International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 01 | Jan-2018 www.irjet.net p-ISSN: 2395-0072

BIG DATA ANALYTICS USING HADOOP TECHNOLOGY

K.Tamilselvi[1], V.Sumithra[2], Mrs.K.Dhanapriyadharsini[3]

[1,2,3] *Dept.of BCA & M.Sc. SS, Sri Krishna Arts and Science College, Tamil Nadu, Coimbatore*

ABSTRACT: *Big data is a technique used to store, distribute and the datasets which are large sized are analyzed with high velocity. Big data can be taken in the form of structured, unstructured or semi-structured which results in incapability of conventional data management methods. Different sources and the system at various rates are used to generate the data's. The term "parallelism" is used for large amount of data which is inexpensive. The paper evaluates the difference in the challenges that is faced by a small organization as compared to a medium or large scale operation and therefore the differences in their approach. The strategies that vary from industries and the products are the examples for implementing BIG DATA. HADOOP is an open source technology that enables the distributing process of large data sets of fault tolerance with a very high degree. HADOOP is the popular tool for implementing BIG DATA. The paper deals with the technology aspects of BIG DATA for its implementation in organizations and the structure of HADOOP with the details of various components.*

KEYWORDS: Big data, Hadoop, map reduce, HDFS, Hadoop components, analytic databases, analytic applications.

1. INTRODUCTION

Big data is the term which is addressed by the data sets so that they are so large or complex data processing application software which is inadequate to deal with them. Big data challenges includes capturing of data, storing the data , analyzing the data , searching, sharing, transferring, visualizing, querying, updating and information privacy. Big data also refers to the use of predictive and user behavior analytics that extract values to data and to a particular size of data set. There enhances a little Doubt about the quantities of data that are available indeed large, but it's not the most relevant characteristic of the new ecosystem. To find the new correlations in spotting business trends, prevent diseases we use the data sets to be analyzed. Data sets grow rapidly when they have increase gathering by cheap and numerous information-sensing internet of things devices like mobile devices , software logs, microphones, cameras, identifying radio frequency readers and wireless sensor networks. There is a rapid increase in the use of mobile phones due to the large amount of data that is being generated for every second and also it is impossible to handle traditional methods. Big data also requires a set of techniques and some of the technologies with new forms of integration to reveal the insights from datasets that are diverse, complex and a massive scale. The term Gartner refers to the data growth challenges for being three dimensional that is used for increasing volume, velocity and variety.

CHARACTERISTICS:

Big data can be described by the following characteristics:

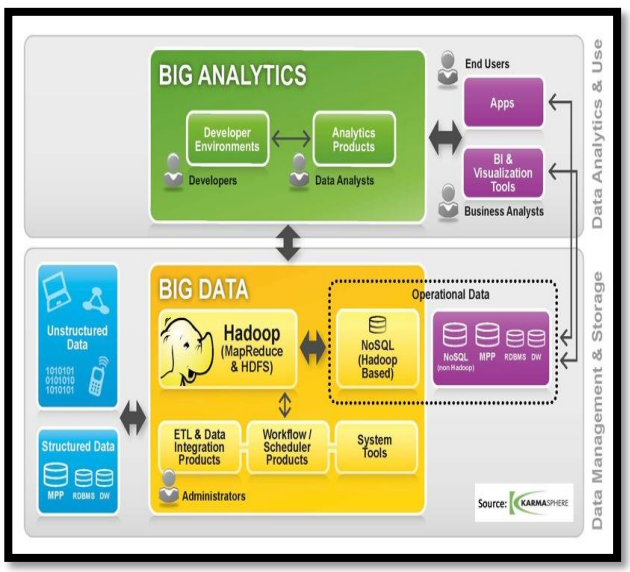
VOLUME: Big data just tracks and observes what is happening. In volume quantity is generated and stores the data. Determining the value and potential insight and whether it can actually be considered big data or not is done by the size.

VELOCITY: In real time big data is often available. Velocity is defined for the speed at which the data is being generated and processed to meet the demands and challenges which lies in the path of growth and development.

