IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

A Review on Big Data Analytics Tools in Context with Scalability

Ajay Kumar Bharti[1], Neha Verma[2*], Deepak Kumar Verma[3]

[1,2]Department of Computer Science, Maharishi University of Information Technology, Lucknow, India
[3]Department of Computer Science, JNPG College, University of Lucknow, Lucknow, India

[*]*Corresponding Author: nehaverma2108@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i2.273277 | Available online at: www.ijcseonline.org

Accepted: 12/Feb/2019, Published: 28/Feb/2019

***Abstract*—** In current scenario the rapid growth in the size of generated data is so huge and complex that traditional data processing application tools and platforms are inadequate to deal with it. Therefore, the big data require suitable analysis mechanisms for data processing and analysis in an efficient and effective manner. Consequently, developing and designing new scalable data mining techniques is very important and necessary mission for researchers and scientists in the last years. Scaling is the ability of the system to adapt to increased demands in terms of data processing. To support big data processing, different platforms incorporate scaling in different forms. We had tried to analyze these platforms on the basis of their performance in different environment.

***Keywords*:** Big data, Scalability, Hadoop

I. INTRODUCTION

Big data is an abstract concept. Apart from masses of data, it also has some other features, which determine the difference between itself and "massive data" or "very big data." At present, although the importance of big data has been generally recognized, people still have different opinions on its definition. In general, big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time. Because of different concerns, scientific and technological enterprises, research scholars, data analysts, and technical practitioners have different definitions of big data. The following definitions may help us to have a better understanding on the profound social, economic, and technological connotations of big data.

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." On the basis of this definition, in May 2011, McKinsey & Company, a global consulting agency announced Big Data as the next frontier for innovation, competition, and productivity. Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software. This definition includes two connotations: First, data sets" volumes that conform to the standard of big data are changing, and may grow over time or with technological advances; second, data sets" volumes that conform to the standard of big data in different applications differ from each other. At present, big data generally ranges from several TB to several PB. From the definition of McKinsey & Company, it can be seen that the volume of a dataset is not the only criterion for big data. The increasingly growing data scale and its management that could not be handled by traditional database technologies are the next two key features.

Figure 1: One Minute in Internet 2017

From a broader perspective, the big data platforms can be categorized into the following two types of scaling:

1.1 Horizontal Scaling:

Horizontal scaling involves distributing the workload across many servers which may be even commodity machines. It is also known as "scale out", where multiple independent

machines are added together in order to improve the processing capability. Typically, multiple instances of the operating system are running on separate machines.

1.2 Vertical Scaling:
Vertical Scaling involves installing more processors, more memory, and faster hardware, typically, within a single server. It is also known as "scale up" and it usually involves a single instance of an operating system.

II. RELATED WORK

Many studies have been conducting on developments and historical reviews in BDA area. A short historical overview of Big Data starting from 1944 introduced by Gil Press based upon Riders search. This work provides the evolution of Big Data starting from 1944 to 2012. The study indicated that the border among the expansion of data and Big Data became unclear. In a comprehensive study covering all BDA and Data Science events until the end of 2013 has been introduced.

The origin of the Big Data established by Frank Ohlhorst in 1880 when the US was doing its 10th census, in the 19th century, The greatest problem faced by the US was the statistics aspects, these aspects were, how to document and survey and 50 million citizens of North-American. On the other hand, Winshuttle consider the foundation of Big Data in the 19th century. They presented the Big Data as the datasets which are so complicated and beyond conventional management and process capability. In addition, Winshuttles review presents estimation for the growth of data upto 2020.

Bernard Marr presents the longest historical survey. This survey concentrates on the historic bases of the Big Data, which are several techniques for people to collect, store, analyze data. Marr showed that Erik as the pioneer of the Big Data. Erik presented an article in the Magazine of Harper and this article reprinted also in the Washington post-1989 presenting afirst description of the Big Data concept.

A great discussion about Big data origin has been presented by Steve Lohr. In this discussion he was trying to identify when the Big Data expression had been used for the first time.

It was difficult to hunt the source of Big Data because the expression of the Big Data is so generic. Instead, the aim was the initial usage of the Big Data term that recommends its current understanding that is, not only the massive quantity of data, but also distinct forms of data managed in modern ways. Cox and Ellsworth presented the origin of Big Data from another point of view. They introduced a relatively precise significance to the current view of Big Data, which they defined datasets are too huge to be managed in the capacities of local disk, basic memory, and even remote disk. This problem is considered as Big Data.

