An overview on Big Data and Hadoop

Shaikh Abdul Hannan
Computer Science and I.T.,
Al-Baha University,
Al-Baha, Saudi Arabia.

ABSTRACT
Big data: Everyone just talking about Big data, but what is meant by big data actually? How is it changing the point of view in different fields such as researchers of the science or at companies, non-profits, governments, institutions, and other organizations are learning about big data that is nothing the world around them? Where this data is coming from, how is it being processed, and how are the results being stored and used for their future work? And why is open source so important to answering these questions? In this paper we will discuss all above points to clear actually what Big data means and how it deals in our day to day life. In today's 21st century, the most important area is social media which shares, search and shares the information and generates huge of data everyday. So the importance of big data is more as millions and billons of peoples are using this media to share and store the information. Nowadays many projects are developing under social media, sensor data, stock exchange, Transport data, and in the field of science where data is most important factor to store and retrieve. So we need new technology which is Big data and Hadoop to handle this huge amount of data which is not possible to handle by RDBMS. Big data has very basic important characteristics such as volume, variety, veracity and velocity. Big data handles the large amount of data with management, analysis, storage and processed data very fast within the time span. In this paper discusses, the important characteristics, types of data which is used in big data, what are the various sources of big data in our day to day life, introduction to big data and Hadoop with explanation, Structure of Hadoop core components, role of Namenode and data node, function of job tracker and task tracker, and Hadoop Ecosystem is explained in detail.

Keywords
Big Data, Hadoop, HDFS, MapReduce, Hadoop Ecosystem, Namenode, Datanode.

1. INTRODUCTION
Data is very important factor in every field of today life. There is no hard and fast rule about exactly what size a database needs to be in order for the data inside of it to be considered "big." Instead, what typically defines big data is the need for new techniques and tools in order to be able to process it. In order to use big data, you need programs which span multiple physical and/or virtual machines working together in concert in order to process all of the data in a reasonable span of time. [1]. Big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. Big data is an evolving term that describes any amount of structured, semi-structured and unstructured data that has the potential to be mined for information [2]

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data creation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. "Big Data refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. Big Data includes business transactions, e-mail messages, photos, surveillance videos and activity logs (see machine-generated data). Scientific data from sensors can reach mammoth proportions over time, and Big Data also includes unstructured text posted on the Web, such as blogs and social media." [3]. Or Big data is a collection of huge amount of data that is larger and complex to process using on hand database management tool or traditional data processing applications like REBMS, DBMS or SQL files.

Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on." Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, complex physics simulations, and biological and environmental research [4].

Normally the data size is like MB, GB for example considering a video which a few GB may 1GB, 2 GB or 5Gb or it can be some GB. An audio file which is 1000 x 1000 x 1000 terabyte, So in social media everyone shares picture, posts, audio file, video file etc. so it is certainly a large amount of data so this is what a big data is.

Every day, the data is created near about 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data. In 1992, the data was produced by the internet was just 100GB per day, 1997 the data produced by the internet was 200GB per hour, in 2002 it was 100 GB per second, and in 2013 it was 28,875GB and it is predicted that in 2018 50,000GB per second data might be generate. The following figure how data is rapidly increasing[5].

