Overview of Big Data and Hadoop

Nupur N Mall, Shikha & Sheetal Rana
Department of Computer Science & Engineering, IIMT College of Engineering, Greater Noida

Abstract: *The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyse large datasets that cannot be processed using traditional computing techniques with high velocity. Big data can be structured, unstructured or semi-structured, and can consist of extensible variety of data. Big Data can be handled by Operational or Analytical classes of system. Hadoop is an Apache open-source framework written in Java that allows distributed processing of large datasets across clusters of computers using single programming model. Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data and scale up degree of fault tolerance.*

1. Introduction

1.1 Big Data: Definition

Big Data is an evolving term that describes any voluminous data structured, unstructured or semi-structured that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily. Big Data Analytics is often associated with Cloud Computing because the analysis of large data sets in real-time requires a platform like Hadoop.

While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes to multiple petabytes. Figure 1 shows an architectural vision of Big Data system.

1.2 3Vs of Big Data

The 3Vs of big data are three defining properties or dimensions of big data.

Volume- Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes to petabytes.

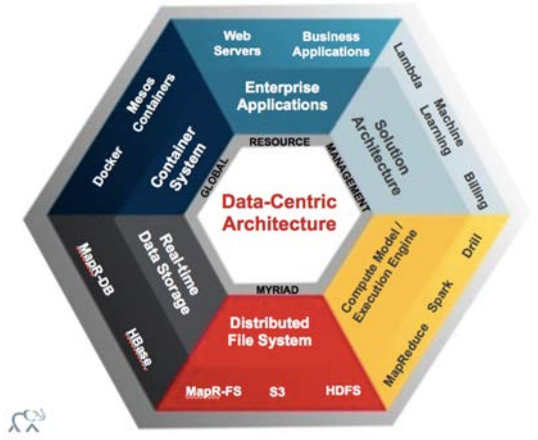

Figure 1: Architectural Vision of Big Data

Variety- Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi-structured, audio, video, XML, etc.

Velocity- Speed of data processing. Big data must be used as it streams into your enterprise in order to maximise its value.

Figure 2 shows the 3Vs characteristics of Big Data. According to the 3Vs model, the challenges of big data management result from the expansion of all three properties, rather than the volume alone.

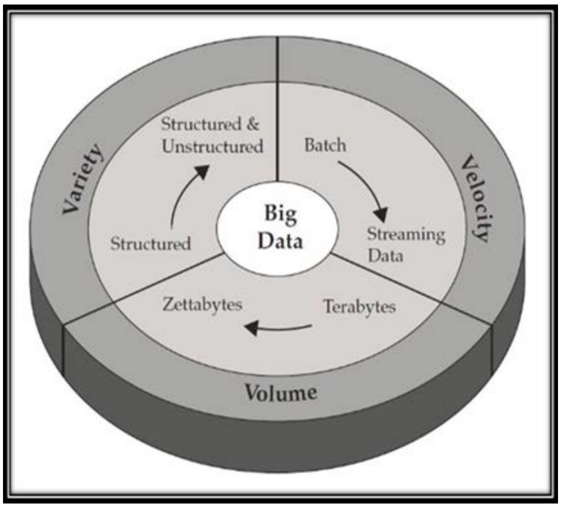

Figure 2: The 3Vs of Big Data

Other big data characteristics are:

Variability- Inconsistency of the data set can hamper processes to handle and manage it.

Veracity- The quality of captured data can vary greatly, affecting accurate analysis.

1.3 Importance of Big Data

The importance of Big Data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyse it to find answers that enable cost reductions, time reductions, new product development and optimized offerings, and smart decision making. When you combine big data with high powered analytics, you can accomplish business-related tasks such as:

- Determining root causes of failures, issues and defects in near-real time.
- Generating coupons at the point of sale based on the customer's buying habits.
- Recalculating entire risk portfolios in minutes.
- Detecting fraudulent behaviour before it affects your organization.

1.4 Benefits of Big Data

Big data has various practical benefits. Some of these benefits are:

- **Re-develop your products-** Big Data can help you understand how others perceive your products so that you can adapt them, or your marketing, if need be. Analysis of unstructured social media text allows you to uncover the sentiments of your customers and even segment those in different geographical locations or among different demographic groups.

- **Perform risk analysis-** Success not only depends on how you run your company. Social and economic factors are crucial for your accomplishments as well. Predictive analytics, fuelled by Big Data allows you to scan and analyse newspaper reports or social media feeds so that you permanently keep up to speed on the latest developments in your industry and its environment.

