UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

A Brief Survey on Big Data

Mohd Saleem[1], Abdul Quyoom[2], Mudasser Nazar[3], Mutasif Ishfaq[4]

Department of CSE, Baba Ghulam Shah Badshah University, Rajouri, India[1]

Department of CS, Baba Ghulam Shah Badshah University, Rajouri, India[2]

Department of CS, Baba Ghulam Shah Badshah University, Rajouri, India[3]

Department of CS, Baba Ghulam Shah Badshah University, Rajouri, India[4]

Abstract: Big data is a huge amount of data that are collected from different sources. Big data refers to large growing datasets which contain heterogeneous formats and roughly ranging from few terabytes to petabytes. Big data management using traditional approach is impossible because it contains different types of data i.e. structured as well as un-structured data. Big data require large amount of storage and one solution is cloud computing technology. Cloud computing provide online storage depending upon the need of the customers. This paper describes what big data is and the technologies like Hadoop, HDFS and MapReduce that are used for handling big data. This paper also discusses big data challenges and benefits.

Keywords: Big Data, Big Data Architecture, Challenges of Big Data, Big data technologies, Benefits of Big Data.

I. INTRODUCTION

A. Big Data

In general big data is a data whose size is very large which contain datasets that are impossible to processed using traditional techniques because traditional computing techniques processed limited size data. Big data contain data that comes from different applications like it capture data from social networking sites, data from share market, data from different companies and data from different databases through search engines. Data warehouse deal with structured data that come under the category of traditional database that organizes data in rows and columns and apply different methods on data but this data is restricted to limited size and include a single server. When the data goes beyond the capacity of the server then problem arises but big data is not restricted to size.

Big data contain large volume of data of different types like structured data, semi-structure data and un-structured data. Big data and Data warehouse have similar goals but they differ in organization of data. Capturing of data is simple in big data but organization of that data and obtained valued information from such data by making a decision is a big challenge in big data. In order to overcome the issue of large space for big data storage companies need to use cloud computing technology.

B. Characterization Of Big Data

The 3V's of big data are Velocity, Veracity and Volume. The velocity refers to data processing speed and veracity tells different data formats like structured and unstructured data and finally volume that include amount of data and measured in petabytes.

C. Architecture

In traditional approach only structured data is involved but for big data both structured and un-structure data are available and data come from different sources.

The various components of big data architecture are as follows:

- Source System: Big data collect data from different sources like social media data, Enterprise data, Survey data and data from various social networking sites.
- Transactional System: These are enterprise backend databases.
- Data Archive: It is a collection of archived data through transactional system in accordance with big data engine.
- ODS: Operational data store is a type of database that contains set of data from various transactional systems. The big data engine uses operational data for providing data to analytical system and data warehouse.
- Big Data Engine: This is the heart of big data architecture. It performs processing of large volume of data ranging from few Megabytes to Petabytes. It handles data of different types that include structured and un-structured data. The Hadoop technology is used for managing heterogeneous data.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

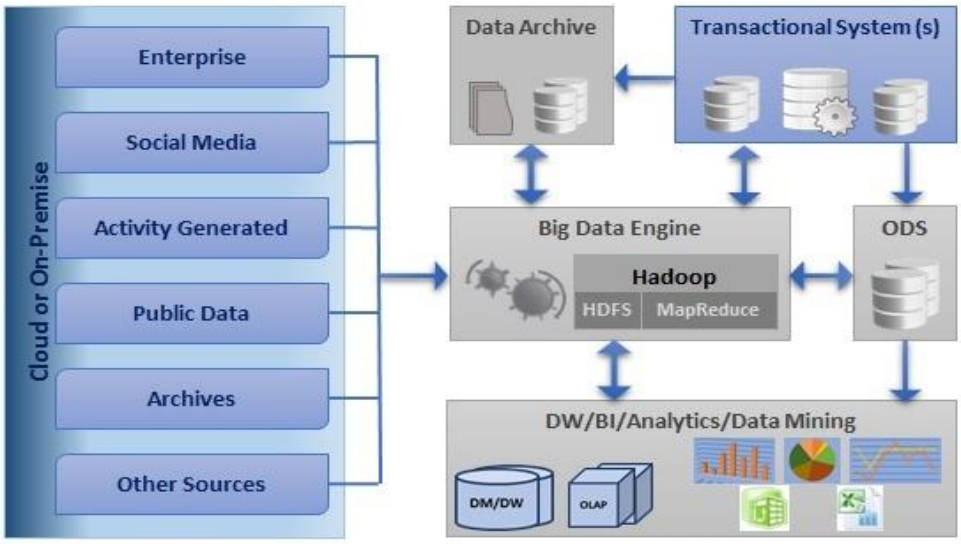

Fig. 1. Big data Architecture

II. CHALLENGES

The various challenges associated with big data are as follows:

- Capturing data is a challenge for big data as data come from different sources.
- Big data contain large volume of data so storage is also a big challenge.
- Managing large and rapidly increasing volume of data is also a big challenge.
- Data analysis is very difficult as it contain different types of data that include structured as well as un-structured data.
- Securing big data is a major challenge for organizations.
- Data validation is also a major challenge for big data.

