BIG DATA & HADOOP A HUGE STARAGE

Prof. Vijaya Chavan

Professor, Computer Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai,Maharashtra, India[1]

ABSTRACT

Big data is a very large volume of structured and unstructured data. An example of big data is petabytes i.e.1,024 terabytes or Exabyte. 1,024 petabytes of data consist of billions to trillions of data of millions of people. Its a process of collecting and processing the huge amount of data. Big companies are using mostly big data for specific surveys. Hadoop is a platform provided that is used for big data. Hadoop stores massive amount of data that have massive power and can process multiple things at a single time. Hadoop Files are called as HDFS. Big data is data that contains large amount in increasing volumes and higher velocity. Hadoop is used to store large amounts of data. Big data consist of very larger, and more complex data sets. These data sets are so large so difficult manage but it is used to solve business problems

Keyword: *Big data, hadoop, HDFS, velociity*

1. INTRODUCTION

Big data includes storing data sets which cannot be stored and used to capture, manage, and process data within a very short time. Large companies are successful in their economy that is totally based on knowledge. Data drives the modern organizations of the world and hence making sense of this data and undo the various patterns and revealing unseen connections within the big sea of data becomes sensitive and a hugely rewards to achieve it.Big data should be correct so that it lead to more positiveand satisfactory decisions resulting in greater effective, cost reduction and reduced risk. Big data consist of 5 V's that are Volume, Velocity and Variety,Veracity,Valuey.Hadoop is a free programming framework that supports the processing of large data sets in a distributed system. Hadoop is an open source software framework which provides huge data storage facility. In hadoop cluster, a number of machines are connected in a network and when a request comes from the client the number of computers or machine will process the data and will provide the output in a very short period of time. Cluster is nothing but group of machines connected in a network. The use of Hadoop makes it possible to run applications on systems with thousands of computers involving thousands of terabytes. Its distributed file system helps in rapid data transfer rates among computers and allows the system to continue process uninterrupted in case of a node failure. Hadoop is providing a platform to the purpose of the Big Data.

2. LITERATURE REVIEW

Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu and Jun Shao says that ,Big data, requires more knowledge for economic growth and technical innovation, has recently received more attention[1], and many research efforts are been done for big data processing due to its high volume, velocity, and variety challenges.

SerefSagiroglu and DuyguSinanc viewed Big data as a large amount of data sets having large group structure [3] .

We can live with many of the doubts of big data for now, with the hope that its benefits will remove its harms, but we shouldn't blind ourselves to the possible irreversibility of changes-whether good or bad-to society[2]. The University of Wollongong discusses the Computer magazine special issue on "Big Data: New Opportunities and New Challenges".

3. METHODOLOGY

3.1 5V's in Big Data

1) Volume –Many companies are now collecting data from various sources, including business transactions, social media and from survey. In the past, storing that huge amount of data was very difficult – but new technologies such as hadoop have made it easy to store.

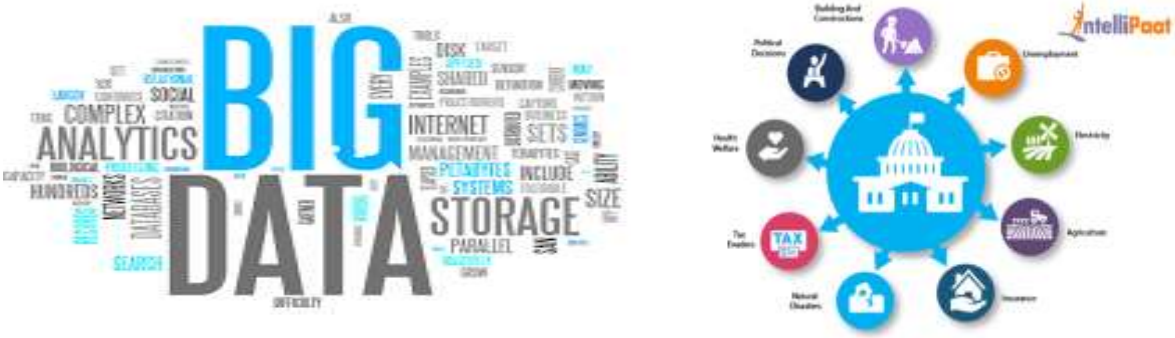
2) Velocity - Data streams is at now unstoppable speed and must be dealt with in a timely manner. RFID tags, smart metering are driving the need to deal with torrents of data in near-real time.

3) Variety - Data is in the form of structured, unstructured, numeric data in traditional databases to unstructured text documents, email, video, audio etc.

