IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Comparative Study of Big Data Technologies and Frameworks

Mayank Tripathi[1*], A. K. Agarwal[2]

[1]Research Scholar, Computer Science &Engineering, Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India
[2]Assistant Professor, Computer Science &Engineering, Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India

[*]*Corresponding Author: tripathimayank100@gmail.com, Mob. No.: +918400980984*

Available online at: www.ijcseonline.org

Accepted: 12/Aug/2018, Published: 31/Aug/2018

Abstract- The organization's hunger for data insights and the adaptation of the World Wide Web has increased exponentially the generation and collection speed of data. There is a challenge to capture, store and analyze this large set of unstructured data, which have taken the shape of Big Data. In this paper, the definition of Big Data is introduced from different aspects to comprehend its concept. The architecture of Big Data is analyzed to study the processing mechanism of Big Data. The various Big Data technologies like Hadoop, HBase, Map Reduce, Pig, Hive, Sqoop, and Flume are studied and compare based on features supported by them. A comprehensive study of frameworks like Apache Spark, Cloudera, and Hortonworks used for execution of Big Data technologies is done by highlighting their important features. This paper also represents how data related to fields like the Stock market, Agriculture, Medical Health Records, and Internet traffic is stored, processed and analyzed using Big Data technologies and frameworks.

Keywords- Big Data; Hadoop; MapReduce; HBase; Sqoop; Flume; Apache Spark; Cloudera; Hortonworks**.**

I. INTRODUCTION

The information propelled by the advent of online social media, Internet and global-scale has led to a formidable challenge to deal with these large-scale datasets. Big data is an innovative form of processing data by extracting key value from data in a cost-effective manner and enabling enhanced data exploration, decision making and process automation. Big Data not only contains traditional relational data but also all patterns of unstructured data sources that are growing at a faster rate. For instance, machine derived data that is growing rapidly contains rich and meaningful content should be extracted to exploit useful information. Another instance, data derived during human communication over social media contains more textual contents than vocal, but the valuable insights are often attached with many possible meanings. To remove significant data from enormous measures of information has turned out to be progressively imperative for associations all-inclusive. Separating important bits of knowledge from such information sources rapidly and effectively is testing errand. Thus, analytics has become inextricably vital to understand the complete value of Big Data and improve the performance of analytics. There are various tools currently available to handle the characteristics of Big Data i.e. volume, velocity, variety, veracity, and value which have grown significantly in recent years. In general, these technologies are affordable and not as expensive as most of the software is available as open source. Hadoop framework is used to integrate all commodity hardware into open source software. Hadoop takes the incoming flow of data and distributes it onto disks. Hadoop is also used as a tool for analyzing the data. Whereas, MapReduce framework is used for parallel and distributed processing of Big Data using Map and Reduce functions. HBase is a column-orientated non-relational database management system built on top of the Hadoop Distributed File System (HDFS). Hive is a data warehouse system expedites querying and managing large datasets residing inside Hadoop distributed storage, which traditional relational database management systems are not able to handle efficiently. PigLatin is a high-level language used for analyzing and evaluating large data sets. Sqoop is used to transfer a large amount of data between Hadoop and relational databases. Flume is a reliable, distributed and easily available service for effectively aggregating, and moving a large amount of weblog data. The platforms like Apache Spark, Cloudera and Horton are currently used for execution of Big Data technologies. Cloudera is the oldest and widely used Hadoop distribution platform. Horton platform is a fully open source product for Big Data Analytics. Apache Spark engine supports parallel processing, distributed processing and real-time data analysis.

Section I contains Introduction, Section II contains Literature review to study Big Data using various Big Data technologies, Section III contains the Overview of Big Data, Section IV contains Big Data technologies, Section V

contains Comparison of Big Data technologies, Section VI contains Big Data frameworks and finally, Section VII contains Conclusion.

