TECHNOLOGIES BEHIND BIG DATA

*Charu Mukhija

*Assistant Professor, Dept. of Computer Science, I.B. (PG) College, Panipat, Haryana, India

ABSTRACT

Data is the raw facts of the information being processed or stored in computer. It is the backbone of the computer. Big Data is a field that treats ways to analyze, systematically extract information from or deal with the data sets that are large complex to be handled by the traditional data processing systems software. Big Data performs data storage, data searching, data analysis, data sharing, data transfer, data updating, data querying. This paper includes why we need the concept of Big Data and what is the source of the collection of data and how it is organized, used, managed and processed. This paper presents an analysis of Big Data techniques. It also provides the improvement of future research.

Keywords: *Types of Big Data, Hadoop, Data Lake, InMemory Computing, Machine Learning and Data Mining, Presto, Examples of Big Data Generation*

INTRODUCTION

Data is the raw facts of the information being processed in computer. It are the backbone of the computer. The data which ranges from Peta byte to yotta byte is termed as Big Data.

NOW THE QUESTION ARISES WHAT IS THE SOURCE OF BIG DATA?

As we know digitization is in trend, and at every second millions billions of users is online at a time and are surfing the online data. Regardless to online users, YouTube generate millions of data per second, Social media generates 5K to 100K per second, Large Hadron Collider experiment generates 40TB data per second and many more.

CHARACTERISTICS

- Volume
- Velocity
- Variety

CHALLENGES IN FRONT OF BIG DATA

Huge amount of data: Our traditional system software cannot handle data which ranges from Tera byte to even Exa byte.

Structures, Unstructured, Semi-structured data: Traditional system cannot handle semi and unstructured data therefore there is a demand of Big Data and its management.

Processing Speed: To processed such a massive amount of data there is a need of Big Data **.**

TECHNOLOGIES

1. Hadoop:

It is an open source software framework used for storing and processing huge amount of data. Hadoop is specially designed file system for storing large data sets with cluster of commodity hardware with streaming access pattern .It started with two guy mainly Doug Cutting and Mike Cafarella.

Hadoop provided two important services:

- HDFS (Storage)
- Yarn (Processing)

HDFS (Hadoop Distributed File System):

The HDFS is the data storage system used by Hadoop applications. It provides a reliable means for managing big data and supporting big data analytics applications[4]

