A Survey Paper on Big Data and Hadoop

Jigisha Trivedi[1]

Assistant Professor, Computer Engineering Department,S.B. Polytechnic, Savli, Vadodara, India

Abstract

We live in an era where data is being generated by everything around us. The rate of data generation is so alarming, that it has engendered a pressing need to implement easy and cost-effective data storage mechanisms. Furthermore, big data needs to be analyzed for insights and attribute relationships, which can lead to better decision- making and efficient business strategies. The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte- or larger-sized datasets with high-velocity and different structures. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.

INTRODUCTION

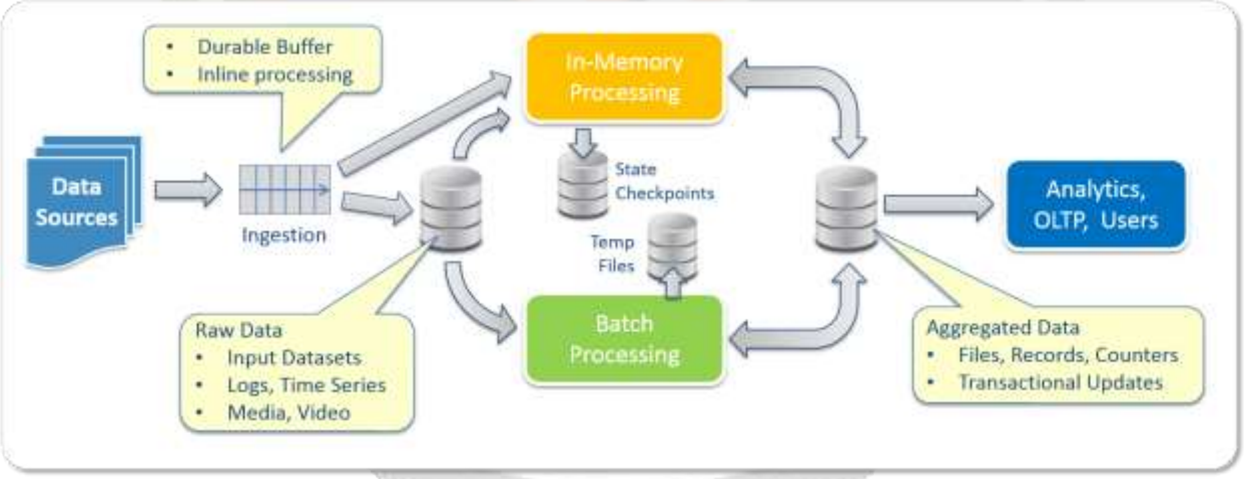

Figure 1: Architecture of Big Data System

Big Data:

Big data is a collection of large datasets- structured, unstructured or semi-structured that is being generated from multiple sources at an alarming rate. Key enablers for the growth of big data are – increasing storage capacities, increasing processing power and availability of data. It is thus important to develop mechanisms for easy storage and retrieval. Some of the fields that come under the umbrella of big data are - stock exchange data (includes buying and selling decisions), social media data (Facebook and Twitter), power grid data (contains information about the power consumed by each node in a power station) and search engine data (Google). Structured data may include relational databases like MySQL. Unstructured data may include text files in .doc, .pdf formats as well as media files.The definition of Big Data contains three different terms.

We can say it Power of 3V's of Big Data which are defined as-

Volume of data: Numerous independent market and research studies have found that data volumes are doubling every year. On top of all this extra new information, a significant percentage of organizations are also storing three or more years of historic data.
Variety of data: Studies also indicate that 80 percent of data is unstructured (such as images, audio, tweets, text messages, and so on). And until recently, the majority of enterprises have been unable to take full advantage of all this unstructured information.
Velocity of data: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Challenges of big data

Big data also has its own unique set of obstacles such as:[15]
Information Growth: Over 80 percent of the data in the enterprise consists of unstructured data, which tends to be growing at a much faster pace than traditional relational information. This massive information threaten to swamp all but the most well-prepared IT organizations.[18]

Processing power: The customary approach of using asingle, expensive, powerful computer to crunch information just doesn't scale for Big Data. As we soon see, the way to go is divide-and-conquer using commoditized hardware and software via scale-out.[6]

Physical storage: Capturing and managing all this information can consume enormous resources, outstripping all budgetary expectations.

Data issues: Lack of data mobility, proprietary formats, and interoperability obstacles can all make working with Big Data complicated.[3]

Cost: Extract, transform, and load (ETL) processes for Big Data can be expensive and time consuming, particularly in the absence of specialized, well-designed software.[8]

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. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. 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.

