Volume: 3 Issue: X Month of publication: October 2015

DOI:

Volume 3 Issue X, October 2015
ISSN: 2321-9653

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Big Data Vs Traditional Data

Raveena Pandya[1], Vinaya Sawant[2], Neha Mendjoge[3], Mitchell D'silva[4]
[1,2,3,4]Department of Information Technology, D.J Sanghvi College of Engineering, Mumbai University

Abstract—* Computing devices have been storing, processing data growingly for decades but the rate of increase of the 'digital universe' has accelerated particularly in recent years, and now manifests exponential growth. Big data is a group of data sets which is very large in size as well as complex. Data sets grow in size because they are increasingly being gathered by cheap and numerous information-sensing devices. Generally size of the data is Petabyte and Exabyte. Traditional database systems are unable to store and analyse this massive amount of data. As the use of internet is increasing, amount of big data continues to grow. Big data analytics provides novel ways for businesses and government to analyse unstructured data. Lately, Big data is one of the most talked subject in IT industry and it is going to play an important role in the future. Big data alters the way the data is managed and used. Some of the applications of Big data are in areas such as healthcare, traffic management, banking, retail, education and so on. Organizations are becoming more lithe and more open. New varieties of data will give new difficulties as well. This paper highlights important concepts of big data.

***Keywords—* Traditional data, Big data, Hadoop, Map Reduce, NoSQL, HDFS**

I. INTRODUCTION

Earlier, the type of information available was limited. There was a distinct set of technology approaches for managing information. Nowadays the amount of data in our world has been increasing. Relational database management systems often have trouble managing big data. Big data can be described using following terms:

Volume: Small sized organizations may have terabytes or gigabytes of data storage. Data volume will continue to expand, irrespective of the organization's size. Many of these organisations datasets are within the terabytes range today but in a little while they could reach to petabytes or even exabytes. Machine generated data is more in volume than the traditional data.

Variety: Various kinds of data are captured. It may be structured, semi structures or unstructured. Earlier we only focused on structured data that effortlessly fitted into relational databases or tables, for example financial data. Nowadays, most of the data that is produced by an organization is unstructured. The extensive variety of data requires a different method to store all raw data. Some more examples of this variety include movies and sound files, images, documents, geo location data, text, web logs, strings, web contents etc[1].

Velocity: The data is arriving continuously as streams of data. Velocity is the speed at which new data is produced and the speed at which data moves around. The difficulty organizations have is to deal with the enormous speed the data is created and used in real-time [2].

Veracity: If the data coming in large volume is not accurate, it can be problematic and of no use. So, the data that is coming should be correct.

There are three types of big data.They are described as follows:

Structured data: Related entities are assembled together in structured data. Objects in the same group have the identical descriptions. Examples are words, figures etc. Relational databases and spreadsheets are examples of structured data.

Unstructured data: It is complicated information. Data can be of any type and does not follow any particular rule. It cannot be evaluated with standard statistical methods. For big data, diverse tools are required. Examples are social media, email, photos, multimedia etc.

Semi structured data: In semi structured data, similar entities are congregated together. Entities in same group may not have equal attribute. Emails, EDI are example of this type of data.

Big data is receiving lot of attention these days**.** It will aid to generate new growth openings and entirely new types of companies. Intelligence is pooled with the production process. Relentlessly refining processing power and innovative ways for data examination indicate that Big Data can be created from range of sources. The creation of Big Data therefore licences organizations to create information about data that were never intended or apparent in the source information. If big data is used accurately, enterprise can get an enhanced view on their business. Some of the evolving applications are traffic management system, healthcare system and many more. The big data can originate from numerous sources. It may be digitally created and can be stored using a series of ones and zeros, and can be handled by computers. It may be mobile phone's position data or call duration time or it may be a result of our daily lives or dealings with digital services. It may be created using unconventional methods outside of data entry like, RFID, Sensor networks etc. Figure 1[3] shows organizations which are implementing or executing big data.

