole concept of big data is divided using five V's: Volume, Velocity, Variety, Value, and *Veracity*. All of these are defined as follows;

2.1 Volume

Volume refers to the large amount of data that companies generate each day and has to save that data for some time but their own processing methods may not be able to handle that larger amount of data. So there big data processing techniques are very useful to handle any volume of data.

2.2 Volume

Velocity refers to the speed of data at which data is coming to the systems; data may come with the difference of milliseconds so database processing systems has to place them accordingly and efficiently.

2.2Variety

Variety refers to the diversity of data types and sources. Data may come in a much unstructured form and does have any relationship with other data sources; the data may be in the form of images, audios, videos or text. So there big data techniques are able to process variant data into sorted and structured form.[3]

2.3 Veracity

Value refers to the quality of data; after a lot of volume of data types and sources came to big data systems data processing systems must make it sure that the data should not lost its quality and credibility. Either many times big data processing techniques are not very efficient that they save hundred percent of the quality but they can be sure of seventy to eighty percent of the quality.

2.4 Value

Value refers to the worth of data that is coming to the system, as there is a lot of data in the system but it will be useless if it is not defined with its importance. Value shows the value of data and its relationship with other data sources; so that it later can be able to collect, monetarize and analyze.[4]

3. Tools of Big Data

As big data is fast growing technology of today's era and following many challenges for analyzing, monitoring and storing of data; so numerous groups and people are working on its improvement and efficiency by introducing newer tools and techniques with enhancement from previously introduced tools. Here we will discuss some of the very common and efficient tools.

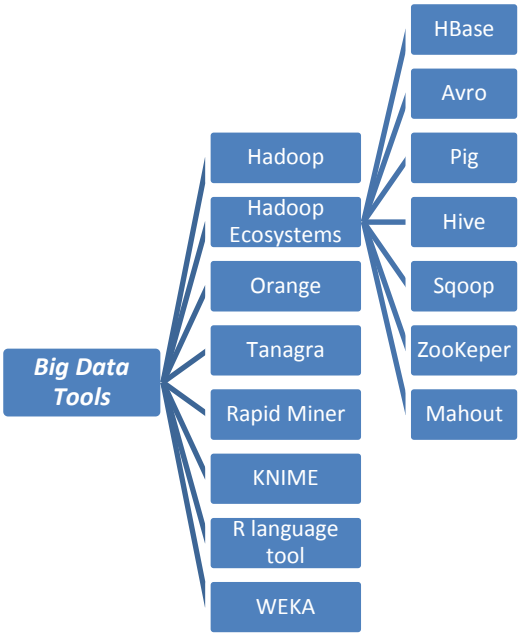

Figure 1: Big Data Tools

3.1 Hadoop

Hadoop refers as a most common and successful open source framework for big data analysts which is used for storage and processing of big data sets on cluster of commodity hardware. Because of the vast efficiency and accuracy of Hadoop it is used to handle and store any type of large data from different sources in distributed computing systems.[7] The storage space for big data could be in peta bytes or terabytes there for the basic purpose of Hadoop framework is to provide an efficient and unstructured large storage framework with realistic cost for both industrial and academic purposes. One of the main features of Hadoop is to provide parallel processing along with distributed storage using cluster computing so that data can be analyzed and monitored in efficient way.[8] Hadoop is a framework that encompasses distributed file system, libraries, implements a version of the MapReduce programming model for large scale data processing, and a resource management platform Hadoop is based on a client/server architecture where file system is handle by a server node called name node and data and storage is handle by data nodes. Name node is used to handle opening, closing and naming conventions of files that are placed in storage in form of block, whereas data nodes are used for file reading and writing requests by clients.

