Big Data, its Issues and Challenges

Kushal Patel
Assistant Professor, GIDC Degree Engineering College, Abrama, Navsari

***Abstract* - Big Data is used to capture, curate, manage, and process within a tolerable elapsed time. A tremendous amount of data about individuals—demographic information, Internet activity, energy usage, communication patterns, and social interactions, to mention a few—are being collected by various organizations such as national statistical agencies, survey organizations, medical centers, and Web and social networking companies. However, there are certain issues associated with Big Data. In this paper, we aim to discuss issues associated with Big Data. In order to realize the use of Big Data, various applications along with their issues and challenges are discussed. In particular, we focus on privacy issues in Big Data.**

***Keywords* - Big Data; Privacy; Security; Voluminous Data**

1. Introduction

Big Data refer to as Data sets with size beyond the ability". These Data volumes are in a range of peta bytes (10^{15}) and exa bytes (10^{18}) and beyond [1]. Data are being produced and accumulated in a order of the exa byte/year range. But, its creation and aggregation are accelerating and will approach the zetta byte/year range (1 zetta=10^{21} bytes) within a few years [1]. Big Data is getting a huge attention in the domains of large database. Such databases are so heavy in size such that processing on such databases can no longer be treated effectively or even completely. Large Databases becomes unwieldy when handling in the conditions of size, speed, processing time. This problem can be classified as a Big Data problem. Big Data problem arise with Big Data include capture, storage, dissemination, search, analytics and visualization [2]. Peta bytes(PB) and exa bytes(EB) of data are produced in the domain of the traditional data-intensive sciences. In addition, domains such as data repository, cyberspace, social web search, economics and business produces tremendous data. In [3], 5 attributes of Big Data viz. volume, velocity, value, veracity, variety are discussed.

Big Data deals with prominent amount of data, but it has challenges and issues as well. Amongst the several issues of Big Data, the storage, data transport, management and processing are technology issues [1]. In addition, privacy is one of the challenging concerns to be preserved with Big Data. When looking at privacy issues in the domain of Big Data we need to make out lot of Big Data application domains.

The application domains such as physics and earth science typically not related to individual information and hence do not cause substantial privacy issues. The privacy critical domains consist of the public social media, life science, marketing, business analytics and public surveillance. In these domains of Big Data, collected information might be used to create and analyze profiles of us, for example for market research, targeted advertisement, workflow improvement or national security [2].

2. BIG DATA

Figure 1 exemplifies 5 attributes of Big Data which is analyze in [3].