II. LITERATURE REVIEW

In previous researches, Hadoop, Map Reduce, Pig and Hive have been used to analyze data related to the Stock market, Agriculture, Medical Health Records, and Internet traffic. The following are presented below:

A. Using Hadoop and Hive to analyze Stock Market data: In this research work, the "New York Stock Exchange" data was stored using the Hadoop framework. The Comma separated files (CSV) that contained stock information such as stock's nominal price, opening price, highest price and daily quotes etc. of the New York Stock Exchange were analyzed using Hive. Using Hive command, a Hive table was created. The table was created, the CSV data was loaded into the Hive Table. By using the Hive select queries, co-variance for the supplied stock dataset for that particular year was calculated.

The co-variance results are used by stock exchange market brokers to predict the possibility of stock prices moving in the upward direction or inverse direction. [1]

B. The identification of crop disease and recommend a solution using Big Data Analytics framework: Due to technological advancements data related to agriculture has gone into the era of Big Data. In this research paper, Big Data analytics framework for agriculture data was developed to identify disease identified based on symptoms resemblance and recommend a solution based on higher similarity with the symptom. The objective was achieved using Hadoop and Hive as tools. Data were obtained from laboratory reports, websites etc. was cleansed by extracting important information from unstructured redundant data. Next, normalization was done; to extract features from cleansed data. Finally, Normalized data was uploaded onto the Hadoop Distributed File System (HDFS) and save in a file supported by Hive. HiveQL is a SQL like a query language and used to analyze and derive results from the data. It helps by identifying disease based on crop disease symptoms similarity and recommends a solution based on historical data of treatment. Results were pictorially represented using graphs treatment based on high symptoms similarity. [2]

C. Electronic Health Record's predictive analysis using Hadoop and Hive: In this research paper, Electronic Health Record data management was developed to present the insights and predict outcomes from the past patient record. In the paper, the author presented an EHR data management system to study and process enormous amounts of healthcare data. The system built using Hadoop and Hive is dynamic and scalable compared to traditional data warehouses. Patient data was uploaded onto HDFS from several sources like flat files, web pages, real-time applications, and databases. The data produced during analysis used to draw graphs and chart, which helped in the easy analysis of data. The graphical charts were helpful for doctors and researchers to study and suggest medications based on evidence from a huge number of past patient records. The predictive analysis was helpful to treat patients using particular medications, based on a number of factors such as standard of living, family history, smoking practice, and health conditions such as blood pressure and diabetes. [3]

D. Internet Traffic Analysis using Hadoop and Hive: In this paper, Hadoop and Hive based traffic analysis system performed analysis of large-sized Internet traffic data using Internet Protocol (IP) and Transport Control Protocol (TCP) in an easy and scalable manner. Hobbits was the first Internet traffic analysis system that integrated Hadoop and Hive to (i) provide a user friendly and easy to use query interface through Hive queries, (ii) enabled more efficient analysis by using the power of MapReduce together with other advantageous formats, and (iii) avoided the boundaries problem caused while partitioning variable length packets.

In Hobbits, traffic traces were uploaded into HDFS and the original large file traces were split into smaller HDFS block sized files with help of packet analyzer (i.e., tcpdump), which ensured that there was no single record split across two files, thus avoided boundary issue generated by varying length records. The small files were uploaded onto HDFS, each of which was stored in one HDFS block to be processed by a map task. Next, packet fields contained in data were extracted using the Hadoop p-cap library and then stored in one or multiple Hive tables. Different file formats including Text File Format, Sequence File Format, and ORC File Format were used within Hobbits. Once the data was loaded onto Hive tables, users of Hobbits have the facility to run their analysis queries by writing SQL like queries. The users were provided with an easy-to-use query interface and freed them from writing complicated and application-specific analysis programs in MapReduce. [4]

III. OVERVIEW OF BIG DATA

A. BIG DATA

Big Data is very diverse in nature. Fundamentally, Big Data characterize as not only a huge volume of data, it basically consists of a set of unstructured, semi-structured and structured which cannot be stored in simple table formats. There are many definitions available for Big Data here in this paper we summarize definitions from attribute, comparative and architectural point of view, which play an important role in modelling how one can view Big Data. We defined Big Data from three aspects as follows:

Attributive Definition: Big Data was defined by several IT companies like IBM, EMC and many more based on Big Data characteristics volume, variety, velocity, and value. EMC supported IDC definition of Big Data in a report out in 2011[5] that "Big Data technologies describe a new generation of technologies and architectures, designed to cost-effectively pull out value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and analysis."