Hadoop provide master slave architecture for fast and parallel processing through MapReduce and HDFS. It uses MapReduce model for parallel processing of large data setsin the form of chunks or blocks on clustered computers. HDFSfile system is also used by Hadoop for data storage;

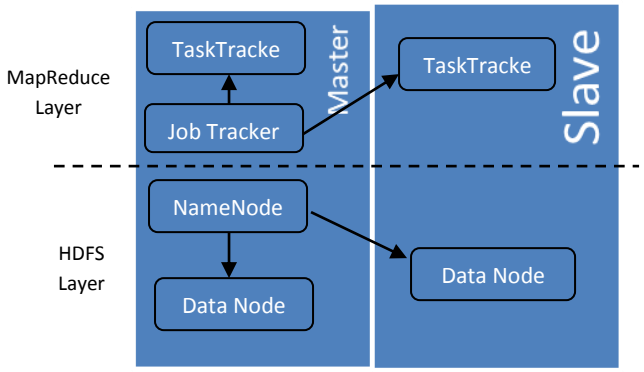

Figure 2: Hadoop architecture

HDFS also can store data files as a whole or can divide them into chunks for rapid processing.

The main features of Hadoop framework is its dynamic nodes handling without changing of data formats, its cost effective parallel computing, and ability to use any type of data sets and fault tolerance. Hadoop has a lot of features but still there are some limitation such as Hadoop cannot be able to update previously added data it can only add new data, because of its model and other features Hadoop cannot be used for many interactive and transactional applications. Hadoop provide more powerful and efficient distributed processing by handling structured and unstructured data in more better and efficient way than a traditional data warehouse.

Key Features

- Hadoop provides flexible data processing for structured and unstructured data to get efficient results [9].
- There would be one or more name node and many data nodes so that data can be read and write accurately and efficiently [10].
- Hadoop provides fault tolerance for data accuracy errors and for data loss.
- Hadoop offer parallel distributed systems that make data processing more efficient than normal storage systems [9].
- When using parallel system, Hadoop provide cost efficient storage by taking lesser amount of servers [9].

3.2 Hadoop Ecosystems

There are many big data tools that are used as Hadoop ecosystems, Hadoop ecosystems are also very efficient system that provide efficient data processing and manipulation. Hadoop architecture may be based on one of these ecosystems to make its work more efficient.

3.2.1 HBase

HBases is an open source distributed column based non-relational database system and use java as a base language; it is useful for data reading and writing on file systems through MapReduce and provide storage for Hadoop distributed computing. It works like Google tables and provide fast lookups for larger tables using hash tables [4].

HBase architecture consist of two main structures: the HBase master for stored data information and some region servers for data manipulation; each region contains several blocks of stored data and every region is separated from other region using start and end key values.[9]

Key Features

- It is integrated in Hadoop as a source and destination of data.
- It is linearly scalable.
- It provides fault tolerance.
- It can read and write files consistently with efficiency.
- It provides user friendly java API.[11]

3.2.2 Avro

Avro is used for compact and fast data serialization format to provide interoperability between components. Avro is a serialized and procedural call framework that is why the main purpose of Avro is to provide data serialization and deserialization very efficiently and to provide storage for persistence data in HDFS. Avro provide rich data types and support Java and Python programming languages [4].

Key Features

- Avro provides API‘s for C, C++, Java, Ruby and

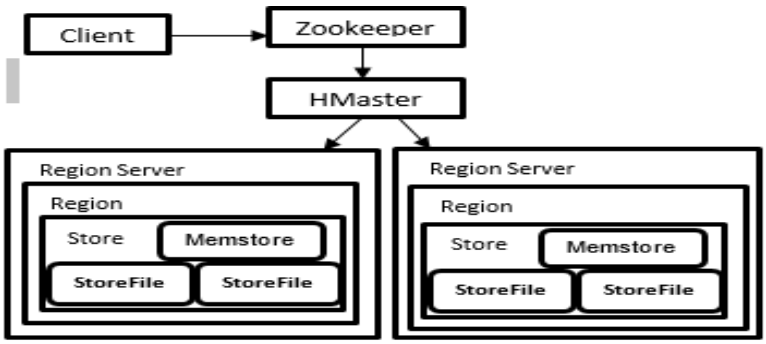

Figure 3: HBase Architecture

- Python.
- It provides fast binary data structure.
- It provides rich data serialization and deserialization.
- Define data types and protocols using JASON.

