Research Paper on Big Data and Hadoop

[1]Iqbaldeep Kaur, [2]Navneet Kaur, [3]Amandeep Ummat, [4]Jaspreet Kaur, [5]Navjot Kaur
[1,2,3,4,5]Dept. of CSE, Chandigarh Engineering College, Landran, Punjab, India

Abstract
'Big Data' describes techniques and technologies to store, distribute, manage and analyze large-sized datasets with high-velocity. 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.. 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 with a very high degree of fault tolerance.

Keywords
Big Data, Hadoop, Map Reduce, HDFS, Hadoop Components

I. Introduction

A. Big Data

It refers to the efficient handling of large amount of data that is impossible by using traditional or conventional methods such as relational databases or it is a technique that is required to handle the large amount of data that is generated with advancements in technology and increase in population. Big data helps to store, retrieve and modify these large data sets. For example with the advent of smart technology there is rapid increase in use of mobile phones due to which large amount of data is generated every second, so it is impossible to handle by using traditional methods hence to overcome this problem big data concepts were introduced. most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple peta-bytes (1015 or 1000 terabytes per peta-byte) as big data. Figure No. 1.1 gives Layered Architecture of Big Data System.

II. 3 Vs of Big Data

A. Data Volume

Data volume refers to the amount of data. At present the volume of data stored has grown from megabytes and gigabytes to peta-bytes and is supposed to increase to zeta-bytes in nearby future.

B. Data Variety

Variety refers to the different types of data-– text, images video, audio, etc and sources of data. Data being produced is not of single category as it not only includes the traditional data but also the semi structureddata from various resources like web Pages, Web Log Files, social media sites, e-mail, documents

C. Data Velocity

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows and aggregated.

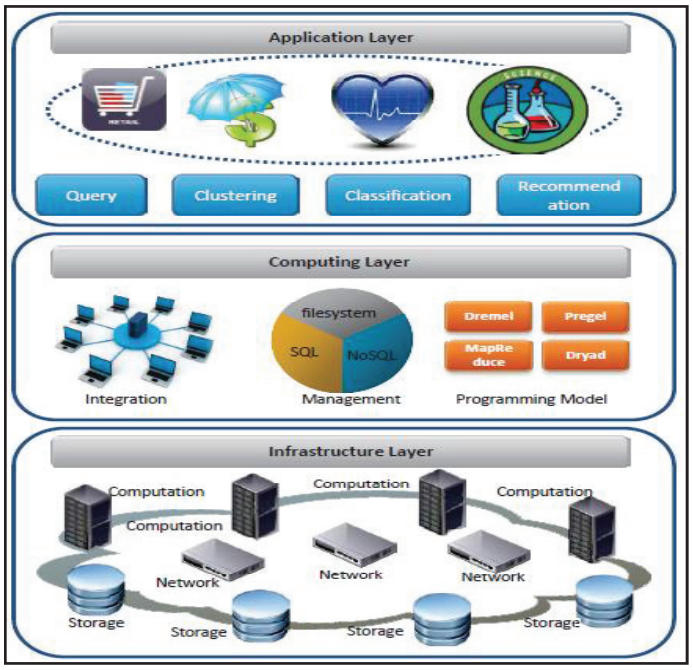

Fig. 1: Layered Architecture of Big Data System

III. Problem or Challenges associated with Big Data Processing

The challenges in Big Data are usually the real hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant results

A. Size

The first thing anyone thinks of with Big Data is its size. The word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by 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.

B. Privacy and Security

It is the most important challenges with Big data which is sensitive.

- The personal information (e.g. in database of social networking website) of a person when combined with external large data sets, leads to the inference of new facts about that person and it's possible that these kinds of facts may be secretive and the person might not want the data owner to know or any person to know about them.
- Information regarding the people is collected and used in order to add value to the business of the organization. Another important consequence arising would be Social sites where a person would be taking advantages ofthe Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.
- Big Data increase the chances of certain tagged people to

suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

C. Data Access and Sharing of Information

Due to huge amount of data, data management and governance process is bit complex adding the necessity to make data open and make it available to government agencies in standardized manner with standardized APIs, metadata and formats.

Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness

D. Analytical Challenges

The main analytical challenging questions are as:

- What if data volume gets so large and varied and it is not known how to deal with it?
- Does all data need to be stored?
- Does all data need to be analyzed?
- How to find out which data points are really important?
- How can the data be used to best advantage?