Comparative Definition: In 2011, Mckinsey [6] defined Big Data as "Datasets, whose size is beyond the ability of typical database software tools to store, capture, manage, and analyze."

Architectural Definition: The National Institute of Standards and Technology (NIST) [7] suggests that "Big Data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of major horizontal scaling for efficient processing."

B. Characteristics of Big Data

Big Data has five characteristics: Volume, Velocity, Variety, Veracity, and Value, as shown in figure 1, are defined below: Volume (a large amount of data): Volume means datasets that are huge. This data can be generated every second Ex. Images, Video, Audio, emails and sensor data share every time. We are talking about zettabytes, but yottabytes or brontobytes of data. Defined datasets those are too large and easily amassed into zettabytes, even petabytes of information. A large volume of datasets is not only an analysis issue but also a storage issue.

Velocity (fast processing velocity): It means fast dataset has been produced and data move around. For example, post comment, image, video, audio file on Facebook; watching and uploading videos on YouTube; Big Data technology now allows us to analyze the data without store data ever putting it into databases.

Variety (different type of data and source): This refers to the different types of datasets that contain structured, unstructured and semi-structured data, such as emails, audio files, documents, video, images, log files, click streams, call records or financial transactions. Many different attributes in multiple dimensions in the datasets provide more and more information for traditional database management tools or application to handle.

Veracity (Correct - meaning useful data): This basically refers to the messiness or trustworthiness of the data. With many forms of data such as Facebook posts containing an asterisk, hashtag, underscore, tiled, smiles, strikers, abbreviation, typos and colloquial speech contain excellence and exactness are less handy. Big Data analytics tools and technology now allow us to work with these types dataset. The huge volumes often make up for the lack of excellence and accurateness.

Value (low - density data value): Big Data tends to have a relatively low-value density, as compared to the data we manage in the traditional system. For Example, the logistics industries have the best mode to transport for goods based on weight and value or a ratio of business relevance to the size of the data.

Figure1: Big Data Characteristics

C. Big Data Architecture

Big Data system architecture provides many functions to deal with different phases of today's data life. The architecture of Big Data system is decomposed into four sequential modules as shown in figure 2. It includes Data Generation, Data Acquisition, Data Storage, and Data Analytics.

C.1 Data Generation

Data generation is the first main phase of Big Data. Data sources such as sensors, social sites, health care centres, satellite, aeroplane, media, business apps, machine log data, generate large, diverse, and complex datasets. Data generation phase shows that the data source contains attribute values, which are mainly from the scientific field, business field, and the networking field. The scientific field produces very low whereas business field produces very high attribute value and the networking field produces a very high data rate.

C.2 Data Acquisition

Data acquisition phase is divided into data collection where data is obtained from various data sources, Data transmission phase and the data pre-processing phase from which useful information is obtained.

C.3 Data Storage

The data storage is always required to keep the data needed for future use hence a data subsystem in a Big Data platform organizes the collected information in a format which can be used for the exploration and value abstraction purpose. The data storage consists of the two parts mainly: Hardware arrangement and for managing data: data management system is required.

C.4 DATA ANALYSIS

Analytical methods or tools are required to inspect, transform, and model data to extract meaningful value. It has certain purposes like to understand the meaningful information from the data and what value-added functions can be added to the given data. Blackett et al. [8] classified data analytics into three levels:

• Descriptive Analytics: Descriptive Analytics on the basis of historical data it is concluded that what has happened.

• Predictive Analytics: Predictive Analytics on the basis of current data and the testing data it predicts the future trends.