HDFS ARCHITECTURE

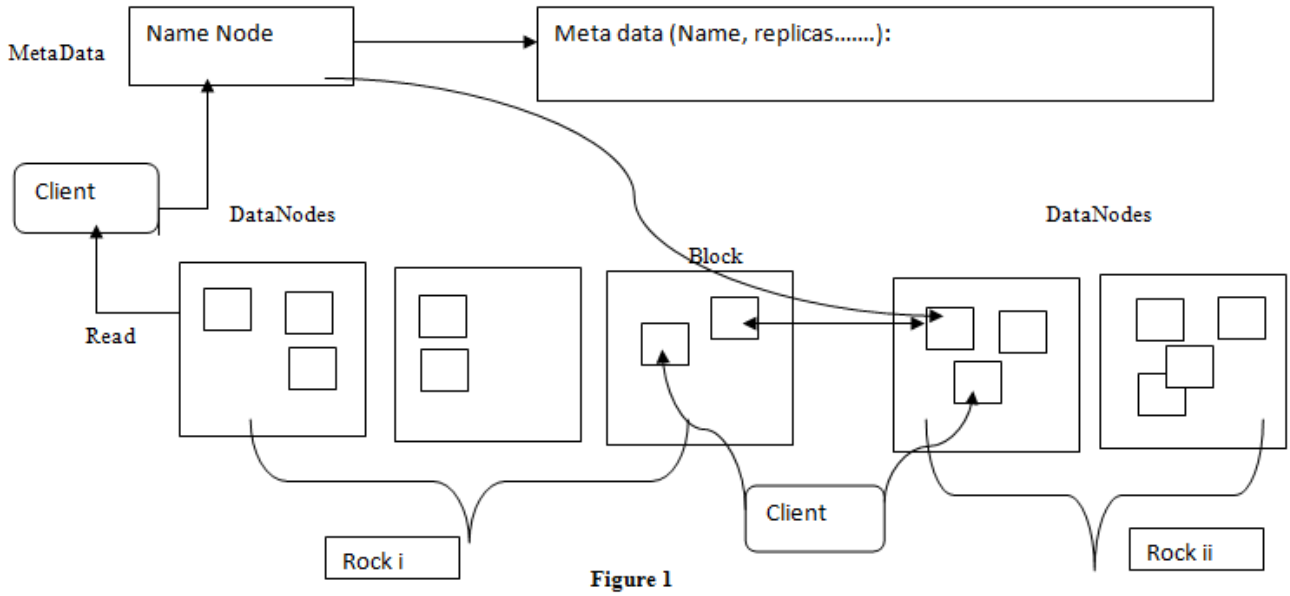


Figure 1

KEY COMPONENT OF HDFS ARE

NameNode and DataNode

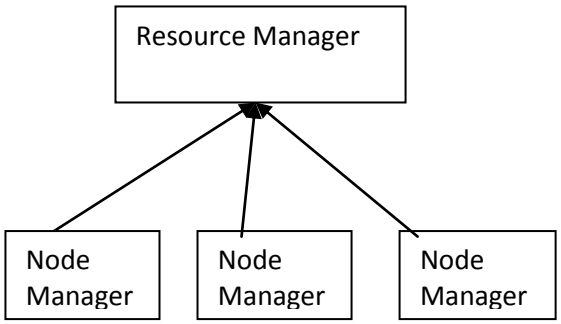


Figure 2

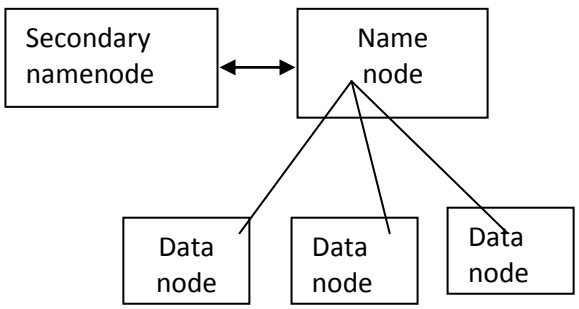


Figure 3

Resource Manager
. receives the processing requests
.passes the aprts of request to corresponding node managers

Node Manager
. installed on every datanode
. responsible for execution of tasks on every single data node

Name Node
. master daemon
. maintains and manages datanodes
. records metadata e.g. location of blocks stored, the size of files, permission etc
.receives heartbeat and block report

Data Node
. slave daemons
. stores actual data
.serves read and write requests

YARN

Yarn is the resource management and job scheduling technology in Hadoop. It stands for Yet Another Negotiator. Yarn comprises of two major components:

- Resource Manager
- Node Manager

1. Spark

In the seacrch technologies of the big data, Spark is a framework in the same way Hadoop open source software of the Apatxhe foundation. Why Spark when Hadoop is already? Hadoop is meant for batch produce processing data but Spark is used for real time processing data i.e it works on the current data gives an immediate processing of the data. The major advantage over Hadoop is the programming in Spark which is 1000 times faster than mapreduce programming which is difficult to use. Spark uses Scala as the programming language which works on the principle of opps , functional programming, inmemory computing and run on JVM. Spark uses RDD as its fundamental and each dataset is divided into logical partition. It makes easy to move operations from the disk to instead use data in memory, they are deterministic function of their input that means RDD i.e partition of data can be created at any time[2].

