ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Confronts of Big Data and its Tools

Jyoti Yadav
Department of Computer Science
Indira Gandhi University
Rewari, India

Monika
Department of Computer Science
Indira Gandhi University
Rewari, India

Muskan
Department of Computer Science
Indira Gandhi University
Rewari, India

Romika
Department of Computer Science
Indira Gandhi University
Rewari, India

Abstract: With the recent advancement technology grows faster in the present era. Technology that provide a new pace of growth which is required to accommodate the upcoming challenges. Now a day data has increased so, it is very difficult to handle data effectively. Recently big data technology having some confronts which need to be analyzed. All the issues are explained through V's of big data. In this paper some other confronts are also explained related to business organizations that are explained by R's of big data. In this paper we also introduced the concept of analytical tools. These tools help to define the applications of data associated with their respective area. We provide a novel approach of big data challenges and their associated analytical tools.

Keywords: Veracity, Variability, Reliability, Visualization, Analysis

I. INTRODUCTION

Big data is a term that refers to combinations of data set whose rate of change (variability), size(volume), growth(velocity), accuracy(Veracity), heterogeneity(variety) makes them complex to process, analyze and capture by traditional tools and technology. Now a day data is growing exponentially in peta byte, exa byte, zeta byte, yotta byte as compare to past years. Every activity of user for example searching of records, blogging surfing generate data. Even if you are connected with internet and your GPS is on generate data about location. However the method for defining the size of big data is not firmly defined. In 1997, NASA scientist defines the term big data because they have a large data set that even cannot be stored in a disk. In 2008, American scientist popularized the term big data because big data analysis can enable unlocking of valuable knowledge and help in decision making in various field like science, industries, agriculture, medical in predicting the patterns. Big data is all about enormous, diversified flood of data which is collected from various heterogeneous data sources and analyzed at high velocity to provide valuable insight. Challenges with big data starts with first phase of big data analysis pipeline that is data acquisition phase. It's a difficult task to determine what data to keep, what to discard and how to efficiently store the data. In this paper these challenges are explained using V's of big data.

RELATED WORK

In recent years a lot of research work has been done on Big Data. Rakesh Ranjan Kumar and Binita Kumari[1] gave the overview of Big Data mining, crisis related to big data and the opportunities. They also include a framework for managing a large data set and discuss various open source software tools. Sameera, Siddique and Deepa Gupta[2] provided a broad review of Big data analytics research, while giving importance to the precise concerns in Big data world. They presented a classification based on the key issues in this area, and discussed the dissimilar methods to deal with these issues. Vinti Parmar, Jyoti, Chanderkant [3] gives the internal detail of big data. They gave a theoretical overview of big data, challenges, opportunities, data analytical tools.

II. V'S OF BIG DATA

The main idea of the V-based characterization is to highlight big data's most critical confronts: the capture, mining, processing, cleaning, transfer, integration, storage, indexing, search, sharing, mining, analysis, and visualization of large volumes of fast-moving highly complex data.

A. Volume

It refers to the large amount of data in peta byte and exa byte. It is the size of data that determine the value and potential and validity of data under consideration. When Big data are taken from some data source it is of no use without filtering and compression by order of magnitude. Our problem is to define such filters that do not discard the required content.

For many decades, hard disk drive is used for storage. After that hard disks are replaced by solid state drive today, and phase change memory are around the corner. Implication of this changing storage system cover every aspect of data processing, query processing, query optimization, query scheduling , recovery methods, concurrency control methods and data base design.

B. Variability

It refers to the inconsistency shown by the data at time interval thus hinder the process to handle and manage the data effectively. It is a challenge to deal with variety of data collected from different sources that arrive at irregular intervals at high velocity.

C. Vagueness

It specifies the confusion over the meaning of big data. Many of the definitions are logically inconsistent, which is one reason for the vagueness of the term big data. A typical flaw is to include both the data and its intended usage in the definition.

