7th International Conference on Research Developments in Applied Science, Engineering & Management
Mahratta Chamber of Commerce, Industries and Agriculture, Tilak Road, Pune (India)
15th -16th June 2019 www.conferenceworld.in
ISBN : 978-93-87793-90-3

Big data : Study of four V's

Amit Bansal [1], **Vipin Babbar** [2]

[1,2] *Assistant Professor of Computer Science*, Department of Computer Science,
Govt. College for Women, Hisar, Haryana, India

Abstract

Today every person has some electronic device always with him/her. Acquiring an electronic device has some advantages like communication, entertainment, social networking, gaining knowledge, quick response, education, e-commerce and a lot of other things also. We have access of the world 24 hours a day and from any place. But can we imagine that our each activity is under surveillance 24 hour a day. Someone is watching the one having any smart device. The data related to our habits, interests, relations, Political affairs, friends, choices, joy, sorrow, day to day activity are capturing by the device and applications we are using for our comfort but on the other side of the coin we cannot totally ignore the technology drawbacks. We must be aware and careful to use the technology. The people can use this data to take better decision related to the work.

Keywords **– Big Data, Volume, Variety, Velocity, Veracity**

1. Introduction

Big data is extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source[1][2]. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*. Other concepts later attributed with big data are *veracity (i.e., how much noise is in the data)* and *value.*

2. Volume

Volume means the quantity of generated and stored data per unit of time. I am not talking about gigabytes of data but this time data may be of size of petabyte, zettabyte or even yottabyte to be considered as big data. The size of the data determines the value and potential insight and whether it can be considered big data or not. Some examples of voluminous big data are

1. People can organize gatherings, manage invitations and send notifications and reminders to their friends on Face book. More than 550 million people use Events every month. People watch more than 100 million

ISBN : 978-93-87793-90-3

hours of video every day on Face book. With more than 350 million photos uploaded each day. 2.23 billion monthly active users on Face book as of June 30, 2018. With this no of user we can think the amount of data they can produce[6].

2. Over 1.9 Billion logged-in users visit YouTube each month, and every day people watch over a billion hours of video and generate billions of views. More than 70% of YouTube watch time comes from mobile devices[5].

3. CCTV cameras are installed everywhere in home, offices, streets etc. are watching each and every happening/activity 24hours a day might produce huge amount of data.

4. Today email is the main media for communication producing huge amount of data containing audio, video, text, images etc.

5. Now a day almost everyone has a mobile with him/her. Your mobile number with your name is captured by any caller id application making a huge data about mobile numbers. GPS system in your mobile always recording your location at various point of time.

6. E-Commerce websites are capturing shopping habits of people, storing data about items, brands, prices, area wise sale etc.

7. Search Engines storing data related to search habits of people.

3. Variety

Variety means **t**he type and nature of the data. We generally deals with structured data (like tables, reports, databases etc. but for big data we have to have deal with lots of unstructured data (like images, audio, video)[3]. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion. Big data technology allows us to harness different types of data, including e-mails, social media conversation, photos, sensor data, video data, and voice recording, and bring them together with more traditional, structured data**.**

3.1 Structured Data

Structured Data is any data or information that is located in a fixed field within a defined record of file, usually in data base or spreadsheets.[7] This data generally organized in rows and columns. In today's era, structured data is

ISBN : 978-93-87793-90-3

decreasing its advantages as the amount of unstructured data that is generating in a very fast rate has big advantage over structured data. If we are using only structured data we cannot make a better decision about the insights i.e. what is going on and why. Therefore we often need to use other data along with structured data. For example, structured data will tell you that no of students applying for a particular course are decreasing, or students are not able to pass a particular course but it will not tell you why it is happening. On the other side, structured data has some advantage that it is very easy to store and analyze.

3.2 Unstructured data

Unstructured data is data that do not have any predefined structure often include text and multimedia content for example e-mail messages, social media conversation, video and audio content, word processing documents, presentations, WebPages, images etc. [3][7] Unstructured data is typically text-heavy, but may contain data such as dates, numbers, and facts as well. Note that while these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database. Most of the data in today's era is unstructured and plays a big role in decision making. It generally tells us the reason of what happened.

3.3 Internal Data

All the data an organization may have comes in the category of internal data. The data related to employee, their salary data, product data, customer data, sales data, customer service data, feedback data, CCTV data, attendance data etc. an organization require lots of effort to store and maintain this data. This data alone do not provide much support to take some strategic decision for the organization but we have to mix this data with external data to take advantage of big data. But on the other hand internal data has capability to initiate a new process and the organizations do not need to pay anyone for this internal data.

3.4 External data

External data is the data that exists outside of your organization. This may be available free of cost or may be owned by some third party. It may be structured or unstructured in format. Some examples of external data are social media data, e-commerce website data, economic data, census data, weather data etc. the main drawback of external data is that you don't own that and you have to pay to owner of that data to use. After paying so many amounts it will not guarantee that that data will help the one for taking any decision.

ISBN : 978-93-87793-90-3

3.5 Activity data

The data produced by your daily activity comes under this category. Records of transactions which indicate habits, interest and intent comes under activity data. You wake up timing can be estimated by first online attempt you made and how often you were online in morning. When I have activated our mobile phone first time in the morning because people generally check their mobile immediately after wake up now a day? If you were online frequently at 1:00 AM then any one can guess about your mental health / business. If you are wearing a fitness band, that will record your movement data. How far you travel? How many calorie you burn? What time you take to cover one KM so that your fitness data may be recorded. Data recorded in CCTV Camera tells your timing of leaving your home and entering in the office, and the route you follow to reach there. If you frequently use e-commerce sites to shop anything than your interest may be recorder on these sites. Then these sites make use of that data to send you the advertisements of the things you have accessed during your visit to site. What types of people you are in touch through social media is also not private. Your comments about social issues on social networking sites are also recoded. Your location is always trapped if you carry a smart phone having GPS with you. [4]

3.6 Audio, Video, Images

Audio, Video, and images are becoming a major component of data. Social networking sites, you tube, emails, messaging application etc are major source of this type of data. Telephonic conversation is the big source of audio data. CCTV camera also stores video data.

4. Velocity

Velocity means the speed at which the data is generated, moves around, and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time.

5. Veracity

Veracity means the messiness or trustworthiness of the data. We generally analyze neat and structured data, which is considered to be accurate. But now we have to deal with completely unruly and unreliable data, like abbreviation, typos, slang, social media posts with incorrect hashtags etc. which may provide wrong information which in turn can be converted into wrong decision. The data quality of captured data can vary greatly, affecting the accurate analysis. I think everything we want to use from internet is not worthy. Before using the data/statistics etc. we have to check the source and its authentication so that correct data may be used for any decision making.

ISBN : 978-93-87793-90-3

6. Conclusion

Big data plays a major role in decision making. Some preventive measure must be taken with respect to Big Data. First source of data must be reliable. Data collected may be huge in quantity; we must separate the data by using some proper analytic techniques to extract the required data. Quick decision making should be there because data is generating with very fast speed and trends may be changed by time.

References:

[1] https://en.wikipedia.org/wiki/Big_data
[2] https://www.dictionary.com/browse/big-data
[3] https://www.webopedia.com/TERM/U/unstructured_data.html
[4] http://www.activitydata.org/What_is_Activity_Data.html
[5] https://www.youtube.com/intl/en-GB/yt/about/press/
[6] https://newsroom.fb.com/
[7] https://www.edureka.co/blog/big-data-tutorial