International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 6 Issue 7, July-2017, Impact Factor: 3.578

Hadoop Handling Big Data

Gaurav Malik

E-Mail: gauravrmalik@gmail.com

ABSTRACT

In this paper I have explained big data afterwards I am throwing light on its characteristics and its importance. And then I explained how Hadoop as technology is going to help in implementation of big data further I explained storage part and processing part of Hadoop.

Keywords- Bigdata, Hadoop and hdfs.

INTRODUCTION

What is Big Data:-Big data is a term that describes the large volume of data – every structured and unstructured – that inundates a business on a day-after-day basis. However it's not the quantity of data that's vital. It's what organizations do with the information that matters. Big data can be analyzed for insights that result in higher choices and strategic business moves.

CHARACTERSITICS OF BIG DATA

1. Volume:-Volume denotes the scaling of data ranging from terabytes to zettabytes.

2. Velocity:-Velocity accounts for the streaming of data and movement of large volume data at a high speed sensors and smart metering are driving the need to deal with torrents of data in near-real time.

3. Variety:- Variety encompasses managing the complexity of data. data comes in all types of formats-from structured,numeric to unstructured text documents, video, audio etc.

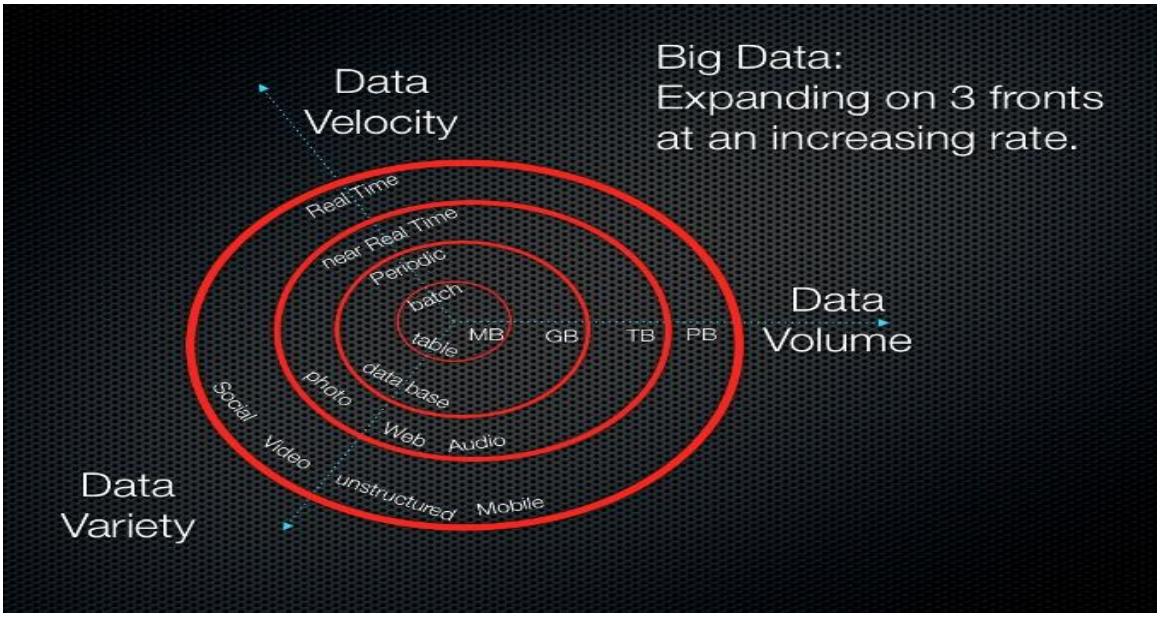

Fig. 1:

Types of Data

1. Structured Data:-Data which is represented in a tabular format
Example: Databases

2. Semi-structured Data :-Data which does not have a formal data model
Example: XML files

3. Unstructured Data:-Data which does not have a predefined data model
Example: Text files

WHY BIG DATA

Over 2.5 Exabyte's (2.5 billion gigabytes) of data is generated every day.

Following are some of the sources of the huge volume of data:

●A typical, large stock exchange captures more than 1 TB of data every day.
●There are around 5 billion mobile phones (including 1.75 billion smart phones) in the world.
●YouTube users upload more than 48 hours of video every minute.
●Large social networks such as Twitter and Facebook capture more than 10 TB of data daily.
●There are more than 30 million networked sensors in the world.

APPEAL OF BIG DATA

Big Data technology is appealing because of the following reasons:-

●It helps to manage and process a huge amount of data cost efficiently.
●It analyzes data in its native form, which may be unstructured, structured, or streaming.
●It captures data from fast-happening events in real time.
●It can handle failure of isolated nodes and tasks assigned to such nodes.
●It can turn data into actionable insights.