Figure 1 V's of Big Data

D. Variety

Here variety represents different types of data and sources. We can say that it is a amalgam of structured, unstructured, semi structured data. In this era generation of unstructured data is more than structured data and analysis of which generate valuable info. So there is a necessity of emergence of new technologies that analyze and manage different type of data. Most of analysis algorithm expect homogeneous data and can not understand nuance. so data must be carefully structured in data analysis phase. Even after data cleansing and error correction there is a possibility of some error and incompleteness in data. Handling heterogeneous and incomplete information is a challenging task.

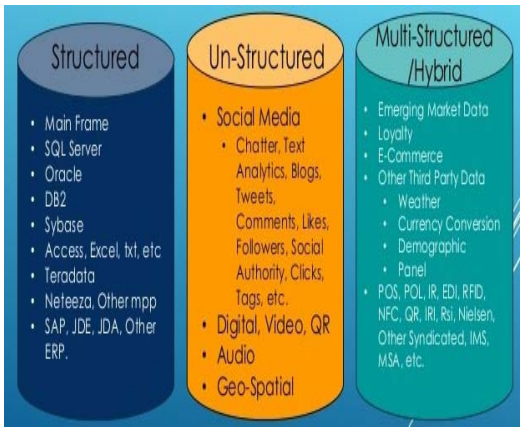

Figure 2 Data Type of Big Data

E. Velocity

Velocity refers to the rate at which the data is processed and produced to meet the requirements and challenges of an organization. As data is growing in high speed organization not only have to find relevant data that is required, they must find it quickly. But the challenge is going through the enormous amount of data and gaining the needed detail at a high speed. As the degree of granularity increases, capturing the relevant data at a very high speed is a major issue.

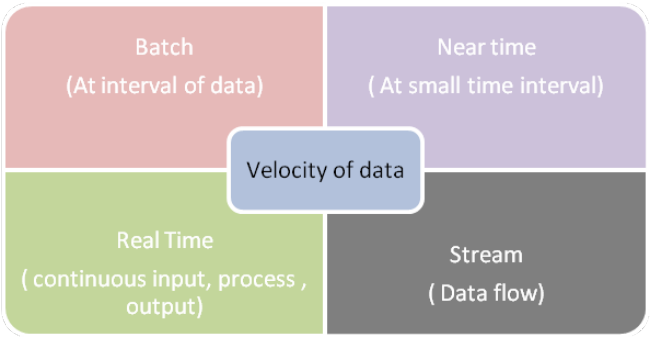

Figure 3 Velocity of Big Data

F. Veracity[4]

Veracity refers to how much data can be trusted and it gives the reliability of its sources or we can say accuracy of data. The quality of data being captured can vary greatly. So accuracy of analysis depends on veracity of the source data.

G. Vulnerability[5]

The major issue which is very critical is the privacy and security. Privacy, in the big data world, indicates “identifiable information chunks” that can be used to establish an individual’s identity. In big data analytics, enormous amount of data involving many variables have a high possibility of displaying correlations or bogus patterns, thereby set up a relationships between variables by analyzing volume of sample data, where such relationships do not exist. These types of results will misguide and mislead decision makers.

H. Visualization

It helps us to understand most ethical way of visualizing our data. With this we can ensure that either we are accidentally or deliberately misleading the users or we are presenting our analysis in fair and honest way.

I. Verbosity

Verbosity refers to questions related to text sources and the problem of machine understanding of the meaning of text. Natural language is often verbose in that it is not firmly concise and is often highly predictable. Does this hinder the analysis of text.

J. Volatility

Volatility refers to the time scale over which data and analysis retain its validity. It means how long data is valid and

how long should it be stored. In reality you need to understand at what point data will no longer relevant to the current analysis. For example weather forecasting of a specific place can change in a short time scale.

K. ***Vocabulary***

Vocabulary refers to schema, data models, semantics, taxonomies, and context based meta data that describe the syntax, data structure, content.

L. ***Valuable***

Valuable refers to the scalability that is the ability to work when size and volume is changed in order to meet user requirements. Managing large and rapid increasing size of data has been a major issue for many decades. In past, this issue was overcome by faster processor but now data volume is scaling faster than computer resources.