A great visualizing historical survey was conducted. This review dedicated to the time-line of the implementation of BDA. This historic discussion is basically specified by actions and events related to Big Data pushed by several IT association like Google, Facebook, Yahoo, Youtube, Twitter, and Apple. Figure 1 shows the history of Big Data and Big Data Analytics tools.

Data Mining refers to the procedure of exploring valuable knowledge like patterns, associations, anomalies, changes, and critical components from huge volumes of data which has been accumulated in data warehouses, databases, or any other data repository. Also, data mining is classified as a very powerful tool to discover hidden relationships in large quantities of data. Now, the modern aspect of Big Data is used to recognize the datasets, which are of a vast size and have higher complexity. So the conventional data mining platforms or methodologies cannot manage, store, and analyze these huge quantities of data. Big data contains composite collections of both unstructured and structured data types. Now, much of the data science efforts are chiefly interested in handling unstructured types of data. Mining of Big Data refers to the ability to extract useful information from these massive data sets, this was very difficult before due to its variety, volume, velocity, and veracity.

Knowledge extraction is very useful and the knowledge mining is the representation of various types of patterns in meaningful way. One of the most important tasks for Data Mining is to analyze data from different perspectives and summarize it into beneficial information that can be utilized as good solutions in business and trying to predict the future direction. Information Mining helps the associations to take better knowledge-driven decisions. Data Mining is known as Knowledge Discovery in Databases (KDD). Data Mining make use of many computational methods from information retrieval, statistics, , machine learning and pattern recognition. Data mining tasks can be categorized into summarization, classification, clustering, and association analysis.

With tremendous technological advances, which led to an increase in data storage capabilities, processing methods and data availability was the basic reason for the growth and emergence of Big Data. Big Data technology attempting to present good business serves and helping in decision making via handling, managing, and processing huge datasets (these datasets may be private, general, and enterprise specific). We can define the process of Mining Big Data as the activity of dealing with big data sets to explore useful information. Big data can be found in various fields such as: natural disaster, atmospheric science, astronomy, social networking sites, life sciences, medical science, government data, web logs, mobile phones, and sensor networks.

III. METHODOLOGY

In this section we have studied and compared the popular Big data analytics tools on the basis of their performance.

3.1 Hadoop

Hadoop is one of the open source frameworks. It had written in Java and it had his debut in October 2003. Hadoop has given a very important advantage in allowing the handling and processing of massive datasets through the use of clusters of computers using simple software models. Hadoop is designed to extend the processing and storage of huge data using thousands of computers rather than using a single node. The essence of the Hadoop consists of two parts: the first is the part of the data storage, which is known as Hadoop Distributed File System (HDFS). The second one, which is responsible for the process of data processing through MapReduce programming paradigm. The nature of the work is based on dividing the large data files into small blocks distributed on the machines in the cluster to be handled in a parallel way by writing programming codes in MapReduce.

The base of Hadoop paradigm consists of the following four modules which are shown in Figure.

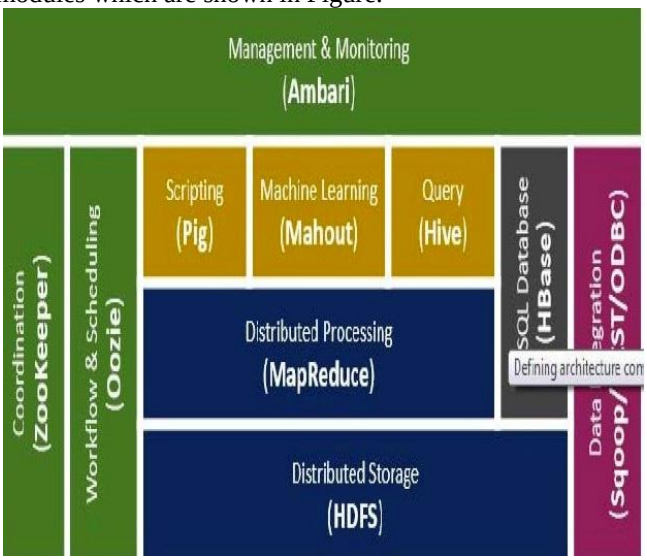


Figure: Architecture of Hadoop

- **Hadoop Common**: Containing utilities and libraries in Java which are required to start Hadoop as it is necessary for the operation of other units such as Hive, HBase etc.

