A SURVEY - BIG DATA USING HADOOP ITS CHALLEGES, TOOLS AND TECHNIQUES

D. Manjula[1]

Abstract

'Big Data' is an technique and technology to detain, store, allocate, manage and analyze huge datasets. it has high velocity and dissimilar structures. Big data mean the datasets that could not be apparent, acquired, managed, and processed by traditional IT and software/hardware tools within a allowable time. The analysis of data becomes difficult because of the targeted threats are increased and also high growth in data. If the database is very enormous the it will be very multifarious. Structured, semi structured and unstructured data are difficult to process by using traditional database like RDBMS. An example of big data may be Exabyte s(1024 terabytes) of data consisting of trillions of records of millions of people from different sources such as websites, social media, mobile data, eb servers, online transactions and so on [2]. Parallelism is an efficient way to process huge amounts of data in an economical and successful way. For analytical purposes Hadoop is the interior platform for structuring big data and to solve the problems. For Big data implementation Haddop has emerged as an popular tool. This paper provides an way of response analysis using Hadoop which will process the gigantic amount of data on a Hadoop cluster faster in real time.

Keywords : Bigdata, hadoop, Security on bigdata., Map reducing

I. Introduction

Nowadays the internet and technology was developed and it was connected by networked system**.** Dueto the development of internet and usage of social sites the world is getting slighter and slighter day by day. Data is growing exponentially as the number of users and activity over the web is increasing rapidly.

Big data :

Structured, semi structured and unstructured statistics are difficult to process by using habitual database like RDBMS. [3].

Big data are difficult to captured, manage and processed. It is a group of Size, variety and Velocity.

II. Characteristics :

Size (Volume) :

Data size was developed from megabytes, gigabytes to peta bytes. Size is nothing but the quantity of data. It is an difficult task to handle large size of data using predictable databases.

Variety :

The data are available in text and tables in earlier versions. But in current versions the data are

[1] Assistant Professor, Department of Computer Science, Karpagam University

available in the form videos, pictures, tweets etc. Data may vary from structured and unstructured data which was stored. Data may also in audio video, XML etc formats.

Velocity :

Velocity is an processing speed of an data. The data which are available in online was continuously changing and it should be available at the correct time in effective[4]. In order to maximize its value, big data must be used.

III. Challenges in Big Data

Understanding about big data

The counting of and data must be properly analyzed. If any change in business environment it require some task to perform in order. Time span plays that role.

New technologies

Many recent technologies are developed each day. The organization must be learned how to use those new technologies in market.

IT specialists

McKinsey's states that : For innovation 1,90,000 workers and 1.5 million data literate managers may need. It is not an easy task to handle.

Privacy and Security:

Privacy and security is also and risky task in big data. Because a company may contain complex data.

IV Tools and Techniques :

Hadoop :

HADOOP is used to store and to process the huge data under big data. Hadoop is build on java platform and it is also an open source tool to develop processing on grouping. The modules like HDFS, PIG, HIVE, Map reduce are compromise to perform processing on large data[5]. Many computers are linked together to perform huge data by using Hadoop. It is efficient way[6]. 1000 single CPU or 250 core machines are better than 1000 cpu machines. Cost is also very huge. Tiny and sensibly prices machines are mainly used in Haddop processing because they are tied with one another for effective process. Hadoop cluster data will be distributed to all the nodes in a group. The dissimilar nodes are handles by HDFS[7]. Hadoop contain two layers:

1. Processing (Map reduce)

2. Storage layer(HDFS)

HDFS:

HDFS is an Hadoop Distributed File System. It will spans the nodes in a group for data storage processing.

The nodes are linked together to make them into an big data file system. File will be put on Hadoop cluster, HDFS will breakdown that into blocks and it will be distributed that across all of other nodes. There is an fault tolerance concept namely HDFS configure replication factor.

The file will be set in Hadoop. Then it will be separated as 3 copy in each block to spread all the node in the clusters. Each node consists of one name node. Name node is also known as an meta data . Meta data will clutch the memory which consists of every block. If it contains multiple rack it may know where the block has been existed, what racks crossways the cluster inside in the network.

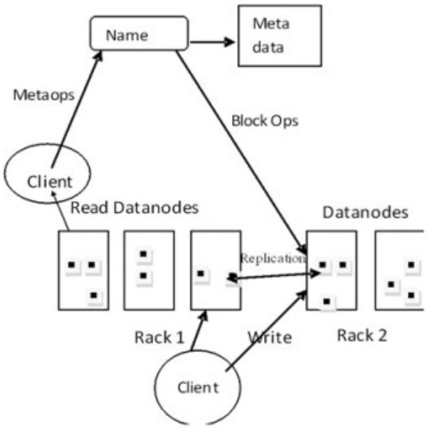
HDFS will run on commodity hardware which have high performance. It is for highly scalable and it is for reproduction of data. Data duplication diagonally three nodes for error tolerance. Hadoop enables a computing solution that is:

Scalable :

Scalable is without any change in format of data, and how the data is overloaded, the applications on peak node should be added.