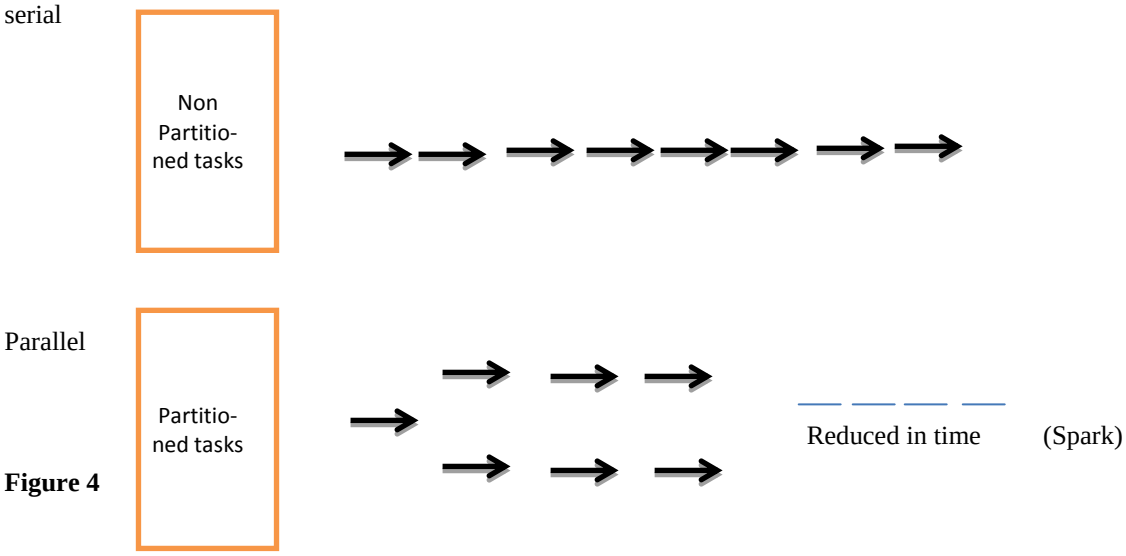


Figure 4

Spark uses cluster computing for its computational power as well as its storage i.e. uses resources from many computer which are networked together for the computation. It works with the system to distribute data across the cluster and process the data in Parallel. It is used by credit card, stock market analysis, Amazon S3, NoSql databases.

Components of Spark[1]:

1 Spark core- it is the base engine for large scale parallel and distributed processing of data. It works atleast with RDD which is used for memory management and fault recovery.

2. Spark streaming: It is used to put query.
3. Spark Sql: It integrates relational processing with Sparks functional programming.
4. MLLib: It contains libraries to do operation on data and make machine learning algorithm .
5. Spark Graphix:It uses graph based processing and data can be represented in network graph.
6. The Spark does not have its storage structure. It relays on storage which may be HDFS, NoSqletc, where we can attach Spark , fetch and process the data.

1) <u>NoSQL</u>

NoSQL stands for “Not Only SQL”. It is an approach to database design that support wide variety of data models. In traditional relational database data is placed in relations and its schema is designed carefully before database is built. NoSQL database are useful for working with large sets of distributed data. It’s Non relational approach and does not require fixed schema. Large scale web organisations such as Google and Amazon used NoSQL database.

EXAMPLE OF HOW TO WORK WITH SQL VS NOSQL

SQL:

Id	Name	Add	Age
123	John	3	30

Join

Add	City	Country
23	Panipat	India

Figure 5

NoSQLSyntax : [3]

ID VALUE

123 {
"Name" : "John",
"Add": {
"id" : 3,
"City": "Panipat",
"Country" : " India",
}
"Age" : 30
},

NoSQL Architecture involves four key elements:

1 Document Database : stores data in documents.

2 Graph Database: Stores data in the nodes and relationships of a graph

3 Key-value Database: uses key values to store data

4 Wide column stores:store data in related rows, but they actually store the data into columns

2) INMEMORY COMPUTING

Inmemory computing has been an active research which focus in the field of big data technologies. The drawback of the disks based databases evolves the invenetion of the inmemory computing .Inmemory computing involves architecture where the data is stored inside the memory i.e DRAM rather than the hard disks which increases the accessing of data . It is used in real time embedded system. It is based on two principles: the way data is stored and scalability i.e. ability of the system to process rapidly and constantly growimg of the data. The computing spotlighted that while accessing data in memory it eliminates that time which is used to locate the area on the disk where the data is to be read is stored , when querying the data.

