Big Data Management: Characteristics, Challenges and Solutions

[1]Iqbaldeep Kaur, [2]Navneet Kaur, [3]Tanisha, [4]Gurmeen, [5]Deepi
[1,2,3,4]Dept. of CSE, Chandigarh Engineering College, Landran, Punjab, India
[5]Dept. of ECE, Chandigarh Engineering College, Landran, Punjab, India

Abstract
Day by day there comes a new technology, devices and communication means which give rise to the rapid growth of data. Now days, data is enormously increasing within every ten minutes and it is hard to manage it and it gives rise to the term BIG DATA. This paper describes the big data and its challenges along with the technologies required to handle big data.

Keywords
Big Data,MapReduce,Hadoop

I. Introduction

Big data is a buzz word that represents the growth of voluminous data of an organization which exceeds the limits for its storage. There is a need to maintain the big data due to:

- Increase of storage capacities
- Increase of processing power
- Availability of data

II. Types of Data for Big Data

- Traditional enterprise data – includes information of customer from CRM systems, transactional ERP data, web store transactions, and general ledger data.
- Machine-generated /sensor data – includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs, trading systems data.
- Social data – includes customer feedback streams posted by the people across the world, micro-blogging sites like Twitter, social media platforms like Facebook
- Stock Exchange Data : The stock exchange data holds information about the 'buy' and 'sell' decisions on mutual funds, shares of the customers are managed by the companies.
- Power Grid Data : The power grid data holds information consumed by
- a particular node with respect to a base station.
- Transport Data: Transport data includes model, capacity, distance and availability of a vehicle.
- Search Engine Data: Search engines retrieve lots of data from different databases.

III. Characteristics of Big Data

A. Volume

Volume refers to the amount of data. Machine-generated data is produced in much larger quantities than non-traditional data. For e.g, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.

B. Velocity

Is the speed of processing the data. The pace at which data streams in from sources such as mobile devices, clickstreams, high-frequency stock trading, and machine-to-machine processes is massive and continuously fast moving.

C. Variety

It refers to the type of data.Big data extends beyond structured data such as numbers, dates and strings to include unstructured data such as text, video, audio, click streams, 3D data and log files.

D. Value

The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis. [June 2013Oracle: Big Data for the Enterprise].

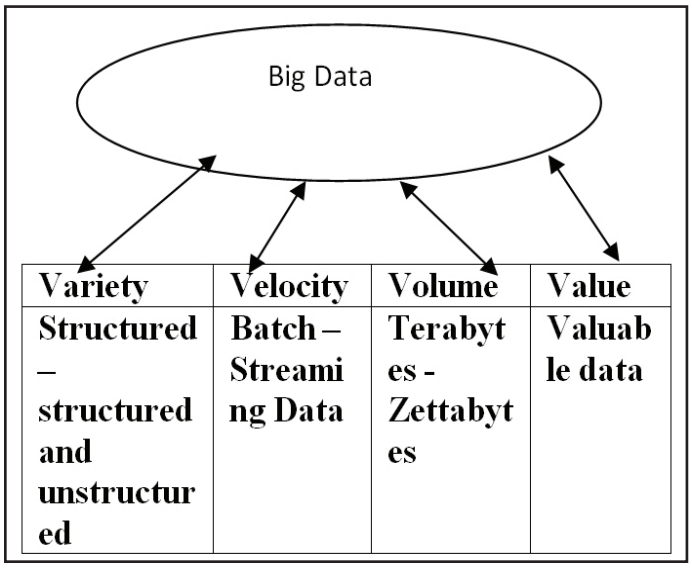

Fig. 1: V's of Big Data

The data has increased since we have started using digital worldwise ,it can include text ,images ,pictures,excel data which is summarised now a days and occupy lots of space.Space has become an issue which creates a problem foe the user to store the data.This is the root of the Big data ,we need to maintain it.
There is an increase in data because we have shifted from analog to digital.the graph represents the change from analog data usage to digital data usage.Large Data is creating a problem for the user to manage it and its memory consumption.