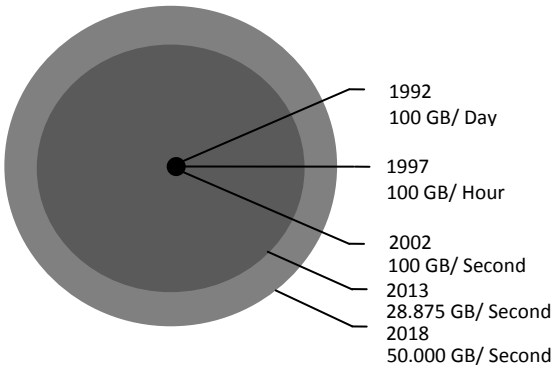

Figure 1: Global internet traffic

2. RELATED WORK

What are the Big data issues, challenges, tools are important studies for the basic things to know for everyone who is related to computer science. In this paper[6] author discussed about basic properties of big data like volume, velocity, variety, complexity, value has been discussed. And what are the important sources from which big data is coming and generated. Big data has a great importance in various fields like social media, government sector, sensor data, and log storage and risk analysis. In social media there is great importance of big data. Facebook is generating Terabytes of data on every day. The authors discussed about the protection and the how to secure data, are very sensitive issue and it can disturb any human beings personal life because the data is so important that has taken for analysis, it is personal data of any individual, in case of social media, after discovery of some pattern may be that person does not want that it should be known to someone. Some technical challenges like fault tolerance and scalability are also discussed. Hadoop and map reduce has been discussed as the main tools and technologies to process and deal with big data. Hadoop is basically a framework on which map reduce works as a programming model. It works in batch processing means it divides the task into smaller units and then executes them parallel. At the end of the paper a comparison between Hadoop and grid computing tools is also shown.[6]

In this paper [7], Big data is growing at an exponential rate so it becomes important to develop new technologies to deal with it. This paper covers the leading tools and technologies for big data storage and processing. Hadoop, Map Reduce and No SQL are the major big data technologies. These technologies are very helpful in big data management. Technologies based on Hadoop called Hadoop Eco system have also discussed. This Paper also throws some light on other big data emerging technologies. There are so many areas from which big data is being generated, this paper covered those areas and provide solutions for dealing with that data.

Ten common Hadoopable problems by Cloudera explained in detail about Hadoop problems. In this, paper, Cloudera Company explained where Hadoop technology can be useful to implement and they also provide a solution to a specific problem with Hadoop. Banks need Hadoop to perform risk analysis on their customer's profile. There is a important and different type of data present in the banking, government, and financial institutes. This data is very sensitive and important due to its privacy and security purpose. So this data needs better management and Hadoop is capable of managing and retrieving data very fast. There are other areas also like advertisement targeting which helps companies to target perfect customer for their products to sell and Hadoop is very much best technology in this area. Point of sale transaction analysis is now in huge demand, which figure out the customer's buying pattern. Fraud analysis, analysing and tracking network data to predict failure, threat analysis or other areas in which Hadoop can be very efficient and helpful[8].

In this paper, [9] discussed about Hadoop and Hadoop distributed file system infrastructure extensions in detail. In this major enhancement in Hadoop is able to store the data so it is nothing but data storage, data processing and placement are done by using MapReduce are also reviewed. It has also shown Comparison of Hadoop Infrastructure Extensions on the basis of scalability, fault tolerance, load time, data locality, data compression, etc. Hadoop is widely accepted in many areas, but its extensions, which are the improvements of Hadoop, can also be very helpful. HadoopDB, Hadoop++, CoHadoop, Hail, Dare, Cheetah, etc. are the main extensions of Hadoop and these are considered for comparison.

Why we use big data and some important Characteristics discussed [10] with unstructured data and how to deal with it. Results shown in this paper, that big data analytics is very much important to make business intelligent. Kapil Bakshi [11] mainly discussed about the analysis of unstructured data. Map reduce and Hadoop are the major tools for the analysis of unstructured data and it is widely discussed in this paper. Demchenko,Y, de Laat, C., Membrey, P. [12] Discussed about the basic definition of big data and also focused on the importance of the Hadoop Eco system. 5 V's Volume, Velocity, Variety, Value and Veracity have been discussed as the main properties of big data. Big data analytics, security, data structure and models are the main components of the Hadoop Eco system. The author reviewed these core architectural components of the Hadoop Eco system. These components are very important in big data challenges.

3. IMPORTANT CHARACTERISTICS OF BIG DATA

The general consensus of the day is that there are specific attributes that define big data. In most big data circles, these are called the four V's: volume, variety, veracity and velocity.

1. Volume:

Big Data implies enormous volumes of data. It used to be employees or the user created data. Now that data is generated and created for different purposes by machines, networks and human interaction on systems like social media the volume of data to be analyzed is huge amount of data [13]. The volume a data which was made earlier come from only the employee of the organization but today data comes from employees, partners, customers, Facebook, Twitter, etc. comes from every place and websites. Now consider a file which might of few KB, be a few GB, now imagine the amount of data that everyday might have. If all that thousands of file put together, so this is volume of data.

