www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 5 May 2016, Page No. 16680-16685

Improvement of HADOOP Ecosystem and Their Pros and Cons in Big Data

S.Suguna[1]***, K.Devi***[2]
PG scholar,
Department of CSE,
Valliammai Engineering college,
Kattankulathur, Chennai603203
kalamanisuguna2014@gmail.com
Research scholar,
Department of CSE,
Valliammai Engineering college,
Kattankulathur, Chennai- 603203.
devii.jeya@gmail.com

Abstract: **-Big data plays a major role in all aspects of business and IT infrastructure. Today many organizations, Social Media Networking Sites, E-commerce, Educational institution, satellite communication, Aircrafts and others generate huge volume of data on a daily basis. This data is in the form of structured, semi-structured and unstructured. So this huge voluminous amount of data is coined as big data. These big data should be stored and processed in the effective manner. But, in the traditional distributed system this data cannot be effectively handled because of lack of resources. So the term Hadoop comes in to the picture. Hadoop stores and process the huge voluminous amount of data with their strong Hadoop ecosystem. It contains many modules for processing the data, storing the data, allocating the resources, Configuration Management, retrieving the data and for providing highly fault tolerance mechanism. In this paper it focuses on big data concepts, characteristics, real time examples of big data, Hadoop Modules and their pros and cons.**

Keywords:Bigdata;characteristics;Hadoop;HDFS; HadoopYARN;Mapreduce;Sqoop;Hive; Hbase.

I. INTRODUCTION

Big data is the accumulation of huge volume of data stored and processed on cloud with the help of internet connection. Even a small amount of data can also be represented as big data depending on the context being used. These massive amount of data may be in the size of GB (Giga Bytes), TB(Tera bytes) ,PB(Peta Bytes), EB(Exabyte).

Where this huge amount of data comes from?

These data comes from the social networking site such as (Face Book, LinkedIn, Twitter, Google+ YouTube etc) E-commerce web sites such as
(Amazon, Flipkart, Alibaba, E-bay), Weather station (Satellite, Indian Meteorological Department), Telecom company (such as Airtel, Vodafone, Aircel, BSNL etc),Share market (Stock Exchange) etc. On a daily basis these networking site, satellite, aircraft station, alone produce huge volume of data. This high volume of data should be processed in the effective manner [1][2][3][4][5][6] .

Real Time Examples are: Some of the real time examples are

E.g: 1. 20GB of data cannot be attached or supported in current email social network. Because this size is larger for current email storage, which can be referred as big data.

E.g:2 Consider 10 TB of image files. We need to process this image in to resize or enhance within the time frame in cloud. Suppose if we use the traditional system, then the processing cannot be done within the reasonable amount of time frame. Because of the computing resources in traditional system cannot be sufficient enough to process the image within the time frame. These 10 TB of data can be referred as big data.

E.g.3 Some of the popular networking sites are Face book, LinkedIn, Twitter, Google +, and YouTube. Each above mentioned sites has huge volume of data on a daily basis. It has been reported on some of the popular sites that FB alone receives around 100TB of data every data, whereas twitter process alone 400 tweets every day, as far as g+ and LinkedIn are concerned each of these sites process 10TB of data every day. And finally coming to the YouTube it has been reported that each minute around 48 hours of fresh video are uploaded. We can just imagine how much volume of data is being stored and processed on these sites. But as the number of users keeps on increasing in these sites, storing and processing this data is now becoming a challenging task. Since this data hold lot of valuable information this data needs to be processed within the short span of time. The companies can boost their sales and generate more revenue by using this valuable information. We would not be able to accomplish this task in the given time frame with the help of traditional computing system. As the computing resources of the traditional system wouldn't be sufficient for processing and storing such a huge volume of data. This is where Hadoop comes in to the picture. Therefore we can term the huge volume of data as big data.

E.g. 4: Another related real world example is airline industry. For instance the aircraft while they are flying they keep transmitting their data to the air traffic control located at the airports. To track and monitor the status and progress of the flight the air traffic control uses this data on the real time basis. Since multiple aircrafts would be transmitting this data simultaneously a huge volume of data gets cumulated at the air traffic control within the short span of time. Therefore it becomes a challenging task to manage and process the huge volume of data using the traditional approach. Hence we can term the huge volume of data as big data.