Benefits of Big Data

Analysis of big data helps in improving business trends, finding innovative solutions, customer profiling and in sentimental analysis. It also helps in identifying the root causes for failures and re-evaluating risk portfolios. In addition, it also personalizes customer and interaction.

Valuable Insights

Valuable insights can be derived from big datasets by employing proper tools and methodologies. This data
includes those stored in the company database, or those obtained from social media and other third party sources. When data is processed and analyzed, one can draw valuable relationships between various attributes that can improve the quality of decision making. Statistics and industrial knowledge can be combined to obtain useful insights.

New Products and Services

Analyzing big data helps the organization to understand how customers perceive their products and services. This aids in developing new products that are concurrent with customer needs and demands. In addition, it also facilitates re- developing of currently existing products to suit customer requirements.

Smart cities

Population increase begets demand. To help cities deal with the consequences of rapid expansion, big data is being used for the benefit of the citizens and the environment. For example, the city of Portland, Oregon adopted a mechanism for optimizing traffic signals in response to high congestion. This not only reduced traffic jams in the city, but was also significant in eliminating 157,000 metric tons of carbon dioxide emissions.

Risk Analysis

Risk is defined as the probability of injury or loss. Risk management is a very crucial process which is often over-looked. Frequent analysis of the data will help mitigate potential risks. Predictive analysis aids the organization to keep up to date with recent technologies, services and products. It also identifies the risks involved, and how they can be mitigated.

Miscellaneous

Big data also aids Media, Government, Technology, Scientific Research and Healthcare in making crucial decisions and predictions. For example, Google Flu Trends (GFT) provided estimates of influenza activity for more than 25 countries. It made accurate predictions about flu activity.

Limitations Of Traditional Approach

The traditional approach consists of a computer to store and process big data. Data is stored in a Relational Database like MySQL and Oracle. This approach works well when the volume of data is less. However, when dealing with larger volumes of data, it becomes tedious to process it through a database server. Hence, this calls for a more sophisticated approach. We will now look into Hadoop – its modules, framework and ecosystem.

Hadoop: Solution for Big Data Processing

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google’s MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points. Apache Hadoop is an open source software framework for storing and processing large clusters of data. It has extensive processing power and it consists of large networks of computer clusters. Hadoop makes it possible to handle thousands of terabytes of data. Hardware failures are automatically handled by the framework.

Apache Hadoop consists of 4 modules:.

Hadoop Distributed File System(HDFS)
Hadoop MapReduce
Hadoop YARN
Hadoop Common

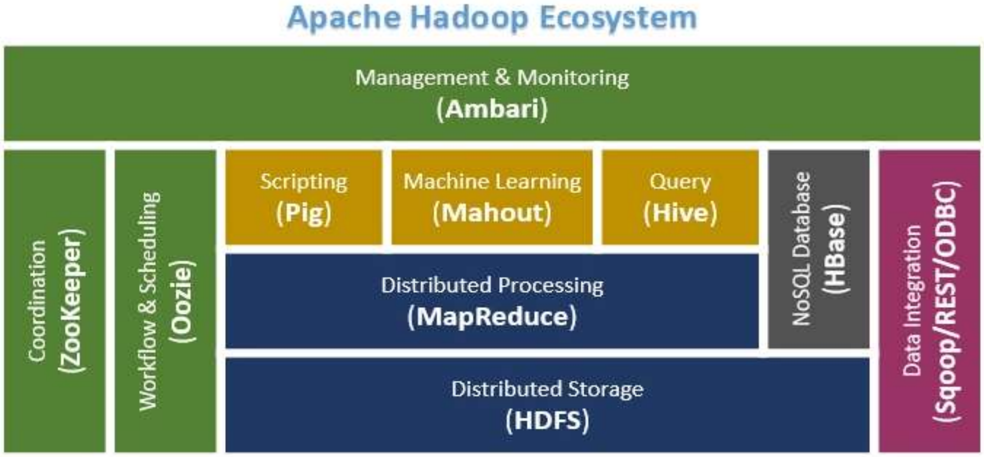
Apache Hadoop Ecosystem
Management & Monitoring (Ambari)
Coordination (ZooKeeper)
Workflow & Scheduling (Oozie)
Scripting (Pig)
Machine Learning (Mahout)
Query (Hive)
Distributed Processing (MapReduce)
Distributed Storage (HDFS)
NoSQL Database (HBase)
Data Integration (Sqoop/REST/ODBC)

Figure 2: Hadoop Architecture

Hadoop Distributed File System(HDFS)

Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data.

Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers.If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster.

HDFS manages storage on the cluster by breaking incoming files into pieces, called “blocks,” and storing each of the blocks redundantly across the pool of servers.