2. Variety:

Variety refers to the many sources and types of data both structured and unstructured. The data is stored in spreadsheets and databases which comes from sources. Now data comes in the form of email, photos, videos, monitoring devices, Pdf. Audio etc. This variety of unstructured data creates problems for storage, mining and analyzing data.

Today data comes in all format for example social networking sites that Facebook, Twitter, LinkedIn, YouTube etc. So this is variety so it has been seen that approximately 400 million tweets are sent per day from 200 million active users on twitter. So this is variety of data that is different forms of data.

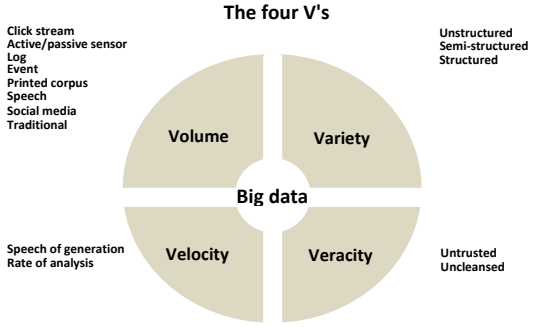

Fig. 3.1 Four V's of Big Data

3. Veracity

Big data Veracity refers to the biases, noise and abnormality in data is the data that is being stored and mined meaningful to the problem being analyzed. Veracity in data analysis is the biggest challenge when compares to things like volume and velocity.

Now biases and noise and abnormally data or uncertainty of data it means the data which is not certain, which does not have any naming for example some peoples uses on the Facebook, like good and good is actually written in Short hand GUD and this GUD has no meaning of its own so this is uncertain data or noisy. Similarly good morning it is written as GM and this GM has no meaning in any English dictionary but still it is accepted and this type of data is called uncertainty of data and it is part of big data

4. Velocity

Big data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices etc. The flow of data is massive and continuous. This real time data can help researchers and businesses make valuable decisions.

Velocity means Analyzing of streaming data. How fast the data is processed for example, the New York stock exchange that is biggest stock exchange in this world. It generates and captures and also deals with 1 Terabyte of data in each trading session so you can just imagine how much fast data has to be processed so this is about the velocity that analyzing of streaming data. So this was about four 'v' of Big data [14].

3.1 Types of data

1. Structured Data
2. Unstructured Data and
3. Semi-structured Data

Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis. Structured data for example RDBMS, DBMS and all so this is structured data. The database which has information stored in such a way that it can be readily used is called structured data. The data is stored in the form of tables and rows and columns and perfect example is RDBMS. So this is structured data.

Unstructured data can be textual or non-textual, this is the variety of data, and the data can be pdf files, audio files, video files, images and all sorts of data that comes from social networking sites. Such as Facebook, Twitter, YouTube other websites which accepts and this data can be in the form of likes, posts, comments then uploading a photo and every sort of this data comes under the unstructured data.

Semi structured data is nothing but the metadata, here under the parent node, the child node further this under child node another child node, so questions might come in mind why it is called as semi structured data. Now since they are organized in a manner but not so organized but they can be stored in databases. For example, consider HTML program, it has tag in that, it has a parent tag then child tag. This type of data is called as semi structured data.

So actually a traditional system that is RDBMS deals with only structured data but now there is need to handle, structured, unstructured and semi structured data. So there was need of technology that deals with all three types of data. So this need had to be fulfilled because reason being approximately total 80% data is either semi-structured or unstructured which cannot be deal by RDBMS that's why we need Big data.[15]

3.2 Sources of Big Data

Sources of big data are first one is social media and networks.

1) Mobile devises
2) Sensor Technology and Networks and
3) Sensor Technology
4) Scientific instruments

Now consider the first one that is social media and network, the social media network including Facebook, then LinkedIn, Twitter, YouTube, Google, all the sources of big data. The big data is in the form of likes, comments, posts, uploads, making pages, making groups this is all big data. So this social media is certainly the sources of big data.