- **Keeping your data safe-** You can map the entire data landscape across your company with Big Data tools, thus allowing you to analyse the threats that you face internally. You will be able to detect potentially sensitive information that is not protected in an appropriate manner.

- **Reducing maintenance costs-** Big Data tools do away with such unpractical and costly averages. The massive amounts of data that they access and use and their unequalled speed can spot failing grid devices and predict when they will give out.

- **Customizing your websites in real time-**Big Data analytics allow you to personalise the content or look and feel of your website in real time to suit each consumer entering your website, depending on, for instance their sex, nationality or from where they ended up on your site.

1.5 Problems with Big Data

- **Heterogeneity and Incompleteness-** When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogenous data, and cannot understand nuance. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure.

- **Scale-** The first thing that pops into mind while thinking about Big Data is size. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by the processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now; data volume is scaling faster than compute resources, and CPU speeds are static.

- **Timeliness-** The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyse. The design of a system that effectively deals with size is likely also to result in a system that can process a given amount of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in context of Big Data, Rather, there is an acquisition rate challenge.

- **Privacy-** The privacy of data is another huge concern, and one of that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations are less forceful. However, there is a great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources.

Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realise the promise of big data.

- **Human collaboration-** In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

2. Hadoop: Solution for Big Data Processing

Doug Cutting, Mike Caferella and team started an open source project, using the solution given by Google for Big Data, called Hadoop in 2005. Doug named it Hadoop after his son's toy elephant.
Now, Apache Hadoop is a registered trademark of Apache Software Foundation.

2.1 Hadoop: Definition

Hadoop is an Apache open source framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework.
The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce.
Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data.

2.2 Hadoop Architecture

The figure 3 shows the four components available in Hadoop framework.

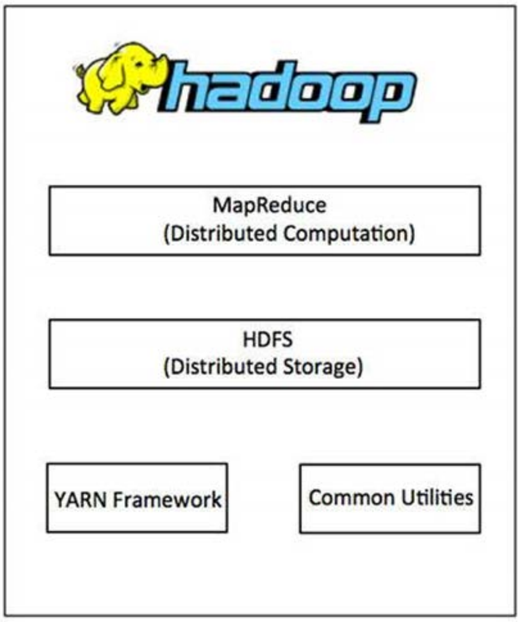

Figure 3: Hadoop Architecture

Hadoop framework includes the following four modules:

- **Hadoop Common-** These are Java libraries and utilities required by other Hadoop modules. These libraries provide file-system and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

- **Hadoop YARN-** This is a framework for job scheduling and cluster resource management.

- **Hadoop Distributed File System (HDFS)-** A distributed file system that provides high-throughput access to application data.

- **Hadoop MapReduce-** This is YARN-based system for parallel processing of large data sets.

2.3 HDFS

HDFS is a Java-based file system, developed using distributed file system design, that provides scalable and reliable data storage. It was designed to span clusters of commodity servers. It is run on commodity hardware.

Unlike other distributed systems, HDFS is highly fault-tolerant and designed using low-cost hardware.

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing. HDFS has demonstrated high production scalability.

- **Features of HDFS**

→ It is suitable for distributed storage and processing.
→ Hadoop provides a command interface to interact with HDFS.
→ Streaming access to file system data.
→ HDFS provides file permissions and authentication.

- **HDFS Architecture**

HDFS follows the master-slave architecture and it has the following elements.

1. **Namenode-** The namenode acts as the master server and performs the task of managing the file system namespace, regulating client's access to files, executing file system operations such as renaming, closing, and opening files, etc.

