Big Data and Hadoop -A Technological Survey

Manika Manwal
School of Computing
Graphic Era HillUniversity,
Dehradun, India
manikamanwal17@gmail.com

Amit Gupta
School of Computing
Graphic Era HillUniversity,
Dehradun, India
amitgupta7920@gmail.com

Abstract: **Generally, when we here a term "BIG DATA" many assumption and queries generate in our mind like when did this term Big Data came into the picture? What is it? Why we need it? How is it beneficial? There are many events which can be taken into consideration starting it from the ancient world to the modern world, data was always stored. Nowadays the enormous amount of data is been produced via many organizations e.g. Twitter, Facebook, LinkedIn etc. so thru Big Data huge amount of data with a varied variety can be stored, managed and processed. Big Data make sure that the accuracy and knowledge from the data are generated rapidly and provides benefit and convenience to the organizations, researchers, and consumers by accounting the properties of Volume, Value, Variety, Veracity, and Velocity. This study comprises of the introduction in which the Big Data is entailed in addition to this is its features and classification. Then this study talks about the hadoop and its architecture plus the components used in hadoop like HDFS, Map Reduce, Pig, Hive, Hbase, and Sqoop. In the conclusion part - a global valuation is tailedup.**

Keywords: Big Data, Hadoop, HDFS, Map Reduce, Pig, Hive, Hbase, Sqoop.

1. INTRODUCTION:

When we explore the term big data in the literature we come across the various definitions related to it. Some of them are mentioned below:

Big Data is a term in which the enormous varied data is generated rapidly, this data is captured, stored and then it is distributed. After distribution, the data is managed for analysis of productive and useful information [1].

Big Data is the amount of data beyond the ability of technology to store, manage and process efficiently [2].

Hashem et.al.defines Big Data after reading many kinds of literature - He collaborated the different definitions to explain what big data is? He mentions that set of methods and technologies have been introduced where there is an integration of new forms to unwind varied high volume and complex form of a hidden value datasets.[3].

Big Data Technologies are new generation technologies and architectures which were designed to extract value from multivariate high volume data sets efficiently by providing high speed capturing, discovering and analyzing[4].

As per the given definitions, we can say that big data is the technology which works on the massive, varied and rapidly generated data that is processed and transformed into the efficientinformation.

1.1 A feature of Big Data:

As we go across the various kinds of literature the most generally talked characteristic of big data is 5 V"s the 3 V which are common to all are volume, variety, and velocity. Where the veracity and value are other two which are not defined in every study. The figure1 below shows 5 V in big data.

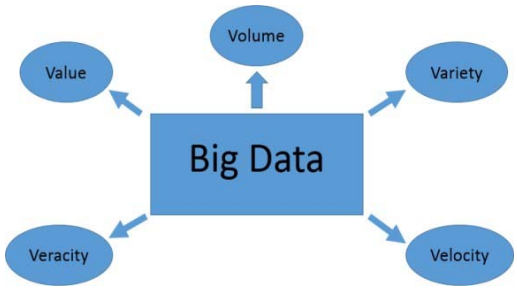

Fig1.Features of Big Data.

These 5 characteristics are described as follows [4][5][6][7].

Volume: The present world is becoming a digital space as the huge amount of data set is generated from the organizations like e.g. social sites, bank transactions etc.

Variety: Innumerable data come to the organization through various means (internal or external). Due to this various type of data sets is produced.

Velocity: It is measured by the rate of data produced per second i.e. how frequently and rapidly the data is being produced e.g. number of transaction done in one minute, therefore a number of data produced are directly proportional to time.

Veracity: Veracity is totally about the accuracy of the data. It should be securely received from the exact resources. The only authenticated person should have the access permission.

Value: The knowledgeable result is produced after all the processing of data is done.

1.2 Classification of the BigData:

The characteristic of big data can be understood better by dividing it into classes. These classes are Data Sources, Content Format, Data Stores, Data Staging and Data Processing[1][3].

1.2.1 Data Sources: The tremendously produced data is collected from the varied organizations, such as web and social sites, machines, sensing, transactions, and IoT. All these resources produce the data at alarming rate e.g. in every 60 seconds 100,000+ tweets are made on Twitter, 695,000+ status updates are made on Facebook, 217+ new mobile user, 168,000,000+ mails are done. Thus millions of people‟s data is being generated every second. This rapidly generated data is being stored in some or the other format in the form of Data Sources

1.2.2 Content Format: As the different type of data is being produced from various sources so, their format structure will also differ from each other, further these content formats are classified into Structured, Semi-Structured and Unstructured type. The data which falls under the structured category is one whose schema or structure is known which is basically in the tabular format. Semi-Structured data is one where the schema is not defined properly like XML, JSON, CSV, TSV, E-mail etc. Unstructured data is one who does not have any specific format like log files, video files, audio files, and images.