• Prescriptive Analytics: Prescriptive Analytics focuses on the how to make an effective decision based on the present scenario.

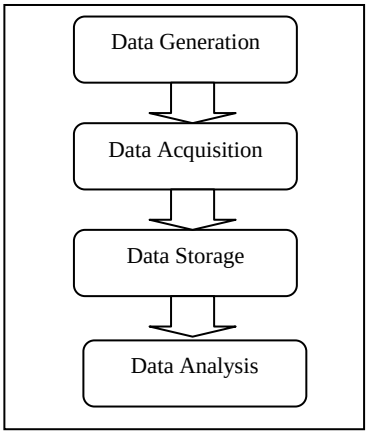

Figure2: Big Data Architecture

IV. BIG DATA ANALYSIS TECHNOLOGIES

The traditional approaches are not an appropriate solution for analysis of Big Data, as many research communities have suggested various solutions for managing various Big Data challenges. Amongst various solutions, Hadoop, MapReduce, Hive, HBase, Sqoop, and Flume are leading ones. Hadoop is an open source framework and provides two important facilities for Big Data i.e. storage and processing which is becoming a mainstay in handling Big Data challenges. MapReduce is a programming framework, to process the very large amount of data in parallel. It provides scalability, fault tolerance, reliability and many more. Hive turns Hadoop into a data warehouse, using Hive Query Language (HQL) data can be filtered out and analyzed. Following are some solutions for Big Data Analysis challenges:

A. HADOOP

The Hadoop framework provides distributed processing of large data sets across Hadoop clusters known as Nodes. It was designed to move data from single servers to thousands of machines with each providing local computation and storage. The basic aim was to allow a single query to find and collect results from all the cluster members and this model was evidently appropriate for Google's replica of search support. In software system to provide a mechanism for storage space, treatment, and information recovery from a large amount of data is the largest technological challenge. Internet and social media today produce together a large amount of data reaching the size of petabytes daily, for example, Facebook, Twitter, Whatsapp etc. These data sometimes contain valuable information, which is not properly extracted by existing systems. Most of this data is stored in an unstructured format using different languages, which is not compatible with existing systems. Parallel and distributed computing, figure 3, has a fundamental role in data processing and information extraction of large datasets. Hadoop framework was developed to take advantage of parallelization and distributed computing using commodity clusters for storing, processing and updating of a large amount of data. The framework was designed over the MapReduce and used HDFS as a file storage system. Hadoop has key characteristics while performing parallel and distributed computing such as data availability, data integrity, scalability, failure recovery and exception handling.

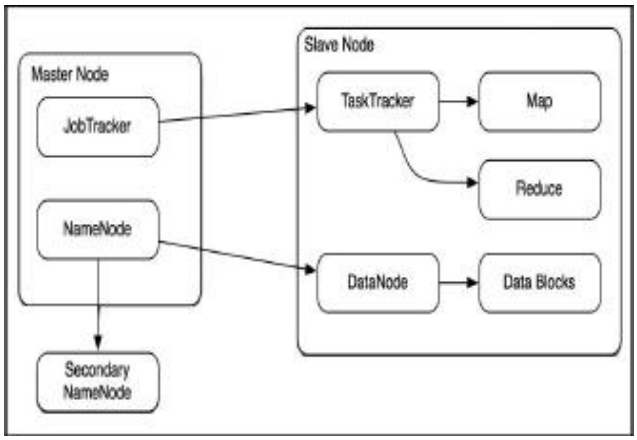

Figure 3: Hadoop and MapReduce Architecture

B. MAPREDUCE

MapReduce was the model of distributed data processing introduced by Google in 2004. Algorithms are used for the processing. MapReduce is formed of Map () and Reduce () procedures, as shown in figure 3. Map () process the data first, it generates the key-value pair and that output then will be sent to Reduce (). Reduce () works, process it and give the final output. All this processing is done in parallel fashion. Firstly Map function is applied to data then Reduce function can be run to combine the results of the Map phases. MapReduce is used for large-scale batch processing and high-speed data retrieval mostly common in web search. MapReduce is the fastest, most cost-effective and most scalable mechanism of returning results processed using distributed computing. Currently, most of the leading

technologies for managing 'Big Data' are developed on MapReduce.

