Big data: Extracting Value from Big data

Suman R. Tiwari
Lecture in Computer Department,
R.C.Techincal Institute
Ahmedabad ,Gujarat,India

Abstract— **we are living in world of internet. People around us use the internet for convenience of services like online banking, online bill payment, ticket booking, merchant banking etc.**

Millions of people from world are daily uploading data on social networking sites such as on facebook,yahoo,gmail, twitter , whatsapp etc. Data available from phones, credit cards, televisions, computers, from the infrastructure of cities, from sensor traffic, trains, buses, planes, bridges, GPS Signal ,factories, Hospital and Banks are in terms of terabyte . As they are available from different sources they are in different format. Volume of data increased so fast that the total accumulation of data in past few year is almost in terms of zeta byte. We can see world as pool of data. It's up to us , Either we can be waste these data or we can utilize these data to predict things in future, which can help us to explore new world and can be very useful to human being.

In this paper I have included various definition used to define big data ,Various Sources from where data can be captured for big data , its characteristics ,big data challenges and I have proposed cyclic method which can be used to retrieve appropriate and useful information from collection of big data.

Index Terms— **Structured data, Unstructured data, GPS, GSM, GPRS**

I. Introduction

Big data does not mean that data is big. Big data can be big or small in volume but its value can be big [6]. Big data are big in sense of value derived from it.

Big data comes in picture when one cannot store, manipulate or retrieve information from database with available resources.

There are various definitions available for big data:

1) Big data is a term used to describe the exponential growth and availability of data, both structured and unstructured [3].
2) *Big data is collection of information from various sources* which can be utilized to retrieve useful information for organization or enterprise.
3) Easiest Definition available for big data that it can be considered as pool of data available from different resources and different format from which we can retrieve hidden gold.
4) "Big data is an all-encompassing term for any collection of dataset of large and complex that it becomes difficult to process them using traditional data processing applications[1]"
5) Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or computer capacity for accurate and timely decision making [7].

II. Big Data sources

Big data can be in structured or unstructured form because it is derived from various sources like sensor, climate, X-Ray, GPRS, GSM, face book, yahoo, twitter, paytm, hospitals, railway ,educational institute, crime branch and many more[4]. As they are derived from various sources they are in different format and in different size.

Data which are captured from sensor network are in different format then data captured from X-Ray or twitter. An example of big data might be Terabyte or Exabyte of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact centre, social media, mobile data and so on)[2].

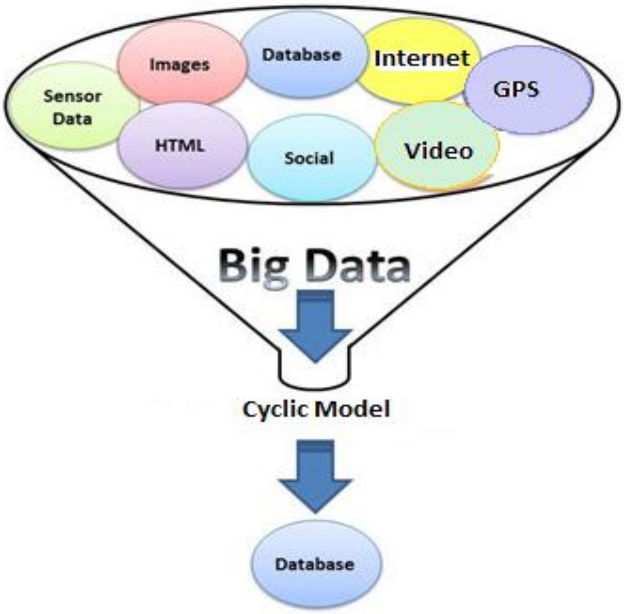

Figure 1 : Big Data Resources

III. CHARACTERISTICS

Big data are derived from different resources so all data have different format then one another one. Generally big data is described by its four basic characteristics these characteristics which are known as 4 V's of big data[5].

1) Volume
2) Variety
3) Variability
4) Velocity

In addition to these four characteristics there are another two characteristics which can be taken in consideration for big data.

5) Veracity
6) Complexity

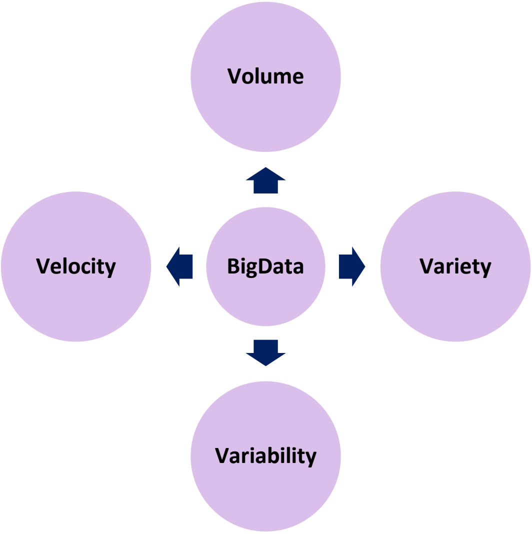

Figure 2 : Big Data Characteristics

Volume – volume means scale or size of data under consideration. The volume of big data can be in form of terabyte, or zeta byte.

