A Review Paper on Big Data and Hadoop Technology

Mrs Pooja
Asst. Prof. in Computer Sc.
S.D. College, (Hoshiarpur).

Abstract: ***The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyse petabyte- or larger-sized datasets with high-velocity and different structures. The word 'Big Data' designates advanced methods and tools to capture, store, distribute, manage and investigate petabyte or larger sized datasets with high velocity and different arrangements. Big data can be organized, unstructured or semi organized, resulting in incapability of predictable data management methods. Put another way, big data is the realization of greater business intelligence by storing, processing, and analysing data that was previously ignored due to the limitations of traditional data management technologies. Hadoop is the main podium for organizing Big Data, and cracks the tricky of creating it convenient for analytics determinations. Hadoop is an open source software project that allows the distributed handling of large datasets across bunches of service servers.***

It is considered to scale up from a single server to thousands of technologies, with a very high degree of fault tolerance. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.

Keywords -Big Data, Hadoop, Map Reduce, HDFS, Hadoop Components

---

A. Big Data: Definition

Big data and analysis are at the center of modern science and business. This data transactions online, e-mails, videos, audio, images, click streams, logs, posts, search queries, health records, social networking, communicating science data, sensors and mobile phones and their applications are created. They are stored in the database and the massive increase, the farm, store, manage, share, analyse and typical database software tools are difficult to see through. Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000gigabytes per terabyte) to multiple petabytes (1015 or 1000terabytes per petabyte) as big data.

B. 5 Vs of Big Data

Volume of data: Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes. Data is ever-growing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.

Variety of data: Data sources are extremely heterogeneous. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way. Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio,video,XMLetc.

Velocity of data: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value. The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organisations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.

Value:

It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

Veracity:

The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

C:Problem with Big Data Processing

i. Heterogeneity an Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In

consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured

ii. Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

iii.Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster.However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge.

iv. Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot bedone. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linkingof data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must beaddressed jointly from both perspectives to realize the promise of big data.

v. Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding.

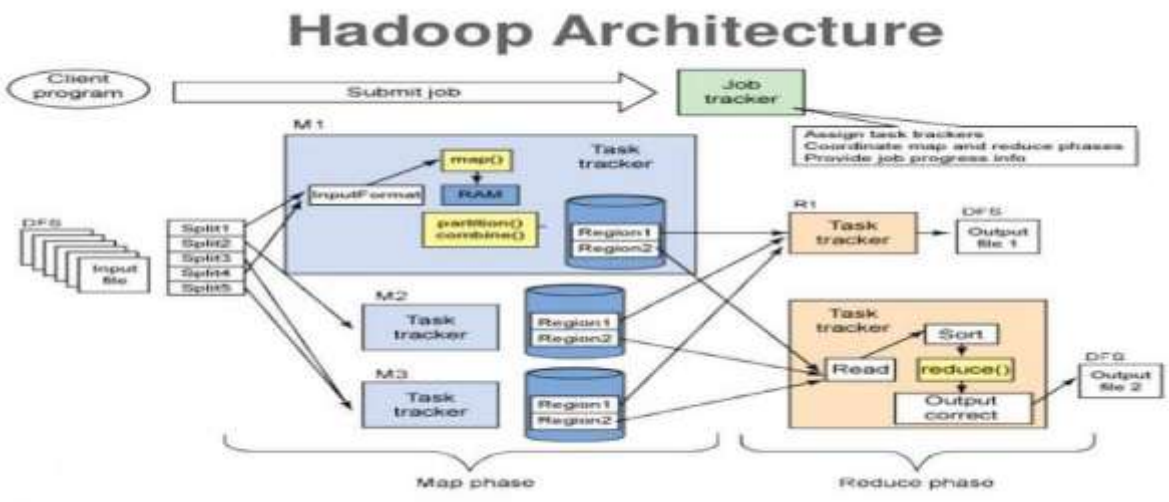
Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

2. Hadoop: Solution for Big Data Processing

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReducethat is a software framework where an application break downinto various parts. The Current Appache Hadoop ecosystemconsists of the Hadoop Kernel, MapReduce, HDFS andnumbers of various components like Apache Hive, Base andZookeeper. HDFS and MapReduce are explained in following points.

