IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 01-05
www.iosrjournals.org

Big Data: How it is Generated and its Importance

Asst. Prof. Mrs. Mugdha Ghotkar[1], Ms. Priyanka Rokde[2]
[1](pradnya2286@gmail.com, Dept. of Computer Science, VMV Commerce, JMT Arts & JJP Science College, Nagpur / R.T.M. Nagpur University, India)
[2](priya.rokde@gmail.com, Dept. of Computer Science, Taywade College (Formally known as Arts, Commerce & Science College), Koradi / R.T.M. Nagpur University, India)

Abstract : *More and more data are being produced by an increasing number of electronic devices surrounding us and on the internet. The amount of data and the frequency at which they are produced are so vast that they are usually referred to as "BIG Data". "Big Data" is the term that describes the large volume of data – both structured and unstructured. This paper discusses the sources and Characteristics of Big Data and the importance of it. Big Data can be analyzed for insights that lead to better decision and strategic business moves.*

Keywords – *Big Data, Human generated, Machine generated, Social media, Structured data, Unstructured data*

1. Introduction

Computer, till World War-II was available only for military use. Post World War-II advanced research in semiconductor technology revolutionized computers. This development enabled computers to be employed for industrial and personal use.

Introduction of computers in civil life brought huge advantages. One of the advantages was it could easily generate information and exchange it between human-to-human, human-to-machine and machine-to-machine, that too in both directions. Every such piece of information is treated as data and when systems and/or humans, individually or combined, generate enormous amounts of information it is called "Big Data". Businesses and organizations utilize this Big Data to provide better service to their customers and increase profits. It is thus important to study different ways of Big Data generation, its storage and its application for commercial use.

This paper is an effort to identify and categorize different ways in which Big Data is generated. The idea is to understand that business and organizations are collecting and using large volume of data to improve their end products whether it is safety, reliability, healthcare or governance.

2. What is Big Data?

"Big Data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." – [1]

"Big Data are 'high–volume', 'high–velocity' and/or 'high–variety' information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" – Gartner, 2012

Big Data is torrent of information generated by machines or humans which is so huge that traditional database failed to process it. To understand the scope of Big Data, let us consider this example:
Twitter processes 1 Petabyte (100 Terabyte) of data daily while Google processes 100 Petabyte data.

- In 2011 alone, mankind created over 1.2 trillion GB of data.
- Data volumes are expected to grow 50 times by 2020.
- Google receives over 2,000,000 search queries every minute.
- 72 hours of video are added to YouTube every minute.
- There are 217 new mobile Internet users every minute.
- Twitter users send over 100,000 tweets every minute (that's over 140 million per day).
- Companies, brands, and organizations receive 34,000 "likes" on social networks every minute.

Big data can be described by the following characteristics

i) Volume – Amount of data
ii) Velocity – speed of data in and out
iii) Variety – range of data types and sources

It is generally known as "3V"s of Big Data.

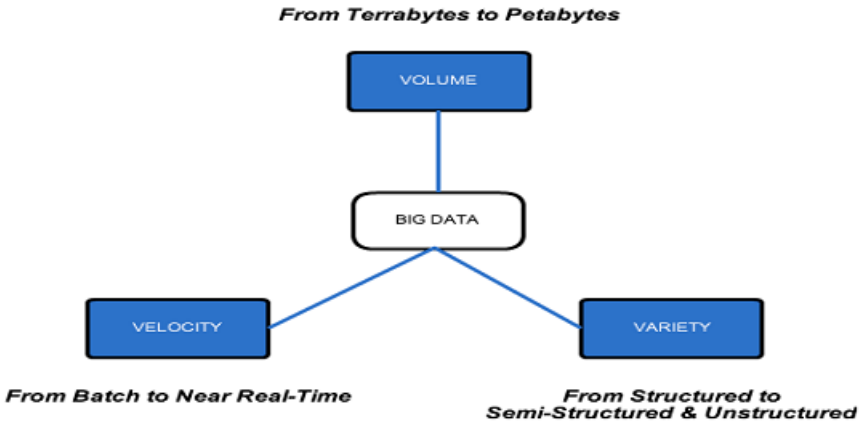

Fig. 1 Characteristics of Big Data

2.1 Volume

90% of all data ever created, was created in the past 2 years. From now on, the amount of data in the world will double every two years. By 2020, we will have 50 times the amount of data as that we had in 2011. The sheer volume of the data is enormous and a very large contributor to the ever expanding digital universe is the Internet of Things with sensors all over the world in all devices creating data every second. The era of a trillion sensors is upon us.