Variety –variety indicate different forms of data. data are available in various forms and formats. For example data available on yahoo, gmail, sensor network, x-ray etc. are available in different form and they are stored in different format .Before doing proper analysis analyst must know about the various format of big data in which they are available.

Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges.[1]

Variability - This refers to the inconsistency which can be shown by the data at times [1].variability indicates that data differ from each other as they are derived from different sources.

We have considered two additional dimensions when thinking about big data:

Veracity - The quality of the data being captured can vary. Accuracy of analysis depends on the veracity of the source data [1].

Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources.

IV. CHALLENGES

Big data is collection of different kind of data available from different sources. These data are available in different format. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data.

Following are the challenges need to consider when we deal with big data.

1) The management of unstructured data i.e., how to Organize GPS Data, Sensor Data, movies, pictures, etc.
2) Big data is messed data then how to retrieve valuable information from it.
3) Data set becomes so large and is in different format and form then how to store them with available database model and technology.
4) Which data should be stored and which data should be discarded to retrieve appropriate result.
5) All the data should be analyzed to retrieve information or part of data should be analysed first as big data is too large.
6) How to recover that this part of data is important and useful to derive another information relevant information.
7) The analysis of big data, both for simple reporting and advanced predictive modelling, as well as deployment [6].

So it is big challenge for us that how to extract information from big data.

.

V. CYCLIC MODEL

In order to understand the hidden value and information of big data, we have to decide first what is requirement of organization or enterprise, in which area they want improvement. According to interest of organization, set analytical objective and try to retrieve analytical result from big data.

In this paper I have proposed Cyclic model to retrieve data from big data which is based on iterative waterfall model. Following Cyclic model shows how to process & Clean messy data afterward Mine & Generate Result from data. To use this model first organization need to decide objective then can go step by step to get solution. We can repeat this cycle number of times till objective is not reached.

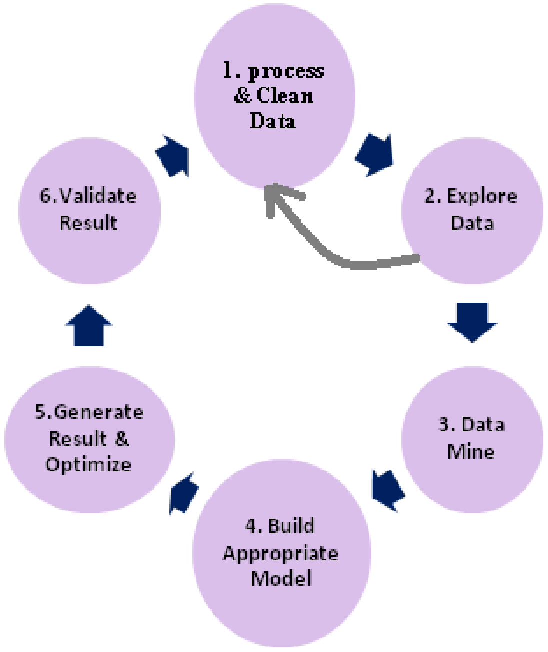

Figure 3 : Proposed Cyclic Model

1) Process & Clean Data

   Collect data from various available sources. First check for all available data then select set of dataset which can be useful. Find out more data which can be related with our selected data set.

2) Explore Data

   Explore the selected data in much more detailed form. If it is required then again include the co-related data from processed & Clean big data by going in backward direction. Select the set of dataset on which detailed work should be done.

3) Data Mine

   This step is used to select relevant information . Explore selected data set in more detailed form. Then find out detailed relationship among related Datasets. Prepare chart and write summary about Data which are co-related with one another and More or less relevant.

4) Build Appropriate Model

   Based on summary and data mine build appropriate model to extract value and give indication of result.

5) Generate Result & Optimize

   Generate the result from model and optimize the result as minimum as possible.

6) Validate Result

   Check generated result with predefined result. That will helpful to check that obtained result is same as goal which was predefined by organization.

7) If predefined result is not obtained then again repeat for same cycle.

VI. CONCLUSION

In this paper we have seen detailed definition of big data and various sources from which it is available. We have seen challenges need to face when we will work with big data.

One word that can be used for Big data is that it is scrap or we can say that it is bundle of data collected from various sources which have different encoding technique. These data may be stored in different format with different data type in different length. They may be supported by different types of organization and can have different storage technique.

Extracting data from big data is same as extracting gold from scrap. Big data can be used to collect the information which can be use for enhancement of business, organization and future prediction of work. The proposed cyclic model may used to derive useful data for its maximum utilization.

REFERENCES

[1] http://en.wikipedia.org/wiki/Big_data
[2] http://www.webopedia.com/TERM/B/big_data.html
[3] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[4] http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal
[5]http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
[6] The Big Data Revolution And How to Extract Value from Big Data by Dr. Thomas Hill
[7] Mark Troester, SAS White paper, Big Data Meets Big Data Analytics
[8] http://watalon.com/?p=722