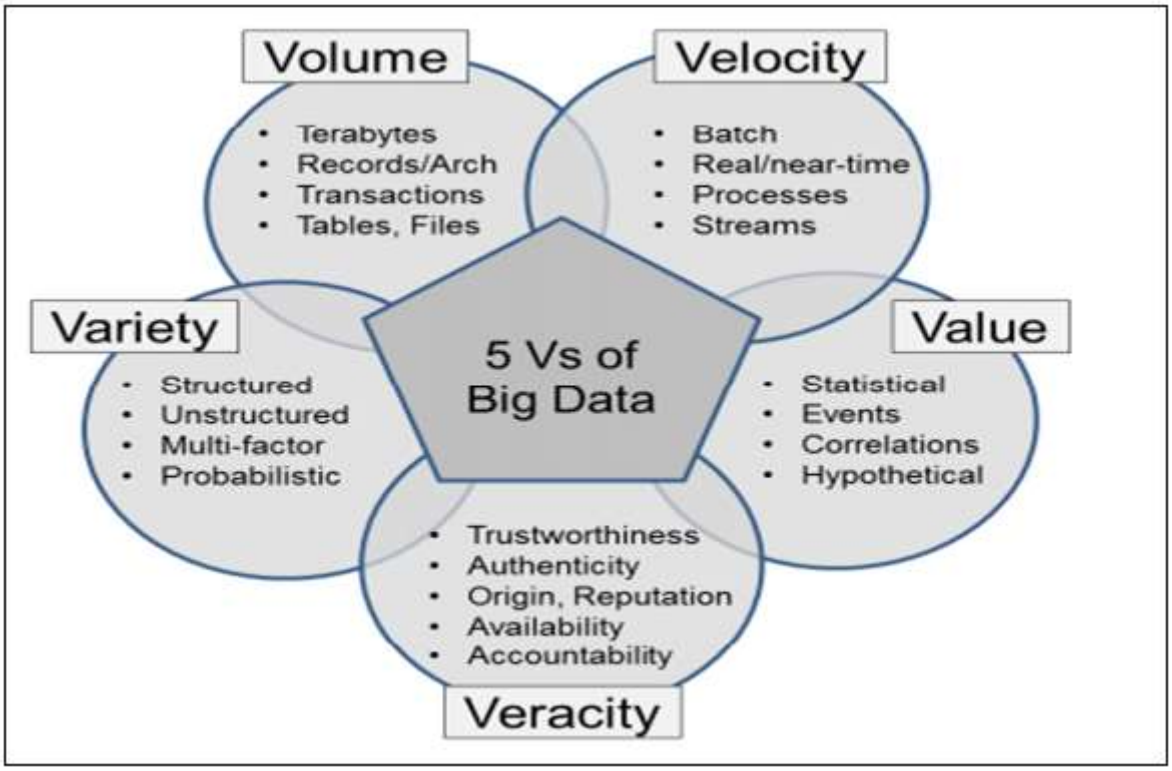

Figure 1. 5 Attributes of Big Data [3]

1) Volume

Huge data being collected from sensors, social activities and medical sciences are. Big Data volume includes features such as size, scale, amount, and dimension [3]. Social Service provider such as Google, Facebook, Twitter, YouTube, Television Media are creating, observing and storing data in vast quantity for providing their services.

According to IBM, it is observed that 90% of overall data in world was generated in last 2 years. In 2000, overall data generated was 5 million of tera bytes, in 2011, it was reached to 2 billion of tera bytes and is estimated that in 2016 it will be reached to 10 billion of tera bytes.

2) Velocity

This attribute of Big Data addresses about the speed at which data being came from diverse data generator places. Velocity assesses the speed of data creation, streaming, and aggregation [1]. E-Commerce has speedily enhanced the affluence of data used for different market proceedings. In addition, nowadays big trend of Social Media generates and transferring media with higher velocity. The traditional storage systems like RDBMS are not adequate to laying in and executing the analysis on the data which is invariably in motion.

Big Data necessitate to be processed in real-time, near real-time or in batch, or as streams (like in case of visualization) [3]. As an example, LHC (*Large Hadron Collider)* ATLAS detector uses about 80 readout channels and gathers up to 1PB (pera bytes) of unfiltered data per second which are diluted by approximately 100MB per second [3].

3) Variety

There are mainly three types of organizing Data: Structured, Semi Structured and Unstructured. Structure Data Include standard spread sheet or RDBMS which has data in Structured format in row and column e.g. Records in SQL, or in Microsoft Excel which has schema and meta data. Semi Structure Data include web Pages, Web Log Files, social media sites, e-mail, documents, and sensor devices data both from active and passive devices. Unstructured Data include Image, Video, audio etc.

Data variety is a measure of the richness of the data [1]. It is very tedious task to analyze large volumes of data of different varieties. In addition, it raises challenges to design of data storage and database with various data format. Data variety increase with various branches of science and societal systems.

4) Veracity

Veracity attribute of Big Data includes two aspects [3]:

1) Data consistency (or certainty)
2) Big Data trustworthiness

It might be required to protect the data collected. Data should be processed and stored at the trustworthy place. In addition, the source of the data should be genuine. Stored data should be available as and when required.

Following aspects should be defined and need to be addressed to ensure data veracity [3]:

- Integrity of data
- Data authenticity and genuine origin
- Computer and storage platform trustworthiness
- Accessibility and timeliness
- Accountability and Reputation

5) Value

Data value measures the usefulness of data in making decisions [1]. Data value will rely on the outcomes of processes they represent such as stochastic, probabilistic, regular or random [3]. Depending on this the demands may be enforced to accumulate all data, store for longer period. Data value is fruitful in getting to recognize about the data stored or produced. User can execute certain queries against the data captured and thus can derive significant consequences from the derived data. In addition, use can rate the data according to the dimensions they expect [4]. For example, the organization can predict and plan their gains from stored and produced data.

2.1 Big Data: Motivation

Here are some points which show, why big data are required?

- **To overcome Limitation of Traditional RDBMS**

Table 1 shows some examples which produce huge amount of data in particular amount of time. Traditional RDBMS cannot handle such a huge amount of data in real time. In addition, RDBMS only support structured data for Storage. Analysis on semi and unstructured data are not possible with RDBMS.