Mobile devices, earlier mobile devices are mostly used for calls and to send a text messages. No this has become a super device which can track the user and this tracking can generate a lot of data this is certainly the source of big data.

Sensor methodology and network in this sensor case, sending of the signal and receiving of a signal then this signal is analyzed in real time and all this generates a lot of data so this certainly the source of big data.

Scientific instruments, now scientific instruments a satellite moving around the earth is programmed to take picture after interval of one min so now imagine the amount of data that how much it will be produced. If it is taking picture after every one min so it is certainly a large amount of data. It is not only important to store the picture but to analyses the picture also. So this is also certainly a source of big data.

4. BACKGROUND

The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks [16].

The main challenge is not to store the data but to read and write the data, to analyze the data and something which distributes the files. So Hadoop is framework which solves this problem and reading from single place is not going to solve the problem. So now need to have some distributed way of storing that data and this HDFS fulfills that requirement which stores the data across the nodes that's why Hadoop technology is needed. Hadoop also helps in processing huge amount of data at a very fast rate.

Hadoop what it does, instead of storing data at a single location it stores data in distributed fashion and it follows the distributed file system where instead of storing data in a single place, it stores data in different nodes and so that data can be retrieve in parallel..

4.1 Hadoop

Hadoop was developed by Daug Cutting and Mike Caferella[17]. Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity commuters using a simple programming model

MapReduce [18] is a framework pioneered by Google for processing large amounts of data in a distributed environment. Due to the simplicity of its programming model and the run-time tolerance for node failures, MapReduce is widely used by companies such as Facebook [19], the New York Times [20], etc.

Distributed processing of large data sets so the data is not stored at one place, input and output is not happened from place that is stored across the cluster on commodity computers using a simple programming model, that programming is called MapReduce.

The way MapReduce is works across the clusters of your computer then there is a framework programming model which takes core of synchronizing the data or reading the data from where to reading the data so all these things are taken care by system. If the job is running, it has to restive this data from multiple clusters all this is taken by Hadoop framework. There is simple model programming which is called as MapReduce programming and it is open source data management.

Hadoop can be said as fault tolerant distributed system for data storage and processing which is open source by Apache. Hadoop provides the reliable shared storage and analysis system. It is designed to scale up from a single server to thousands of machines with a high degree of fault tolerance. Reliable means it creates a replica of data three times, means a data corrupt or bad or anything happen to access with data, the replica created with cluster, so no need to think about data to loose.

Hadoop is technology, data warehouse is where data store, mining means analyzing and Hadoop is what data mining is done on data warehousing.

Data store in warehouses is from operational systems all the data for example data center and data is stored in clusters and cluster is where that data is stored, that in combination of racks or racks are a combination of data nodes. When all these combines a data center is made for example data warehouse is made. Data warehouse stores the current and historical data in order so that data mining can be done on it. Actually data mining is analyzing the data there has to be a technique which we can do it. This technique is called Hadoop.

4.2 Hadoop Core Components:

A Hadoop cluster is composed of two parts: Hadoop Distributed File System and MapReduce. A Hadoop cluster uses Hadoop Distributed File System (HDFS) [21] to manage its data.

HDFS is Hadoop Distributed File System: This HDFS is Hadoop Distributed File System is used for storing and processing is done by MapReduce. Now Hadoop Distributed File System is a distributed file system which holds large amount of data across multiple nodes in a cluster which is in contrast with primitive server which has limited storage and do not store data on multiple nodes. In HDFS the file is broken into a small blocks with default size of 64 MB and these blocks are replicated across various clusters which is three by default. Replica is for durability, high availability and through put. Basically there are some features of HDFS which are very important as it is very highly fault tolerant because the data is retrieved from different multiple nodes not from one single node because the data is replicated by default on three different nodes. And high through put means amount of time to read the data is very high because it reads the data from different machines [22].

HDFS allows to put/get/delete files. It also follows the policy for write once and read multiple times. For example upload a picture in Facebook,, it can be seen whenever the user want to see it, and user can see the picture after 5 years. So user is writing, user uploading file once and reading it multiple times.

