International Journal of Computational and Applied Mathematics. ISSN 1819-4966 Volume 12, Number 1 (2017)

A Panorama of Big Data Analytics with Hadoop

K.Madasamy
Research Scholar (Part-Time)
Department of Computer Applications
Madurai Kamaraj University, Madurai-625 021
madasamyk63@rediffmail.com

Dr.M.Ramaswami
Associate Professor
Department of Computer Applications
Madurai Kamaraj University, Madurai-625 021
mrswami123@gmail.com

Abstract

In recent times, data analytics is becoming a part and parcel of our day to day activities that provides big insights about stored data and help us to take better decisions at the right time in every walk of life. The need for data analytics has been augmented significantly in all spheres. The traditional data processing techniques are obsolete and inefficient in handling such a huge volume of data of sizes varying from Peta byte to Zeta bytes. It is become necessary to invent or develop new techniques for efficient processing of data which are originated from different sources like social media, IoT, satellite, high bandwidth networks etc.,. The data received from such sources are inherently unstructured or in semi-structured format. Recently Big data analytics has become flourishing as a new technology for handling such huge volume of data with Apache Hadoop framework. Hadoop offers many methods for effective processing of big data in parallel and distributed environment with ease. This studyreviews the various aspects of Big data analytics along with the growth of Hadoop and its related ecosystems.

Keywords: *Big Data, Hadoop, Hadoop Distributed File System(HDFS), MapReduce, Ecosystems, Business Intelligence(BI).*

I. INTRODUCTION

Today, the digital universe is flooded with images and videos on mobile phones uploaded to YouTube, digitalmovies populating the pixels of our high-definition TVs, banking data swiped in an ATM, securityfootage at airports and major events such as the Olympic Games, subatomic collisions recorded bythe Large Hadron Collider at CERN, transponders recording highway tolls, voice calls zipping throughdigital phone lines, and texting as a widespread means of modern communications [19]. In recent years, these large volumes of highly detailed data from the various strands of a business provide the opportunity to deliver significant financial and economic benefits to firms and consumers. More data can lead to more accurate analysis, which can lead to better decision making, greater operational efficiencies, cost reductions and reduced risk [18]. Data is considered as the heart of all organizations and it can be used in different ways to run the business, market to customers, forecast sales, measure performance, gain competitive advantage, and discover new business opportunities. Recently, a synergy of new technologies and market dynamics has opened a new frontier for information management and analysis. This new generation of computing includes data with too greater volume, variety and velocity, than ever before [15].This huge amount of the data is known as “Bigdata”.The term “Big data” was first introduced to the computing society by Roger Magoulas from O’Reilly media in 2005. Big

data is a buzzword or catch-phraseto describe a massive volume of both structured andunstructured data that is so huge and much complicated toprocess using traditional relational database management systems (RDBMS) and its associated techniques [13].Also, RDBMS is not thought of as a scalable solution to meet the demanding criteria for big data applications. Table 1 shows the comparision between the traditional database system and Hadoop environment [20].To cope up with big data adaption, a flexible database management system and Extract-Transform-Load (ETL) tools are indispensable.

Table 1. Comparison of Traditional Systems and Hadoop.

RDBMS	FACTORS	HADOOP
Structured	Data Types	Multi and Unstructured
Limited, No Data Processing	Processing	Processing coupled with Data
Required On Write	Schema	Required On Read
Reads are Fast	Speed	Writes are Fast
Software License	Cost	Support Only Open Source
OLTP Complex ACID Transactions Operational Data Store	Best Fit Use	Data Discovery Processing Unstructured Data Massive Storage / Processing

NoSQL (commonly known as “Not Only SQL”) databases are becoming more popular for big data storage. NoSQLdatabases features flexible modes, support for simple andeasy copy, simple API, eventual consistency, and supportof large volume data. NoSQL databases are becomingthe core technology for big data. There are three different NoSQL databases are currently in use . Table 2 depicts the various types of NoSQL databases and their associated Extract-Transform-Load (ETL) tools.

Table 2 : Different types of NoSQL databases

Sl.No	Type	Example
1	Key-Value	Dynamo, Voldemort
2	Column-oriented	BigTable,Cassandra, Hbase and HyperTable
3	Document-oriented	MongoDB,SimpleDB and CouchDB

Big data can unlock significant value by making information transparent and usable at much higher frequency.Frameworks like Hadoop and MapReduce are used to store and process large scale of data.