- **Hadoop Distributed File System (HDFS)**: Is considered to be the secret to the success of the Hathoub system, as it represents the class responsible for data storage. It first divides the large data files and then distributes them to the different machines within the system, allowing high speed to access and collect data. It produces many copies of the data files and keeps them on a number of nodes in the Hadoop cluster until the process of processing finished.

- **Hadoop YARN**: It is a model for managing the resources and job scheduling in Hadoop cluster.

- **Hadoop MapReduce**: A Java-based programming model has been created by Google to handle the data stored in the HDFS. MapReduce disaggregates the task of processing big data into smaller tasks. MapReduce aggregates data and the results from the various machines in the Hadoop cluster to end the analysis process in a parallel model.

3.1.1 Hadoop Data Access Modules

- **Pig:** Pig is designed by Yahoo to make the process of analyzing large data easier and more efficient. Its main mission is to provide a high level of data flow in the system of Hadoop and it is characterized by ease of use and scalability. The most important feature of Pig is its open structure, making it easy to handle large data in parallel.

- **Hive:** Hive was developed by Facebook and is a data warehouse that has been placed on the top of the Hadoop, providing a simple language such as SQL called HiveQL, used to analyze, query and summarize big data. The most important feature is that it makes the process of querying big data faster because of its use of the indexing process.

3.1.2 Hadoop Data Integration Modules

- **Sqoop:** The main task of Scoop is to transfer data from its external sources and place it in the designated places in the Hadoop system, such as HDFS also used to transfer data from inside the Hadoop to store in external sources of data. Scoop is a data transfer that works in a parallel way, making it more effective in data analysis and faster copying of data.
- **Flume:** The main use of the Flume is to collect and aggregate huge data from its main sources and send it to any proper layer in the Hadoop system which is HDFS. The task of data flow is accomplished by Flume 3 basic structures, which are channels, sinks, and sources. The process of data flow by Flume is defined as a factor, as are the Bits of the data transmitted by the Flume is known as events.

3.1.3 Hadoop Data Storage Modules

- **HBase:** The HBase component is a column-oriented database that stores data in HDFS. HBase Supports random reading and batch computing by using MapReduce. We can make huge tables containing

millions of rows and columns and stored in the Hadoop system using HBase NoSQL. The best use is when you need to read or write randomly to access large data.

3.1.4 Hadoop Monitoring & Management Modules

- **Oozie:** The Oozie component is a process scheduler that uses a directed graph to create workflow maps. Oozie stores the current tasks, their cases and their changes with the workflow map to control the implementation of the running tasks in the Hadoop system. Workflow maps extracted by Oozie depend on time and data dependencies.
- **Zookeeper:** Zookeeper is the ultimate coordinator and provides fast, simple, reliable and ordered services in the Hadoop system. Its primary task is synchronization service and distributed configuration service to provide a list of distributed system naming.

3.2 Apache Spark

Apache Spark is one of the latest and fastest data analysis tools based on cluster computing technology.

It is designed and developed based on Hadoop MapReduce system where the MapReduce framework has been developed to become more effective for use in more types of computations such as processing data streams and interactive queries. The main advantage of Spark is that it uses the method of in-memory computation, which makes it faster in processing the application. Spark has been developed to cover a wide range of workloads in iterative algorithms, interactive queries, and batch computation. Its basic structure was the resilient distributed dataset (RDD) which enabled it to deal with the analysis of data in parallel, in a cluster of a large number of computers and in fault tolerance way.

3.2.1 Apache Spark Features

Apache Spark has many important features:

- **Speed:** The Apache Spark is very fast in carrying out tasks in the Hadoop system, which is 100 times faster when used in-memory and up to 10 times when using disk storage. The speed of Spark is due to the fact that reading and writing data from and to memory has decreased. spark stores current operations in memory.
- **Supports the use of many programming languages**: Spark is characterized by the presence of built-in APIs in many programming languages such as Scala, Java, Python, or R. This has made it easier for users to write their applications in any of the available programming languages. Spark's new releases are supported by more than 80 high-level interactive searches.
- **Advanced Analysis:** Spark supports the use of Streaming data, SQL queries, Graph algorithms, and Machine learning (ML), as it also supports the use MapReduce model.