II. 5 V'S OF BIG DATA

The 5 main characteristics of big data are: Volume, Velocity, Variety, Variability and Value.

Volume

It represents the amount of data to be generated per second. Today, every organization generates petta bytes of data every second. [1]

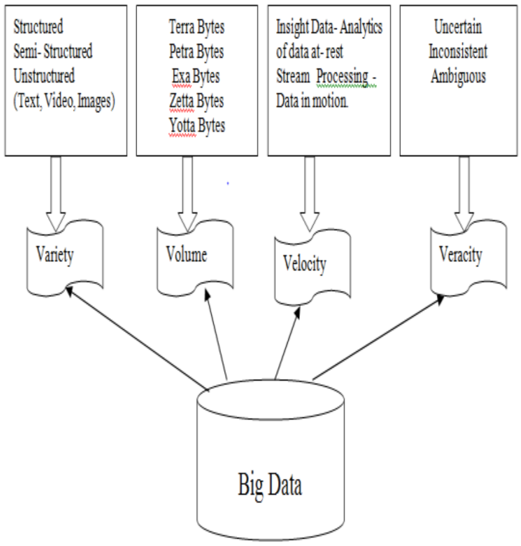

Fig . 1. Big Data Characteristics

Velocity

It represents the speed of data to be delivered to the end users. Any information can be delivered all around the world within a fraction of second[1][2].

Variety

It represents the various kinds of data. In the past only structured data was used. It can be stored only in the form of tables. But now all the structured and unstructured data (image, audio, video and sensors data) can also be stored and processed in the big data [1][2][3].

Veracity

It represents the accuracy, Trust worthiness and reliability of data [1][2][3][4].

Value

It represents the statistical value of data [1,2]

III. BIG DATA TECHNOLOGIES AND ITS IMPROVEMENT

Big data contains the massive amount of data. This massive amount of data is in the form of Relational Data base, (Transaction/Tables/Legacy data) ,Semi –structured data (XML, Text data(web),Graph data(Social Networking, Semantic web),Structured data(the data which resides in the traditional form of rows and columns), unstructured data(E-mail messages, word processing documents, videos, photos, audio files, presentations, web pages and many other kinds of business documents) and Streaming data .This data should be processed and handled in the effective manner[1][2][3][4][5][6].There are lots of tools and technologies are applied in the big data to process and produce the meaningful data to the end user. In this concern Hadoop plays a major role in order to process the huge volume of data. Here comes the technologies and their improvement and its advancement.

IV. HADOOP ECOSYSTEM AND THEIR MODULES

HADOOP

It was developed in the year of 2005 byDoug Cutting and mike caferella. It is the Apache open source software which allows to store and process the huge volume of data in a distributed environment and it is written in java. Hadoop is also called MR1.The major social networking sites such as Face book, Yahoo, Google, Twitter and LinkedIn uses the Hadoop technology to process their huge volume of data [2][3][4][5][6].

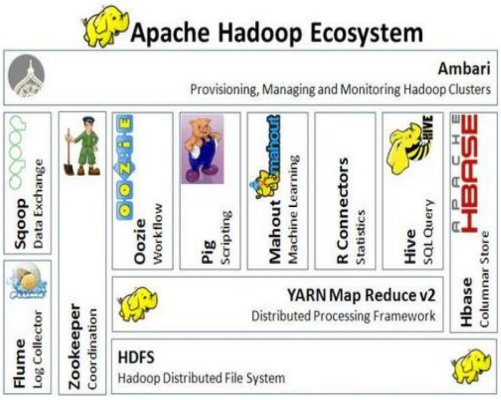

Fig 2.Hadoop Ecosystem

It is mainly designed to scale up from a single machine to thousands of machine; each offers the local computation and storage. Here, a single name node manages the whole namespace in the hadoop cluster. The major uses of this technology are Fast, scalable, Inexpensive Hardware and resilient to failure. Hadoop consist of two main frame work MAP REDUCE layer and HDFS layer. Map reduce layer is used for processing the big data (where the user application executes) and HDFS is used for store the big data (where the user data resides) [1][2][3][4][5][6].