1.2.3 Data Stores: These are the stores where the data generated from various sources is collected and stored. The data is stored in some categories like Document-oriented, Column-oriented, Graph-based and Key-value[1].

1.2.4 Data Staging: This is the process in which three major steps - Cleaning, Normalization and Transformation. In cleaning the data which is being stored in the data stores is taken and cleansing of the data is done, that is, unwanted and useless data is being removed from the stored data. Further, this cleaned data goes under the normalization process in which data is rearranged in the database so that there should be no redundancy and inconsistency in the database. After the normalization process data is transformed, which means, source data format is transformed into the human understandable format or the format in which it can be easily used for analysis.

1.2.5 Data Processing: In this, the tremendously large and efficient data is being processed by using the Batch and Real-time data processing system. In the batch processing system the data is gathered, processed and output is produced (Hadoop uses batch processing) but in this separate algorithms or programs are mandatory for input, process, and output e.g. billing system. Where in Real-Time Data Processing System there is no necessity for creating the distinct programs, the continual processing of the input, process, and output is done. Data should be processed in minimum time e.g. radar systems, bank ATM's.

2. HADOOP:

Hadoop (Highly Archived Distributed Object Oriented Programming) Hadoop actually came into the picture in 2005 which was the result of Doug Cutting plus Mike Cafarellapractice. Hadoop name was baptized by Doug Cutting as it was the name of his lad's toy elephant. Initially, hadoop was formed to backing distribution designed for Nutch, which is a search engine venture. It's an open-source software which facilitates dependable along with is it scalable and also provides distributed computing on the groups of economical servers[9]. This software knobs huge amount of different data from different resources like Images, Videos, Audios, Folders, Files, Software, Sensor Records, Communication data, Structured Query, Unstructured Data, E-mail & conversations, and any kind of different format[10]. Hadoop comprises of many components like Flume, HBase, Hive, Lucene, Pig, Oozie, Sqoop, Zookeeper, Avro, Chukwa. Hadoop is also a package that offers Documentation, source code, location awareness, Work scheduling etc. As hadoop work on the master and slave architecture so, its master node comprises of Data Node, Name Node, Job Tracker and Task Tracker whereas the slave node performances as a Task Tracker and Data Node. Slave Node is only responsible for computation of data and is referred to as worker node. The Job Tracker copes up with the Job Scheduling. The Hadoop comprises of two parts namely Hadoop Distributed File system (HDFS) and Map Reduce[11].

The basic and simplest architecture of Hadoop is shown in the figure.