II. BIG DATA ANALYSIS PIPELEINE

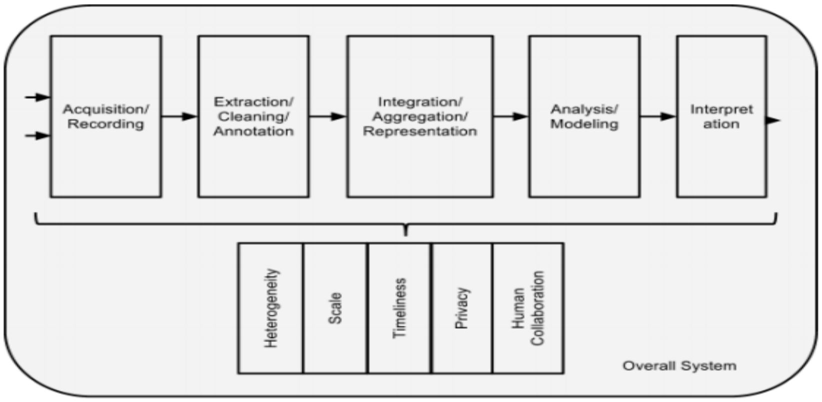

Fig 1:Big Data Analysis Pipeline

Phases in the Processing Pipeline are as follows:

A. Data Acquisition and Recording

Big Data does not come out of a vacuum: it is logged from some data producing source. For example, ruminate our capability to sense and perceive the world around us, from the heart rate of an aged citizen, and existence of toxins in the air we breathe, to the prearranged square kilometre array telescope, which will yield up to 1 million terabytes of raw data every day. Likewise, scientific tests and models can easily yield petabytes of data today.

B. Information Extraction And Cleaning

Commonly, the information gathered will not be in a format that is prepared for analysis. For example, consider the assortment of electronic health accounts in a hospital, including transcribed notations from quite a few physicians, structured data from sensors and measurements. We cannot effectively analyse data in this manner. Rather we need an information abstraction process that pulls out the essential information from the primary sources and presents it in a structured form

C. Data Integration, Aggregation, Representation

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. We need to integrate, aggregate and represent data effectively

D. Query Processing, Data Mining, Analysis

Mining in Big Data necessitates cleaned, integrated, trustworthy, and efficiently accessible data, , scalable mining algorithms ,declarative query and mining interfaces, and big-data computing environments.

E. Interpretations

Having the skill to analyse Big Data is of partial significance if users are not able to understand the analysis. Generally, it encompasses examining of all the conventions made and reviewing the analysis.

III.DIFFERENCE BETWEEN TRADITIONAL AND BIG DATA ANALYTICS

Big data analytics can be discerned from traditional data-processing architectures. In traditional data, sources are structured. In Big data analysis data quality and data normalization take place and the data is moulded into rows and columns. The modelled data is then consigned into an enterprise data warehouse. Big data is data that is excessively outsized to process using traditional methods. As the capacity of data bursts out, organizations will require analytic tools that are consistent, vigorous and proficient of being automated .Traditional data warehouse is unable to handle dispensation of big data as data is approaching from various sources like video etc. This type of data raises at very high speed. The database necessities are very diverse in big data. With big data analytics data can be in anyplace and in huge volume. Big data analytics delivers useful information. Concealed patterns are learnt. It emphases on unstructured data. Some technologies like Hadoop, NoSQL and Map Reduce are vital for the analytics of big data. In big data analytics, the Hadoop system apprehends datasets from diverse sources and then implements functions such as storing, cleansing, distributing, indexing, transforming, searching, accessing, analysing, and visualizing .So the unstructured data is transformed into structured data. The operational norm behind Hadoop and big data is to transfer the query to the data to be handled not the data to the query processor. Numerous languages used in the big data analytics are Oracle, Java, and JavaScript etc. Big data necessitates several approaches to analysis, traditional or advanced, reliant on the problem. It is subjected to the type of that individual problem. Some analytics takes into account traditional data warehouse theory. But some involves more advanced techniques. The IT tools to implement big data processing are new, very vital and exhilarating. Big data mechanism is faster in comparison to traditional data warehousing techniques. Figure 2. [4] Shows comparison of big data and traditional data.