III.BIG DATA TECHNOLOGIES

In order to handle a large volume of different types of data that contain structure as well as un-structure data we need a different technology. The traditional approach uses centralized server but this approach is not useful in big data.

A. MapReduce

Google provide a solution for handling big data by using an algorithm called MapReduce. It is a technique for distributed computing. MapReduce perform some operations on large dataset and divide it into smaller parts and assign it to different computers and run it in parallel. MapReduce have two processes i.e. mapping and reducing. Firstly mapping is performed that transform a piece of data into some key/value pairs and then sorting is performed according to the key. Secondly, reducing process takes mapped output as input and gives smaller tuples. MapReduce implement some other algorithms for mapping process to work. The algorithms are sorting, searching and indexing.

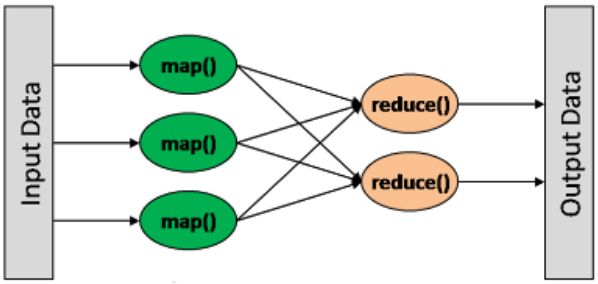

Fig. 2. Map Reduce

B. Hadoop

The most popular technology that is used for sorting is Hadoop. Hadoop is open source software and is mostly used for handling Big Data. Facebook and Google are also using Hadoop technology. Hadoop uses HDFS (Hadoop Distributed

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

File System) and it breaks data into smaller blocks and spread it to different systems. In order to assist Hadoop, Facebook developed a software system called Hive. Hive is basically a "SQL-like" bridge that connects with Hadoop in order to allow conventional applications to run queries.

Hadoop includes four modules:

- Hadoop common: It includes java libraries and utilities for Hadoop modules.
- Hadoop Distributed File System (HDFS): It is a distributed file system that store huge amount of data and it includes fault-tolerant storage system.
- Hadoop YARN (Yet Another Resource Negotiator): It is a cluster management technology that provide job scheduling and resource management.
- MapReduce: It performs parallel processing on large dataset using key value pair. It includes two processes i.e. mapping and reducing.

Hadoop uses a master/slave architecture designed for data storage and distributed data processing. The master node for data storage is the NameNode and the master node for parallel processing of data using Hadoop MapReduce is the Job Tracker. The other machine in the cluster of Hadoop is the slave node that stores the actual data.

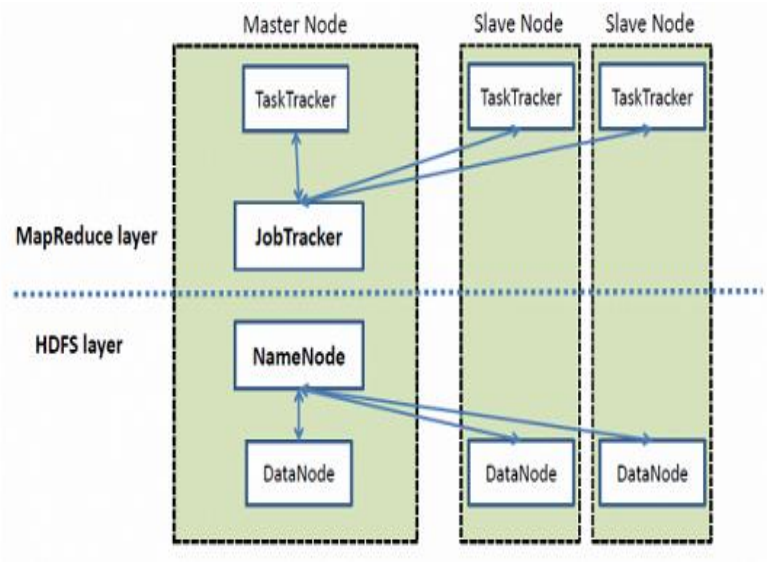

Fig. 3. Hadoop Architecture

The various benefits of Hadoop are as follows:

- Hadoop is cost-effective as compare to traditional approach.
- Easily scalable.
- Hadoop distributed file system replicate data across the cluster. Due to this it can stand against failure. If one node fails then data is obtained from other node.
- Fast data retrieval.
- Hadoop is an open source and it is compatible on all platforms.