2. **Datanode-** Datanode acts as a slave to the namenode. It performs read-write operations on the file systems, as per client request. They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

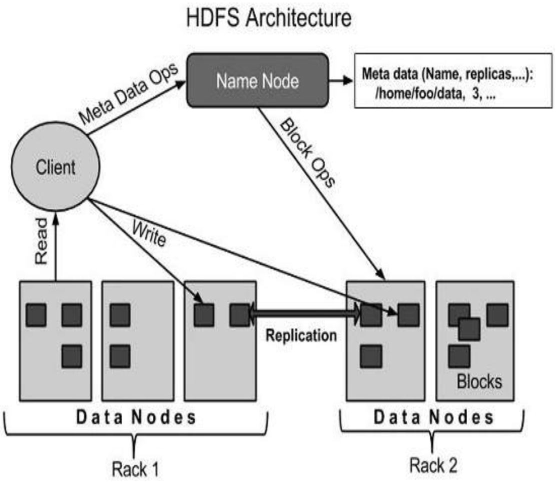

Figure 4: HDFS Architecture

3**. Block-** Generally, the user data is stored in files of HDFS. The file in a file system is divided into one or more segments and/or stored in individual data nodes. These file segments are called blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block.

- **Goals of HDFS**

→ Fault detection and recovery
→ Huge datasets
→ Hardware at data

2.4 MapReduce

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data in- parallel on large clusters of commodity hardware in a fault-tolerant and reliable manner. It is the processing pillar in the Hadoop ecosystem. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse. There are two functions in MapReduce as follows:

map- the function takes key/value pairs as input and generates an intermediate set of key/value pairs.

reduce- the function which merges all the intermediate values associated with the intermediate key.

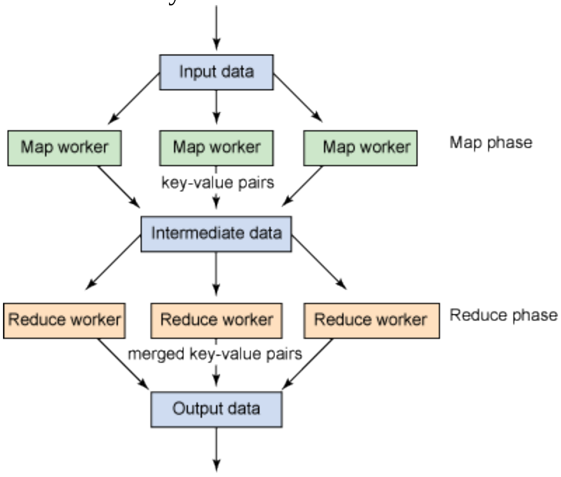

Figure 5: MapReduce Architecture

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 to slaves whereas the slave TaskTracker executes the tasks as directed by the master and provides information to the master periodically.

2.5 Advantages of Hadoop

- **Scalability-** Hadoop is a highly scalable platform, largely due to its ability to store as well as distribute large data across plenty of servers. Contrary to the traditional relational database management systems (RDBMS), Hadoop enables business organisations to run applications from a huge number of nodes that could involve the usage of thousands of terabytes of data.

- **Cost-effective solution-** Hadoop's scale-out architecture with MapReduce programming allows the storage and processing of data in a very affordable manner. It can be used in later times. In fact, the cost savings are massive and the costs can reduce from thousands figures to hundreds figures for every terabyte of data.

- **Flexibility-** Hadoop offers support for numerous languages that can be used for data processing and storage.

- **Fast-** The tools used for data processing, such as MapReduce programming, are generally located in the same servers as the files in a distributed file system, which allows for faster processing of data. Hadoop takes minutes to process large volumes of unstructured data.

- **Parallel processing-** One of the primary aspect of the working of MapReduce programming is that it divides tasks in a manner that allows their execution in parallel. Parallel processing allows multiple processors to take on these divided tasks, such that they run entire programs in less time.

3. Conclusion

We entered an era of Big Data. The paper describes the concept of Big Data along with 3Vs, Volume, Velocity and Variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualisation, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data. The paper also focuses on Hadoop architecture i.e. HDFS and MapReduce along with the advantages of Hadoop.

4. References

[1] S.Vikram Phaneendra & E. Madhusudan Reddy "Big Data- Solutions for RDBMS problems- A Survey" in 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Apr 19{23 2013}).

[2] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).

[3] Jonathan paul Olmsted: "Scaling at Scale: Ideal Point Estimation with 'Big Data'" Princeton Institute for Computational Science and Engineering 2014.

[4] Niketan Pansare1, Vinayak Borkar2, Chris Jermaine1, Tyson condie "Online Aggregation for Large MapReduce Jobs" August 29 September 3, 2011, Seattle, WA Copyright 2011 VLDB endowment, ACM.

[5] Tom White "Hadoop: The Definitive Guide".