II. LITERATURE SURVEY

The survey described here is about the core themes of research articles published in various journals extend over the field of big data analytics. The big data analytics is a hot topic and recently vast amount of research is in progress by incorporating all aspects of this flourishing domain.Hadoop is a framework for dealing with big data. Hadoop has two main components, namely HDFS and MapReduce. In this literature survey, an attempt is made tofocus the research articles covering the various issues related with big data analytics have been presented.

Dawei jiang et.al., presented a study on the performance of MapReduce.They identified the five factors that affect the performance of MapReduce: I/O mode, indexing, Data parsing, grouping schemes and block-level Scheduling. By carefully turning the above factors, the overall performance of MapReduce can be improved by a factor of 2.5 to 3.5. Further, they suggested that MapReduce-based systems are not inferior to parallel database systems in terms of performance; instead they can offer a competitive edge as they are elastically scalable and efficient. Benchmarking results were given for quantifying the impact of the factors they considered[1].

MapReduce was proposed by Dean.J et.al.,[3] as a programming model for processing and generating large data sets. MapReduce is recongnised as a possible means to perform elastic data processing in the cloud. The most successful system is the Google's MapReduce framework, which hides the complexity of data distribution, communication and task scheduling and offers a simple programming model for writing analytical applications, while also providing strong fault-tolerance guarantees.

M.Zaharia et-al investigated the scheduling algorithm of Hadoop and proposed a LongestApproximate Time to End(LATE) scheduling algorithm which improves Hadoop response times by a factor of two[4].

Vasiliki Kalavri et.al.,[5] have extensively studied the limitations, optimizations and many open issues related with MapReduce. Several implementations of the MapReduce model have been investigated. Among the many variations of MapReduce model, they suggested Apache Hadoop framework being the most widely adopted.

Christopher Garcia [6] presented a study on MapReduce, which enables computational processes can be scaled up to very large sizes with the advent of Cloud computing. Also, the author have shed light on the types of problems for which MapReduce is well-suited, as well as those for which it is not. Further, the author suggested that, MapReduce is well-suited for processing tasks where relatively few processing steps depend implicitly or explicitly on one another.

Navya Francis et.al [7] have studied on data processing for big data applications using Hadoop framework. The authors nicely presented the main idea of Hadoop distributed file system and MapReduce framework. They suggested that Pig and Hive are the two languages which helps us to link MapReduce framework within short period of time.

In this paper the chapters are organized as follows. Section II discusses a brief literature survey. Section III gives an overall view about big data analytics. Section IV provides the details about Hadoop framework and its related applications. A brief note on Hadoop Ecosystem is given in section V. Finally, section VI outlined the conclusion of this present study.

III. BIG DATA ANALYTICS

Recently "Big Data" is a buzzword that describes voluminous of data that is structured, semi-structured andunstructured data that has the potential to be mined for precise information. Even though, big data does not refer any specificquantity of data, this term is often used when speaking about the Petabyte to Yottabyte of data.Due to its recent popularity, the definition of big data is rather assorted, and reaching a consensus is difficult. Basically, big data means not only a huge volume of data but also otherfeatures that differentiate it from the concepts of "massivedata"' and "very large data" [14].The major goal of big data analytics is to provide more informative business related decisions and thus byenabling data scientiststo analyse huge volumes oftransactional data, that maybe untapped by more conventional Business Intelligence(BI)programs. The primary sources for data flooding includes web server logs, Internetclick stream data, social media content, social networkactivity reports, text from customer emails and surveyresponses, mobile phone call records and machinedata captured by sensors which are connected to the Internet ofThings(IoT) [21].It is estimated that, in the field of IoT, more than 30 million networked sensornodes are now actively functioning in the transportation, automotive,industrial, utilities, and retail sectors. The number of thesesensors will increase at a rate of more than 30 percent peryear [23].

3.1. BIG DATA CHARACTERISTICS

The amount of the data which can be characterized to be big data is everchanging factor and newer ETL tools are continuously being developed to handle this big data. In order to make sense out of this overwhelming amount of data, it is often broken down using five V's namely Velocity, Volume, Value, Variety, and Veracity [2]. These five v`s shown in figure 3.1 describes the term "Big Data".