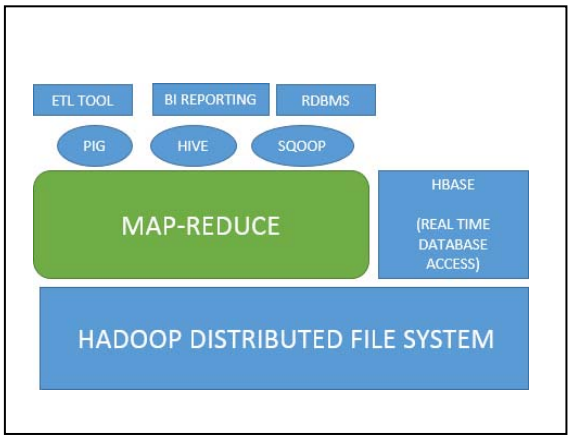

Figure 2: Hadoop Architecture

- **Components of Hadoop:**

2.1. HDFS (Hadoop Distributed File System): The HDFS is a special file system (like) other operating system. It is basically a framework which is written in Java. It consists of Name node and a group of Data nodes. As HDFS works on the master-slave architecture, therefore, name-node acts as a master and stores all the information and data node act as a slave node. The Name Node comprises information like Meta Data, Attributes, File Location, Job Tracker, Task Tracker, Active Node and Passive Node, Free Space and data they store, replication of data etc. It keeps a record of all the metadata, attribute, the location of files and data block into the Data node[13]. Attribute looks after the features which are file authorizations, file alteration, its access time and namespace. These file and directories arranged in the form of hierarchy. The mapping of namespace hierarchy and file block is done in the Data node. This work is done by the name node. If a client reads a file in HDFS then the master node (i.e. name node) is contacted first with respect to collect the position of the data blocks related to required files. Name node always stocks the doppelgänger and periodical logs of a system, keeps them up to date and it always keeps in the knowledge that where the data blocks, replicas of the files and free blocks are in the system. Data node is a slave node saves data and comprises of a task tracker which tracks the active work of a data node plus it tracks job impending of the name node [14]. Each and every block report track is kept by the data node. The data node hourly sends its blocks activity log to the master node or name node so that master node should know in which data node the block replica is kept in the cluster and remain sup to date. When a usual operation takes place in HDFS, in every ten seconds the data node sends its heartbeat to the name node to specify its availability and operating accuracy. If within ten minutes heartbeat is not received by the name node it starts producing replicas on the other nodes of the lost blocks of a data node. It receives thousands of heartbeats every second but it does not affect the other name nodes [13].

2.2 MR (MapReduce): MapReduce is software program whose agenda is to practice distributed processing on huge data sets of the clusters in the computer. It is proposed by Google. MapReduce accelerates plus simplifies, huge amount data handling in a parallel fashion. It correspondingly works on huge groups of service hardware's, in a dependable and fault-tolerant manner [15]. Hadoop map reduces engine counts on the map-reduce algorithm is used to allot work all over the cluster. The MapReduce and HDFS layers are incorporated in typical hadoop cluster. The job or the work tracker in MapReduce layer allocates work to the task tracker whereas job tracker of Master node likewise allocates work to the slave node of the task tracker. MapReduce follows the Master-Slave architecture. Therefore the Master node comprises of Job tracker node and Task tracker node in MapReduce layer, where the Name node and Data node in HFDS layer. Similarly, the Slave node comprises of Task tracker node and Data node in HFDS layer. Where MapReduce layer comprises of both job and task tracker nodes, therefore in HFDS layer name and data nodes are found[13].

Solo Job Tracker for each master controls the scheduling activity of jobs components tasks upon the slave, observers slave advancement and again executes the botched task.On the other hand, the single Task Tracker per slave executes according to the masters given directions. Map Reduce basic functionalities are grounded in the Map phase and in the reduce phase too. Java language is generally used to write the code however it is not imperative though any another language can be used to inscribe code with Hadoop Streaming API[13].

MapReduce is divided into two parts i.e. Map and Reduce:

2.2.1 Map step: In this Map stage the giant problem is taken as the input by the master node after that the problem is divided into subproblems and these subproblems are further allocated to the worker nodes. The worker node possibly will divide the subproblems which form a multi-level tree-like structure. Data node works on subproblem and later arrows it back to the master. The transformation of a map is done to convert a row of input data key plus the value of it into the output key as well as value:

a. M(K1,V1) -> L<K2,V2> therefore, M stands for a map where L stands for list, the "K" stands for key and "V" stands for value. In this expression, for an input the list comprising zero or further pairs of (K, V) isyielded**:**

i. Input key and the output key can bedifferent.

ii. With a key, multiple entries can be made by the output.

2.2.2.Reduce Step: In this reduce stage answers of the subproblems are collected by the master node and then these sub-answers are integrated in a predefined manner in such a way that the answers obtained are for the problem specified initially. The reduce transmute takes the entire values of the definite key and produce a new reduced output list.

i. Reduce (K2, list<V2>) -> list<V3>.

ii. "Map" Step: Every worker node practice the "map()" function to the local data, and writes the output to temporary storage. A master node executes only one input copy out of the redundant input data copies.

iii. "Shuffle" Step: Worker nodes reallocate data which is based on the output keys (by using "map ()"), such that all data belonging to one key is positioned under the same workernode.

iv. Simultaneously the Worker nodes process every group of output data, perkey.

2.3 PIG: Yahoo! is the one who initially developed Pig, to let persons using hadoop concentrate largely on analysis of huge dataset rather than wasting time on writing mapper and reducer programs. Alike Pigs eats nearly everything, the Pig programming language is designed to knob any kind of data. Two major components which are imperative for building pig were first is Pig Latin(language) and the runtime environment. The runtime environment provides the platform for the execution of the pig Latin programs [16]. Steps in Pig programming language is asfollows:

1. Pig program LOAD the data which is to be operated fromHDFS.

2. In this step, data undergoes the set of transformation (i.e. converted to a fixed mapper and reducer task).

3. Lastly, data is DUMP to the screen or the result is STORE in afile.

2.3.1 LOAD: The basic feature of Hadoop is that objects on which work is done are stored in HDFS.

So, the data to be accessed by the pig program is the first Load from the HDFS on orders of the program. Which articulate Pig files (or what file) it"s going to use, and this is completed thru command i.e. LOAD 'data file'. (Therefore „data_ file „state is either directory or HDFS file). In case the directory gets selected then every file of the directory is placed inside the program. If it's not in pig accessible format then the USING function can be added to LOAD statement which interprets the read in data into user-defined function[16].