4) Data veracity-It is used to check accuracy and truth of a data set . Accuracy means not only quality but also truthfulness of the data sources type, and processing of it. To increase the accuracy it is needed to remove bias, abnormalities or inconsistencies, duplication, and volatility.

5) Value – It is data which is whether structured or unstructured consist of boundless process with endless options.

3.1.1 Big data in Government industry-



Governments, requires very huge amount of data daily because they have to keep track of various records and databases regarding the citizens in the country. All these records requires big data to store all these records.It is also used to store records of political programs.It is used in agriculture field for keeping track of all the land and livestock.To overcome national challenges such as unemployment, terrorism, energy resource exploration and more.Governments are also uses of big data in catching tax evaders.

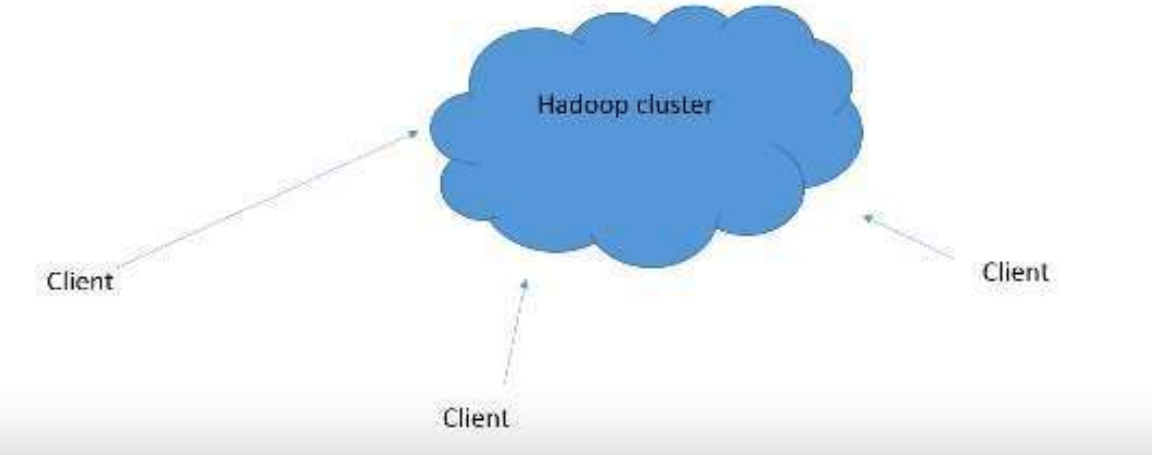
3.1.2 Where Big data is used?

- In Banking Sector
- In Transportation Industry
- In Weather patterns
- In Media and Entertainment industry etc.

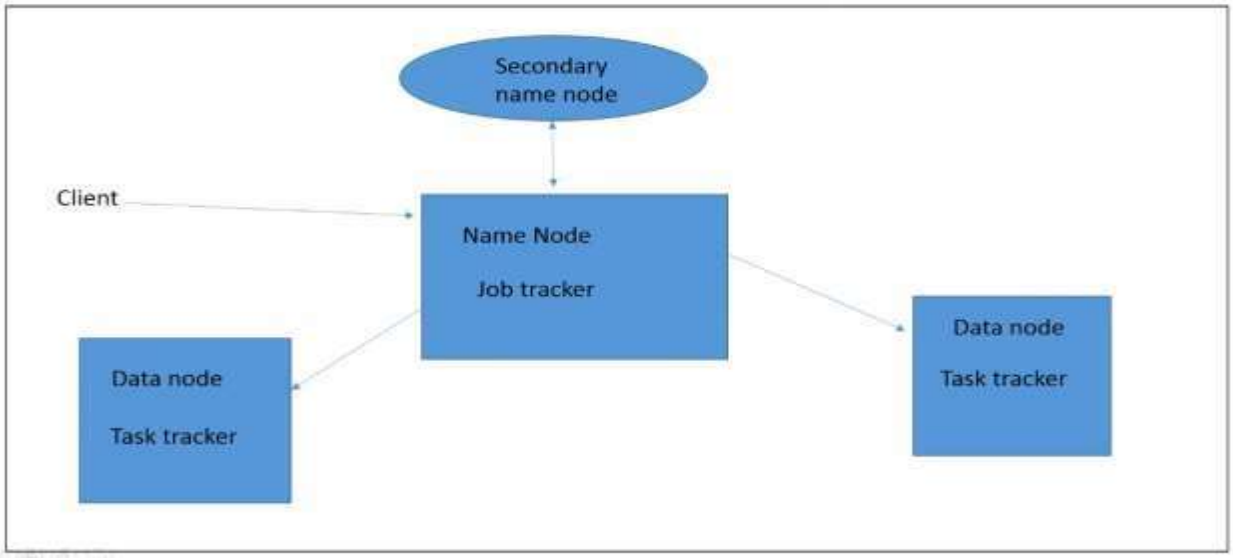
3.2 Hadoop

Hadoop is used for storing large amount of data across distributed group of servers and then running it from "distributed" analysis applications in each cluster. It's designed to be strong, in that your Big Data applications