In this diagram two things are important to note and just remember Name node is nothing but Admin node it's a master slave type of configuration. So there is something called name node or admin node which is master then there are slave which is called data nodes and then there is something called job tracker which is associated with name node and there is something called task tracker which is associated with data node.

Basically a HDFS cluster is the name given to whole configuration masters and slaves where the data is stored and MapReduce engine is the programming model which is used to retrieve and analyze the data.

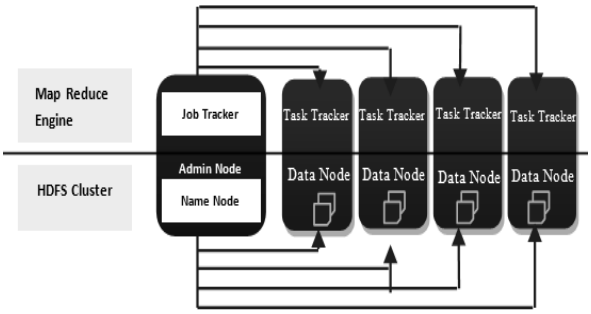

Fig 4.1 Core Components of Hadoop

MapReduce

Hadoop MapReduce is the computation framework built upon HDFS. There are two versions of Hadoop MapReduce: MapReduce 1.0 and MapReduce 2.0 (Yarn [22]). The MapReduce takes a set of input key/value pairs and produce a set of output key/value pairs. When a MapReduce job is submitted to the cluster, it is divided into M tasks and R reduce tasks, where each map task will process one block (e.g., 64 MB) of input data. MapReduce is a distributed programming paradigm used to analyse the data in HDFS and it is made up of two procedures. In map phase, data is mapping and sorting is done and in Reduce phase which performs logic operations. So this is about core components.

4.3 HDFS Components

Basically there are two major components of HDFS. One is Name node and Data Node. Namenode is the master node on which the job tracker runs and Data node is slave on which task tracker runs. Namenode is the master system with high reliability machine. It does not store any data it maintains and manages the blocks which are present on the datanodes. It is the datanode which actually keeps the data. Datanodes are slaves are deployed on each machine and provide the storage facility. These are responsible for serving read and write request for the client. Read and write directly happen from the data node requested by the client and the task tracker runs on each data node.

A Datanode periodically reports its status through a heartbeat message and asks the Namenode for instructions. Every Datanode listens to the network so that other Datanodes and

users can request read and write operations. The heartbeat can also help the Namenode to detect connectivity with its Datanode. If the Namenode does not receive a heartbeat from a Datanode in the configured period of time, it marks the node down. Data blocks stored on this node will be considered lost and the Namenode will automatically replicate those blocks of this lost node onto some other datanodes [23].

The datanode is place where the actual data is stored i.e. structured and unstructured data is stored. The Namenode is the node which contains the metadata around this datanode. Any information regarding the data, where it is available is in the data node.

YARN is a complete framework which is a resource manager that resides in Namenode which takes care of the all over resource management. Then there is something called Node manager which resides on data node which manages the data inside or jobs which are running inside the datanode [24].

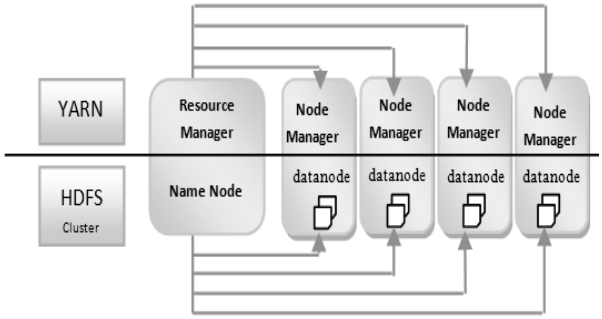

Fig. 4.2 HDFS cluster and YARN

One is HDFS side which is storage side and another is YARN side which is framework side. Under Namenode, the datanode and Namenode interacts with Node manager where the program is run in datanode. The node manager is running on the datanode and resource manager interacts with node manager to make sure which job is allocated to the node manager inside it runs the job and on the data node and gives the signal back to resource manager [25].