HDFS:

HDFS is the Java based distributed file system used by Hadoop. So, it is called Hadoop distributed file system. HDFS is mainly designed to store the huge volume of data in a reliable and fault tolerant manner. It consist of Blocks, Name node (master) and the Data node (slave). Block is the minimum amount of data that it can either perform read or write operation. The default size of HDFS blocks are

128Mb.The files which are stored in the HDFS are split in to multiple block that are called Chunks which are independent of each other.**(i.e)** If the file size is 50 Mb then the HDFS blocks takes only 50Mb of memory space within in the default 128 Mb. [1][2][3][4][5][6].Name node is responsible for storing the Meta data which means it contains all the information about on which rack the data nodes data are stored. It contains the directory and location of data. The data nodes contain the actual user's data. On a single hadoop cluster there is only a single name node and multiple numbers of data nodes are present.

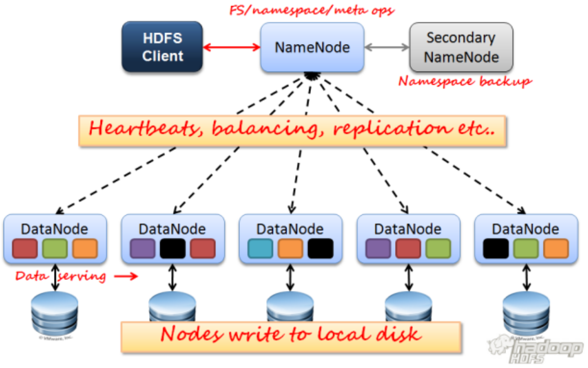

Fig 3.HDFS STRCUTURE

Advantages:

- ➢ It has very high bandwidth to support map reduce jobs.
- ➢ It is very less expensive.
- ➢ We can write the data once and read many times.

Disadvantages:

- ➢ Cluster management is hard.
- ➢ Join operation for multiple data base is slow and tricky

MAP REDUCE : Through map reduce the Hadoop process huge volume of data parallel on large number of clusters. Map reduce consists of Job tracker (Master) and task tracker (Slave) node. The job tracker is responsible for allocating the resources and scheduling the job on to the slaves.

The slaves task tracker is responsible for execute the task which was allocated by the job tracker and intimates its status periodically to the Job tracker with the help of heart beat message. [1][2][3][4][5][6].If the heart beat message is not received by the job tracker, then it assumes that the task tracker is fails in execution at that moment it reallocates the execution of failed task tracker jobs to another task tracker that are still alive in execution.

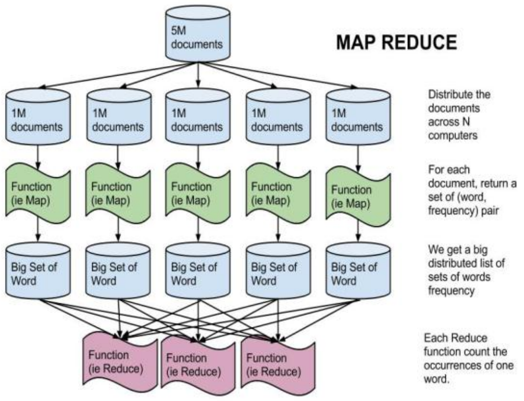

Fig 4.MAP-REDUCE STRCUTURE

Map reduce contains two major function MAP () and REDUCE () function.

MAPPER ()

- ➢ The Mapper function takes the collection of data and split in to multiple data where the individual elements are broken in to tuples (Key/value).

REDUCER ()

- ➢ The reducer function get the input from the map() and combines the data tuples in to small number of tuples. After the map () function, the reducer () initiates to get the final output.

Advantages:

- ➢ It supports wide range of language Java.
- ➢ It is a platform independent.

Disadvantage:

- ➢ It is applicable only for batch oriented process.
- ➢ It is not applicable for interactive analysis.
- ➢ It is not works for intensive algorithm, machine learning and graph.
- ➢ To overcome the limitation of Hadoop 1.0 we move to hadoop 2.0

Hadoop YARN

YARN (Yet Another Resource Negotiator). In hadoop YARN multiple name node server manages the whole namespace in the hadoop cluster. YARN is the heart or OS of the hadoop 2.0.It consist of four components such as Client, Resource Manager, node manager and map reduce application master. Client submits their job to the resource manager for processing[1][2][7].