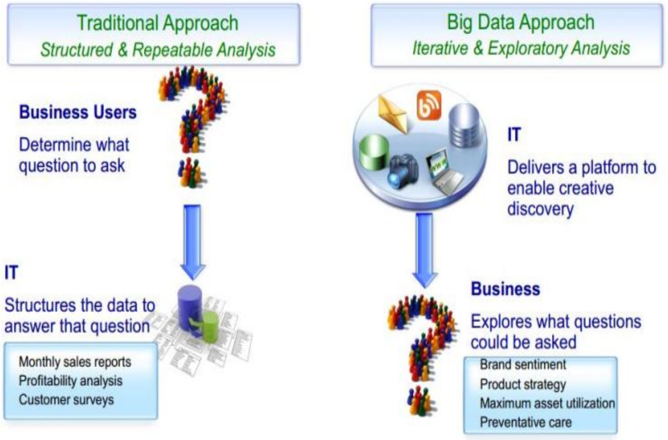
Traditional Approach
Structured & Repeatable Analysis
Business Users
Determine what question to ask
IT
Structures the data to answer that question
Monthly sales reports
Profitability analysis
Customer surveys
Big Data Approach
Iterative & Exploratory Analysis
IT
Delivers a platform to enable creative discovery
Business
Explores what questions could be asked
Brand sentiment
Product strategy
Maximum asset utilization
Preventative care

Fig 2: Traditional Data vs Big Data

IV.TECHNOLOGIES USED FOR BIG DATA ANALYTICS

A. NoSQL Database

NoSQL database is able to handle unpredictable and unstructured data. The data warehoused in a NoSQL database is usually of a great variety. A NoSQL database offers a mechanism for packing and recovery of data that is demonstrated in means excluding the tabular relations that is used in relational databases. NoSQL data models is different than relational Models. The relational model collects data and splits it into many interconnected tables containing rows and columns.

But NoSQL database takes the data into documents using the JSON format. JSON is JavaScript Object Notation. Another major variance is that relational technologies have inflexible schemas while NoSQL models don't have schemas. Numerous NoSQL databases have exceptional integrated caching capabilities. So, the regularly used data is reserved in system memory. NoSQL database[6] types are :

1) Document Database: pair each key with composite data structure known as document. Document may contain nested document. This kind of database stores unstructured (text) or semi-structured (XML) documents which are typically hierarchal in nature.

2) Graph Stores: Graph database is grounded on graph theory. It collects information about network.

3) Key Value Stores: Every single item is stockpiled as an attribute name together with its value.

4) Wide Column Stores: They are enhanced for queries over huge datasets and stock column of data together instead of rows.

B. HDFS Architecture

Apache Hadoop is a fast-growing big-data processing open source software platform. Hadoop is able to handle all kind of data like unstructured, structured, and audio. It runs OS/X, Linux, Solaris, and Windows. Hadoop is flexible, scalable and fault tolerant. It comprises of HDFS.Hadoop HDFS is distributed and scalable file system which is written in Java. HDFS has master/slave architecture. An HDFS cluster comprises of a single Name Node. It is able to handle the file system namespace. The name node is the equal to the address router for the big data application. Moreover, there are a many Data Nodes, typically one per node in the cluster, which manages storage connected to the nodes that they run on. Figure 3 explains [7] HDFS architecture. Data Nodes also executes functions like block deletion, creation and replication as per the instruction of Name Node .Hadoop forms *clusters* of machines and organises work among them. If any of the clusters is unsuccessful, then Hadoop carries on the operations on the cluster without losing data.