III. R'S OF BIG DATA

According to business perspective there are some other confronts that must be considered while dealing with big data .these challenges are explained by term R's

A. ***Reliability***

Consider the situation where missing data points are vital to performance the unreliable information inclusion can sweep out results, and have a negative impact on user experiences or, in most of the cases it may be a big problems for business. In today's world all aspects of our life are changing faster and more often. This affects reliability of data by increasing the speed at which data goes out of date and no longer be relevant to the requirements or accurate.

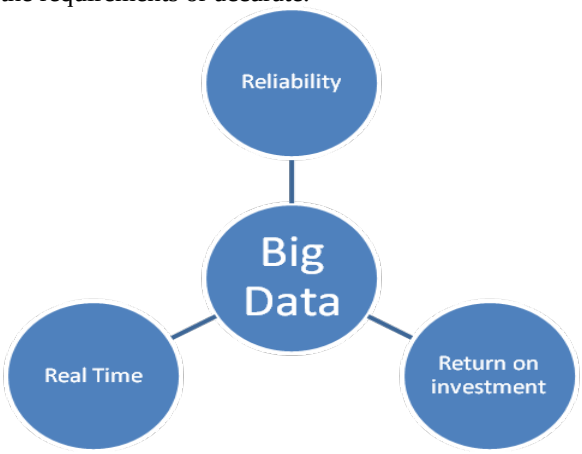

Figure 4 R's of Big Data

B. ***Real Time***

Big data is often about responding in events now but not later. The organization analyses real-time information from its transport networks such as bus locations, and passenger movements, road traffic volumes, to manage traffic light timing so buses run as close as possible to their timetable. This leads to requirement of combining big data with real time analytics. Most of the data are analyzed offline, but insight is required in real time when data is in use. When you deal with customers in real time, you may need to get information to channels quickly to turn a call into a sales offer. For example If you have access to what they have tweeted, then you can use that additional data during the telephonic conversation."

C. ***Return on Investment***

Return on investment considers data as an asset. Here the major challenge is to manage and analyze data to make the best business decisions to maximize return on investment.

IV. OPEN SOURCE BIG DATA ANALYTIC TOOL

Big data analytics enables organization to analyze a mix of structured, semi structured and unstructured data in search of valuable business information and insights. Big data analytics refers to the process of analyzing, accumulating, organizing large sets of data (big data) to discover patterns and other required information

A. ***Hadoop:***

Hadoop[6] is `a popular option when you need to filter, sort, or pre-process enormous amounts of new data in place and distill it to generate denser data that contains more "information". Pre-processing is the process of filtering new data sources to make them available for additional analysis in a data warehouse. Hadoop is useful for pre-processing data to recognize macro trends or find nuggets of information, such as out of-range values. It enables businesses to find potential value from new data using inexpensive commodity servers. Organizations primarily use Hadoop as a precursor to advanced forms of analytics.

B. ***Apache Drill and Dremel***

Apache Drill and Dremel[7] make it possible to execute large-scale, ad-hoc queries, with lower latencies. These tools can scan PB of data in terms of few seconds. Apache Drill and Dremel is useful for both data engineers and business analysts. The business organizations adopt Apache Drill and Dremel tools but still it has not had a strong development community's attention. Hadoop has some disadvantages as it uses batch processing for all workflows. The Hadoop team has worked very hard to incorporate ad hoc analytics. Many interface layers such as swazall, have been developed on top of Hadoop to make it more user friendly and business-accessible. In contrast to workflow based analysis, most of analytics queries and business-driven Business Intelligence are interactive, ad hoc, and low-latency analyses. Writing Map Reduce workflows becomes limited for many business analysts. In interactive applications it is not preferable to wait for job to start and end for several minutes. As Apache Drill and Dremel can execute ad hoc queries with low latency, it was argued that Apache Drill and Dremel are better than Apache Hadoop and may be better option for users and business organization instead of Apache Hadoop.