C. HIVE

The hive was developed at Facebook in the year 2006 to handle a large amount of data which had increased from a few gigabytes to terabytes. Hive is a data warehouse system built inside the Hadoop file system. It is used to study and analyze large datasets which cannot be handled by traditional RDBMS. It provides a user-friendly platform where they can easily use queries similar to SQL but is named differently called HiveQL. HiveQL also helps in managing structured data. It hides the various complexity of Hadoop like now there is no need to learn Map-Reduces which is very important in Hadoop. Apart from this, no need to learn Java and Hadoop APIs. All in all, it is very useful but with just one constraint that it can be just used for structured data, it cannot handle unstructured and semi-structured data. Hive can be used for log processing. In this logs get partitioned and bucketed in the forms of tables and then can be easily analyzed. Indexing of huge documents can be easily using Hive. Hive is stored inside Hadoop in the form of hive tables from which data is accessed. It is stored inside the Hadoop file system because of its properties like scalability on various type of commodity hardware. The workflow of Hive is shown in figure 4.

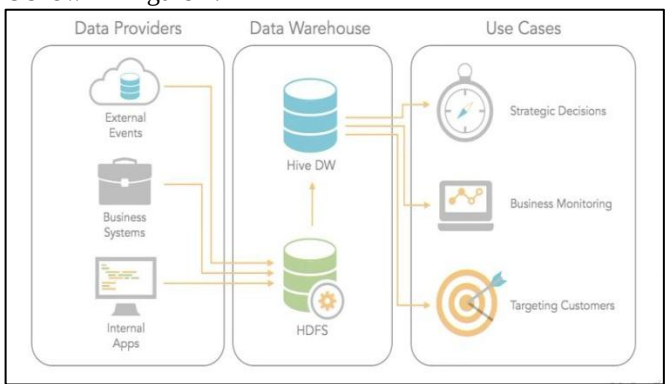

Figure 4: Hive Workflow

D. HBASE

HBase was started at Facebook in the year 2010. It is an acronym for Hadoop database which is built on the top of HDFS. HBase basically is a column-oriented based on the relational database system. Unlike Hive, HBase does not provide SQL interface. The architecture of HBase is shown in figure 5. It is used to handle a huge amount of data, the data having millions of rows and column in an efficient way to get more throughputs. HBase provides real-time read and writes operations. There is no need to define a particular schema, schema-free. It provides fault tolerance and is highly available. HBase is also very fast.

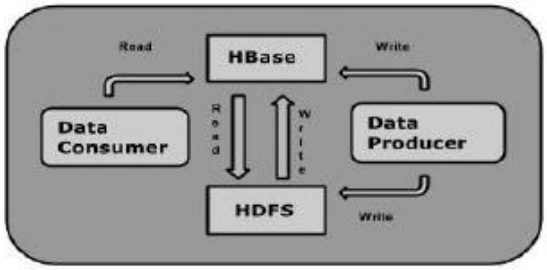

Figure 5: HBase Architecture

E. PIG

The Pig was initially developed at Yahoo in 2006 then moved into the Apache Software Foundation in 2007. Pig comes from the language Pig Latin. Pig Latin is a procedural programming language and adopts intrinsically into the pipeline paradigm. Pig is strongly recommended when queries become complex using Joins and Filters. It helps in the processing of large data sets present in the Hadoop cluster. Pig is a substitute to Java for creating MapReduce programs which help the developers to spend less time in writing Mapper and Reducer programs and focuses more on analyzing their datasets.