3.2.2 Spark Modules

The Figure depicts the distinct modules of the Spark which are given as the following.

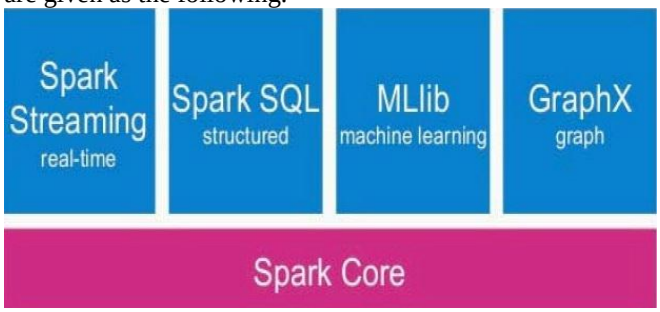


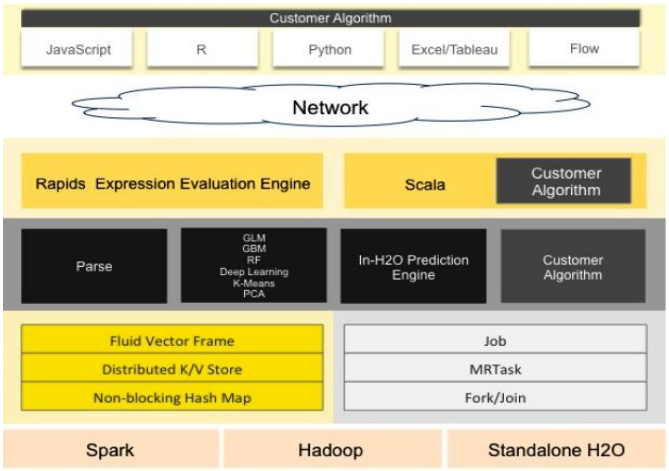
FIGURE: Spark Modules

- **Spark Core**: Spark Core is the main component of Apache Spark execution engine and all other Spark components have been built on it. It provides the in-memory computation model and it contains the referencing of datasets in the external storage system.
- **Spark SQL**: Spark SQL is an Apache module built on Spark Core. It provides a type of data known as Schema RDD. It also provides support for analyzing structured and semi structured data through the query process.
- **Spark Streaming:** Spark streaming is the component that has increased the speed of SparkCor fast scheduling capabilities to enable analysis of data streams. It receives data streams and analyzes them in real time.
- **Machine Learning Library (MLlib):** MLlib is a framework for distributed machine learning build on Spark Core and contains an implementation for the majority of the machine learning techniques. It takes the place Mahuot in the Hadoop system and is characterized by the super-speed as it relies on in-memory computation.
- **GraphX**: GraphX is a paradigm for distributed graph-processing build on Spark Core. It provides an excellent programming interface that allows the use of Pregel abstraction API to write applications of the of the graph computation executed in an ideal time.

3.3 H2O

H2O is one of the big data analytics frameworks open source and is characterized as distributed, in-memory, scalable, and fast machine learning. H2O allows users to develop predictive and machine learning models on big data. In 2014 that the mathematical core of H2O was built and developed by his Amo Candel as apart of Fortune's 2014 "Big Data All Stars".H2o can be run from Scala, Python, R, and Java. In addition, users can make use H2O to be integrated with Spark streaming and Spark paradigm through the Sparkling Water project. This method is based on breaking the main job into many small tasks that are executed in parallel. The use of this method has resulted in a significant improvement in the distribution of increasing workloads in big data

analysis tasks. The next diagram shows the basic components of H2O. As shown in Figure 2.5 it is divided into two bottom and top parts separated by the part of the cloud network. The bottom part contains the parts that work with the core to perform different tasks. The top part provides a lot of APIs for users to write their applications easily.