3.2.3 Pig

Pig is also an ecosystem of Hadoop that provides platform to data processing tasks for filtering, transforming and maintaining data more efficiently. This platform handles a very complex and larger amount of data with its defined algorithmscalled Pi. Pi is based on Pig Latin language that can be able to execute on JVM and Hadoop clusters [4].

As shown in the architecture figure 4; pig Latin offers scripted algorithm to the system and then system uses

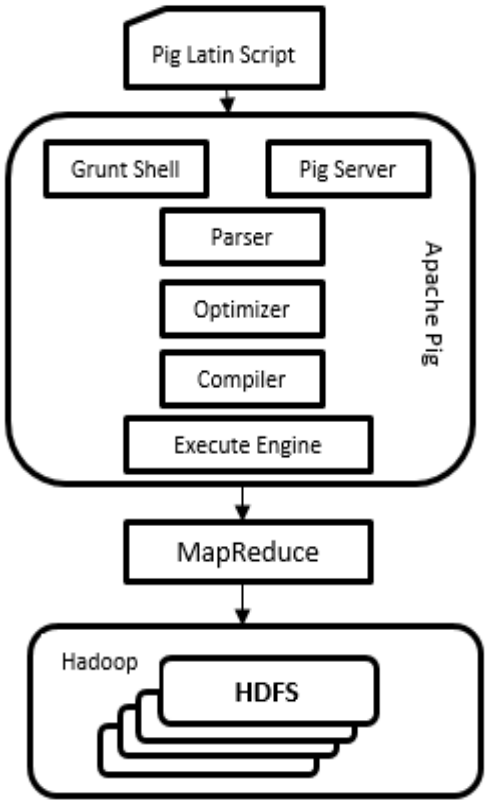

Figure 4: Pig Architecture

MapReduce and HDFS for more efficient data processing.[11]

Key Features

- It provides data processing into multiple functions through multi-steps data flow applications.
- It uses SQL-like for data manipulation.
- It provides simplified data processing for decision making [12].

3.2.4 Hive

Hive is another ecosystem for Hadoop; it provides framework of data ware housing that helps to write SQL queries for the processing and analyzing of big data systems; it acts as an intermediate between HDFS and user. Hive provides efficient tasking for larger immutable data sets using batch jobs [4]. Figure 5 shows the Hive architecture where hive provides user interfaces along with programming interfaces that are JDBC and ODBC. Then there is a thrift server API for query execution.The Driver helps thrift server by compiling and executing query; and send the result to next level for further processing using MapReduce. Another basic component of this architecture is Metastore that contains information about data tables.[13] Figure 5 shows the Hive architecture where data is processing parallel for each client server and after passes from driver it goes to job tracker and name nodes for clustering and classification.

Key Features

- Allow users to write Hive query language with SQL standards.
- It is designed as a scalable, extensible and fast query language.
- It provides support for command line, web UI and