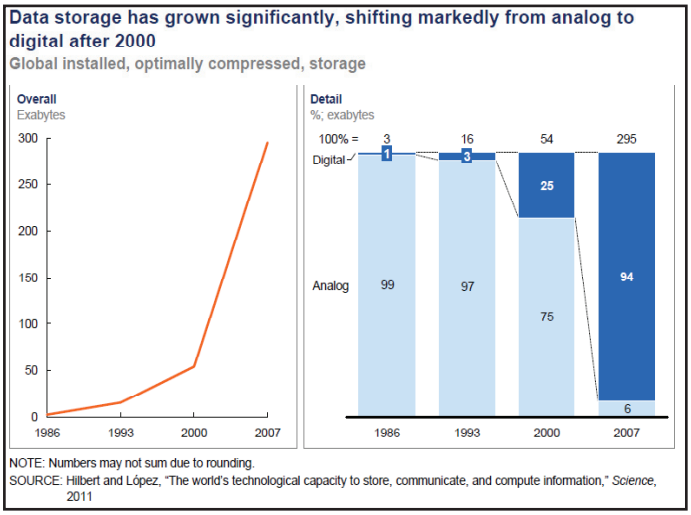

Fig. 2: Variation of Data from Analog to Digital

IV. Benefits of Big Data

Big Data is significant to our life and it provides some benefits to us like:

- Using the information kept in the social network like Facebook, the marketing agencies are guiding themselves about the response made by the people for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are scheduling their production.
- Using the data regarding the previous medical history of patients, hospitals are providing the expert systems for better and quick service.

V. Big Data Technologies

Big data is an umbrella term describes the collection of datasets that cannot be processed using traditional computing techniques. In order to manage the big data various tools, technqiues and frameworks are required.Big data technologies are important to process huge volumes of structured and unstructured data in realtime in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. The two classes of technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data are:

A. Operational Big Data

1. System like MongoDB

Provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.

2. NoSQL Big Data systems

Designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

3. NoSQL Systems

Provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

B. Analytical Big Data

Massively Parallel Processing (MPP) database systems and MapReduce provide analytical capabilities for retrospective and complex analysis.

MapReduce provides a new method of analyzing data that is opposite to the capabilities provided by SQL, and a system based on MapReduce can be scaled up from single servers to thousands of high and low end machines.

These two classes of technology are complementary and frequently deployed together.

Table 1: Operational vs. Analytical Systems

	Operational	Analytical
Time period	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce,MPP Database

VI. Big Data Challenges

The major challenges associated with big data are as follows:

- Capturing data
- Curation
- Storage
- Searching
- Sharing
- Transfer
- Analysis
- Presentation

Organizations normally take the help of enterprise servers to fulfill the challenges.

A. Traditional Approach

In this approach, an enterprise will have a computer to store and process big data. In this data stored in Oracle Database, MS SQL Server or DB2 and sophisticated softwares can be written to interact with the database, process the required data and give it to the users for analysis purpose.

B. Limitation

This approach works well on less volume of data that can be supported by standard database servers, or up to the limit of the processor which is processing the data. It is not suitable to huge data sets.

C. Google's Solution

Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small partsand assigns those parts to many computers connected over the network, and collects the results to form the final result dataset.

D. Hadoop

Doug Cutting, Mike Cafarella and team took the solution provided by Google and startedan Open Source Project called HADOOP in 2005 and Doug named it after his son's toy elephant. Now Apache Hadoop is a registered trademark of the Apache Software Foundation.

Hadoop uses the MapReduce algorithm to run its applicationswhere the data is processed in parallel on different CPU nodes. In short, Hadoop is capable enough to develop applications which can run on clusters of computers and they could perform complete statistical analysis for a huge amounts of data.

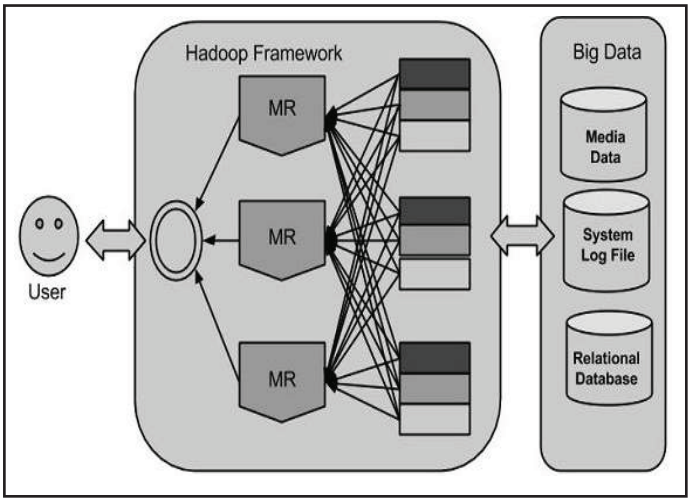

Fig. 3: Hadoop Framework

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. Hadoop is designed to to scale up from single to many many machines providing storage and computation.

E. Hadoop Architecture

Hadoop framework includes following four modules:

- **Hadoop Common:** It is the Java libraries and utilities which are required by other Hadoop modules. These libraries provide filesystem, abstractions at OS level, containsthe necessary Java files and scripts required to start Hadoop.
- **Hadoop YARN:** It is required for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS™):** A distributed file system provides high-throughput access to application data.
- **Hadoop MapReduce:** This is for parallel processing of large data sets and it is YARN-based system.