will continue to run even when individual servers or clusters fail. And it is also designed to be more efficient, because it doesn't require your applications to transfer huge amount of data across your network. In hadoop, there are several numbers of client and in hadoop cluster number of machines or computers are connected through a network which process the job given by the client to the hadoop. It is divided in to small parts which is explained in the physical architecture of the hadoop.



3.2.1Physical Architecture of Hadoop



3.2.2 Components present in Physical Architecture of hadoop

1) Name Node (NN)

i) The job from the client goes to the name node and it's on the name node that accepts the job given by the client.
ii) It is the master of HDFS i.e. Hadoop File System.
iii) It has job tracker which keeps track of files distributed to data nodes.

iv) Name node is the only single point to failure.
If the name node fails then the whole structure is crashed

2) Data Node (DN)

i)The job is given to the data node by the job tracker. Data node will then further pass the job to the task tracker to be performed.
ii) It is the slave of HDFS.
iii) It takes client block address from name node.
iv) For replication purpose it can communicate with other name node.
v) Data node informs local changes/ updates to name node.
vi) Each node in the structure can communicates with each other.

3) Job Tracker (JT)

i)Job tracker divides the job in to number of small parts that are easy to be executed and are passed to the data nodes.
ii) It determines the files to process,
iii) Only one jobtracker per hadoop cluster is allowed.
iv) It runs on a server as a master node of cluster.

4) Task Tracker (TT)

i)Task tracker does the job given by the data node.
ii) There is a single task tracker per slave node.
iii) It may handle multiple taskparallelly.
iv) Individual tasks are assigned by job tracker to task tracker.
v) Job tracker continuously communicates with task tracker and if anytime it fails to reply then it assumes that the task tracker has crashed.

5) Secondary Name Node (SNN)

i) State monitoring is done by SNN.
ii) Every cluster has one SNN.
iii) SNN resides on its own machine.
iv) On that machine or server no other daemon (DN or TT) can work.
v) SNN takes snapshot of HDFS metadata at constant intervals.

4. FUTURE SCOPE

Big data is used for storing huge amount of data in a single cluster which is used to store data in small units. This is a huge revolution in the world of storing data. It is used by big companies for storing huge data. Hadoop is very useful and famous in future as it is developing day-by-day and is allowing a platform for huge amount of data in a single cluster resolving the pervious issue of storing the data

5. CONCLUSIONS

The Big data consist of 5 important V's Volume ,Velocity, Variety, veracity, Value.Physical architecture of Hadoop consist different components like Name Node(NN), Data Node (DN), Job Tracker (JT), Task Tracker (TT), Task Tracker (TT).Thus Big data and Hadoop is used for storing huge amounts of data

6. REFERENCES

[1] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu and Jun Shao "Toward efficient and privacy-preserving computing in big data era"
https://ieeexplore.ieee.org/abstract/document/6863131/
[2] Katina Michael and Keith W. Miller "Big Data: New Opportunities and New Challenges [Guest editors' introduction]"
https://ieeexplore.ieee.org/abstract/document/6527259/
[3]SerefSagiroglu and DuyguSinanc "Big data: A review"
https://ieeexplore.ieee.org/abstract/document/6567202/
[4] AvitaKatal, Mohammad Wazid and R. H. Goudar "Big data: Issues, challenges, tools and Good practices" https://ieeexplore.ieee.org/abstract/document/6612229/
[5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding "Data mining with big data" https://ieeexplore.ieee.org/abstract/document/6547630/
[6] Sachchidanand Singh and Nirmala Singh "Big Data analytics" https://ieeexplore.ieee.org/document/6398180/
[7] Ibrahim AbakerTargioHashema, IbrarYaqooba, Nor BadrulAnuara, SalimahMokhtara, AbdullahGania and SameeUllahKhanb "The rise of "big data" on cloud computing: Review and open research issues" https://www.sciencedirect.com/science/article/pii/S0306437914001288?via%3Dihub JyotiNadimath, Ekata Banerjee, AnkurPatil, Pratimakakde, SaumitraVaidya and DivyanshChaturvedi "Big data analysis using Apache Hadoop" https://ieeexplore.ieee.org/document/6642536/