C. *Apache Hive*

Apache Hive[7] is Apache Hadoop's data warehouse system or we can say that it is a warehouse infrastructure tool to process structured data. Hive has following features: ad-hoc queries, analysis of massive datasets and data summarization, and Hive use SQL like query language HiveQL which provides a mechanism to query the data. Initially Hive was developed by facebook later on taken by Apache Software Foundation and developed as an open source under the name Apache Hive. Traditional MapReduce programmers are also allowed in Apache Hive when it is ineffective or inconvenient for them to express custom mappers and reducers in ApacheHiveQL. D.

D. *Cloudera Impala*

Cloudera Impala is open source MPP query engine that runs on the top of Apache Hadoop. Cloudera Impala brings scalable and parallel database to Apache Hadoop giving users to experience fast, interactive SQL queries for data stored in Hadoop Distributed File System and Apache Hbase without need for data alteration or movement. With Cloudera Impala, data scientists and analysts can perform real-time analytics by using BI tools and structured query language on data stored in Hadoop. In addition to using same unified storage platform, Impala uses the meta data, SQL syntax, Open Data Base Connectivity driver and user interface similar to Apache Hive and provide real time and batch oriented queries

E. *Giraph*

Giraph[8] analytical tool enables graph analysis. These tools are often binded with graph DB's like Infinite GraphorNeo4j.

It is an interactive graph processing framework built with Apache Hadoop. Another tool for graph base project is Golden Orb. Graph DB's are pretty cutting edge. Graphs do a great job in social networks, computer networks, mapping, and geographic pathways for calculating optimal routes or we can say anything that bind the data together. Graphs are also used in physics and bioscience. Graph databases, big picture and analysis languages and frameworks are examples of how the world is started to understand that Big Data is not having one programming or one database framework. Graph DB's based techniques are a killer applications, more specifically, for analysis of large networks with many linked pathways in the network. This type of analysis allows individual recommendations across multiple channels, maximizing the value of every customer interaction. Oracle Advanced Analytics scores can work with operationalize complex predictive analytic models and create a decision processes in real time. All of these technique have a role in determining uncovering hidden relationships.

V. CONCLUSION AND FUTURE WORK

Big data analysis enable in uncovering of valuable knowledge and help in decision making in various field like science, industries, agriculture, medical in predicting the patterns. But big data confronts are increasing day by day and it becomes a necessity to handle and discuss those effectively. Because now a day data has increased so, it is very difficult to handle data effectively In this paper we conclude that big data analysis give rise to opportunities in designing of competitive offer packages for customers, configuring network to provide reliable services but analysis must be accurate and timely for successful decision making. A review of various confronts faced with big data has been outlined in this paper and these challenges must be addresses in order to realize full potential of big data. Also all the challenges outlined are not domain specific, they are amalgam of varieties of domain. In future research must be done to address outlined challenges related to V's and R's of big data

VI. REFERENCES

[1] Rakesh Ranjan Kumar, Binita Kumari, " Visualising Big Data Mining : Chellenges, Problem and Opportunities" , IJCSIT, vol-6, pp-3933-3937, 2015 .

[2] Sameera, Siddique and Deepa Gupta, "Big data process analytics",International Journal of Emerging Research in Management & Technology, 2014.

[3] Vinti Parmar, Jyoti, Chanderkant, " Innards of Big Data", International Journal of Engineering Research Online, vol.-4, pp.-216-222, 2016.

[4] Xindong Wu , Xingquan Zhu ; Gong-Qing Wu ; Wei Ding, "Data Mining with Big Wu ; Wei Ding, "Data Mining with Big Issue:1, P 97 – 107, 2014.

[5] Mr. Mahesh G Huddar, Manjula M Ramannava, "A Survey on Big Data Analytic Tools", IDEAS-2013.

[6] Bernice Purcell, "Emergence of big data technology and analytics",Journal of technology research, 2012.

[7] www.dremel.com/

[8] http://giraph.apache.org/