In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.Apache Hadoop uses the Hadoop Distributed File System.
It is highly fault tolerant and uses minimal cost hardware. It consists of a cluster of machines, and files are stored across them. It also provides file permissions and authentication, and streaming access to system data.

The following figure depicts the general architecture of HDFS

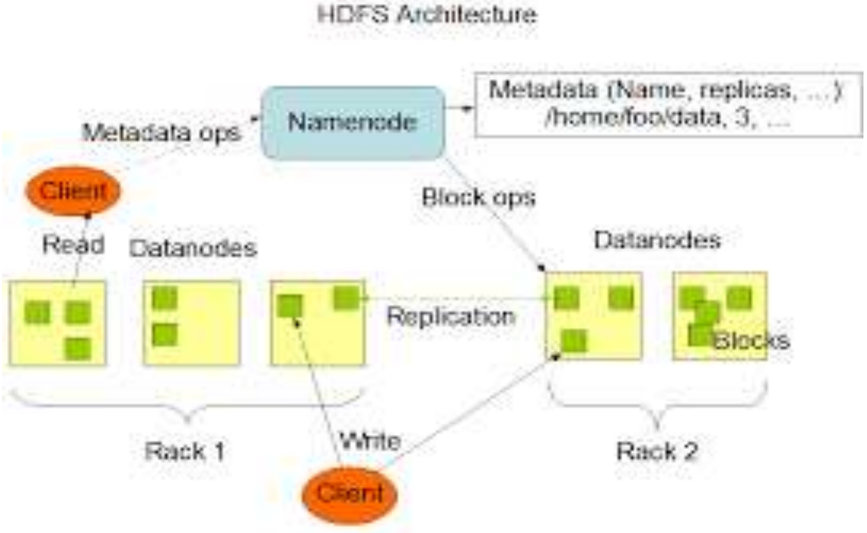

Figure 3: HDFS Architecture

HDFS follows the Master- Slave Architecture. It has the following components.

Name node

The HDFS consists of a single name node, which acts as the master node. It controls and manages the file system namespace. A file system namespace consists of a hierarchy of files and directories, where users can create, remove or move files based on their privilege. A file is split into one or more blocks and each block is stored in a Data node. HDFS consists of more than one Data Nodes.

The roles of the name node are as follows:

Mapping blocks to their data nodes.
Managing of file system namespace
Executing file system operations- opening, closing and renaming of files.
Data node

The HDFS consists of more than one data node. The data nodes store the file blocks that are mapped onto it by the Name node. The data nodes are responsible for performing read and write operations from file systems as per client

request. They also perform block creation and replication. The minimum amount of data that the system can write or read is called a block. This value however is not fixed, and it can be increased.

Conclusion

We have entered an era of Big Data. The paper describes the concept of Big Data along with 3 Vs, 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 visualization, at all stages of the analysis pipeline from data acquisition to result interpretation.This paper starts off by giving a formal definition to Big Data. Then, the challenges of handling big data are examined, followed by the limitations of using the traditional big data processing approach.

REFERENCES

1. Bakshi, K.,(2012),” Considerations for big data: Architecture and approach” Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012)
2. “Shared disk big data analytics with Apache Hadoop” Harshawardhan S. Bhosale1, Prof. Devendra Gadekar, JSPM’s Imperial College of Engineering & Research, Wagholi, Pune,a review on Big Data Aditya B. Patel, Manashvi Birla, Ushma Nair,(6-8 Dec. 2012)
3. ” Big Data Processing in Cloud Computing Environments” Garlasu, D.; Sandulescu, V; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013)
4. ”A Big Data implementation based on Grid Computing”, Grid Computing Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013)
5. “Big Data Analytics using Hadoop”, International Journal of Computing Applications, Volume 108, No.12, December 2014 Ms. Gurpreet Kaur,Ms. Manpreet Kaur
6. “Review Paper on Big Data using Hadoop”, International Journal of Computing Engineering and Technology, Volume 6,Issue 12, Dec 2015, pp. 65-71
7. Harshwardhan S. Bhosale et al, “Review paper on Big Data using Hadoop”, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014
8. Poonam S. Patil et al. “Survey Paper on Big Data Processing and Hadoop Components”, International Journal of Science and Research, Volume 3, Issue 10, October 2014.
9. Abhishek S, “Big Data and Hadoop”, White Paper
10. Konstantin Shvachko et.al, “The Hadoop Distributed File System”

11. S.Vikram Phaneendra & E.Madhusudhan Reddy “Big Data- solutions for RDBMS problems- A survey” In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).
12. Kiran kumara Reddi & Dnvsl Indira “Different Technique to Transfer Big Data : survey” IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}