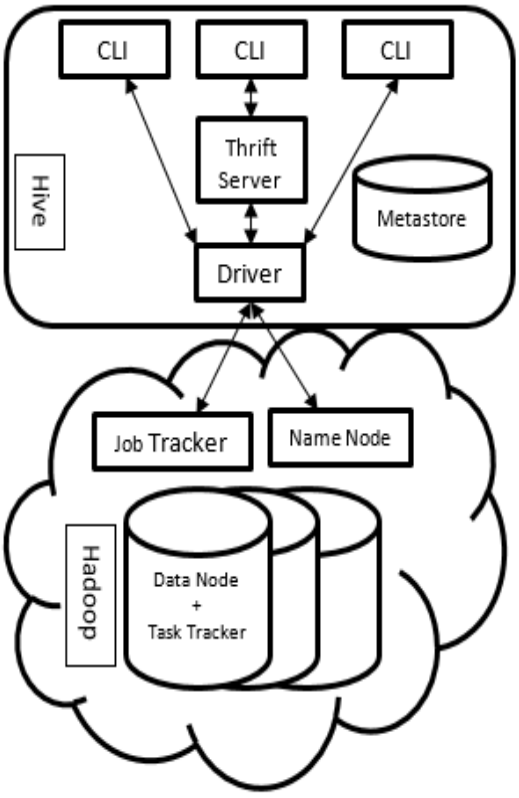

Figure 5: Hive Architecture

HD insight as user interface [13].

3.2.5 Sqoop

Sqoop ecosystem works as a data transfer tool to export and import required data between relational databases and Hadoop systems. It can provide data storage though some external source that makes data processing more efficient. It can automate data import and export by dividing the processing into different sub tasks [4].

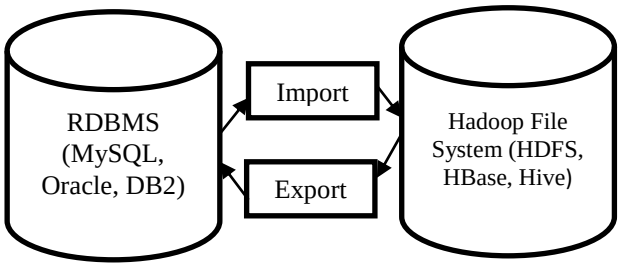

Figure 6: Sqoop Architecture

Key Features

- It uses cloud computing techniques for data import and export.
- It can provide a moderate amount of storage to other systems and processing.
- It uses JDBC base implementation.
- It can be able to integrate with Hive.
- It provides data security for external data processing [14].

3.2.6 ZooKeper

Zookeeper systems provide distributes services for Hadoop clustering, this system offers efficiency by point out the defaulted nodes as if one block or node is not working properly then Zookeeper will track the node which is not working properly and gets the information of this node to required protocol[4]. It can provide faulted information to original systems synchronously.

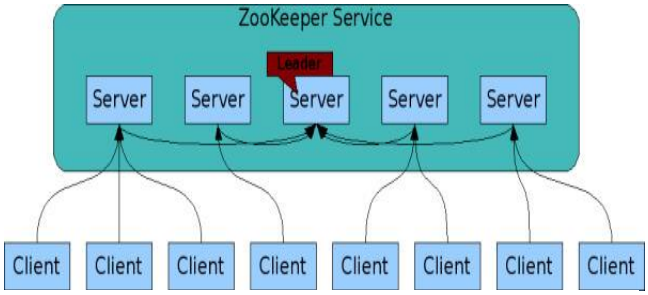

Figure 7: ZooKeeper Architecture

Figure 7 shows the zookeeper architecture where it shows the parallel processing of different servers against different clients for fast and efficient services.

Key Features

- It supports distributed systems and group services.
- It can be accessible by different users concurrently at same time.
- It can act synchronously or asynchronously

3.2.7 Mahout

Mahout is a collection of machine learning functions that can act as a library or application which is used by Hadoop architecture; and is used for collectiveness, filtering, categorization, clustering and mining of parallel data blocks through Map Reduce [4]. Figure 8 shows Mahout Architecture for distributed processing of large data sets from different users to give subsequent recommendations to every data storage for data clustering and data classification.