Table 1 Example of Big Data generated by Various Application/System

Data Set/ Domain	Description
Internet Communications (Cisco)	667 Exa bytes in 2013 [1]
Facebook	500 TB/Day in 2012 [5]
Twitter	12+ TB of tweets every day and growing. Average re-tweets are 144 per tweet [1]
British Airways (Flight)	Flight from London to New York in 30 Minutes it generates 10 TB data (Logs) per Engines and it will generates total 650 TB of data in its overall flight [1]
US Library of Congress	235 TB data collected by April 2011 [6]
Large Hadrons Collider/Particle Physics (CERN)	13-15 peta bytes in 2010 [3]

- **High volume storage hardware**

The demand for hardware is that, it can store, process, analyses prominent amount of data. Such a huge data cannot be stored on single machine because today's maximum hard disk available in the range of almost 16-20 TB. In addition, such devices are very expensive. Furthermore in certain domains, data required to be processed in parallel for faster response in small amount of time.

- **Failures of Hardware and/or Network**

Failure at the points of storage disk and network can be quite problematic. In addition, it is required that data should be available as and when it is required and at any place (Availability). By creating multiple copies (Replication) of data, it should be available when failure of one or more location.

2.2 Source of Big Data

Main source of Big Data are the data generated by Users, Application, Services, System, Sensors/ Device as shown in figure 2.

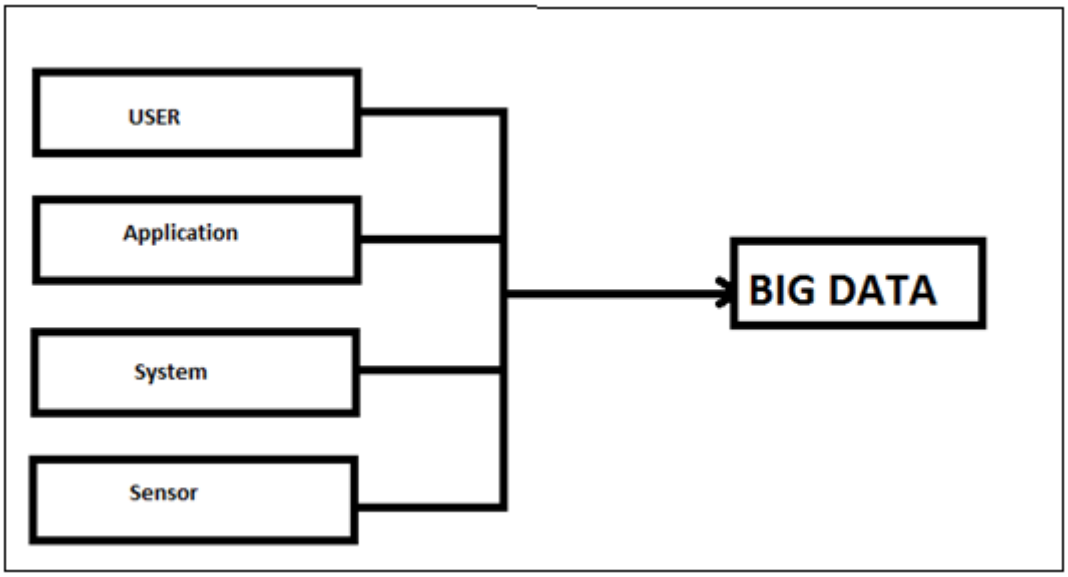

Figure 2 Sources of Big Data

2.3 Challenges

- Design, analytical and Maintenance

It is require to understand the needs of users and technology that the design for the systems and components that work with big data [1]. Such analysis can be used to solve the problem being investigated. In addition, not all big data and its requirements are the same. For example, requirement for field of science, social science, markets and elsewhere are different.

There may be multiple sources of data that may be stored either on single or multiple sources. These data must be store based on some rules and regulations. For example, in America, insurance and health organization can keep individual data under HIPPA rule. These rules and regulations include how long these data should keep? Which data should be store? In addition, it is challenging task to decide whether collected data should kept even after its lifetime or after its usage. For Example, to decide after some experiment whether given observed result should keep or discard and how much result should be keep for future use.