2.3.2 TRANSFORM: In this, the interpreted or manipulated activity is performed on the data like FILTER out unwanted data, JOIN two data file sets, GROUP data, ORDER result etc.[13].

2.3.3 DUMP and STORE: If DUMP or STORE command is not stated then results of a Pig program are not produced. The DUMP command is responsible for sending output to the screen while debugging is practiced on Pig programs. When a DUMP call is a converse into the STORE call the outcomes from running programs are deposited in a file for advance succession or examination. The DUMP command can be used at any place in the program which dumps intermediary outcome sets to display, which helps in correcting bugs.

2.4 **HIVE:** In hadoop Hive is a data warehouse infrastructure which shows data summary, query, and analysis. The hive was introduced by Facebook and later Apache bought Hive. Hive is used by Netflix, Amazon etc[13].

Apache Hive rations analysis of huge datasets stored in HDFS. It can also analysis similar in temperament file systems (e.g. Amazon S3)[17]. Hive comprises of HiveQL language which is similar to the SQL language with schema on read and clearly translates queries to map or reduce. To quicken queries, it offers indexes plus bitmap indexes. By default, Hive„s metadata in is stored in embedded Apache Derby database, and other client-server databases like MySQL can optionally be used [17]. Currently, Hive support four file formats which are TEXTFILE, SEQUENCEFILE, ORC, and RCFILE. Hive features Indexing to offer acceleration, index type including compaction and Bitmap index as of 0.10, other index types are also intended. It also consists of diverse storage natures example RCFile, ORC, HBase, plain text etc. As in RDBMS metadata storing is their due to which time relaxation during query check is there. Algorithms like gzip, bzip2, snappy, etc. are used to operate on compressed data in hadoop storage. HiveQL queries are indirectly transformed into MR jobs. As grounded on SQL, the HiveQL doesn't severely shadow the complete SQL-92 standard. Also, HiveQL compromises of extensions which SQL don't consist, it can insert multi-tables plus generate the table as select. Though it merely provides simple provision used in indexes [18]. Hive shell is command line interface in which hive quires can run. HQL understands restricted commands though it is quite useful. Hive service fragments the HQL statements into MapReduce jobs and executes it

through a Hadoop cluster.

2.5 **HBASE:** Hbase is ACME Apache project. Apache HBase was initiated as a project by the corporation. It is authorized to work on gigantic data for the commitments of natural language search. Facebook in November 2010 used HBase to bring in practice its new messaging platform. Hbase database is managed in column-oriented formats and its base is HDFS on which it runs. It works best for light data sets, commonly cast in numerous big data use cases. Structured query language (like SQL) cannot be practiced by HBase. HBase is never considered as a relational data store. It uses Java to write an application somewhat similar to MapReduce application. It supports Avro, REST, and Thrift platform for writing applications[13].

The HBase structure consists of tables sets. Every single table comprises of rows plus columns, alike old-fashioned database. Every single table should consist of the element such as Primary Key, it is used by each and every access shots to Hbase tables. The column in HBase signifies the object quality[13][19].Sample, a table consists of rows and columns in which rows keeps track of the diagnostic logs from the server, where each row will be log record and column comprises of the time stamp of log or possibly the name of log generating the resource. Hbase permits to group various attributes into column families so that elements of a column family are entirely stored together. Hbase is dissimilar from a row-oriented relational database, in which rows all columns are stored together. With help of HBase one can predefine schema for a table and column families are specified. The fresh columns can be introduced at whichever point of time in the families this adds flexibility to the schema which becomes accustomed to the altering requirements of an application. In comparison to the MapReducethe HDFS comprises of Name Node and data nodes(slave nodes), whereas MapReduce comprises Job Tracker and Task Tracker, similarly, in HBase the master node manages the cluster and region servers store portions of the tables and perform the work on the data[19].Hbase is delicate towards the loss of its master node.