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It distributes thefile among the nodes and allows to system continue work in case of a node failure. This approach reduces the risk ofcatastrophic system failure.In which application is broken into smaller parts (fragments or blocks).Apache Hadoop consists of the Hadoopkernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive.Hadoop Distributed File System consists of three Components: the Name Node, Secondary Name Node and Data Node.The multilevel secure (MLS) environmental problems of Hadoop by using security enhanced Linux (SE Linux) protocol.

In which multiple sources of Hadoop applications run at different levels.This protocol is an extension of Hadoop distributed file system. Hadoop is commonly used for distributed batchindex building; it is desirable to optimize the index capability in near real time. Hadoop provides components for storageand analysis for large scale processing. Now a day's Hadoop used by hundreds of companies.The advantage of Hadoop is Distributed storage & Computational capabilities, extremely scalable, ptimized for highthroughput, large block sizes, tolerant of software and hardware failure.



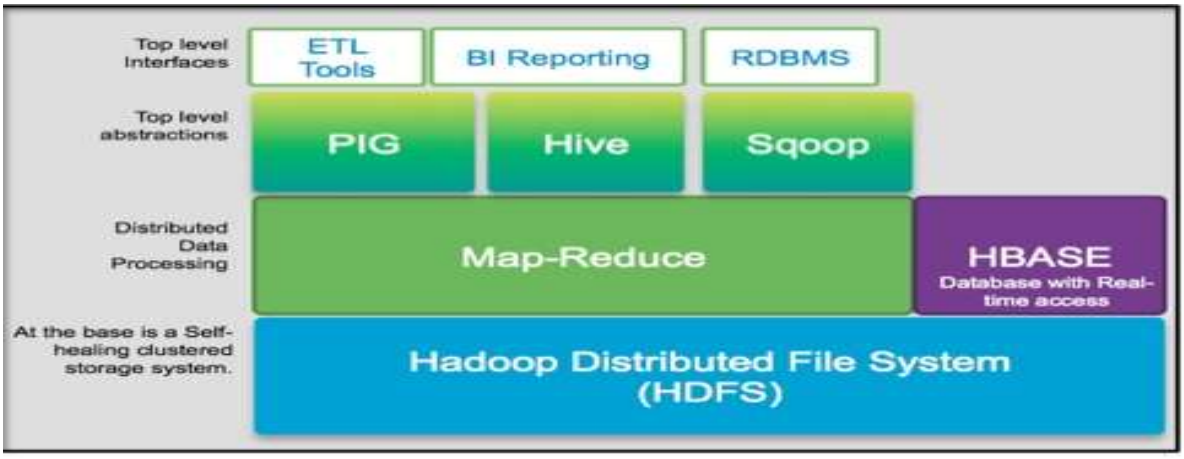


Fig: Components of Hadoop

Components of Hadoop:

HBase: It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well-mannered structure.

Oozie: Oozie is a web-application that runs in ajava servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. It manages the Hadoop jobs in a mannered way.

Sqoop: Sqoop is a command-line interface application that provides platform which is used for converting data fromrelational databases and Hadoop or vice versa.

Avro: It is a system that provides functionality ofdata serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

Chukwa: Chukwa is a framework that is used fordata collection and analysis to process and analyze the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.

Pig: Pig is high-level platform where the MapReduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language.

Zookeeper: It is a centralization based service that provides distributed synchronization and provides group services along with maintenance of the configuration information and records.