F. SQOOP

Sqoop is a command line interface tool made to move data between Hadoop and relational databases. Sqoop is used for importing data from RDBMS such as MySQL or Oracle Database into HDFS and then exporting the data back into RDBMS after data has been transformed using MapReduce. Sqoop also has the ability to directly import data into HBase and Hive, shown in figure 6. Sqoop connects to RDBMS through its JDBC connector and relies on the RDBMS to describe the database schema for data to be imported. Both importing and exporting is utilize MapReduce, which provides parallel operation and fault tolerance. Sqoop, during importing reads the table, row by row, into HDFS. Because importing is done parallelly, therefore, the output using HDFS is in multiple files.

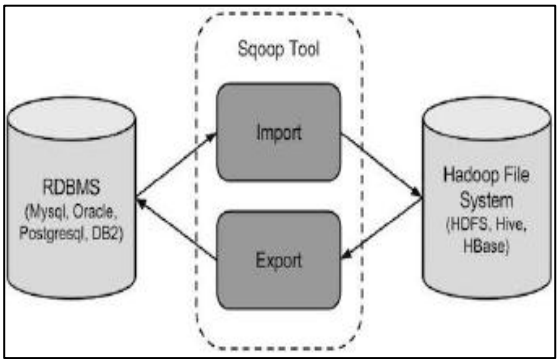

Figure 6: Sqoop Architecture

G. FLUME

Flume is an application built over Hadoop which is used for moving of huge amounts of streaming datasets into the Hadoop Distributed File System (HDFS) [9]. Sources of stream data are machine data, sensors, logs and social media. The component of Apache Flume includes Source as an entity through which data enter into Flume. The sink is used for delivering the data to the destination. Channel is the medium between source and sink. The agent is the physical Java virtual machine on which Flume runs, shown in figure 7.

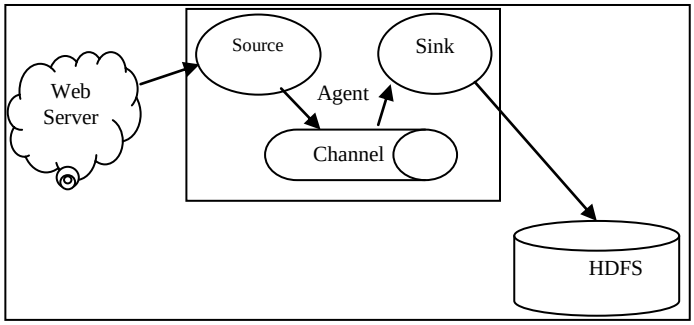

Figure 7: Flume Architecture

V. COMPARATIVE STUDY OF BIG DATA TECHNOLOGIES

Comparative study of Big Data technologies helps to classify which technology should be used when and where. Table 1, helps to distinguish the features of technologies currently used to study Big Data.

VI. BIG DATA FRAMEWORKS

A. APACHE SPARK [10]

Apache Spark is an open source platform, supports distributed batch processing, real-time, and advanced analytics using flexible in-memory data processing on the Apache Hadoop platform. Apache Spark has the feature to develop applications in Java, Scala, Python, as well as languages like R.

Apache systems component, Mahout is currently used by Apache Spark as a processing engine instead of MapReduce. With help of Apache Spark SQL, it is possible to use Hive context to have the Spark applications process data directly to and from Apache Hive. Apache Spark system contains four main sub-modules: MLlib, SQL, GraphX, and Streaming since these modules are interoperable; therefore data can be passed among them.

The streamed data can be passed to SQL, and a temporary table can be created. Apache Spark Streaming is used for performing data movement using Apache Kafka and Flume. The MLlib or the machine learning module helps in the implementation of machine learning paradigm. The Spark SQL can use as Hive Context. Thus, Spark application is used developed to generate Hive oriented objects, and run Hive QL against Hive tables and stored them in HDFS. The Spark GraphX module is used to process Big Data scale graphs, and these Big Data scale graphs can be stored using the Titan graph database. Titan also allows Big Data scale graphs to be stored and queried as graphs. It shows how Titan can use both, HBase and Cassandra as a storage medium. Using HBase, it shows how internally Titan used HDFS as a cheap and reliable medium of distributed storage. Spark is an in-memory processing system. It can also be used along with the Hadoop toolset, and the associated ecosystem.