Why is big data important

You can take data from any source and analyze it and it reduces cost, time of a project. when you combine big data with high-powered analytics, can accomplish business relatedtasks such as

. You can determine root causes of failures, issues and defects in real time
. You can generate coupons at time of sale on customer's buying habits
. It can detect fraudulent behaviour before it affects your organization

Difference between hadoop and traditional Database

Table 1:

Table: Comparison between RDBMS and Hadoop		
Characrteristics	**RDBMS**	**Hadoop**
Basic Description	Traditional row-column databases used for both transactional systems, reporting, and archiving.	An open-source approach to storing data in a file system across a range of commodity hardware and processing it utilizing parallelism (multiple systems at once)
Manufacturers	Sql Server, MySql, Oracle, etc	Hadoop implementations by CloudEra, Intel, Amazon, Hortonworks
Best for applications	Reads & Writes, "reasonable" data sets (< 1B rows)	Inexpensive storage of lots of data, structured & semi-structured
Strength, Weakness	Massive data volumes, unstructured & semi-structured data	Complex, code-based, incompatible approaches in market, writes (one at a time)
Scalability	Challenging to "scale-out"	Strong bias to the open-source community & Java

Hadoop:- Hadoop is an open-source framework used for distributed storage and process of dataset of massive data using the MapReduce programming model. It consists of laptop clusters engineered from artifact hardware. All the modules in Hadoop are designed with an elementary assumption that hardware failures are common occurrences and will be mechanically handled by the framework. [3].The core of Apache Hadoop consists of a storage half, referred to as Hadoop Distributed file system (HDFS), and a processing half that may be a MapReduce programming model. Hadoop splits files into giant blocks and distributes them across nodes in a cluster. It then transfers prepackaged code into nodes to method the information in parallel. This approach takes advantage of data locality [4] wherever nodes manipulate the information they need access to. this enables the information set to be processed quicker and a lot of with efficiency than it would be in supercomputer architecture design that depends on a parallel file system.[5][6].
It is based on Google file system

HDFS (Hadoop distributed file system):-HDFS stores large files (typically in the range of gigabytes to terabytes[65]) across multiple machines. It achieves reliability by replicating the data across multiple hosts.

Daemons of Hdfs:-

1. **Name node:-**a master server that manages the file system namespace and regulates access to files by clients.

- It is a Master daemon for hdfs.
- We use enterprise hardware for Name node(Raid)
- It is one per cluster
- It is a single point of failure
- It stores metadata of hdfs

2. **Secondary Name node:-** it is used to recover metadata in case Name node fails

- It is master daemon
- It is used for check pointing process
- It uses enterprise hardware

3. **Data Node:-** It stores the data and responds to name node for file system operations

- It is a slave daemon
- It is used for storing data
- It uses commodity hardware(jbod)
- Wecan have any number of datanodesin a cluster

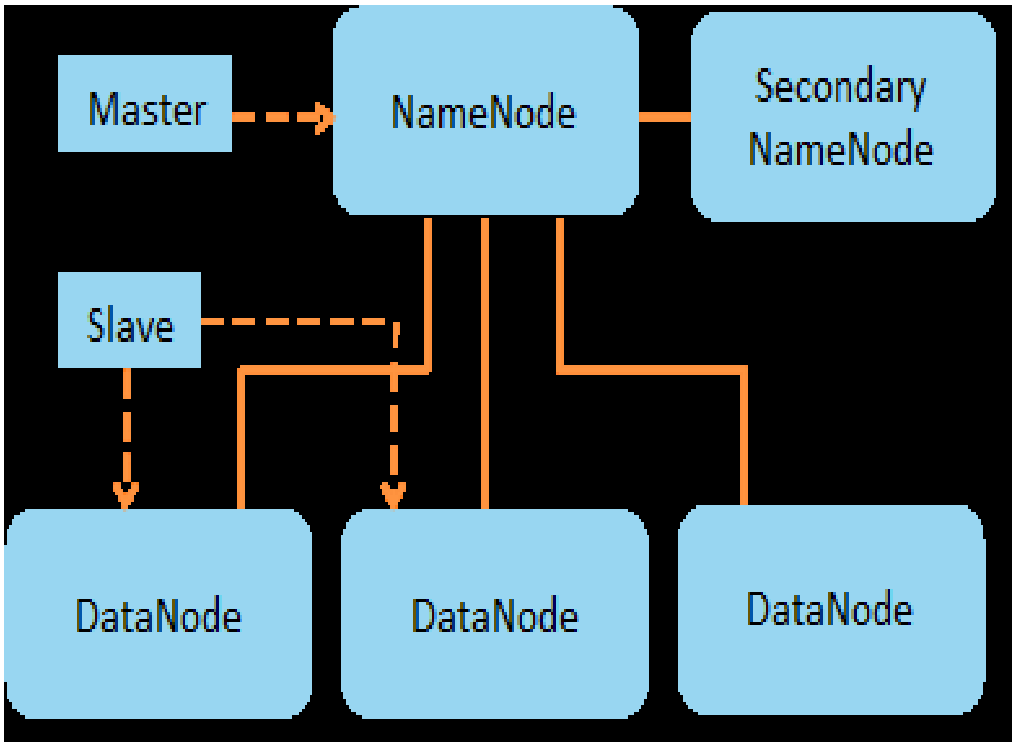

Fig. 2:

HDFS Cluster

Daemons of Map Reduce:-

1. **Job Tracker:-**it provides resources for job and monitors the job

    - It is a master daemon
    - It runs on enterprise hardware
    - It is one per cluster
    - It runs on a separate node and not usually on a Data Node.

2. **Task Tracker:-**it executes the job

    - It is a slave daemon
    - It runs on commodity hardware
    - Task Tracker runs on Data Node. Mostly on all Data Nodes.
    - It is one per job

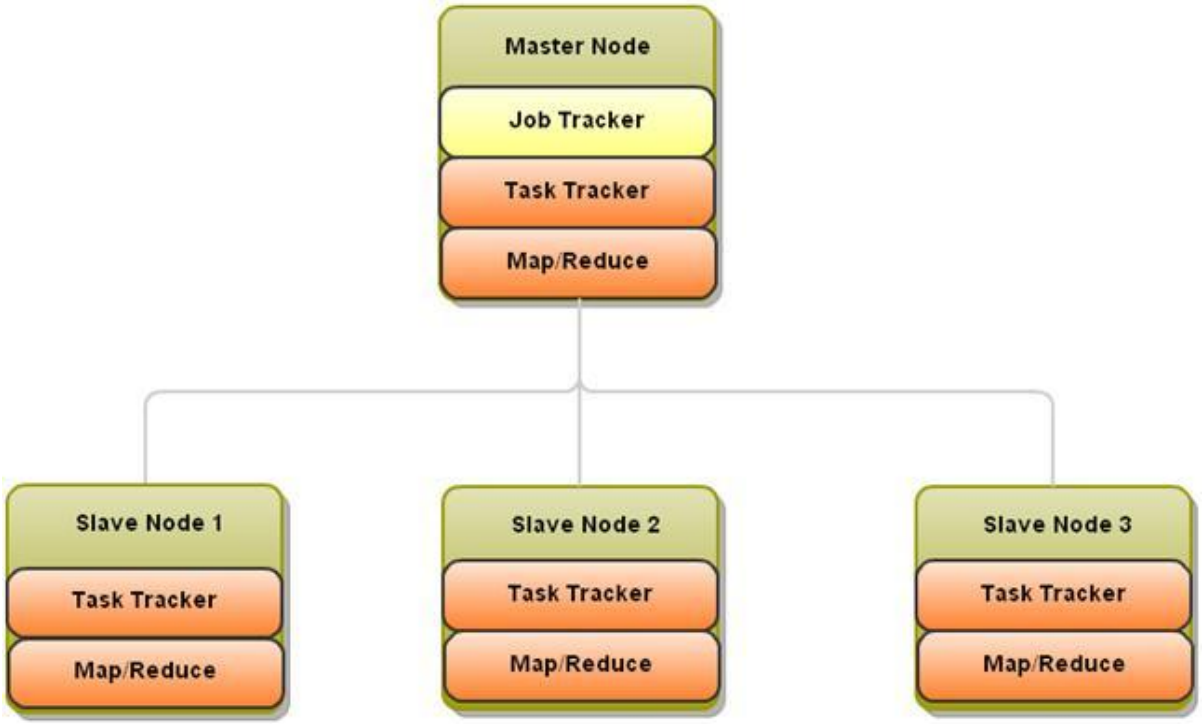

Fig 3: Map Reduce processing

REFERENCES

[1]. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html#dmhistory
[2]. http://searchcloudcomputing.techtarget.com/definition/Hadoop
[3]. "Welcome to Apache Hadoop!". hadoop.apache.org. Retrieved 2016-08-25.
[4]. What is the Hadoop Distributed File System (HDFS)?". ibm.com. IBM. Retrieved 2014-10-30
[5]. Malak, Michael (2014-09-19). "Data Locality: HPC vs. Hadoop vs. Spark". datascienceassn.org. Data Science Association. Retrieved 2014-10-30.
[6]. Characterization and Optimization of Memory-Resident MapReduce on HPC Systems"(pdf). IEEE. October 2014.
[7]. https://www.linkedin.com/pulse/hadoop-vs-rdbms-thiensi-le