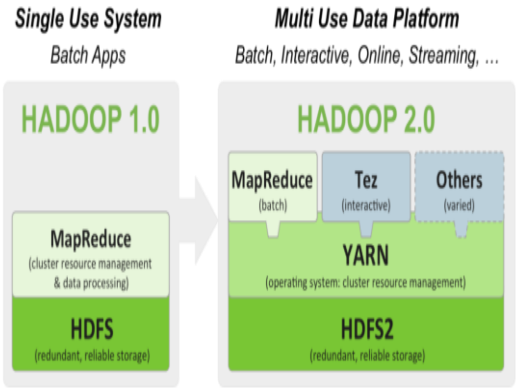
Single Use System
Batch Apps
Multi Use Data Platform
Batch, Interactive, Online, Streaming, ...
HADOOP 1.0
HADOOP 2.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
MapReduce
(batch)
Tez
(interactive)
Others
(varied)
YARN
(operating system: cluster resource management)
HDFS2
(redundant, reliable storage)

Fig 4.HADOOP 1.0 TO HADOOP2.0

The resource manager manages all the resources across the cluster. The node manager is responsible for maintaining the meta data information, monitors and keep track of user application. Map reduce application master is responsible for executing the application. It makes more interactive to multiple applications, manages all the resources management, job scheduling, provide security controls and high availability of data. Here it can execute the both non-map reduce and map reduce application.

Advantages:

- It provides efficient utilization of resources.
- It can run the application which doesn't follow
- It provides high availability of data.

HBase

It is an open source Apache frame work built on top of the HDFS and it is written in java. All the data are organized in the form of tables (i.e) rows and columns . The data stored in the form of Key/Value Pairin the columnar fashion. It supports structured data storage for large number of tables and the data can be retrieved easily. It offers the benefits such as we can insert, update and delete the data at any time. It is scalable, column oriented, non- relational and distributed data base. HDFS doesn't support random read /write operation but whereas HBase supports it. HBase offers storing of large amount of sparse data (a missing value does not need any space)and it support fast look up of data. The data structure used in Hbase is Log structured Merge (LSM) tree. The LSM provides indexed access of files and it gives high volume of insertion of data in to the tables [1][2][8][9][11].

Advantages:

- It is highly Fault Tolerant.
- It provides low latency access to small amounts of data from within a large data set.
- It is highly flexible data model.
- Strongly consistent

Disadvantage:

- It cannot be applicable to complicated data access patterns(Such as joins)
- It is not applicable for transactional applications or relational analytics.

Hive

Hive Hadoop was founded by Jeff Hammerbacher who was working with Face book. When he was worked with Face book he realized that huge amount of data is generated in the daily basis and that data needs to mined and analyzed in the well-defined manner. So that he gives birth to Hive. It supports queries expressed in the language called HiveQL or HQL. It is similar to SQL. We can query the data up to peta bytes .It automatically converts the normal SQL query in to map reduce jobs. It supports all kinds of data types such as integer, Float, double and strings as well as complex data types such as array, list and maps. Hive can also supports custom map reduce scripts in to queries. Hive provides Data ware house Infrastructure where we can summarize, manage and analyze the huge volume of data .The data which are stored in the Hbase can be processed and accessed through hive[1][2][21][22].

Advantages:

- The users need not to write their jobs in to Map Reduce programs.
- It provides tools for easy extraction, Transformation and loading of data from the large data ware house.
- It offers infrastructure for storing the data.
- Any SQL developer can easily write hive query.
- It can be integrated with Hbase for easy querying and retrieving of data.
- Map reduce programs needs more lines of code than HiveQL.

Disadvantages:

- It doesn't support for processing unstructured data.
- The complicated jobs cannot be performed using Hive.
- The output of one job can to be used to query the input for another jobs

Pig (Programming Tool)

It was developed by Yahoo in the year of 2006. It provides parallel computation with the help of high level data flow language and the powerful execution engine (i.e Pig engine) framework. It is written in the language Pig Latin. Pig is a scripting language for writing complex map reduces application such as joining, aggregation of huge volume of data. 10 lines of pig latin code is equal to 200 lines of java code. Pig reduces the number of lines of code and also save time. It converts the pig latin scripts in to map reduce tasks. We can have the flexibility that we can combine the java code with pig. The structured, semi structured and unstructured data can be processed by pig. It is mainly used for programming the data and operates on client side of any cluster [1][2][10].