B. CLOUDERA BIG DATA SOLUTION [11]

Cloudera was formed in 2008 to help enterprise companies use Hadoop to the get the more valuable output of all of their data. Cloudera is an open source platform. Cloudera is the most popular distributed Big Data technologies. The biggest contribution of Cloudera is that it provides abstraction over different Big Data technologies such as Hadoop, Hive, HCatalog etc. It also provides ready to use technologies without going into technical details and also provides the system management, deployment, configuration, security management, diagnostics, reports creation etc.

C. HORTONWORKS [12]

This technology was funded and developed in 2011 by the engineers worked in Yahoo. HortonWorks is a similar concept to Cloudera which works to provide the same functionality as different Big Data technologies for enterprise computing. HortonWorks data platform was built for enterprise computing which builds on Hadoop and also provides a different type of analysis like real-time, batch and interactive. It also supports different type of technologies for data integration and data flow control. Thus technologies such as Apache Falcon, Sqoop, and Flume, which are an integral part of system's platform, provides an easy and systematic way of accessing data in and beyond Hadoop. When HortonWorks data platform is deployed and implemented along with other technologies it requires substantial expertise and training to use it.

VII. CONCLUSION

In this paper, we discussed how data associated with various fields like the Stock market, Agriculture, Medical and Traffic of Network is stored in a structured manner, processed to refine and analyzed to predict the outcome using Big Data technologies. Next, we discussed the definitions of Big Data based on the Attributive, Comparative, and Architectural behaviour of Big Data. The architecture of Big Data presented different stages: Data generation, Data acquisition, Data storage and Data Analysis. Then, we discussed Big Data technologies like PigLatin, Hive, HBase, Sqoop, and Flume that are used to enhance the performance of basic

Hadoop and MapReduce framework. PigLatin is a procedural scripting language used to decrease the execution time of MapReduce program by using nested data type feature and as a result number of code lines also decreased. Hive is similar to SQL; hence it assists developer by providing a familiar environment like SQL language for MapReduce programming. HDFS is used as a storage component, which can perform read and write operation to Big Data with the help of HBase. Data exchange between Hadoop and RDBMS can be performed with the help of Sqoop. Flume is used for transferring stream of the weblog to HDFS. Next, we compared these Big Data technologies based on their availability, usage, language supported, when to use, the data structure on which it operates and which are companies using them. Finally, various Big Data framework like Apache Spark, Cloudera and Horton are available to implement Big Data technologies are discussed. Spark is currently the best platform to implement real-time data analytics for Big Data.

REFERENCES

[1] 3pillarglobal.com, How to Analyze Big Data with Hadoop Technologies [Online], Available: http://www.3pillarglobal.com/ and http://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies (accessed on 11 April 2018)

[2] Er. Rupinder Kaur, Raghu Garg, Dr Himanshu Aggarwal, Big Data Analytics Framework to Identify Crop Disease and Recommendation a Solution, IEEE, International Conference on Inventive Computation Technologies (ICICT), volume 2, 2016.

[3] Haritha Chennamsetty, Suresh Chalasani, Derek Riley, Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive, IEEE, International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2015.

[4] Abdeltawab M. Hendawi, Fatemah Alali, Xiaoyu Wang, Yunfei Guan, Tianshu Zhou, Xiao Liu, Nada Basit, John A. Stankovic, Hobbits: Hadoop and Hive Based Internet Traffic Analysis, IEEE, International Conference on Big Data (Big Data), 2016.

[5] J. Gantz and D. Reinsel, Extracting value from chaos, in Proc. IDC iView, pp. 1–12, 2011.

[6] J. Manyika et al, Big Data: The Next Frontier for Innovation Competition, and Productivity, San Francisco, CA, USA: McKinsey Global Institute, pp. 1–37, 2011.