Big data brings along with it some huge analytical challenges. The type of analysis to be done on this huge amount of data which can be unstructured, semi structured or structured.

E. Human Resources and Manpower

Since Big data is an emerging technology so it needs to attract organizations and youth with diverse new skill sets. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive andcreative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations. Moreover the Universities need to introduce curriculum on Big data to produce skilled employees in this expertise.

F. Technical Challenges

1. Fault Tolerance

With the incoming of new technologies like Cloud computing and Big data it is always intended that whenever the failure occurs the damage done should be acceptable.

Fault-tolerant computing is extremely hard, involving intricate algorithms. Thus the main task is to reduce the probability of failure to an "acceptable" level.

Two methods which seem to increase the fault tolerance in Big data are as:

- First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation.
- Second is, one node is assigned the work of observing that these nodes are working properly. If something happens that particular task is restarted. But sometimes it's quite possible that that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the output of the previous computation of task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

2. Scalability

The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters.

This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks. There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having the same performance between sequential and random data transfer. Thus, what kinds of storage devices are to be used; is again a big question for data storage.

3. Quality of Data

Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

4. Heterogeneous Data

Unstructured data represents almost every kind of data being produced like social media interactions, to recorded meetings, to handling of PDF documents, fax transfers, to emails and more. Working with unstructured data is a cumbersome problem and of course costly too.

Converting all this unstructured data into structured one is also not feasible.

Structured data is always organized into highly mechanized and manageable way. It shows well integration with database but unstructured data is completely raw and unorganized.

IV. Hadoop: Solution for Big Data Processing

Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. Hadoop is an open source project It consists of many small sub projects which belong to the category of infrastructure for distributed computing.

Hadoop mainly consists of:

- File System (The Hadoop File System)
- Programming Paradigm (Map Reduce)

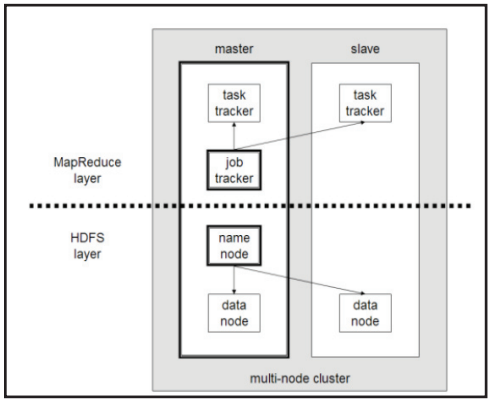

Fig. 2: Hadoop Architecture

A. HDFS Architecture

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.

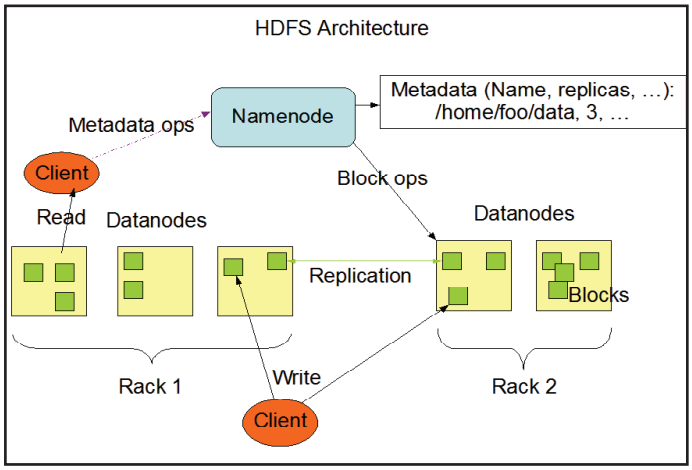

Fig. 3: HDFS Architecture

B. Map Reduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional data warehousing scenario, this might entail applying an ETL operation on the data to produce something usable by the analyst. 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 same intermediate key

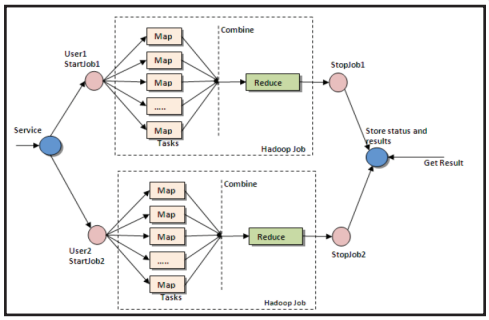

Fig. 4: MapReduce Architecture

C. Big Data Advantages and Good Practices

The Big Data has numerous advantages on society, science and technology. Some of the advantages (Marr, 2013) are described below:

1. Understanding and Targeting Customers

This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models.