Advantages:

- It decreases the Duplication of data.
- It reduces the number of lines of code and save the development time.
- The user defined functions can be easily programmed for read and write operations.
- It supports nested data models.
- The programmer who knows SQL language can easily able to learn and write pig scripts.

Disadvantages:

- It doesn't provide JDBC and ODBC connectivity.
- There is no dedicated Metadata data base.
- It doesn't offer web interface.

Sqoop

Sqoop is mainly used to transfer the huge amount of data between Hadoop and relational database. Sqoop refer " **SQ**L to Had**oop** and Hadoop to SQL". [11]It imports the data from the relational database such as Mysql, oracle, postgreSQL to the Hadoop (HDFS, Hive, HBase) as well as export the data from HDFS to relational data base. The non-Hadoop data store can also be extracted and transformed to Hadoop data store The Extraction, Transformation and Loading (ETL) can be performed by using Sqoop. It is the open source framework of cloudera Inc. [12]The data can be imported and exported in a parallel manner [1][2]

Advantages:

- It offers the migration of heterogeneous data.
- It offers easy integration with Hive,HBase and oozie.
- We can import the whole data base or the single table in to HDFS.

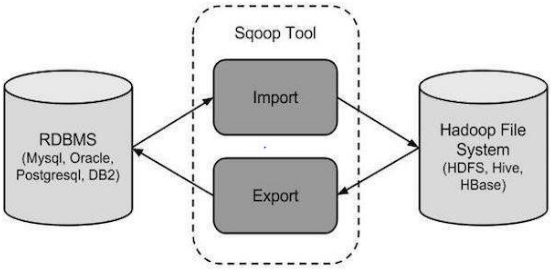

Fig 5.SQOOP Transformation

Zookeeper:

In a traditional distributed environment coordinating and managing the task is a complex and complicated one. But the zookeeper over comes this problem with the help of simple architecture and its API. In the cluster (group of nodes) to maintain the shared data and coordinating among themselves they use zookeeper as a service to provide robust synchronization. It provide services such as naming service(identify the name of the node in the cluster), configuration management(Up to date information is maintained), cluster management(the status of the node leaving or joining in the cluster), leader election(for coordinating among themselves they elect a single node as leader in the cluster), Locking and synchronizing service(when any data can be modified in the cluster it locks that particular data to provide consistency) etc. The inconsistency of data, race condition and deadlock problem in the traditional distributed environment can be solved easily with the help of Zookeeper mechanism such as Atomicity, serialization property and synchronization respectively[1][2][13][14][15].

Advantages:

- It provides reliability and availability of data.
- It offers high synchronization and serialization.
- The atomicity eliminates the inconsistent of data among clusters.
- It is fast and simple.

Disadvantages:

- The large number of stacks needs to be maintained.

Cassandra

It was initially developed by face book to handle large volume of data and later it was acquired by Apache Software foundation in the year of 2009.It provides no single point of failure, high availability, high scalability, highly fault tolerant and high performance of data. Many servers are interconnected with each other and if the node goes down, another node can provide service to the end user .To allow all the nodes to communicate with each other and to detect the faulty node a gossip based protocol is used[16][17].

Advantages:

- The biggest companies such as Face book, Twitter, Rack space and Cisco uses the Cassandra to handle their data.
- It contains Dynamo- Style Replication model to provide no single point of failure.
- It provides high throughput and quick response time if the number of nodes in the cluster increases.
- The ACID property is supported for transaction.
- It is a column oriented data base.

Disadvantages:

- It does not support sub query and join operation.
- Limited support for data aggregation.
- Limited storage space for a single column value.

Mahout:

It is the Apache open source software framework and it provides data mining library. The processing task can be split in to multiple segments and each segment can be computed on the different machine in order to speed up the computation process. The primary goals of the mahout are data clustering, classification, regression testing, statistical modeling and collaborative filtering. It provides scalable data mining and machine learning (it makes the decision based on the current and previous history of data) approaches for the data[1][2][18].

