2 Days International Conference on Recent Innovations in Engineering, Science, Humanities and Management
Dev Samaj College for Women, Ferozepur, Panjab (India) (ICRIESHM-17)
18th -19th March 2017, www.conferenceworld.in ISBN: 978-93-86171-33-7

BIG DATA: ARCHITECTURE AND ISSUES

Rupinder Pal Kaur

Assistant Professor in Computer Science, Guru Nanak College for Girls,
Sri Muktsar Sahib, Punjab (India)

ABSTRACT

Big Data technology is getting very famous now days because of its utility in storing very large amount of data. In this paper I had tried to discuss what is Big Data in actual. Basic three Vs characteristics and architecture are explained. Traditional analysis tools cannot be used for knowledge discovery. So I had discussed the analysis tools for data mining. In last some challenges in Big Data are discussed and paper is concluded in last.

Keywords:Big Data, Architecture, Characteristics, Challenges

I. INTRODUCTION

Organizations store their data in large databases known as data warehouses in which information is stored in very structured way. ETL process that is extraction, loading and cleaning is performed before loading data into warehouse. Analytical tools like OLAP, ROLAP etc. are used for fetching data from warehouse. All is to say that in data warehouse data is stored in very structured way on which SQL queries are implemented to extract any knowledge. Data in warehouses are stored in form of dimensions, each dimension represent a particular attributes of data. Now if we talk about Big Data, Data stored in warehouses is very important data related to organizations, then where the data should stored which is not as important as data

In warehouse , but size of data is very large. Data which is difficult to process is stored in Big Data. Big Data store structured data, semi structured data, un structured data from which any information can be fetched which help in decision making. For example records of millions of people that too from different sources like from sales, from customer care, social media etc. Big Data is used by companies to gain profit but large amount of data is to be analyzed before fetching any knowledge , this analysis is not possible with traditional tools. Big Data analysis is done to collect, organize and analyze large volume of data to find out the favorable patterns and useful information. This analysis is done by specialized software tools for data mining, prediction and analysis.

Best example of big data can be taken from facebook. As we upload photographs on facebook, people comment on photographs and many upload even photographs in comment section. This is irrelevant data for anyone to store. But it should be stored to detect the future interest or trend in users and to make plan for next features to be provided to users. This data , which is not structured but can give some knowledge can be stored in big data.

The development of big data [1] In the late 1970s, the concept of “database machine” emerged, which is a technology specially used for storing and analyzing data. With the increase of data volume, the storage and processing capacity of a single mainframe computer system became inadequate. In the 1980s, people proposed

"share nothing," a parallel database system, to meet the demand of the increasing data volume [2]. The share nothing system architecture is based on the use of cluster and every machine has its own processor, storage, and disk. Teradata system was the first successful commercial parallel database system. Such database became very popular lately. On June 2, 1986, a milestone event occurred when Teradata delivered the first parallel database system with the storage capacity of 1TB to Kmart to help the large-scale retail company in North America to expand its data warehouse [3]. In the late 1990s, the advantages of parallel database were widely recognized in the database field.

II. CHARACTERISTICS OF BIG DATA

Characteristics are defined by three Vs that are Volume, Variety and velocity.

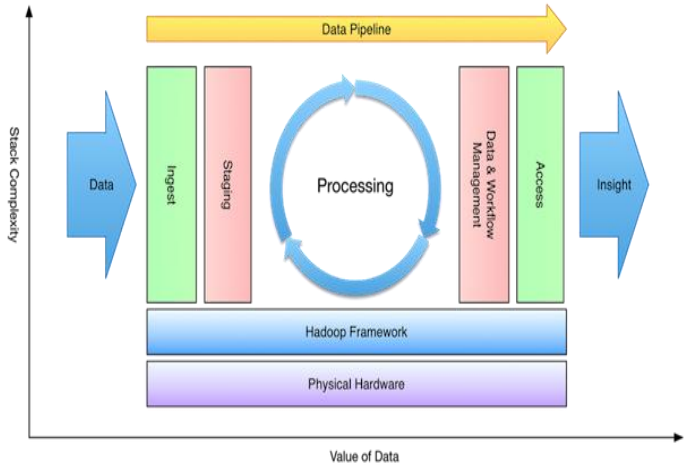
Volume : Volume refers to the amount of data. Large amount of data can be stored . We can find data in the format of videos, music's and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As we had discussed the examples of facebook data for volume. According to sources facebook is generation around 1TB of data /sec. we can imagine the volume of data which facebook have to manage.

Variety: variety refers to the number of types of data. Data can be stored in multiple format. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world have data in many different formats and that is the challenge we need to overcome with the *Big Data*. This variety of the data represent represents Big Data.

Velocity: velocity refers to the speed of data processing. The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data.

III. ARCHITECTURE:

Big Data architecture [4] is premised on a skill set for developing reliable, scalable, completely automated data pipelines. That skill set requires profound knowledge of every layer in the stack, beginning with cluster setting up the top chain responsible for processing the data. The following diagram shows the complexity of the stack, as well as how data pipeline engineering touches every part of it.