VARIETY: Big data draws from the text, audio, images, videos and through data fusion it completes missing pieces.

Variety is defined as the inconsistency of data set in which hamper processes to handle and manage.

1.1 ARCHITECTURE OF BIG DATA



Big data repositories have existed in many forms with a special need which is often built by corporations. In beginning parallel database management system is offered for commercial vendors. Terabyte data is stored and analysed by the Teradata systems. The relational database management system is based on Teradata.
Data storage and query are used for distributed file sharing framework. The relational database management system stores and distributes structured, unstructured or else semi-structured data in which it across multiple servers. ECL is an application schema on read method which the structure of stored data is queried instead of when it is stored. High speed parallel processing platform is used to acquire the queries. The MAP REDUCE concept uses the similar architecture. The map reduce concept provides a parallel processing model, and released the huge amounts of data which is implemented. Queries are being spited and distributed across the parallel nodes and processed parallel with the help of map step. Then the results are being gathered and delivered by the reduce step. To replicate the algorithm, framework is very helpful. An apache open source is used for implementing map reduce framework. Big data solution is a term which is further used for identifying the big data implications and another feature called MIKE is an open approach for information management. The useful permutations of data sources and complexity in interrelationships are used for handling big data and also for finding difficulty in deleting individual records. To address the issue we use a multiple layered architecture as an option. By using distributed parallel architecture we can access the data in multiple servers. These parallel execution environments can improve data processing speed dramatically. A parallel database management system is inserted into the data which is used for implementing the use of map reduce and Hadoop frameworks. The Hadoop framework is to make the processing power looks transparent to the end user by using front end application server. The 5C architecture (configuration, connection, conversion, cyber, cognition) is for manufacturing the big data analytics application. The data lake is a concept used for organizing the focus from centralized control to shared model used to respond the dynamics changing of information management.

1.2 TECHNOLOGIES USED IN BIG DATA

Business intelligence and cloud computing are the techniques used for big data. Visualization can be represented in the form of charts and graphs. Tensor is a term used for representing multidimensional big data which can be handled efficiently by tensor based computations. Multi linear subspace learning is a technique used for handling tensor based computation. In addition parallel processing databases like search based application, data mining, distributed file systems, cloud based infrastructure and HPC based infrastructure are being applied. Even though some technologies are used and developed still it remains the difficulties to carry the machine learning with big data. To use and optimize the large tables in relational database management system it should have the ability of loading, monitoring, back up by using implicit function. The fundamental structure of massive data sets is been programmed by topological data analysis. The big data analytics is processed for slower shared storage which prefers for direct attached storage in different forms to have high capacity buried inside the parallel processing nodes from solid state drive. There are two types of storage architecture. They are Storage area network (SAN) and Network attached storage (NAS).The two types of storage architecture are relatively slow, complex and expensive. The qualities are not consistent with big data analytics systems which are for system performance, low cost and commodity infrastructure. The big data analytics can be characterized by real or near real time information delivery. Whenever and wherever needed it is possible to avoid latency.

1.3 PROBLEMS IN BIG DATA PARADIGM

The major problem in big data is that we do not know about the under lying empirical micro processes which leads to emergence of typical network characteristics. Future development can predict through some of the algorithms that is fed by large number of data based on past experiences. To make predictions in the changing environment it requires the theory of systems dynamic. Big data approaches with computer simulations are complex systems and agent-based models. Agent-based models are used for predicting the outcomes of social complexities of unknown future scenarios through computer simulations that are based on collection of mutually interdependent algorithms. Factor analysis and cluster analysis are the two multivariate methods that probe the latent structure of data. The analysis of smaller data sets is compared to big data. The challenges is to extract, load part of data processing in big data projects seems to be difficult because there is no large data analysis. The bias problem cannot be solved by adding more data but the other sources such as twitter and Google translate are not represented by overall population and results from sources that may lead to wrong conclusions. The results may be skewed dramatically. Multiple comparison problems are the new term introduced by bi data. These problems are used for testing a large set of hypothesis that is likely to produce many false results that appear mistakenly significant.