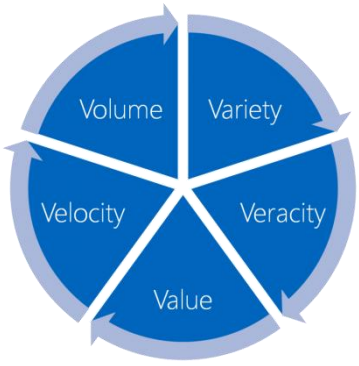

Figure 3.1. The five V's of Big Data

Volume

Big data is the term which indicates large volumes of data. Now, the data is being generated and obtained by machinaries, networks and human interactions with social media. The volume of data to be analyzed is ever increasing and accumulated massively[15]. With big data technology, we can now store and process these huge data with the help of distributed systems, where parts of the data is stored in remote locations and brought together by software.

Variety

Variety tells us the existence of many types and sources of data both which are unstructured and structured. The data is being stored with the help of various sources such as database as well as spreadsheets. Recently data is available in varied formats like photographs,mails, as well as videos, PDF files, audio tracks or clips, etc. Hence, large and variety of unstructured data creates problems for mining,storage and analysis of data [15].

Velocity

Velocity of big datarefers to the speed at which new data is generated and the speed at which data moves around.The flow of data is massive and continuous. This data which is real-time will help two important fields i.e., businesses and also researchers to make important decisions which will provide strategic comparative advantages [15]. The recent advances in big data technology allows us to analyse the data while it is being generated, without ever storing it into databases[9].

Veracity

It refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable. The data which is being stored, and searched is meaningful and relevant to the problem statement [15].

Value

Big data must have value. Since, the potential in big data is huge,make it into big business by recognizing the value in the data. It is important to ensure that the insights that are obtained are based on accurate data and lead to measurable improvements at the end.

3.2. APPLICATIONS OF BIG DATA ANALYTICS

Recently, the applications of big data analytics havea wide scope in all fields and trades. It has emerged as an advanced analytical technology for making better decision.It is very difficult to narrate all the domain specific applicationsof big data. There are many frameworks and platforms have been developed for big data analysis and each one has its own pros and cons . The Cloud computing technologies are widely used on these platforms and frameworks for the demanding computing resources and storage [17].In this section, a brief applications of big data in the core fields have only been discussed.

Enterprises

In enterprises, BI and OLAP can be regarded as the forefather of big data applications. It enhances their production productivity and competitiveness in many aspects. In sales planning, after comparison of massive data, enterprises can optimize their commodity prices. In finance sector, the application of big data in enterprises has steady swift. There has been innumerable applications in e-commerce and on-line shopping portals[16]. Today the banking sector performs its own creditscore analysis for their existing customers using a wide range of data collected from savings, credit cards, mortgages, and investment data [21]. Also, the banking sector is deploying big data analytics for churn prediction and fraud detection.

IoT

The emerging IoT is the main source of big data. Due to the huge variety of sensor objects, the applications of IoT also evolve endlessly. Logistic enterprises have profoundly experienced with the application of IoT big data. In many countries, the Smart city initiative is a hot research area based on the application of IoT data[16].

Health Care

Healthcare and medical data are continuously and rapidly growing complex data, containing abundant and diverse information values. Big data has vast potential for effectively storing, processing, querying, and analyzing medical data. The application of medical big data extremely influenced all spheres of health care business. HealthVault, an Android mobile application of Microsoft, which was launched in 2007 is an excellent application of medical big data. Its goal is to store and manage individual health information in individual and family medical devices [16].

Agriculture

The potential for big data applications in agriculture is attracting a growing interestfrom food and agribusiness industry players, researchers, and policy makers. Potential gains in agricultural productivity and supply chain efficiencyfrom big data based solutions could significantly enhance the ability of global agri-food systemsto face the challenge of doubling the food supply by 2050. Most of the research in this arearevolves around the big data applications in commercial agricultural production in developedcountries [22].Sjaak Wolfert et.al., [24] presented a review study which describes the use of big data analytics for Smart Farming. In the near future, IoT and Cloud computing will be expected to leverage this growth and the deployment of robots for farming activities

IV. HADOOP FRAMEWORK

Because of the great success of Google's distributed filesystem and the MapReduce computation model in handlingmassive data processing, its clone, Hadoop, has attractedsubstantial attention from both industry and academic scholars alike.In fact, Hadoop has long been the mainstay of the bigdata movement. Apache Hadoop is an open-source softwareframework that supports massive data storage and processing.Hadoop was created by Doug Cutting and Mike Cafarella

in2006. Doug Cutting, who was working at Yahoo! at thetime, named it after his son's toy elephant [27]. It was originallydeveloped to support distribution for the Nutch searchengine project.Instead of relying on expensive, proprietary hardware to storeand process data, Hadoop enables distributed processing oflarge amounts of data on large clusters of commodity servers. Hadoop framework has two main components, namely Hadoop Distributed File System (HDFS) and MapReduce as shown in figure 4.1.