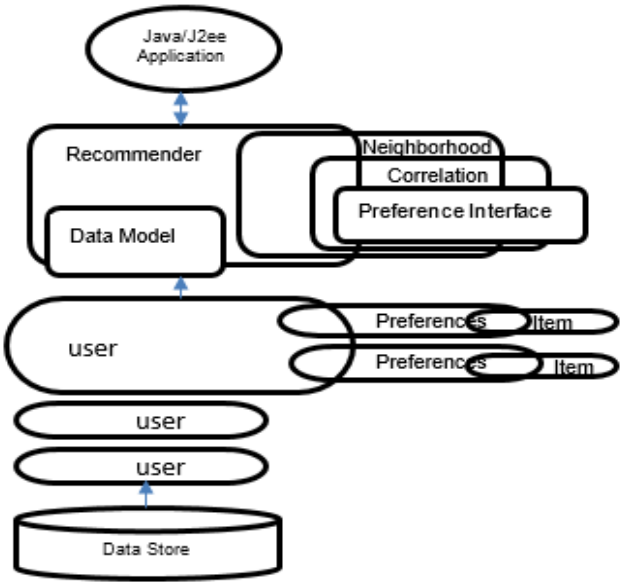

Figure 8: Mahout Architecture

Key Features

- It can be able to build data mining frameworks and machine learning tools.
- It can evaluate different combination of parameters very efficiently.
- It provides parallel processing for efficient data processing.
- It can provide applications for recognitions.

3.3Orange

Orange is a component based open source tool for processing and mining big data. Main responsibilities of this tool are graphic visualization, classification of data, regression, evaluation, data association and unsupervised of data. All these functionalities are engaged as widgets that support most common tasks of data science and are visually available; all of these widgets can be combined together for the process of data science. Languages that can be used by this tool are python and C++.[15]

Key Features

- Orange is a component-based data mining and machine learning software suite.
- It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques.
- Data mining in Orange is done through visual programming or Python scripting
- Open source data visualization and analysis for novice and experts.
- It contains components for machine learning and add-ons for bioinformatics and text mining. Along with

this, it can also provide features for data analytics. [11].

- Specialized for data visualization along with mining.

3.4 Tanagra

Tanagra was built for research and study purposes. This tool can perform visualization, descriptive statistic, instance selection, feature selection, feature construction, regression, factorial analysis, clustering, supervised learning, meta-spy learning, learning assessment and associate rules. This tool is based on multiple implementations of various algorithms whereas it is developed in Delphi [15]. TANAGRA acts more as an experimental platform and its main purpose is data management. Another purpose is for the direction of the novice developers, consists in diffusing a possible methodology for building this kind of software. The programmers took advantage of free access to source code, to look how this sort of software is built; this will help the programmers to avoid the main steps of the large project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques [16].

Key Features

- It provides a large amount of unit- and multivariate parametric and nonparametric tests.
- It provides efficient algorithms for clustering, feature selection, supervised learning assessment and feature selection.
- It provides easy data accessibility through data warehouses and databases.
- It also provides the techniques for data cleansing and shared utilization of data.[17]

3.5 Rapid Miner

Rapid Miner is also a tool for big data which is not fully available as open source except some basic parts, such as k-mean clustering and some other algorithms are available as open source. The main features of this tool are performance process, import, export data, data transformation, classification, regression, evaluation and connect to a repository. The open source version of this tool gets data from CSV and MS excel, only full version of it will be able to access databases. [15]

Key Features

- It is a machine learning based processing tool.
- It uses modular operator to design complex and large number of problems.
- It proceeds with XML techniques for tree model knowledge processes.
- It provides flexible input and output processing.
- It provides efficient regression, classification and clustering.
- It supports different learning algorithms.
- It can be able to read and write data from different files and databases.
- It provides rapid actions against any error.[3]

3.6 KNIME

Another tool of big data is called KNIME, pronounced "naim", which is available in various versions but the most advanced versions of this tool, like big data extension, are not open sourced and is featured with very advanced sub tools. Its main features are univariate and variate statistics, data mining, time series analysis, image processing, web analytics, text mining, network analysis and social media analysis. KNIME is an eclipsed platform based tool that provides open API for extended functionalities. [15]