C. HDFS (Hadoop Distributed File System)

Hadoop uses a distributed file system that is based on Google file system and it run on large clusters. HDFS split files into multiple blocks and distribute it to different Hadoop cluster. HDFS store huge amount of data by replicating it to different servers and survive against storage part failure.

The two important components of HDFS architecture is the NameNode and Data Node. NameNode is the master node that contains all the information and manages file system namespace. The operations that are performed by the NameNode on file system are opening, closing and renaming files. The NameNode provide location of file data blocks whenever client wants to read a file. The DataNode creates various blocks and saves data and it performs block deletion and replication operations. On start-up firstly handshake is performed between NameNode and DataNode after that DataNode matches Namespace-id and if there is a mismatch then DataNode shutdown automatically. The DataNode track running jobs are other jobs that are coming from NameNode with the help of task tracker. The DataNode keep track of its blocks and replicas locally and send this report to NameNode.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

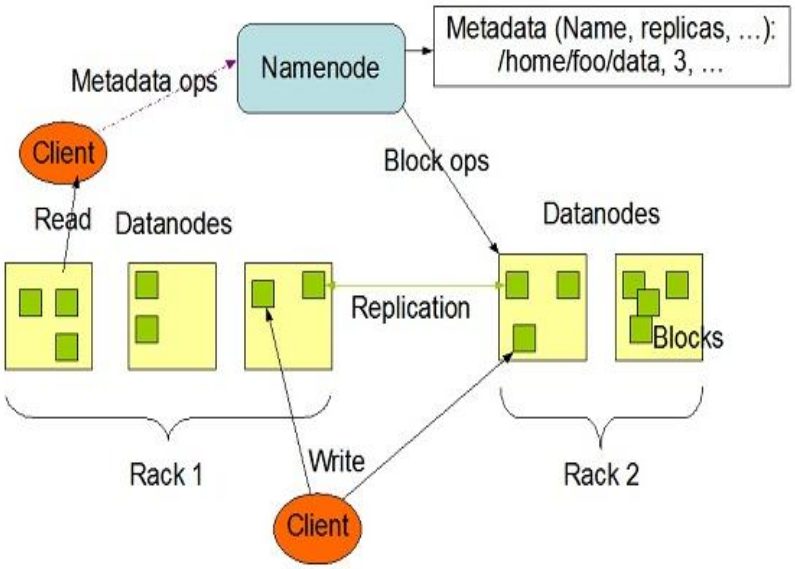

Fig.4. HDFS Architecture

IV. BENEFITS OF BIG DATA

The various benefits of big data are:

- Big data is accessible everywhere.
- Big data is secure as it uses secure infrastructure.
- The information in social media helps different companies to sell their product with good response because social media play a big role in product awareness to various customers.
- With the help of information stored in social networking sites a better goal can be achieved.
- Different offers are generated for customers depending upon his recent activities.
- Collecting information through different surveys help in providing better product quality.
- Managing large volume of data using traditional approach is expensive but big data simplify it by performing processing at high speed.
- Big data contain different data types so it requires different protocols and interfaces for integration of data.

V. CONCLUSION

In today's environment various types of data i.e. structured, semi-structured and un-structured data are collected through various digital platforms, social media or generated through surveys and simulation. Using traditional database methods analysing and understanding such heterogeneous type of data is not possible. This survey paper discusses the big data with its architecture and various challenges of big data. This paper also discusses the 3V's and the technology that is used for managing such type of heterogeneous data which include Hadoop, MapReduce and Hadoop distributed file system.

REFERENCES

[1] Jonathan Stuart Ward and Adam Barker "Undefined By Data: A Survey of Big Data Definitions" Stamford, CT: Gartner, 2012.
[2] Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.
[3] 2013 Big Data Survey Research Brief, SAS, The power to know, 2013.
[4] SAS, The power to know, Five Big Data Challenges And how to overcome them with visual analytics
[5] Apache Hadoop, http://hadoop.apache.org.
[6] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, "Survey Paper on Big Data" International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.
[7] Suman Arora, Dr.Madhu Goel, "Survey Paper on Scheduling in Hadoop" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014..
[8] Jonathan Stuart Ward and Adam Barker "Undefined By Data: A Survey of Big Data Definitions" Stamford, CT: Gartner, 2012..
[9] Kaisler S, Armour F, Espinosa J.A and Money W. Big data: issues and challenges moving forward, in: Proceedings of the 46th IEEE Annual Hawaii international Conference on System Sciences (HICC 2013), Grand Wailea, Maui, Hawaii, January 2013, pp. 995–1004.