1.4 APPLICATIONS

Usage of big data in government

To work in collaboration and to create new and innovative Process the often requirement is data

analysis. In terms of cost, productivity, and innovation the big data is used and adopted within the governmental processes.

Usage of big data in the manufacturing field

In manufacturing industry big data provides an infrastructure for transparency and performs inconsistency and availability.

To begin the data acquisition a conceptual framework is used. Acoustics, vibration, pressure, current, voltage and controller data are the different types of sensory data to predict the manufacture. To construct the big data vast amount of sensory data is needed.

Usage of big data in healthcare

Big data analytics helps the healthcare to improve personalized medicine and prescriptive analytics, clinical risk intervention, waste and care variability reduction, automated external and internal reporting of patient data. Healthcare systems are not trivial and generated in level of data. Mhealth, ehealth and wearable technologies are the technologies used for increasing volume of data. The volume of data includes electronic health record, imaging data, patient generated data, sensor data, and other forms which feel difficult to process the data.

1.6 IMPORTANCE OF BIG DATA ANALTICS

To help the organizations harness their data and to identify the new opportunities big data is being effectively used. Big data analytics leads to move the smart business moves, high profits and to satisfy the customer. To examine the large amount of data's that is uncovered by hidden pattern, correlations and other insights there enhances the term big data analytics. In today's technology the data can be analyzed and we can get the answers immediately. In traditional business the effort made is slower and less efficient.