Hive: It is application developed for dataware house that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open- source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers. Hadoop is:

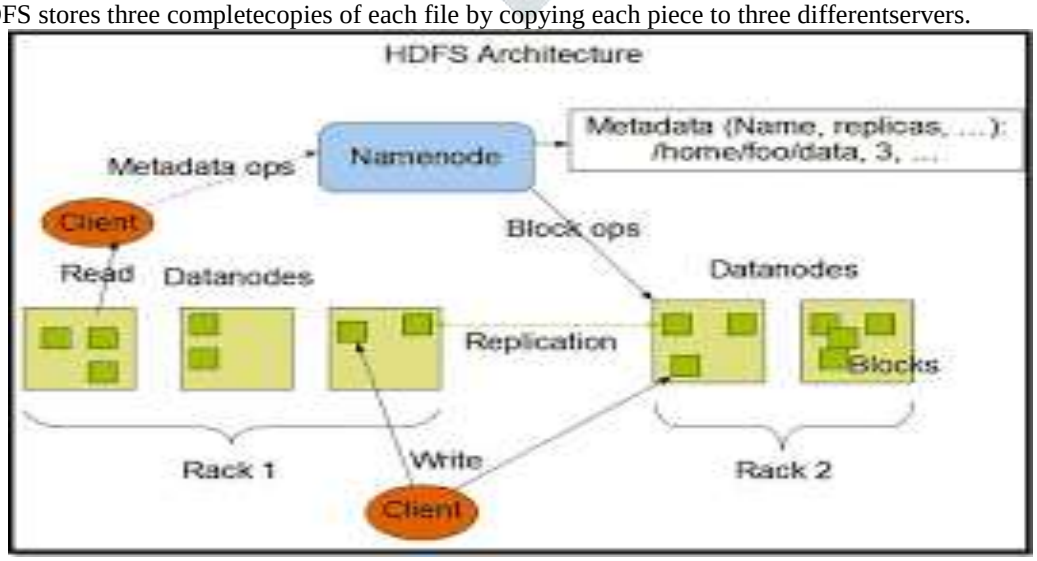
Reliable: The software is fault tolerant, it expectsand handles hardware and software failures

Scalable: Designed for massive scale ofprocessors, memory, and local attached storage

Distributed: Handles replication. Offers massivelyparallel programming model, MapReduce.

A)HDFS Architecture

Hadoop includes a fault-tolerant storage system called theHadoop Distributed File System, or HDFS. HDFS is able tostore huge amounts of information, scale up incrementally andsurvive the failure of significant parts of the storageinfrastructure without losing data. Hadoop creates *clusters* ofmachines and coordinates work among them. Clusters can bebuilt with inexpensive computers. If one fails, Hadoopcontinues to operate the cluster without losing data orinterrupting work, by shifting work to the remaining machinesin the cluster. HDFS manages storage on the cluster bybreaking incoming files into pieces, called “blocks,” andstoring each of the blocks redundantly across the pool ofservers. In the common case, HDFS stores three completecopies of each file by copying each piece to three differentservers.



B. MapReduceArchitecture

The processing pillar in the Hadoop ecosystem is theMapReduce framework. The framework allowsthespecification of an operation to be applied to a huge data set,divide the problem and data, and run it in parallel. From ananalyst's point of view, this can occur on multiple dimensions.For example, a very large dataset can be reduced into a smallersubset where analytics can be applied. In a traditional data warehousing scenario, this might entail applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse. There are two functions in MapReduce as follows:

map – the function takes key/value pairs as input and generates an intermediate set of **key/value pairs reduce** – the function which merges all the intermediate values associated with the same intermediate key

Conclusion

we have entered an era of Big Data. The paper describes the concept of Big Data along with Volume, Velocity and variety, value, veracity of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.

REFERENCES

[1] H. S. Bhosale and P. D. P. Gadekar, "A Review Paper on Big Data and Hadoop," vol. 4, no. 10, pp. 1–7, 2014.
[2] Y. Demchenko, C. Ngo, and P. Membrey, "Architecture Framework and Components for the Big Data Ecosystem,"2013.
[3] W. Fan and A. Bifet, "Mining Big Data : Current Status , and Forecast to the Future," vol. 14, no. 2, pp. 1–5.
[4] S. Sagiroglu and D. Sinanc, "Big Data : A Review," pp. 42–47,2013.
[5] E. Sivaraman and R. Manickachezian, "High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing Using Hadoop," 2014 International Conference on Intelligent Computing Applications, pp. 32–36, Mar. 2014.
[6] C. Kaewkasi, "A Study of Big Data Processing Constraints on a Low-Power Hadoop Cluster," 2014.
[7] Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24.
[8] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―"Survey Paper On Big data "International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014