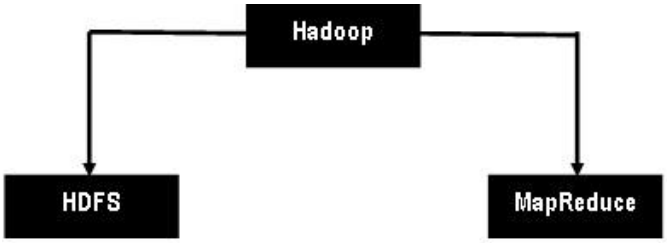

Figure 4.1 Main Components of Hadoop

Hadoop has many advantages, and the following featuresendorsed Hadoop will be the best choice for big data managementand analysis [14].

4.1. SALIENT FEATURES OF HADOOP

Scalability: Hadoop allows hardware infrastructure to bescaled up and down with no need to change the data formats.It will automatically reallocate data and computationjobs to accommodate hardware changes [14] .

Cost Efficiency: Hadoop provides massive parallel computationto cheap commodity servers, leading to diminution in cost per terabyte of storage, which makesmassively parallel computation affordable for the evergrowing volume of big data [14].

Flexibility: Hadoop is free of schema and able to absorbany type of data from any number of sources. Moreover,different types of data from multiple sources can beaggregated in Hadoop for further analysis. Thus, manychallenges of big data can be addressed and solved [14].

Fault tolerance: Missing data and computation failuresare common in big data analytics. Hadoop can automatically recoverthe data and computation failures caused by node breakdownor network congestion [14].

4.2. HADOOP DISTRIBUTED FILE SYSTEM

Hadoop distributed file system (HDFS) provides high-throughput access to large scale data sets. HDFS is built using the Java programming language. All HDFS Communication protocols are build on the TCP/IP protocol. In HDFS, the processing logic is very near to the data, rather than the data nearer to the processing logic. It is portable across heterogeneous commodity hardware and operating systems. It is highly scalable to reliably store and process large amounts of data. It is veryinexpensive by distributing data and processing across clusters of commodity personal computers. Hadoop distributed file system (HDFS) has a Master-Slave architecture. A HDFS cluster has a single Name node and a single or multiple Data node(s). The Name node acts as the Master node. It manages the file system namespace, providing supervisory access to files by a

client. It executes all operations on file or directory given by the client, such as reading, writing, renaming, updating, and so forth. Hadoop maps files into a set of blocks that are being stored in the Data node or Slave nodes. Data node not only stores data, but also is responsible for creating or deleting or updating a block. As HDFS is being entirely built using the Java language, it can be freely deployed on a wide range of commodity machines.Figure 4.2 illustrate the simple architecture of Hadoop distributed file system[11].