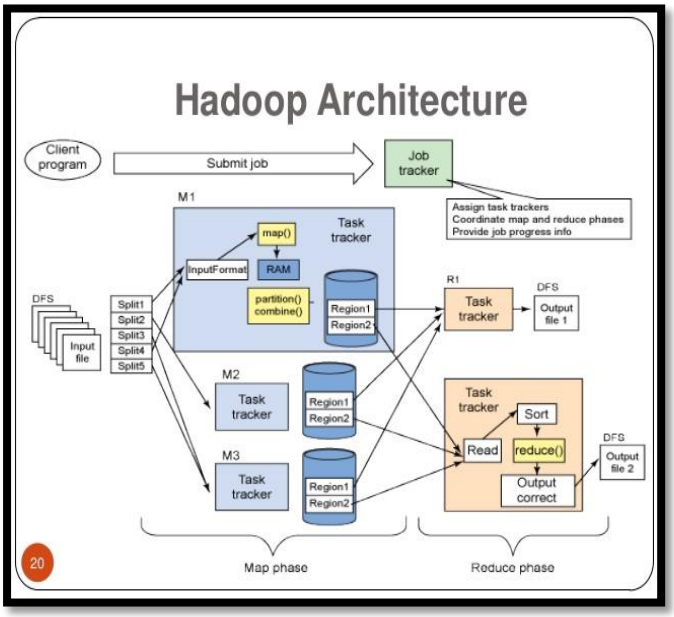
2. HADOOP TECHNOLOGY

INTRODUCTION

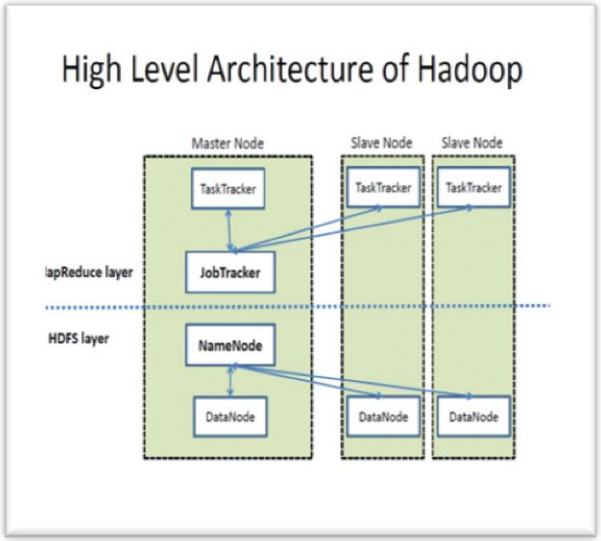
Hadoop is an opens source software framework which is used for distributed storage and it can process the datasets of big data by using Map reduce programming model. The model consists of computer clusters which are being built from commodity hardware. In Hadoop the modules are designed with fundamental assumption in which the hardware fails with common occurrence and it is automatically handled by framework. The Hadoop distributed file system consists of a storage part in the core of apache Hadoop. Hadoop is spited up into large blocks and among clusters it is distributed. The code which is packaged is transferred into nodes to process the data parallel. The nodes are manipulated to access the data which takes data locality as the advantage. The process becomes faster and efficient when it uses conventional supercomputer architecture where computation and data are distributed through high speed networking. Map reduce programming model is an implementation process which is used for large-scale data processing.

2.1 HADOOP ARCHITECTURE

The package in Hadoop contains the java archive files and some of the scripts which help to start the Hadoop. The package provides file system and operating system level abstractions and a Mapreduce engine and a Hadoop distributed file system.



To work with effective scheduling every Hadoop file system provides awareness on location and name of the rack where the worker node exist. Hadoop application uses this information which is provided by Hadoop file system to execute the code on the node where the data is and fails when it uses same rack or switch that reduces backbone traffic. This method is also used in Hadoop distributed file system to replicate the data across multiple racks. To reduce the impact of rack power outage or switch failure this approach is used. A single master and multiple worker nodes have been included in the small Hadoop cluster. In master node job tracker, task tracker, data and name node has been included. In data node and task tracker, it is possible to have only data and can compute inly worker nodes. These nodes are usually used in non-standard applications. The most requirements for Hadoop are java runtime environment. The term Secure Shell (SSH) can be set up between two nodes which will be used for standard start up and shutdown scripts.

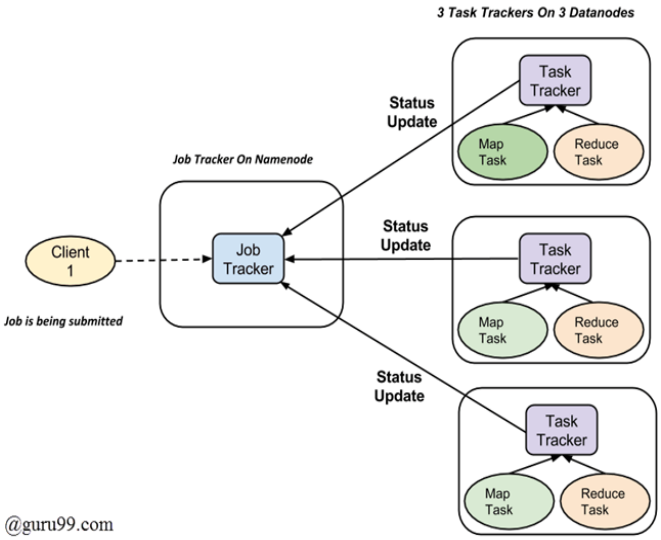


Hadoop distributed file system nodes are managed through a name node server which is used to host the file index system and a secondary name node that can generate snapshots of the name node's memory structures and also to prevent file system corruption and loss of data. Job scheduling across nodes can be managed by job tracker server. An alternate file system is used in Hadoop Mapreduce called as the name node, secondary node and data node architecture which can be replaced by the file system specific equivalents.