Based on the computing, [2]HANA (High Speed Analytical Appliances) developed by SAP uses a technique called data compression to store data in DRAM and makes HANA a 10,000 fatest accessing of the data than the standard disks. The main drawback of the HANA is the storing multiple amount of data but it doesnot scale to manage truly big data scenarios. The research supports that HANA supports real time analysis on large structured data sets and the Hadoop which facilitates on processing and storing of large amount of data even the historical data also.

3) **DATALAKES**

Regardless to the traditional database management system and datawarehouse, big data evolves into a new technology data lakes. It works on the principle that all the data incorpating to its format is stored all togethered and from these sources it can be combined and processed by search and analytics techniques. The content which is stored in the data lake is in the normalized and enchriched form. If the user wants to work on the content , the data is prepared as needed.

CHARACTERISTICS

1 Centralised content

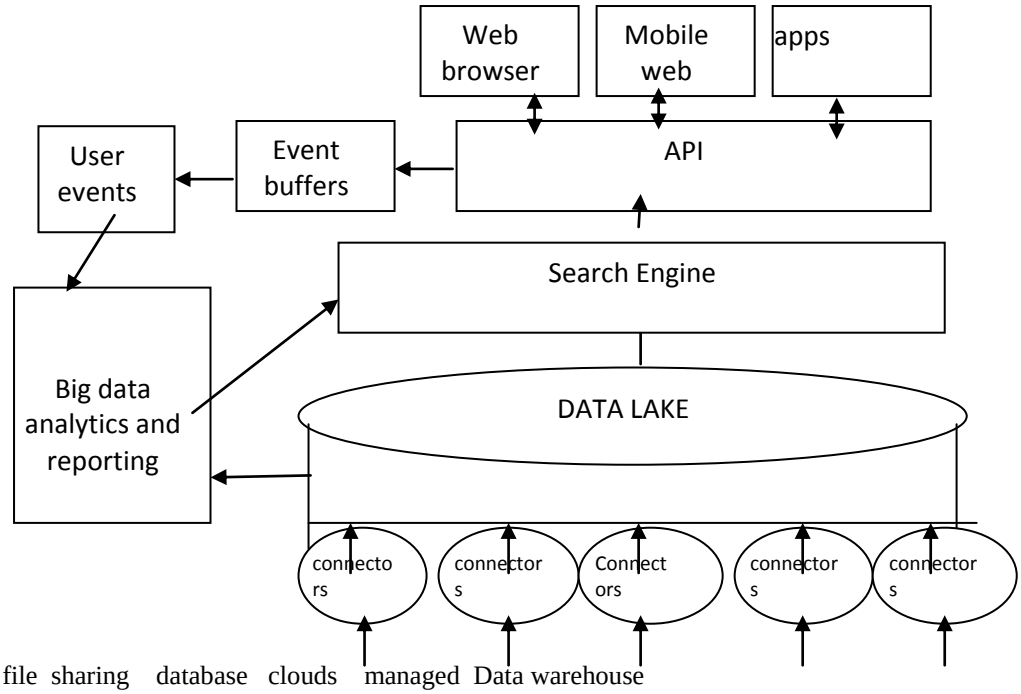
2 Agility

3 Scalability

4 Distributed access control

Biotech and health research are using data lake technique in which manufacturing, research, public and customize support data is managed[2]

ARCHITECTURE



file sharing database clouds managed Data warehouse

document repository

Figure 6

CONCLUSION

Big Data provides Data analytics on low commodity hardware. These trend have the capabilities required to analyze massive data sets quickly. It provides efficiency productivity and profitability.

References:

1 bigdata.ieee.org

2 journals.elsevier.com

3 journalofbigdata.springeropen.com

4 researchgate.net