Data should be kept about organization or individual secured from outsider. So all security related concern like privacy, confidentiality, authentication, and authorization should be maintained. For analysis perspective, main challenges are handle large scale of data. Out of such huge data, how much data is important? In other words it is very challenging task to mine useful data. For example, for government, it is tedious task to observe sensitive information of data from large amount especially data exchanging to/from outside the country. Also from analysis, crime investigation is also challenge for government.

2.4 Issues in Big Data

2.4.1 Storage and Transport

Current disk technology confines to 16-20 terabytes per disk. So, 1 exabytes of data would require 50,000-60,000 disks. Even if an 1 exa byte of data could be processed on a single machine, it would be ineffective to immediately attach the required number of disks. Access to that data would deluge current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 800 megabytes [1]. Thus, transferring an exa byte would take about 40-50 years. The transmission of data from source location to storage or processing location may take long time than analysis.

Management

People contribute Big Data in various form: documents in various formats, images, sound and videos, software, designs, etc. with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance [1]. Security, privacy, availability, validity are management issues.

2.4.2 Security

Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media. Data represents a severe security pertain, especially when many individuals so willingly surrender such information. IDC (International Data Corporation) suggested five points regarding information threat: privacy, compliance-driven, custodial, confidential, and lockdown [1]. According to Symantec Internet Security Threat Report [7], the total number of breaches for phishing host in 2015 was 50.8 %, which is 4.2% greater than 46.6 % in 2014.

Privacy

The traditional Big Data applications such as astronomy and other e-sciences doesn't disclose personal identity hence do not have significant privacy issues. Demesnes such as social website, consumer and business analytics and governmental surveillance are some area in which privacy is very important issue. According to Symantec Internet Security Threat Report [8], 1 in 3 consumers acknowledged that they offer false information (credit cards, passwords, address details etc) to achieve privacy. In addition, 8 of the breaches in 2013 exposed more than 1m identities each [9]. Aggressor attack a customer's personal details and attempt to gain access to accounts. For example, through password reset features on websites or Email. Depending on the slipped information, attackers could employ these data to authorize bank account (e.g. Paypal) transfers to accounts under their control.

Privacy Issue

- **Personal Privacy**

Personal privacy includes Individual's name, signature, address and telephone number, date of birth and commentary or opinion about a person.

- **Information Privacy**

It includes about someone racial or cultural origin, political beliefs, membership and affiliation, religious beliefs and affiliations, membership of a professional or business affiliation, membership of a trade union, criminal record if any, and health information, genetic information. In certain domains, such as social media i.e. facebook, twitter, whatsapp etc. and health information, data is accumulated at greater extent about individuals. As a result certain establishments recognize overmuch detail about individuals.
The delicate information of an individual when aggregated with extraneous prominent data sets extends to the illation of new conceptions regarding that individual. In addition, it is possible that such conceptions regarding the individual are Arcanum and an individual might deny the information possessor or any other person to utilize such information.
Privacy may be breaches in following area of Big Data:

- Social Media
- Location Based Service or Application
- In cloud Storage
- Government Website
- Science and Technology Research especially in medical science
- Business /e-commerce and few more

Conclusion

By analyzing Big Data, we can say that though Big Data enables organizations to analyze, curate, manage, and process the data with a high speed, it requires preservation of privacy at either individual level and/or at corporate/organization level. To preserve personal privacy, one should determine the amount of information disclosed to others. In addition, the organizations which accumulate such information should preclude others to access such data.

References

[1] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data : Issues and Challenges Moving Forward," *46th Hawaii Int. Conf. Syst. Sci. IEEE*, pp. 995–1004, 2013.
[2] M. Smith, C. Szongott, B. Henne, and G. Von Voigt, "Big Data Privacy Issues in Public Social Media," *IEEE*, 2013.
[3] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," *IEEE*, pp. 48–55, 2013.
[4] A. Katal, "Big Data : Issues , Challenges , Tools and Good Practices," *IEEE*, pp. 404–409, 2013.
[5] Eliza Kern, "Facebook is collecting your data — 500 terabytes a day," *GAGAOM*, 2012. [Online]. Available: https://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/.
[6] A. C. M. Webinar and W. H. Guide, "Simplifying Big Data … with Hadoop."
[7] "Internet Security Threat Report | Appendices," vol. 21, no. April, 2016.
[8] "Internet security threat report," vol. 20, no. April, 2015.
[9] "INTERNET SECURITY THREAT REPORT," vol. 19, no. April, 2014.