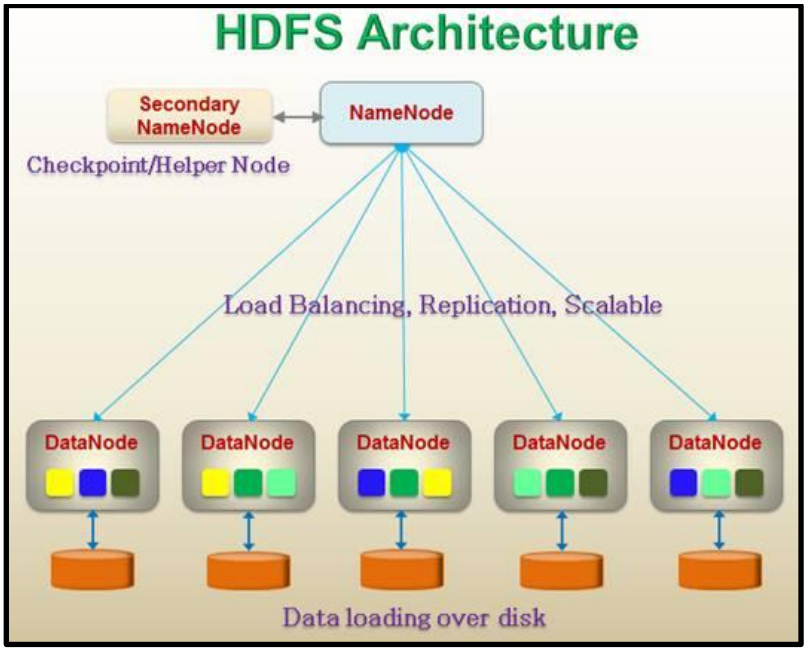

Figure 4.2 Architecture of Hadoop Distributed File System

4.3.HADOOP'S MAPREDUCE PARADIGM

Hadoop supports a MapReduce model of computation which was introduced by Google as a method ofsolving Petascale problems with large clusters of inexpensive machines. MapReduce uses two distinct phases for solving any big data related problems.

• ***Map:***It is the first phase of the MapReduce process. It is an initial ingestion and transformation step, in which the individual input records can beprocessed in parallel.

• ***Reduce:***It is the second phase of the MapReduce process. It is an aggregation or summarization step, in which all the associated records must beprocessed together by a single entity. The block diagram of MapReduce framework is shown in figure 4.3.

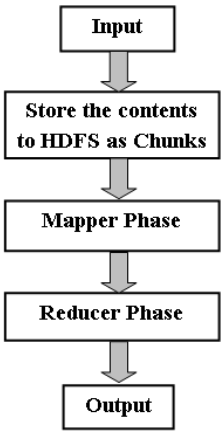

Figure 4.3 Block Diagram of MapReduce framework

The core concept of MapReduce in Hadoop is that, the given input will be split into logical chunks, andeach chunk may be initially processed independently by a map task. The results of theseindividual processing chunks can be physically partitioned into distinct sets, which are thensorted. Each sorted chunk is then passed to a reduce task. Figure 4.4 shows the detailed working method of MapReduceparadigm .

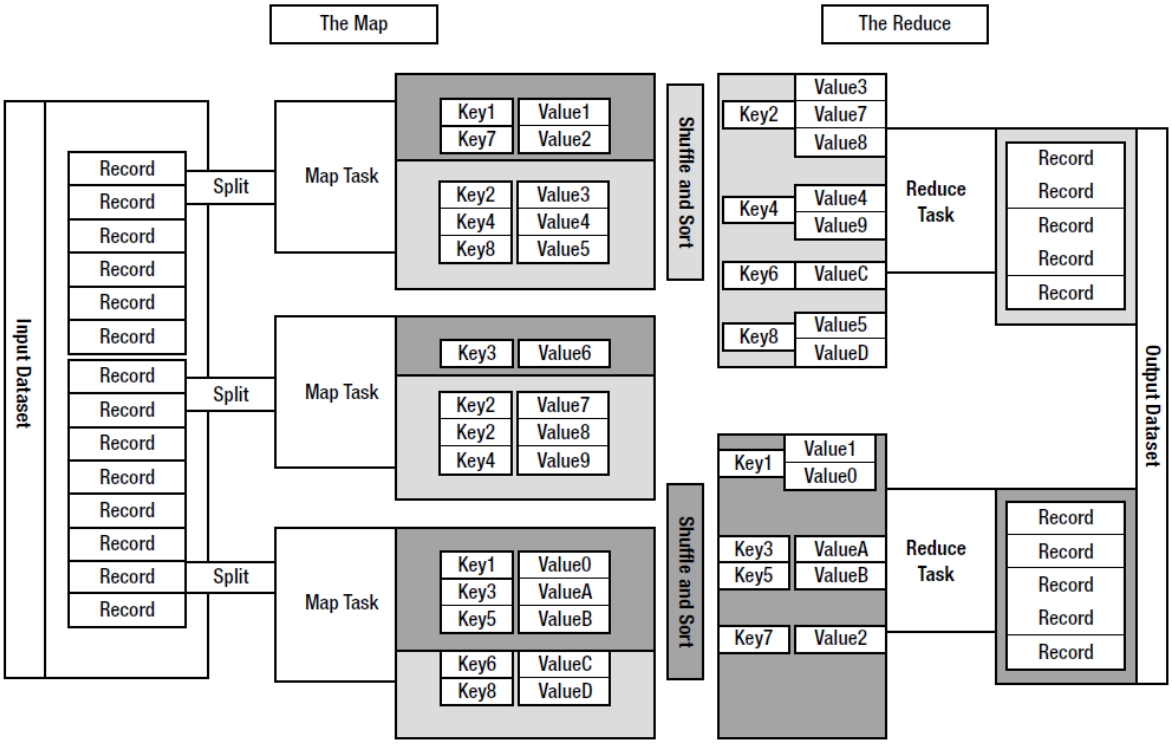

Figure 4.4. MapReduce Working Model

A map task may run on any compute node in the cluster, and multiple map tasks may be running in parallel across the cluster. The map task is responsible for transferring the input records into