2.6. SQOOP:

It's a tool in Hadoop developed for effective export or import of data in a huge amount between Hadoop and structured data stores. Currently, Sqoop is epitome apache project

command line interface application in java. The Sqoop is planned professionally for the cause of transporting a massive quantity of data between hadoop and structured data stores such as relational. It also replicates data rapidly from external systems to hadoop. It permits data imports from external data stores and enterprise data warehouses into hadoop. Fast performance of Sqoop is due to parallelizing data transfer and optimal system utilization. Sqoop also provisions analyses of data proficiently. It even all deviates extreme heaps to external systems [20].

Sqoop runs in hadoop cluster. It has right of entry to hadoop core. The mappers are used to cut the incoming data. Sqoop will establish a connection with the database store for obtaining data called meta-data from the relational data store, for instigating the java class meta-data is used by the Sqoop. The source from which metadata is obtained is database store. Internally a JAVA class is formed by Sqoop using JDPC API. The compilation of the java class is done using JDK and it compares the jar files. Sqoop again establishes connection with database store, the jar files are formed in a direction to discover the split column which facilitates Sqoop to fetch data from the database. Lastly, Sqoop will put the recovered data into HDFS[20].

3. CONCLUSION:

The above study provides knowledge about the big data, its characteristics, features, and classifications. This paper possesses the basic and technical information about hadoop and its architecture. It elaborates the components of hadoop like HDFS (hadoop distributed file system), Map-Reduce, Pig, Hive, Hbase and Sqoop. This paper mentions the feature and working of the hadoop components.

References:

[1] Hakan Özkösea, Emin Srtac Ari, Cevriye Gencerb, World Conference on Technology, Innovation and Entrepreneurship Yesterday, Today and Tomorrow of Big Data, Procedia - Social and Behavioral Sciences 195 (2015) 1042 –1050, ELSEVIER.

[2] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

[3] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems,47, 98-115.

[4] Gantz, J., &Reinsel, D. (2011). Extracting value from chaos. IDC review, (1142),9-10.

[5] Gartner IT Glossary, What is Big Data?, URL: http://www.gartner.com/it-glossary/big

[6] Elragal, A. (2014). ERP and Big Data: The Inept Couple. ProcediaTechnology,16,242-249

[7] López, V., del Río, S., Benítez, J. M., & Herrera, F. (2015 Cost-sensitive linguistic fuzzy rule-based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets and Systems, 258, 5-38.

[8] Fadiya, S. O., Saydam, S., &Zira, V. V. (2014). Advancing big data for humanitarian needs. Procedia Engineering,78,88-95.

[9] Dhole Poonam B, GunjalBaisa L, "Survey Paper on Traditional Hadoop and Pipelined Map Reduce",

International Journal of Computational Engineering Research Vol 03, Issue12

[10]"Leveraging Massively Parallel Processing in an Oracle Environment for Big Data", An Oracle White Paper, November2010.

[11]Jeffrey Dean and Sanjay Ghemawat, "Map Reduce: Simplified Data Processing on Large Clusters", Google, Inc.

[12] B. Saraladevia, N. Pazhanirajaa, P. VicterPaula,M.S. Saleem Bashab, P. Dhavachelvanc,"Big Data and Hadoop-A Study in Security Perspective" ,2nd International Symposium on Big Data and Cloud Computing (ISBCC,15), ELSEVIER.

[13] Ms. VibhavariChavan, Prof. Rajesh. N. Phursule," Survey Paper On Big Data, "(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014,7932-7939".

[14]Suman Arora, Dr. Madhu Goel, "Survey Paper on Scheduling in Hadoop", International Journal of Advance Research in Computer Science and Software Engineering Volume 4, Issue 5, May2014.

[15] Wang, F. et al.,"Hadoop High Availability through Metadata Replication".ACM (2009).

[16] Apache Pig. Attained fromhttp://pig.apache.org.

[17] Apache Hive. Attained fromhttp://hive.apache.org.

[18] Vishal S Patil, Pravin D. Soni, "Hadoop Skeleton & Fault Tolerance in HadoopClusres", International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 2, Issue 2, February 2013 ISSN 2319 – 4847

[19] Apache HBase. Attained from http://hbase.apache.org

[20] Mr. S. S. Aravinth , Ms. A. HaseenahBegam , Ms. S. Shanmugapriyaa, Ms. S. Sowmya , Mr. E. Arun ,"An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing", IJIRST –International Journal for Innovative Research in Science & Technology| Volume1 Issue 10 , March 2015 ISSN (online): 2349-6010.