2.2 HADOOP DISTRIBUTED FILE SYSTEM

The HDFS file system is written in java framework which helps to distribute, scale and portable. Instead of storing the data in data store due to the lack of POSIX compliance, it does it by providing shell commands and java application programming interface methods that are similar to other file systems. A single name node and a cluster of data nodes can hold a Hadoop cluster that requires redundancy options and are available for the name node due to its critical situations. Each data node can be served by block protocol which is specific to HDFS. The file system uses sockets named as TCP/IP sockets used for communication. Remote procedure calls (RPC) is used for clients to communicate with each other. Across multiple machines the large files can be maintained in the range of gigabytes to terabytes. By replicating the data across multiple hosts it achieves the reliability. It does not require redundant array of independent risks for storing the data in hosts. To increase the Performance in input and output some raid configurations is used. For the default replication value 3, data is been stored on three nodes. First and second node the data will be stored in same rack and third node will be stored on different rack. To keep the replication of data high the data nodes can talk to each other to rebalance the data and to move the copies around. HDFS is not fully POSIX compliant, because the requirements for a POSIX file system gets differ from the target goals of a Hadoop application. The data throughput can improve the performance by not having a fully POSIX compliant file system but it supports for non-POSIX operations. Append is the one of the non-POSIX operations.

2.3 WORKING AND ORGANIZING MAP REDUCE



Map reduces works or organizes in two phases that is map and reduce. The map splits and processes the data. Reduce shuffles the data. There are two entities which complete the process namely, job tracker and multiple task trackers. Job trackers acts like the masterpiece which is responsible for the complete execution of all submitted jobs. Multiple task trackers act like the slaves. The multiple data nodes can run into a cluster when a job is divided into multiple tasks. To run on different nodes job tracker has to coordinate with the activity which is been scheduled by tasks. The job of a task tracker is to send the progress to job tracker. In the event of task failure, the job tracker can reschedule the jobs on different task tracker.

2.4 HADOOP MAPREDUCE

Map reduce is an application which is used to write large number of data in parallel or on large clusters of commodity hardware in a reliable manner. It is based on a distributed computing which is a processing model and a programming model. Map and reduce are the two important tasks for the mapreduce algorithm. A set of data will be converted into another set of data by using map and the every element will be broken down into tuples. The tuples can be key and value parts. The task of a reduce takes the output from a map as an input and combines the data's tuples as a smaller set of tuples. The task of a reduce is performed when the map completes its job.

2.5 ADVANTAGES OF MAPREDUCE

The mapreduce is easy to scale the data processing over many multiple computing nodes. Mappers and reducers are the term which used for data processing primitives. In the concept of map reduce model, it will be a non-trivial process when the decomposition of data processing application takes place in Mappers and reducers. The configuration changes when we write an application in the mapreduce form and scaling the application over hundreds or thousands of machines in a cluster. This is the process which attracts many users' to use this model.

CONCLUSION

For analyzing the large amount of data there should be some technical advancement. To analyzing the huge data need of some scientific potential. In large variety of application domains the process is cost effective and faster method should be implemented.

REFERENCES

[1]M.A.Beyerand D.Laney,"The importance of "big data": A definition," Gartner, Tech. Rep., 2012.

[2]T.C.Havens, J.C.Bezdek, C.Leckie, L.O.Hall, and M.Palaniswami, "Fuzzy- Means Algorithms for Very Large Data, "IEEE Trans. On Fuzzy Systems, vol.20,no. 6, pp. 1130-1146, December 2012.

[3]Z.Liu, P.Li, Y.Zheng, etal.,"Clustering to find exemplar terms for key phrase extraction,"inProc.2009 Conf. On Empirical Methods in Natural Language Processing, pp. 257-266,May2009.

[3]Z.Zheng,J.Zhu,M.R.Lyu."Service-generated Big Data and Big Data – as – a - Service: An Overview," in Proc.IEEEBigData,pp.403-410,October2013.A.Bellogín, I.Cantador, F.Díez, etal., "An empirical comparison of social, collaborative filtering, and hybrid recommenders, "ACM Trans .on Intelligent Systems and Technology, vol.4, no. 1,pp. 1-37,January2013.

[4]W.Zeng,M.S.Shang,Q.M.Zhang,etal., "Can Dissimilar Users Contribute to Accuracy and Diversity of Personalized Recommendation?, "International Journal of Modern Physics C,vol.21 ,no.10 ,pp. 1217-1227, June 2010.