If we look at airplanes they generate approximately 2.5 billion Terabyte of data each year from the sensors installed in the engines. Self-driving cars will generate 2 Petabyte of data every year.

2.2 Velocity

The Velocity is the speed at which the data is created, stored, analyzed and visualized. In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers and servers required substantial time to process the data and update the databases. In the big data era, data is created in real-time or near real-time. With the availability of Internet connected devices, wireless or wired, machines and devices can pass-on their data the moment it is created.

The speed at which data is created currently is almost unimaginable: Every minute we upload 100 hours of video on YouTube. In addition, every minute over 200 million emails are sent, around 20 million photos are viewed and 30.000 uploaded on Flickr, almost 300.000 tweets are sent and almost 2.5 million queries on Google are performed.

2.3 Variety

In the past, all data that was created was structured data, it neatly fitted in columns and rows but those days are over. Nowadays, 90% of the data that is generated by organization is unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

There are many different types of data such as images, text, network data, geographical data, maps, computer generated simulations, etc. and each of those types of data require different types of analyses or different tools to use. Social media like Facebook posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are.

3 How is Big Data Generated?

Big Data is not new, it existed even before the phrase "***Big Data***" was invented. Big Data became important only after emergence of Social Media in 2008.

Big Data can be generated by humans, machines or humans-machines combines. It can be generated anywhere where any information is generated and stored in structured or unstructured formats. It can be generated in industries, in military units, on internet, in hospitals or anywhere else.

Big Data can be broadly categorized on basis of the sources:

i. Machine Generated
ii. Human Generated
iii. Organization Generated

2.1 Machine Generated

It is the biggest source of Big Data. With machine generated data, we refer to data generated from real time sensors in industry machinery or vehicles. Data comes from various sensors, cameras, satellites, log files, bio informatics, activity tracker, personal health care tracker and many other sense data resources.

To make the idea more clear, consider an example of submarine. Almost every parts of submarine generate data constantly. This data comes from sensors attached to it such as radio antenna, rudder, etc. Activity tracker is another example of machine generated Big Data. It tracks the body temperature, distance walked, heartbeat, quality of sleep, and so on. If everyone uses activity tracker, it will generate large amount of personalized data. So in short, machine collects the data from sensors on personal as well as industrial level.

Machine generates the data at real time and normally requires real time action. Real time data often requires in-situ processing. In traditional RDBMS model, data is moved to the computational space. In-situ, we bring the computation where the data is generated.

These sensors & instruments were always generating information but were never considered as source of Big Data. Now when industry has analyzed & accepted its importance, Machines are emerging as largest source of Big Data.

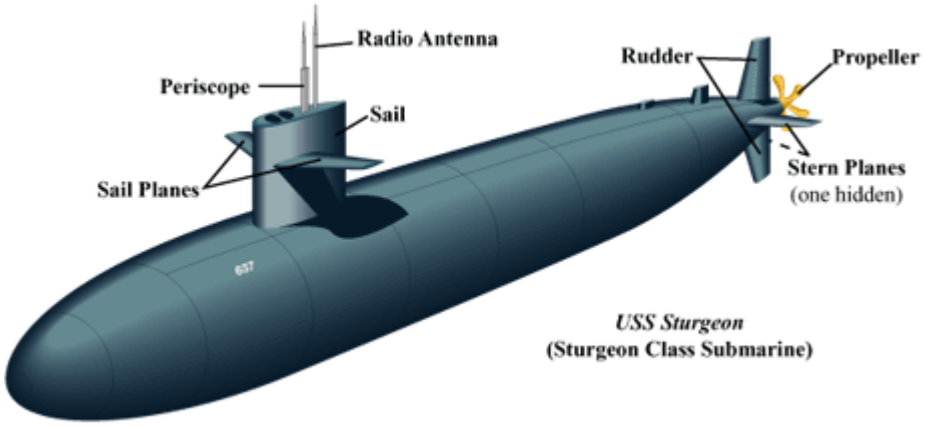

Fig. 2 A Submarine with its sensors

2.2 Human Generated

With human generated data, we refer to the vast amount of social media data as status update, twits, photos, videos, etc. The data generated in this category are normally much unstructured.