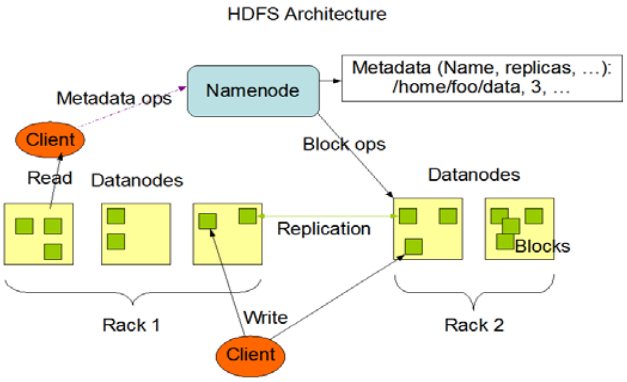

Fig 3: HDFS architecture

C. Map Reduce

Map Reduce is a programming model and software framework first established by Google. It works similar to a UNIX pipeline. A Map Reduce job splits the input dataset into independent subsets that are managed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result. MapReduce framework consists of a single master Job Tracker and one slave Task Tracker per cluster node. The master is responsible for scheduling the job’s component tasks on the slave, re executing the failed task. The slave executes the task as directed by the master. Some of the Hadoop related projects are described as:

1) Pig: It is a Scripting language and run time environment. It allows users to execute MapReduce on a Hadoop cluster. Pig's language layer currently consists of a textual language called Pig Latin.

*2) Hive***:** It provides SQL access for data in HDFS. Hive's query language, HiveQL, compiles to MapReduce. It also allows user-defined functions.

3) HBase: A scalable, distributed database that supports structured data storage for large tables. It is column based rather than row based.

4) Mahout: Library of machine learning and data mining algorithm. It has four types of algorithm.

5) Oozie: Oozie is a Java Web-Application that runs in a Java servlet-container – Tomcat. It is job coordinator and workflow manager.

6) BigTop: It is used for packaging and testing the Hadoop ecosystem.

V. CHALLENGES IN BIG DATA

A. 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 homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress.

B. Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, 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 up 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.

C. Timeliness

There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion The larger the data set to be processed, the longer it will take to analyse. It is difficult to design a structure when data is growing in very high speed.

D. Privacy

Privacy The privacy of data is another huge concern, and one 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, particularly in the US, are less forceful. However, there is 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 realize the promise of big data.

VI. CONCLUSION

We have arrived in an era of Big Data. Through improved analysis of the huge volumes of data that are becoming accessible, there is the possibility for making faster improvements in numerous scientific disciplines and refining the profitability and success of many enterprises. On the other hand, many technical challenges called in this paper must be addressed beforehand. The challenges comprises of 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. These technical tests are common across a large variability of application domains, and hence not cost-effective to talk in the context of one domain alone. Additionally, these challenges will need transformative solutions, and will not be addressed logically by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFRENCES

[1] Payal Malik, Lipika Bose,” Study and Comparison of Big Data with Relational Approach”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013 pp 564-570

[2] Wei Fan, Albert Bifet,” Mining Big Data: Current Status, and Forecast to the Future”, SIGKDD Explorations, Volume 14, Issue 2

[3] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole Velicanu,” Perspectives on Big Data and Big Data Analytics”, Database Systems Journal vol. III, no. 4/2012

[4] Big Data Survey Research Brief”, Tech. Rep.SAS, 2013

[5] Srinivasan,”SOA and WOA Article, Traditional vs. Big Data Analytics,“Why big data analytics is important to enterprises”,[Online].Available: http://soa.sys-con.com/node/1968472

[6] Available: http://www.mongodb.com/learn/nosql

[7] Dhruba Borthakur, The Hadoop Distributed File System: Architecture and Design

[8] https://hadoop.apache.org

[9] http://www.revelytix.com/?q=content/hadoopecosystem

[10] http://www.mckinsey.com

[11] NESSI-Big Data White Paper,” Big Data –a new world of opportunities” December 2012