key-value pairs. The output of all of the maps will be partitioned, and each partition will be sorted. There will be one partition for each reduce task. Each partition's sorted keys and the values associated with the keys are then processed by the reduce task. There may be multiple reduce tasks running in parallel on the cluster. MapReduce is oriented around key-value pairs. The framework will convert each record of input into a key-value pair, and each pair will be input to the map function once. The map output is also a set of key-value pairs. The map output pairs are grouped and sorted by key. The reduce function is called one time for each key, in sort sequence, with the key and the set of values that share that key. The reduce method may output an arbitrary number of key-value pairs, which are written to the output files in the job output directory. If the reduce output keys are unchanged from the reduce input keys, the final output will be sorted.

V. HADOOP ECOSYSTEMS

Hadoop ecosystem is a collection of ETL tools and are built with different components on top of Hadoop framework to store and to ease the data processing needs of different users. These Hadoop ecosystem components can handle and analyze the data coming from different data sources like database systems ,OLTP systems, OLAP Systems and Web data written in scripting languages like Python, JavaScript, and Pearl etc.,Figure 5.1 shows a wide collection of Hadoop ecosystem ETL tools and each tool is meant for specific applications. In this present study we discuss the features of very limited tools like Sqoop, Flume, Pig, Hbase, and Hive only[20].

Figure 5.1 A compendium of Hadoop Ecosystem Tools
[Source: Uweseiler-Codecentric]

SQOOP

Apache Sqoop is anopen source tool for efficiently moving data between Hadoop file system and relational databasesystems. It is justa tool for data transfer. Sqoop efficiently imports individual database tables or entire database into Hadoop file system. It can work only on Linux operating system. It is provided by the Apache Software Foundation. The import tool helps to import individual tables from RDBMS to HDFS. Sqoop works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB. Each row in a table will be treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files. The export tool will be used to export a set of files from HDFS to RDBMS[8]. Figure 5.2illustrate the workflow of Sqoop environment.

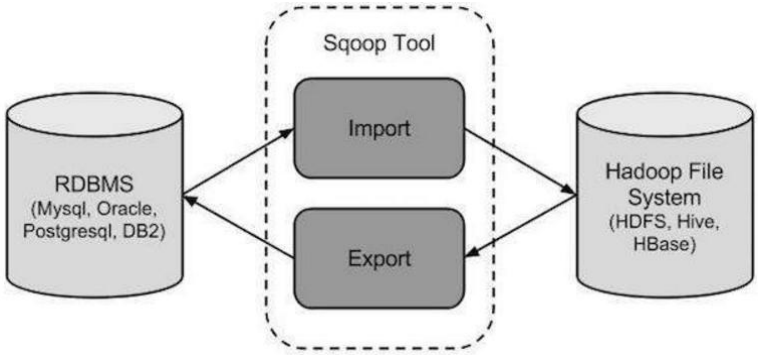

Figure 5.2 Data transfer between RDBMS and Hadoop

FLUME

Apache Flume**[20]**is a Hadoop ecosystem component which provides services for handling the data from different data streaming sources like Twitter, Facebook,Cloud and WebServers efficiently and to load this stream of data into HDFS for analysis.

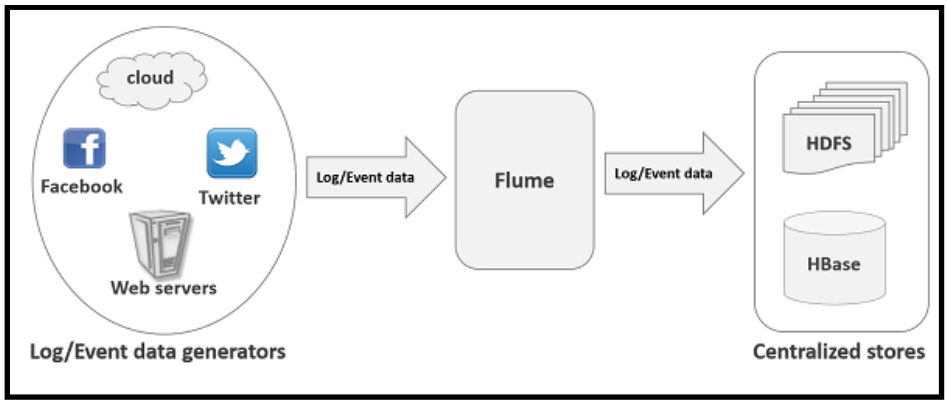

Figure 5.3. The working Scheme of Flume

Apache Flume is a data ingestion mechanism for collecting aggregate and transporting large amounts of streaming data, such as log files, events etc., from various sources to a centralized