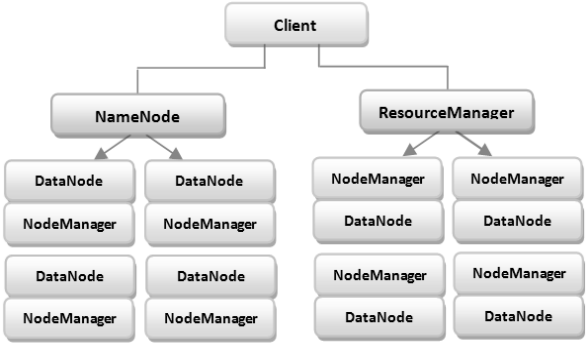

Fig. 4.3 Namenode and Resource Manager

5. HDFS ARCHITECTURE

Client is the application software which will run on machine which is used to interact with datanode and Namenode. The client, read and write the data from the datanodes [23].

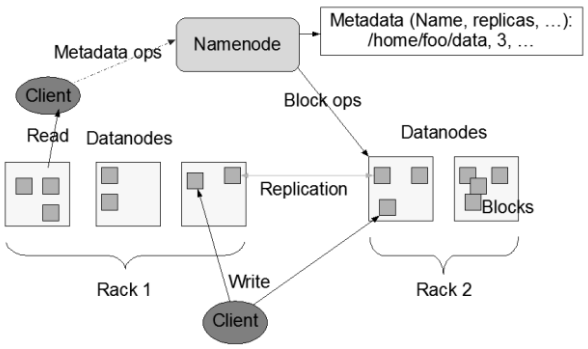

Fig5.1HDFS Structure. Source: http://hadoop.apache.org

It contains Namenode and client. Now there is something called Rack so rack is the storage area where multiple data nodes put together physically all these data nodes may be of different places. Basically rack is a physical collection of data nodes which are stored at single location and there can be datanodes in different places. Basically the diagram shows the client which interacts with Namenode. Here term called replication so as to maintain the data from the fault tolerance in Hadoop system, the same data is distributed in multiple copies and replicate in multiple data nodes and minimum number of replicas required by HDFS and it can be defined by the user. So basically when the data writes on HDFS it replicates in three data nodes by default on different nodes. And there is something called Block Ops and there are the operations are performs by using blocks and the default size of each block in HDFS is 64 MB.

Metadata is the actual data, it is the data over data where the block are stored which are racks available on which rack which datanodes is available or it is nothing but block map of Namenode. Namenode this data is stored in RAM to have fast access and actual data is stored in datanode.

5.1 Job Tracker and Task Tracker

MapReduce is the programing model to retrieve and analyze the data. And client is actually application which is used to interact with both the Namenode and datanode which means to interact with Job Tracker and Task Tracker. A client is application software which will be running on your machine which is used to interact and give commands and look status of Job Tracker or Task Tracker so it is an interaction between user and Namenode and datanode is done through a client and it is also called HDFS client [26].

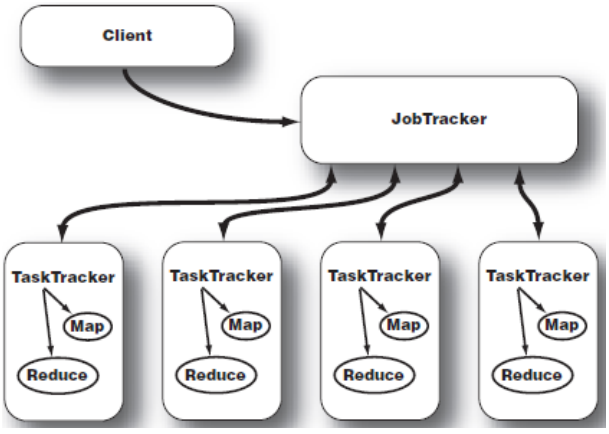

Fig. 5.2 JobTracker

6. HADOOP ECOSYSTEM

The Hadoop Ecosystem consists of tools for data analysis, moving large amounts of unstructured and structured data, data processing, querying data, storing data, and other similar data-oriented processes. These utilities each serve a unique purpose and are geared toward different tasks completed through or user roles interacting with Hadoop [28].