IV. CONCLUSION AND FUTURE SCOPE

In this paper we have analyzed various big data analytics tools in context with scalability and realized that when number of available options increases, the task of selecting big data analysis tools becomes difficult as the available tools have their own advantages and disadvantages. Because the world is developing rapidly, the ordinary tools of data mining have become ineffective, which led to the tendency to use data processing tools that can be scalable and distributed. In this paper we have reviewed the Hadoop structure and a look at the projects accompanying it. Spark was also highlighted as one of the most important tools in-memory computing and its various components were reviewed. In the end, a brief overview of H2O was presented as one of the most important tools in the process of analyzing the big data.

REFERENCES

[1] Shao, H., L. Rao, Z. Wang, X. Liu, Z. Wang and K. Ren., "*Optimal Load Balancing and Energy Cost Management for Internet Data Centers in Deregulated Electricity Markets*", IEEE Trans. Parall. Distr. Syst., Vol. 25, No. 10, pp. 2659–2669 , 2014.

[2] SWDS Li, J., Bao, Z. and Z. Li, *"Modeling Demand Response Capability by Internet Data Centers Processing Batch Computing Jobs"*, IEEE Trans. on Smart Grid, Vol. 6, No. 2, pp. 737–747, 2015.

[3] Liu, X., N. Iftikhar and X. Xie, *"Survey of Real-Time Processing Systems for Big Data"*, 18th Int. Database Engineering and Applications Symposium, New York, pp. 356–361, USA, 2014.

[4] Singh, K. and R. Kaur, *"Hadoop: Addressing Challenges of Big Data*", 2014 IEEE Int. Advance Computing Conf., Navi Mumbai, pp. 686-689, India, 2014.

[5] Liu, X., N. Iftikhar and X. Xie, *"Survey of Real-Time Processing Systems for Big Data"*, 18th Int. Database Engineering and Applications Symposium, New York, pp. 356–361, USA, 2014

[6] Shao, H., L. Rao, Z. Wang, X. Liu, Z. Wang and K. Ren., "*Optimal Load Balancing and Energy Cost Management for Internet Data Centers in Deregulated Electricity Markets*", IEEE Trans. Parall. Distr. Syst., Vol. 25, No. 10, pp. 2659–2669 , 2014.

[7] Singh, K. and R. Kaur, *"Hadoop: Addressing Challenges of Big Data*", 2014 IEEE Int. Advance Computing Conf., Navi Mumbai, pp. 686-689, India, 2014.

[8] Sun, D., G. Fu, X. Liu and H. Zhang, "*Optimizing Data Stream Graph for Big Data Stream Computing in Cloud Datacenter Environments"*, Int. J. of Advancements in Computing Technology, Vol. 6, No. 5, pp. 53–65, 2014.

[9] K. Parimala, G. Rajkumar, A. Ruba, S. Vijayalakshmi, *"Challenges and Opportunities with Big Data*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.5, pp.16- 20, 2017

[10] Sun, D., G. Zhang, S. Yang, Zheng W., S. U.Khan and K. Li, *"Re-stream: Realtime and Energy-efficient Resource Scheduling in Big Data Stream Computing Environments*", Information Sciences, No. 319, pp. 92-112, 2015.

[11] Mantripatjit Kaur, Anjum Mohd Aslam, "*Big Data Analytics on IOT: Challenges, Open Research Issues and Tools"*, International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.3, pp.81-85, 2018

[12] V.K. Gujare, P. Malviya, "*Big Data Clustering Using Data Mining Technique"*, International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.9-13, 2017.

[13] Shilpa Manjit Kaur, "*BIG Data and Methodology- A review"* ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

Authors Profile

Dr. Ajay Kumar Bharti, MCA, Ph.D., *is working as Professor, Computer Science & Engineering at Maharishi University, Lucknow(U.P.) India. His research interest is e-Governance, Service Oriented Architecture and Knowledge Based System. He has published number of research papers in reputed national/international journals and conferences.*

Ms. Neha Verma*, is pursuing M.Tech (CSE) in Department of Computer Science & Engineering at Maharishi University of Information Technology, Lucknow (U.P.) India. Her research interest is Big Data and Data Mining.*

Dr. Deepak Kumar Verma, MCA, Ph.D., *is working as Assistant Professor, Department of Computer Science, JNPG College, University of Lucknow, Lucknow(U.P.) India. His research interests are IOT, Artificial intellegence and Software Engineering. He has published number of research papers in reputed national/international journals and conferences.*