Key Features

- It provides processing under the development environment of IBM's eclipse.
- It provides platform to visually create data flows, execution of some tasks and experimental results.
- It provides combination of hundreds of nodes for data reading and writing, data preprocessing, analysis and modeling.
- It provides user friendly views like scatter, parallel coordinates and other interfaces.[3]

3.7 "R" Language Tool

"R" language is also an open source tool available for almost all operating systems, which is used for statistic and graphing, as R language tool is used for big data so it is also used for many known sectors that are dealing with big data. The best feature of R language tool is that it can be able to integrate with many other tools to organize more advanced routines. There are many featuring tasks that R tool can perform but some of main features are clustering, regression, time series analysis, text mining and statistical modeling. The R tool is considered more a language than a tool. This tool also provides big data processing using RHadoop. [15]

Key Features

- It provides platform for data analysis, user interfaces and software development tasks for big data and related activities.
- R language tool is an open source framework that provide command line driven and statistical packages along with the support to get merge with other platforms.
- It provides machine learning, data mining and statistical techniques.
- It can process efficiently without depending on the features of a computer system [3].

3.8 WEKA

Another tool described is Weka which is used for knowledge analysis and is licensed under GNU and

provides a combination of machine learning and java programming packages. Open source API of Weka is available specifically for academic and business purposes. There is a high amount of users who are using Weka API for their own purposes and they also contributed for many add-in packages of Weka for data mining and decision trees. [15]

Key Features

- Data preprocessing, classification and clustering.
- Interfaces provided by Weka are very user friendly and efficient. [3]
- Weka is a java based open source tool that provides efficiency.
- It is based on three tier graphical user interface that includes preprocessing analyzer, attribute selector and an experimenter for testing.
- It provides another interface of user's inputs and commands.

4. Challenges for Big Data

As big data has to analyze and monitor large amount of data whereas continuous data comes in unstructured form, so there are a lot of challenges that big data techniques are overlooking.

4.1 Data Filtering

As every big data system is getting a large amount of data continuously so it is important for big data techniques to filter the useful data from redundant data.

4.2 Data Ownership

Data comes from different servers and systems and that data may also be same as other data came from the other systems so big data techniques has to take care of the data owner and make one's data separate from others.

4.3 Security

As there are lots of data sources and variety of data comes to big data and the data is coming from different places and all sources want their data to be secure and save. So it is important for big data techniques to make all the data secure. [5]

4.4 Data Complexity

As large amount of data comes continuously in unstructured form as it could be a video, image or text form so it is difficult for data interpreted tools to analyze, store and interpret very large amount of data in lesser time and more efficiently. There are also many challenges that big data is facing nowadays but it may be overcome by improving big data techniques day by day.[6]

5. Conclusion

In this paper, we studied some valuable characteristics of big data along with some challenges of big data; whereas there are also 5 V's of big data that are introduced to overcome many challenges but still there are some challenges that a user or organization must consider for a system when using big data technique; we also get briefed about many different big data tools that are introduced differently for many different purposes and all have their own features.

Table 1: Characteristic of some Big data Tools

	Orange	Tanagra	*Rapid Miner*	*KNIME*	*R Language*	*Weka*
Developer	Univ. of Ljubljana, Slovenia	Lumiere University Lyon 2	Rapid Miner, Germany	KNIME.com AG, Switzerland	Worldwide Development	Univ. of Waikato, New Zealand
Programming language	C++, Python, Qt frame	Delphi	Java	Java	C, Fortran, R	Java
License	Open source	Open source	Open source	Open Source	Free software	Open source
Latest version	2.7	1.4.50	6	2.9.1	3.02	3.6.10
GUI/ command line	Both	GUI	GUI	GUI	Both	Both
Purpose	General data mining	General data mining	General data mining	General data mining	Computation and statistics	General data mining

References

[1] Bogdan OANCEA, Raluca Mariana DRAGOESCU, "Integrating R and Hadoop," *Revista Română de Statistică,* 2014.