[7] M. Cooper and P. Mell (2012), Tackling Big Data [Online], Available: http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf (accessed on 13 May 2018)

[8] G. Blackett (2013), Analytics Network-O.R. Analytics [Online], Available: http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx (accessed on 13 May 2018)

[9] Palanisamy, B. Singh, & Liu, "cost-effective resource provisioning for MapReduce in a cloud," IEEE Transactions on Parallel and Distributed Systems, pp: 1265-1279, 2015.

[10] Mike Frampton, Mastering Apache Spark (ed.) 2015, Packet publication ltd., U.K.

[11] Cloudera, Cloudera Platform 2018, [Online] http://cloudera.com/ (accessed on 15 January 2018)

[12] Hortonworks, Discussion about Horton Platform working,[Online] http://hortonworks.com/hdp/ (accessed on 15 June 2018)

Authors Profile

Mr Mayank Tripathi is currently pursuing M. Tech. in Computer Science and Engineering from Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India. He received B. Tech. degree in Computer Science and Engineering from Dr A.I.T.H., Awadhpuri, Kanpur, Uttar Pradesh, India in 2016. His research interests include the study of Big Data technologies and implement it to analyze data of social relevance. The author has published a paper in Global Scientific Journals. The author has also written a paper which is under a review process in the Journal "Big Data Research, Elsevier."

Dr A. K. Agarwal is currently working as Assistant Professor in Computer Science and Engineering Department at Kamla Nehru Institute of Technology, Sultanpur, Uttar Pradesh, India. He received his PhD in 2017 from Dr APJAKTU Uttar Pradesh, India. He received his M. Tech. degree in 2006 from Samrat Ashok Technological Institute (SATI), Vidisa and B.Tech. degree in 1999 from BIET Jhansi, Uttar Pradesh, India. His research interests include the study of parallel computing, data mining, data warehouses and Big Data analytics. The author has published about 30 papers in International/National Journals.

Table 1: Comparison of Big Data Technologies

Features	Hadoop	MapReduce	HBase	Pig	Hive	Sqoop	Flume
Developed by	Yahoo	Google	Apache Software Foundation	Yahoo	Facebook	Cloudera	Cloudera
Available	Open-source	Open-source	Open-Source	Open-Source	Open-Source	Open-Source	Open-Source
Language supported	Java	Any language	Java	PigLatin	HQL(Hive Query Language)	MYSQL, Microsoft SQL Server, Postgre SQL, IBM DB2	Java
When to use	Real-time Analytics, Multiple datasets of	Word count, Inverted Index, Grep etc.	When there is a need for random read and write access to our Big Data	For data processing on Hadoop cluster	For analytical purposes	When there is a need to import and export data from RDBMS to Hadoop	For moving a large amount of data to a centralized data store
Data structure it operates on	Binary Files	File-based data structure	NoSQL	Complex, Nested	Apache Derby Database	Simple	Simple
Schema	Not supported	Schema-free	Required	Optional	Required	Optional	Required
External file support	No	No	No	Yes	Yes	Yes	Yes
Required software	JDK 1.6 or above supported	JDK 1.7 or above supported	JDK version 1.7 recommended	Java 1.6 or above supported	Hive version 1.2 or above, Java 1.7, Hadoop version 2.X	No such requirement	Java 1.7 recommended, sufficient memory and disk space
Event Driven	No	No	No	No	No	No	Yes
Companies using	Amazon web services, Microsoft etc.	Google	eBay, Yahoo, and Facebook etc.	Yahoo	Facebook, Netflix	Yahoo, Amazon	Yahoo, Google etc.
Used for	Processing of structured and unstructured data	Parallel processing and Distributed computing	To provide quick random access to huge amount of structured data	For the processing of large data set present in Hadoop clusters	Used for effective data aggregation method, ad-hoc querying, and analysis of the huge volume of data	To transfer data between Hadoop and relational databases	For moving, streaming web log data into HBase