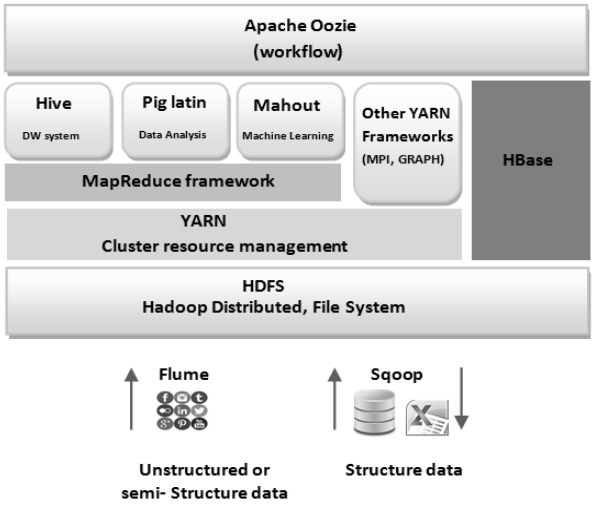

Fig. 6.2 Hadoop Ecosystem

So Hadoop is not one tool, its ecosystem its framework so that are the various things in Hadoop.

The place where actually get data into Hadoop so Hadoop has HDFS which is called as Hadoop Distributed File System which is nothing but the file system which is on top of cluster that is Hadoop Distributed File System. All the data is stored in Hadoop Distributed File System or the HDFS. Now how do you store more data into to Hadoop?

Flume is a framework for harvesting, aggregating and moving huge amounts of log data or text files in and out of Hadoop. There are multiple ways for that there is tool called Flume which is mostly used for moving unstructured or semi structured data into Hadoop. To store structured data into Hadoop it doesn't stop for storing structured data. Anything which is coming from the web such as Facebook, LinkedIn, Twitter or any social media, the loading of method that data is done through Flume so what Flume does is it is a tool from the channel which HDFS it can send the data inside the Hadoop then the data such as in RDBMS from my SQL or there is a connector for oracle, the tool called Sqoop.

Apache Sqoop efficiently transfers bulk data between Apache Hadoop and structured data stores, such as relational databases. Sqoop imports individual table or complete dataset to HDFS. Sqoop can also be used to extract data from Hadoop and export it into external structured data stores. Sqoop works with relational databases such as Teradeta, Oracle, MySQL, etc.

The full form of Sqoop is SQL to Hadoop. The structured data can move into Hadoop the tool Sqoop is used and the data from RDBMS tool the connectors are available on top of it. Assume this is your cluster. On the top of this cluster, need to have a system or a framework which can do resource management or to run a Job on Hadoop or program on Hadoop there has to be a framework where should this job be done, where the data is available, where it stored, where it should move data, this all that is done by a framework called YARN and it is stands for 'Yet Another Resource Negotiator'. It manages the framework which manages the complete resource manager. Then it contains something called MapReduce framework. The MapReduce framework is the framework which is used on top of Hadoop to process the job. So the programs can be written in MapReduce then that program is broken and sent to the different nodes on the cluster where the actual data resides and to collect your results and come back.

Hive is a data warehousing package built on top of Hadoop that is used for complex data analysis and exploration. Hive is a tool which essential internally used MapReduce but it gives you the flexibility, those who are not from the programming background and this was developed by Facebook.

Pig is an open-source; high-level dataflow system that sits on top of the Hadoop framework and can read data from the HDFS for analysis [29]. Pig Latin which was developed by Yahoo and mostly it is used for data analysis.