[2] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, "A Review Paper on Big Data and Hadoop," *International Journal of Scientific and Research Publications,* vol. 4, no. 10, 2014.

[3] Kalpana Rangra , Dr. K. L. Bansal, "Comparative Study of Data Mining Tools," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 4, no. 6, 2014.

[4] V. B.Bobade, "Survey Paper on Big Data and Hadoop," *International Research Journal of Engineering and Technology,* vol. 3, no. 1, 2016.

[5] Priya P. Sharma, Chandrakant P. Navdeti, "Securing Big Data Hadoop: A Review of Security," *International Journal of Computer Science and Information Technologies,* vol. 5, 2005.

[6] M. K. a. S. Maniyam, Hadoop Illuminated, GitHub, 2013.

[7] K. Gauraw, "Big Data And Hadoop – Features And Core Architecture," Data Integration Ninja, 2015.

[8] S. P, "Features of Hadoop HDFS – An Overview for Beginners," *HDFS Tutorials,* 19 Aug 2016.

[9] Ainhoa Azqueta-Alzuaz, Marta Pati ´ no-Martinez,Ivan Brondino, Ricardo Jimenez-Peris, "Massive Data Load on Distributed Database Systems over HBase," *17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing,* 2017 .

[10] A. P. V. S. Nikita Bhojwani, "A SURVEY ON HADOOP HBASE SYSTEM," *International Journal of Advance Engineering and Research,* vol. 3, no. 1, 2016.

[11] Z. A. Swarna C, "Apache Pig - A Data Flow Framework Based on Hadoop Map Reduce," *International Journal of Engineering Trends and Technology (IJETT),* vol. 50, 2017.

[12] Satya S. Sahoo, Annan Wei, Joshua Valdez, Li Wang, Bilal Zonjy, Curtis Tatsuoka, Kenneth A. Loparo and Samden D. Lhatoo, "NeuroPigPen: A Scalable Toolkit for Processing Electrophysiological Signal Data in Neuroscience Applications Using Apache Pig," *Neuroinform,* 6 June 2016.

[13] Ashish Thusoo , Joydeep Sen Sarma , Namit Jain , Zheng Shao , Prasad Chakka , Suresh Anthony , Hao Liu , Pete Wyckoff , Raghotham Murthy, "Hive- A Warehousing Solution Over a Map-Reduce Framework," in *PROCEEDINGS OF THE VLDB ENDOWMENT*, 2009.

[14] Veershetty Dagade, Mahesh Lagali, Supriya Avadhani, Priya Kalekar, "Big Data Weather Analytics Using Hadoop," *International Journal of Emerging Technology in Computer Science & Electronics* , vol. 14, no. 2, 2015.

[15] R. Baker, "https://intellipaat.com/tutorial/hadoop-tutorial/sqoop-impala/".

[16] H. Wimmer, "A Comparison of Open Source Tools for Data Science," in *Information Systems Applied Research* , Wilmington, North Carolina USA , 2015.

[17] R. PRIY, "AN ANALYSIS ON DATA MINING TOOLS FOR BUSINESS," *INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN DATAMINING AND CLOUD COMPUTING*, vol. 3, no. 8, 2015.

[18] Asha Rajkumar, Mrs. G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm," *Global Journal of Computer Science and Technology,* vol. 10, no. 10, 2010.

[19] Seref SAGIROGLU and Duygu SINANC, "Big Data: A Review," *IEEE,* 2013.

[20] Min Chen · Shiwen Mao · Yunhao Liu, "Big Data: A Survey," *Springer Science+Business Media New York,* 2014.

Amna Asif received the B.S degree in Computer Science from the Lahore College for Women University, Lahore, Pakistan, in 2014, and the M.S. degrees in Computer Science from the University of South Asia Lahore, Pakistan, in 2018. In 2011, she joined the Department of Databases in a reputed software house. Her current research interests include database administration, Big data Tools, Hadoop architecture and integration of Big Data with Python.