Four components in Hadoop framework are:

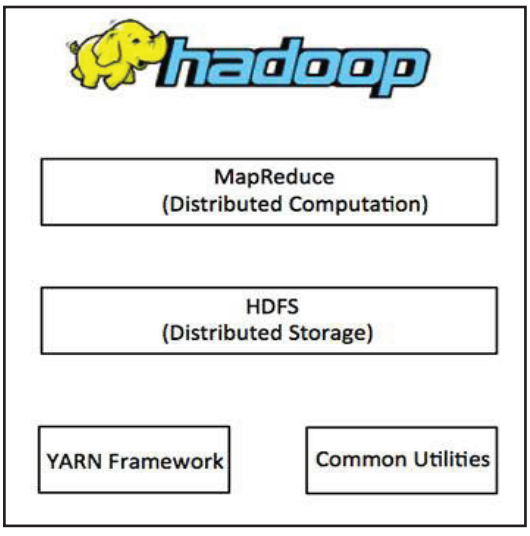

Fig. 4: Components of Hadoop

Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

F. MapReduce

It is a software framework used for writing the applications for processing large amount of data on large clusters in parallel mode to provide reliability and fault tolerance.

Tasks of MapReduce that Hadoop programs perform:

- **The Map Task:** It is the initial task which converts the input data into a set of data, which comprises of of breaking the individual elements are into tuples (key/value pairs).
- **The Reduce Task:** It takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

Both the input and the output are stored in a file-system. The framework consists of scheduling tasks, monitoring them and re-executes the failed tasks.

The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves TaskTracker execute the tasks as directed by the master and provide task-status information to the master periodically.

The JobTracker is a single point of failure for the Hadoop MapReduce service in which if JobTracker goes down, all running jobs are halted.

G. Hadoop Distributed File System

Hadoop can work directly with any mountable distributed file system such as Local FS, HFTP FS, S3 FS, and others, but the most common file system used by Hadoop is the Hadoop Distributed File System (HDFS).

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner.

HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data.

A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.

HDFS provides a shell like any other file system and a list of commands are available to interact with the file system. These shell commands will be covered in a separate chapter along with appropriate examples.

H. Hadoop's Working

Stage 1

A user/application submits a job to the Hadoop (a hadoop job client) for required process by specifying the following items:

- The location of the input and output files in the distributed file system.
- The java classes in the form of jar file containing the

implementation of map and reduce functions.
- The job configuration by setting different parameters specific to the job.

Stage 2
The Hadoop job client then submits the job (jar/executable etc) and configuration to the JobTracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

Stage 3
The TaskTrackers on different nodes execute the task as per MapReduce implementation and output of the reduce function is stored into the output files on the file system.

References
[1] [Online] Available: https://www.Big Data Analytics Tools Overview & Benefits _ Qubole.htm
[2] Marko Grobelnik marko.grobelnik@ijs.si Jozef Stefan Institute Ljubljana, Slovenia Stavanger, May 8th 2012
[3] An Oracle White Paper June 2013 Oracle: Big Data for the Enterprise
[4] [Online] Available: http://www.tutorialspoint.com/hadoop//hadoop_introduction.htm.
[5] [Online] Available: http://www.simplilearn.com/Big-Data_Analytics.htm.
[6] Viktor Mayer-Schönberger, Kenneth Cukier, Big Data: A Revolution that Will Transform how We Live, Work, and Think, 2013.
[7] Andrew McAfee and Erik Brynjolfsson ,Big Data: The Management Revolution, October 2012 Harvard Business Review
[8] Huawei Noah's Ark Lab, Hong Kong Science Park, Shatin, Hong Kong, Mining big data: current status, and forecast to the future, Vol. 14, Issue 2, pp. 1-5, December 2012.
[9] Meng Xiaofeng, Ci Xiang (School of Information, Renmin University of China, Beijing 100872)Big Data Management: Concepts, Techniques and Challenges,Meng Xiaofeng and Ci Xiang (School of Information, Renmin University of China, Beijing 100872) in Journal of Computer Research and Development 2013-14.
[10] Amit Verma et al.,"Software Reusability and Maintenance", National Conference on Emerging Trends in Communication held at SVIET-BANUR(District Patiala, PUNJAB) pp. 72, on 20th -21st February, 2009.
[11] Amit Verma et al.,"Risk of Passport Single Sign on Protocols", All India Seminar on Emerging Trends in Wireless Communication-Vision 2020 Sponsored by Institute of Engineers, pp. 73-81, Held at DIET-Kharar (PUNJAB) on 13th -14th, March 2009.