Cost effective :

In Hadoop Technology the data may be simultaneously calculating to creation of servers. The cost of data storage is very low even if the model contains cost per terabyte.



Flexible :

Hadoop can be work in any form of data i.e. ordered or unordered. Any form of data can be enabled to any data and it produce the result of analyzing any other system.

Fault tolerant :

The data can be redirect to another location if the problem is occurred in that data.

In this HDFS a huge volume of data is stored and that data provides fast and easy access. The data can be stored across numerous machines. The data can be retrieved from multiple machines and that data cannot be loss or failure while copying. The data can be stored in superfluous machines and it can be free the system for data losses in the way of occurrence of failure of data for the distributed storage and calculating.

So that hadoop will manage the interface to perform operations with HDFS. The cluster position can be checked by using Name node and Data node for the authentication of files. In Fig.1 HDFS structural design can be described. That structural design can be follows by master-slave architecture with the subsequent elements.

V. Name Node

The GNU/Linux operating system is used in name node and it is also called as commodity hardware. To run on commodity hardware this name node is also work as an software. If the name node in the system perform as an master server and it follows the works like, the files can be regularly accessed by clients and the file system operation can be performed.

VI. Data Node

The data node will manage the data storage of each hardware in a system. According to client request read and write operations will be performed like creation, deletion and replication of a block where the information is given by name node.

VII. Block

HDFS contains the user data. Each data node contains one or more segments. The configuration of a block size will be increased up to 64MB. Each segments are called as blocks.

VII. Map Reduce

Map reduce provides the data. there are 2 step process which involves mapper and reducer. Map function will be written by programmers and intimate the cluster what data to be retrieve. The reducer make the data aggregation. Map reduce is working on all of data inside clusters. The myth behind is one need to understand the java to complete the clusters. HIVE is a small project for the engineers of face book In face book the engineers use a sub project called HIVE, which is in SQL. Instead of forcing the people to learn java the HIVE reduce the complexity[7].

Pig is another one built by yahoo,. In yahoo they uses PIG built in, it use to draw data out of clusters. Pig and hive are under the Hadoop. Job will be finalized to cluster by map reduce. The advantage of open source framework, user can build, add and area which increase the hadoop technologies and projects[8]. Map reduce is easy to process data. It is an big advantage in map reduce.

Data processing primitives are called mappers and reducers. The scaling of an application can be run over thousands of machines. It was enthralled by many programmers to optimize map reduce model. Hadoop transfer the Map Reduce tasks [9] to the appropriate servers in the group. Hadoop will manage the task like certifying, completion and copying the data between the nodes in a cluster. Local disk will reduce the traffic in a network. The data will be reduced to form the exact result and it will send back to the server.

Data Mining For Big Data

Data mining is used to extort the data from huge data. And it will be send to one form to another form. The classification is risk if the database is larger. Many techniques are used in data mining to process the data[10]. Clustering is an important technique in data mining applications to group the set of data. Clustering is an important technique mainly used in big data analysis.

VIII Conclusion And Future Work

Nowadays, Data is generated from various areas and it arrive from various rate. To process the huge amount of data is a risky task and it is high issue today. This paper describes the concept of Big Data along with three V's, Volume, Velocity and Variety of Big Data. It also focus on Big data processing. Hadoop is a platform for structuring big data. Hadoop is an open source software used for processing big data. It is useful for analytical purposes. The challenges must be find out for capable and fast processing of Big Data. In future we Can use some clustering techniques and check the performance by implementing it in hadoop

References

1. S.Vikram Phaneendra & E.Madhusudhan Reddy *"Big Data-solutions for RDBMS problems-A Survey"* In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka,Japan, Apr 19{23 2013).

2. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N *" Analysis of Bidgata using Apache Hadoop and Map Reduce"* Volume 4, Issue 5, May 2014" .

3. X. Wu, X. Zhu, G. Q. Wu, et al., *"Data mining with big data,"* IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2012.

4. W. Zeng, M. S. Shang, Q.M. Zhang, et al., *"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.

5. http://www.edupristine.com/courses/big-data-hadoop-program/big-data-hadoop-course/

6. Apache Hadoop Project http://hadoop.apache.org/ [7] *"Hadoop Tutorial from Yahoo!",* Module 7: Managing a Hadoop Cluster

7. Yamashita, H. Kawamura, and K. Suzuki, *"Adaptive Fusion Method for User-based and Item-based Collaborative Filtering,"* Advances in Complex Systems, vol. 14, no. 2, pp. 133-149, May 2011.

8. D. Julie, and K. A. Kumar, *"Optimal Web Service Selection Scheme With Dynamic QoS Property Assignment,"* International Journal of Advanced Research In Technology, vol. 2, no. 2, pp. 69-75, May 2012.

9. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, *"Shared disk big data analytics with Apache Hadoop",* 2012, 18-22.