data store. Figure 5.3 depicts the working model of Flume.Flume is a reliable, distributed, and easily configurable tool. It is primarily designed to copy streaming data (log data) from various web servers to HDFS. Flume is used to move the log data generated by application servers into HDFS at a higher speed [12].

Here are the advantages of using Flume:

- Data can be stored in the centralized stores like HBase, HDFS etc.,
- When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a moderator between data producers and the centralized stores and offers a smooth flow of data between them.
- Flume provides the feature of **contextual routing**.
- The transactions in Flume are channel-based, where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery.
- The important characteristics of Flume are reliable, fault tolerant, scalable, manageable, and customizable.

Some of the notable features of Flume are as follows:

- Flume ingests log data from multiple remote web servers into a centralized hub (HDFS, HBase) efficiently.
- Using Flume, we can get the data from multiple servers immediately into Hadoop.
- Along with the log files, it is also used to import high volumes of event data generated through social networking sites like Facebook, Twitter and e-commerce portals like Amazon and Flipkart.
- Itaccepts varieties of source and destination data formats.
- It supports multi-hop flows, fan-in and fan-out flows, contextual routing, etc.
- It can be scaled horizontally.

PIG

PIG [25] was developed by Yahoo!, and it is an open source tool. It is a high-level declarative language similar to SQL and can accept any kind of data like structured, semi structured and unstructured data. As it accepts any kind of data it is named as Pig. Pig is extended with UDF *(User Defined Functions)* feature that accepts the code written in other languages like Python, JavaScript, Ruby, PHP and Perl. Pig is mainly used to reduce the development time of programs and also simplifies the coding task. Pig has its own scripting language called "***Pig Latin"***, which can store all commands and can be executed sequentially. Pig can be accessed in 2 modes or shells: *local mode* or *HDFS mode.*

HIVE

Hive [26] is a data warehouse package built on top of Hadoop. It allows conventional business intelligence (BI) applications to run queries on a Hadoop cluster. Hive is developed by Facebook to analyze several Terabytes of data. As most of the ETL developers are used to programming in SQL background, they came up with a language called HQL (Hive Query Language) which looks similar to SQL. Hive uses a data model for efficient processing of data with the components like Hive Tables, Partitions and Buckets.

Hive table: Data can be stored using hive tables.

Partitions: Partition is dividing the hive table data into groups based on columns which makes it faster to access the data and to do querying on hive data.

Buckets: Partitions are further sub divided into Buckets. Bucketing by using User-id or column makes it easier and fast access to data.

HBASE

HBase [10] is an open source, distributed database, developed by Apache Software foundation. Initially, it was Google Big Table, afterwards it was re-named as HBase and is primarily written in Java. HBase can store massive amounts of data from Tera bytes to Peta bytes. It is built for low latency operations. Most of the online analytics applications uses HBase. HBase storage model is different from other NoSQL models. HBase stores data in the form of key-value pairs in a columnar model, also known as tabular data store . In this model, all the columns are grouped together as Column families. HBase on top of Hadoop, will increase throughput and performance of distributed cluster environment. In turn, it executes faster random reads and writes operations. It supports automatic and configurable sharding of tables. Further, it provides automatic failover supports between the Region Servers. It is built on CAP theorem (that uses 3 characteristics like Consistency, Availability and Partition tolerance).

VI. CONCLUSION

Due to data deluge, the applications and the use of big data analytics is being felt in many domains. Selecting the right choice of big data tools and finding the right solution strategy depends on the problem being undertaken. There is no common solution methodology is available for all the problems in the big data domain . Also selecting the right choice of hardware to handle the big data problem is a cumbersome task. Recently, many Cloud service providers offering elastic computing resources which attracts increasing interest from both academic researchers and industry that drastically reduce the cost of hardware for dealing with big data problems.Big data analytics helps business people to make better decisions at the right time and researchers to identify new avenues. The huge volume of big data can be efficiently processed andstored with the help of Hadoop ecosystem ETL tools. It supports manytools that are integrated with each other. Each toolprovides many sophisticated features for processing, storing andretrieving the big data in an effective manner. This paper briefly touches the fundamental concepts of Bigdata, its characteristics, Hadoop and its constituent components like HDFS, MapReduceand finallyHadoop ecosystem ETL tools.