The main detail here is that data pipelines take raw data and convert it into insight (or value). Along the way, the Big Data engineer has to make decisions about what happens to the data, how it is stored in the cluster, how access is granted internally, what tools to use to process the data, and eventually the manner of providing access to the outside world. The latter could be BI or other analytic tools, the former (for the processing) are likely tools such as Impala or Apache Spark. The people who design and/or implement such architecture I refer to as Big Data engineers.

In the remainder of this post, you'll learn about the various components in the stack and their role in creating data pipelines.

IV. BIG DATA ANALYSIS TOOLS:

Traditional tools are not capable of handling such a large amount of data (actually the unstructured form of data). So, some specialized tools are required to analyze data in Big Data.

Hadoop is core platform for structuring Big Data. The Hadoop platform was designed to solve problems where you have a lot of data — perhaps a mixture of complex and structured data — and it doesn't fit nicely into tables. It's for situations where you want to run analytics that are deep and computationally extensive, like clustering and targeting Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

NoSQL: NoSQL, which encompasses a wide range of technologies and architectures, seeks to solve the scalability and big data performance issues that relational databases weren't designed to address. NoSQL is especially useful when an enterprise needs to access and analyze massive amounts of unstructured data or data that's stored remotely on multiple virtual servers in the cloud.

R: R, an open source programming language and software environment, is designed for data mining/analysis and visualization.

KNMINE: KNIME (Konstanz Information Miner) is a user-friendly, intelligent, and open-sourcerich data integration, data processing, data analysis, and data mining platform [113]. It allows users to create data flows or data channels in a visualized manner, to selectively run some or all analytical procedures, and provides analytical results, models, and interactive views. KNIME was written in Java.

ISBN: 978-93-86171-33-7

Weka/Pentaho: Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java. Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule, and visualization, etc.

V. CHALLENGES IN BIG DATA:

Data representation: Variety is explained with the multiplication of data sources where comes the explosion of data formats, ranging from structured text to free text. The necessity to collect and analyse non-structured or semi-structured data goes against the traditional relational data model and query languages [7]. This reality has been a strong motivation to create new kinds of data stores able to support flexible data models.

Data Processing: Data processing captures the growing data production rates. More and more data are produced and must be collected in shorter time frames. The daily addition of millions of connected devices (smart phones) will increase not only volume but also velocity. Real-time data processing platforms are now considered by global companies as a requirement to get a competitive edge.

Redundancy reduction and data compression: Due to huge amount of data, redundancy is the major issue in storing data in Big Data. Due to different dimensions in data, relational databases cannot be maintained with high degree of redundancy.

Analytical mechanism: Big data contain both structured and un structured data. Structured data can be analyzed by traditional tools available. But un structured data need special tools like hadoop etc. its hard to manage and query unstructured data.

Volume: lots of data (which is labeled as "Tonnabytes" by Dr. Kirke Borne, to suggest that the actual numerical scale at which the data volume becomes challenging in a particular setting is domain-specific, but we all agree that we are now dealing with a "ton of bytes".

Data confidentiality: 'Big Data' (BD) is best understood as a more powerful version of knowledge discovery in databases or data mining [6]. Privacy and confidentiality protections are threatened in this brave new world of data because the traditional role of data producers has become less relevant. This means the standard sets of confidentiality protections that were applied to data collected by statistical agencies, the traditional data producers, are also less relevant.

In the new paradigm represented by Big Data, each individual is his or her own data producer, the data are housed in businesses or administrative agencies, and researcher access is largely unregulated.

Energy management: Energy management includes planning and operation of energy production and energy consumption units. In this era of rapid transformations, Energy Management is more of a necessity than a choice. Retail businesses spend billions of dollars each year on energy. The industry faces the dual challenge of not only reducing carbon footprint, but also imbibing sustainable strategies that balance business objectives with environmental responsibilities [8].

VI. CONCLUSIONS

Big Data is a new revolution in technology. It is highly useful for organizations or other companies in storing every type of data in large space from where knowledge can be fetched for future use. Big Data provide

provides a great helping hand in decision making. Important predictions can be made by sorting through and analyzing big data. 80% of data is unstructured, thus it must be formatted or structured in a way to make it suitable for data mining and subsequent analysis.

REFERENCES

[1] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." Mobile Networks and Applications 19.2 (2ing 014): 171-209.

[2] DeWitt D, Gray J (1992) Parallel database systems: the future of high performance database systems. Commun ACM 35(6):85–98 Mobile Netw Appl (2014) 19:171–209 205 20.

[3] Walter T (2009) Teradata past, present, and future. UCI ISG lecture series on scalable data management.

[4] http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture.

[5] Beyond the PC. Special Report on Personal Technology (2011).

[6] Rubinstein, Ira S. "Big data: the end of privacy or a new beginning?."International Data Privacy Law 3.2 (2013): 74-87.

[7] Siddiqui, Sameera, and Deepa Gupta. "Big data process analytics: a survey." Int J Emerg Res Manag Technol 3.7 (2014): 117-23.

[8] George, Gerard, Martine R. Haas, and Alex Pentland. "Big data and management." Academy of Management Journal 57.2 (2014): 321-326.