Advantages:

- It supports complementary and distributed naive bayes classification.
- It mines the huge volume of data.
- The company's such as Adobe, Twitter, Foursquare, Face book and LinkedIn internally uses mahout for data mining.
- Yahoo specially used for pattern mining.

Disadvantages:

- It doesn't support scala version in the development.
- It has no decision tree algorithm.

Oozie

It was initially developed at yahoo for their complex workflow search engine. Later it was acquired by open source Apache incubator. It is a workflow scheduler for managing Hadoop jobs. There are two major types of oozie jobs are available i.e oozie workflow and oozie coordination. In the oozie workflow it follows Directed Acyclic Graph (DAG) for parallel and sequential execution of jobs in the Hadoop. It contains the control flow node. The control flow node controls the beginning and end of the workflow execution. In the oozie coordination, workflow jobs are triggered by time [1][2][19][20].

Advantages:

- It allows the workflow of execution can be restarted from the failure.

- It provide Web service API (i.e we can control the jobs from anywhere)

Disadvantages:

- It is not a resource scheduler.
- It is not suitable for off grid scheduling.

V. CONCLUSION

The huge volume of big data can be processed and stored with the help of Hadoop ecosystem. It supports many modules that are integrated with each other. Each module provides many services for processing, storing and retrieving the big data in an effective manner. This paper brings a brief note on big data concepts, their characteristics, hadoop ecosystem, HDFS, Map reduce concepts, their improvement modules and their pros and cons

REFERENCES

[1] Ishwarappa, Anuradha “A Brief Introduction on Big data 5Vs Characteristics and Hadoop Technology” Ishwarappa and J. Anuradha / Procedia Computer Science 48 (2015) 319 – 324

[2] Manoj Kumar Singh and Dr.Parveen Kumar “Hadooo: A Big data framework for Storage, Scalability, Complexity,Distributed Files and Processing of massive Data Sets” International Journal of Engineering Research and General Science Volume 2, Issue 5, August – September 2014.

[3] Amrit pal, Pinki Agrawal,Kunal Jain and Sanjay Agrawal “A Performance analysis of map reduce task with large number of files data sets in big data using Hadoop” 978-1-4799-3070-8/14 $31.00 © 2014 IEEE

[4] Shakil Tamboli and Smita shukla patel “A survey on Innovative approach for improvement in efficient of caching Technique for Big Data applications” -1-4799-6272-3/15/$31.00(c)2015 IEEE

[5] Shankar Ganesh Manikandan and Siddarth Ravi “Big Data Analysis Using Apache Hadoop” 978-1-4799-6541-0/14/$31.00 ©2014 IEEE

[6]Ankita Saldhi, Abniav Goel, Dipesh Yadav, Ankur Saldhi ,Dhruv Saksena and S.Indu “Big Data Analysis Using Hadoop Cluster” 978-1-4799-3975-6/14/$31.00 ©2014 IEEE.

[7] http://www.ibm.com/developerworks/library/bd-yarn-intro/

[8]http://wikibon.org/wiki/v/HBase,_Sqoop,_Flume_and_More:_Apache_Hadoop_Defined

[9]http://stackoverflow.com/questions/16929832/difference-between-hbase-and-hadoop-hdfs

[10]http://www.tutorialspoint.com/apache_pig/apache_pig_overview.htm

[11]http://wikibon.org/wiki/v/HBase,_Sqoop,_Flume_and_More:_Apache_Hadoop_Defined

[12]http://www.dummies.com/how-to/content/hadoop-sqoop-for-big-data.html

[13]https://www.dezyre.com/hadoop-tutorial/zookeeper-tutorial]

[14] http://hortonworks.com/hadoop/zookeeper/

[15]https://cwiki.apache.org/confluence/display/ZOOKEEPER/Index

[16] http://www.edureka.co/blog/apache-cassandra-advantages

[17]https://www.quora.com/What-are-the-advantages-and-disadvantages-of-using-MongoDB-vs-CouchDB-vs-Cassandra-vs-Redis

[18] http://hortonworks.com/hadoop/mahout/

[19] http://hortonworks.com/hadoop/oozie/

[20]https://www.dezyre.com/hadoop-tutorial/oozie-tutorial

[21] http://www.tutorialspoint.com/hive/hive_introduction.htm

[22]http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/