REFERENCES

[1]. Dawei Jiang, Beng Chin Ooi, Lei Shi, Sai Wu, "The Performance of MapReduce:An In-depth Study", Proceedings of the VLDB Endowment, Vol.3, No.1,2010.
[2]. Jenn Cano, The V's of Big Data, blog post on March 11, 2014. https://www.xsnet.com/blog/bid/205405/The-V-s-of-Big-Data-Velocity-Volume-Value-Variety-and-Veracity.
[3]. J.Dean and S.Ghemawat, "Mapreduce: Simplified data processing on large clusters" In OSDI, pages 137-150,2004.
[4]. Matei Zaharia, Andy Konwinski, Anthony D. Joseph, Randy Katz, Ion Stoica, "Improving MapReduce Performance in Heterogeneous Environments", In OSDI, pages 29-42, 2008.
[5]. Vasiliki Kalavri and Vladimir Vlassov, "Mapreduce : Limitations, Optimizations and Open Issues", In: Proceedings - 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2013, IEEE , 2013, pages.1031-1038.
[6]. Christopher Garcia, "Demystifying MapReduce", Procedia Computer Science, 20(2013),484-489.
[7]. Navya Francis and Sheena Kurian.K, "Data Processing for Big Data Applications Using Hadoop Framework", International Journal of Advanced Research in Computer and Communication ngineering, Vol.4, issue 3, March 2015.
[8]. https://www.tutorialspoint.com/sqoop/sqoop_introduction.html.
[9]. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know
[10]. http://www.guru99.com/hbase-tutorials.html
[11]. http://www.developer.com/db/getting-familiarized-with-the-hadoop-distribution-file-system.html
[12]. https://www.tutorialspoint.com/apache_flume/apache_flume_introduction.htm
[13]. Vibhavari Chavan and Rajesh N.Phursule, " Survey Paper on Big Data", International journal of computer science and Information technology, Vol.5(6), 2014.
[14]. H. Hu, Y. Wen, T.-S. Chua, and X. Li, ``Toward scalable systems for big data analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652687,Jul. 2014.
[15]. Ankita Karale, Bharathi Patil, "A Survey on Big Data", IJCSIT, Vol 2, Issue 4, August 2015.
[16]. Min Chen , Shiwen Mao, Yunhao Liu, "Big Data: A Survey", Springer Science and Business Media, New York, 2014.
[17]. Tsai, Chun-Wei, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V. Vasilakos. "Big data analytics: a survey." Journal of Big Data 2, no. 1 (2015): 21.
[18]. Report for SAS, "The value of Big Data and the Internet of Things to the UK Economy", February 2016.
[19]. John Gantz and David Reinsel, IDC iView"Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East"December 2012, sponsored by EMC.
[20]. Narayana Bhagavatulal.V.S, Srinadh Raju.S, Sudhir Varma.S and Jose Moses.G, "A Survey Of Hadoop Ecosystem As A Handler Of Bigdata", International Journal of Advanced Technology in Engineering and Science, Vol.4(08), August 2016.

[21]. Kuchipudi Sravanthi, Tatireddy Subba Reddy, "Applications of Big data in Various Fields", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 6 (5) , 2015, 4629-4632.
[22]. Iuliia Protopop and Aleksan Shanoyan, "Big Data and Smallholder Farmers: Big Data Applications in the Agri-Food Supply Chain in Developing Countries", International Food and Agribusiness Management Review, Special Issue - Volume 19 Issue A, 2016.
[23]. J. Manyika et al., "Big data: The Next Frontier for Innovation, Competition, and Productivity", San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1-137.
[24]. Sjaak Wolfert, Lan Ge, Cor Verdouw and Marc-Jeroen Bogaardt, " Big Data in Smart Farming-A review", Agricultural Systems, 153 (2017), 69-80.
[25]. Pig- Apache Software Foundation project home page : http://pig.apache.org
[26]. Hive - Apache Software Foundation project home page : http://hive.apache.org
[27]. https://en.wikipedia.org/wiki/Apache_Hadoop