Mahout is machine learning tool. Mahout is a scalable machine learning library that implements various different approaches machine learning. At present Mahout contains four main groups of algorithms:

1. Recommendations, also known as collective filtering
2. Classifications, also known as categorization
3. Clustering
4. Frequent item set mining, also known as parallel frequent pattern mining

There is other framework graph job, to run a real time job which is again and in Hadoop 1.0, there is only MapReduce it was supported but now YARN coming to have a lot of tasks then there is no SQL data base which is called HBase. HBase is a distributed, column oriented database and uses HDFS for the underlying storage. As said earlier, HDFS works on write once and read many times pattern, but this isn't a case always. We may require real time read/write random access for huge dataset; this is where HBase comes into the picture. HBase is built on top of HDFS and distributed on column-oriented database. Apache HBase is a column-oriented, NoSQL database built on top of Hadoop. This can be stored data in column then there is tool called Apache Oozie which is for work flow management for example thousands of job running and to manage the work flow then this tool is used.

7. CONCLUSION

This paper has presented all the overview and basic introductory discussion which is more important for the researcher those who are joint as a beginners. Time has been started when the world is capable of generating data in terabytes and petabytes every day, every hour and every minute. Big data and Hadoop provides the more facilities for the technology and related to open source tools which is in developing stages given and becomes more important in future. In the recent years, related to all the fields such as social media, government projects, sensor data all are giving more important to Big data and its technologies. This Paper also throws some light on other big data emerging technologies.

Although this paper clearly has not resolved and covered all the points so the researcher can extend this topic with their need in the subject of big data for different topics. The

researcher can frame a framework and use this important overview for their basic research. The researcher can use this paper and can extend their work for reading and writing the data in Hadoop.

8. REFERENCES

[1] https://opensource.com/resources/big-data

[2] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[3] http://www.opentracker.net/article/definitions-big-data

[4] http://studymafia.org/wp-content/uploads/ 2015/05/CSE-Big-Data-Report.pdf

[5] http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

[6] Avita Katal, Mohammad Wazid, R H Goudar "Big Data: Issues, Challenges, Tools and Good Practices". In IEEE, Contemporary Computing (IC3), Sixth International Conference, pages 404-409, Noida, 2013.

[7] Jaskaran Singh and Varun Singla "Big Data: Tools and Technologies in Big Data", International Journal of Computer Applications, Volume 112, No. 15, Feb. 2015.

[8] Cloudera White paper,"Ten Common Hadoop able Problems", 2011.

[9] Kala Karun. A, Chitharanjan. K, "A Review on Hadoop – HDFS Infrastructure Extensions". In IEEE, Information & Communication Technologies (ICT), pages 132-137, 2013.

[10] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics". In IEEE, International Conference on Communication, Information & Computing Technology (ICCICT) pages 1-4, 2012.

[11] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach". In IEEE, Aerospace Conference, pages 1-7 2012.

[12] Demchenko,Y, de Laat, C., Membrey, P.," Defining architecture components of the Big Data Ecosystem".In Collaboration Technologies and Systems (CTS),pages 104-112,2014.

[13] http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

[14] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[15] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[16] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System". In IEEE, Contemporary Computing (IC3), Sixth International Conference, pages 404-409, Noida, 2010.

[17] http://searchcloudcomputing.techtarget.com/definition/Hadoop

[18] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," in Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6, ser. OSDI'04. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10.

[19] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, "Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling," in Proceedings of the 5th European conference on Computer systems. ACM, 2010, pp. 265–278.

[20] D. GOTTFRID, "Self-service, prorated supercomputing fun!" http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-supercomputing-fun/.

[21] Apache, "Hdfs," http://apache.hadoop. org/hdfs/

[22] A. Foundation, "Yarn," https://hadoop.apache.org/docs/r0.23.0/hadoopyarn/hadoop-yarn-site/YARN.html

[23] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System" Yahoo! Sunnyvale, California USA.

[24] Jia-Chun Lin, Ingrid Chieh Yu, Einar Broch Johnsen , "ABS-YARN: A Formal Framework for Modeling Hadoop YARN Clusters ?, Ming-Chang Lee Department of Informatics, University of Oslo, Norway.

[25] Khalid Adam Ismail Hammad, et. al. Big Data Analysis and Storage, Proceedings of the 2015 International Conference on Operations Excellence and Service Engineering Orlando, Florida, USA, September 10-11, 2015.

[26] https://hadoopinku.wordpress.com/category/hadoop-2/