2. Understanding and Optimizing BusinessProcess

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web searchtrends and weather forecasts.HR business processes are also being improved using big data analytics.

3. Improving Science and Research

Science and research is currently being transformed by the new possibilities big data brings. Take, for example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of our universe – how it started and works - generate huge amounts of data. The CERN data centre has 65,000 processors to analyse its 30 petabytes of data. However, it uses the computing powers of thousands of computers distributed across 150 data centres worldwide to analyse the data. Such computing powers can be leveraged to transform so many other areas of science and research.

4. Improving Healthcare and Public Health

The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns.

5. Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

6. Financial Trading

High Frequency Trading (HFT) is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make buy and sell decisions in split seconds.

7. Improving Security and Law Enforcement

Big data is applied heavily in improving security and enabling law enforcement. The revelations are that the National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us). Others use big data techniques to detect and prevent cyber-attacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data use it to detect fraudulent transactions.

Some other advantages and uses includes improving sports performance, improving and optimizing cities and countries, personal quantification and performance optimization.

8. Good Practices for Big Data

- Creating dimensions of all the data being store is a good practice for Big data analytics. It needs to be divided into dimensions and facts.
- All the dimensions should have durable surrogate keys meaning that these keys can't be changed by any business rule and are assigned in sequence or generated by some hashing algorithm ensuring uniqueness.
- Expect to integrate structured and unstructured data as all kind of data is a part of Big data which needs to be analyzed together.
- Generality of the technology is needed to deal with different formats of data. Building technology around key value pairs work.
- Analyzing data sets including identifying information about individuals or organizations privacy is an issue whose importance particularly to consumers is growing as the value of Big data becomes more apparent.
- Data quality needs to be better. Different tasks like filtering, cleansing, pruning, conforming, matching, joining, and diagnosing should be applied at the earliest touch points possible.
- There should be certain limits on the scalability of the data stored.
- Business leaders and IT leaders should work together to yield more business value from the data. Collecting, storing and analyzing data comes at a cost. Business leaders will go for it but IT leaders have to look for many things like technological limitations, staff restrictions etc. The decisions taken should be revised to ensure that the organization is considering the right data to produce insights at any given point of time.
- Investment in data quality and metadata is also important as it reduces the processing time.

V. Conclusion

There is a potential for making faster advancements in scientific descipline for analysing the large amount of data .The technical challenges are most common across the large variety of application domains, therefore new cost effective and faster methods must be implemented to analyse the big data.

References

[1] S.Vikram Phaneendra, E.Madhusudhan Reddy,"Big Data-solutions for RDBMS problems- A survey", In 12thIEEE/ IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[2] Aveksa Inc. (2013). Ensuring "Big Data" Security with Identity and Access Management. Waltham, MA: Aveksa.

[3] Hewlett-Packard Development Company. (2012). Big Security for Big Data. L.P.: Hewlett-Packard Development Company.

[4] Kaisler, S., Armour, F., Espinosa, J. A., Money, W. (2013). Big Data: Issues and Challenges Moving Forward. International Confrence on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Soceity.

[5] Katal, A., Wazid, M., Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

[6] Marr, B. (2013, November 13). The Awesome Ways Big Data is used Today to Change Our World. Retrieved November 14, 2013, from LinkedIn: https://www.linkedin.com/today /post/ article/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-tochange-our-world

[7] Amit Verma et al.,"Cross Layer Feedback Design: Optimization For Energy Efficient Mobile Devices Protocols Stacks Over Wireless Sensor Networks", National Conference on Emerging Trends in Communication held at SVIET-BANUR(District Patiala, PUNJAB) pp. 90, on 20th -21st February, 2009.

[8] Amit Verma et al.,"Care 2000: Analysis and Re- Engineering ", National Conference on Emerging Trends in Communication held at SVIET-BANUR(District Patiala, PUNJAB) pp. 73, on 20th -21st February, 2009.