Now a day, all most everyone is active on social media. People are generating large amount of data on social networking sites like Facebook, Twitter and Linked in, online photo sharing sites like Instagram, Picasa and video sharing site like YouTube. Large scale data is generated using blogging sites, email, mobile text messages and personal documents. Most of this data is majorly text. So it is not stored in well-defined format. Hence it is known as unstructured data. The problem in processing of this data is its rapid growth, its multiple formats like images, pdf, ppt, xml, web pages and its volume. Here is some analysis of data generated per day.

According to survey done in the year 2012 by Jarmo Roksa, in every minute,

- 2 million searches are sent.
- 100,000 tweets are made.
- 2, 84,500 content elements are shared on Facebook.
- 3600 photos are shared.
- 48,000 apps are downloaded.

Company	Data Processed Daily
eBay	100 PB
Google	100 PB
Facebook	30 PB
Twitter	1PB

Processing of such a large amount of unstructured data has lot of challenges like data acquisition, storage and cleaning.

Social Media gained popularity from 2008 and gave birth to the term Big Data. Due to popularity & success of different social networks, many more sites are launched and contribute to Big Data.

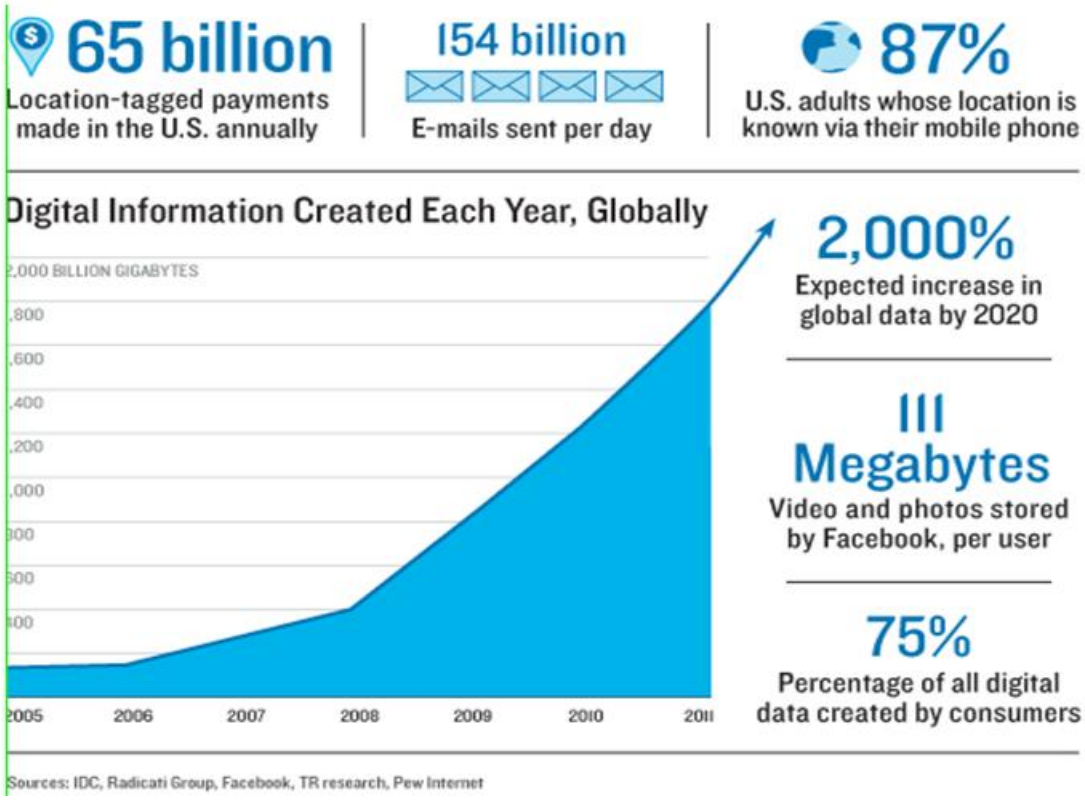

Fig. 3 Digital information created each year globally

2.3 Organization Generated

Organization generated data, is highly structured in nature and trustworthy. The structured data is any data in the form of records located in a fixed field or file. Relational databases are widely used to store this type of data. Traditionally, IT has managed and processed organization generated data in both operational and business intelligence system. Organization stores the data for current and future use as well as analysis of past.

Consider an example of an organization that collects sales transactions. This transaction records can be used to detect correlated products, estimate demand and capture illegal activity. Using proper analytics, organization can build inventories to match predicted growth and demand.

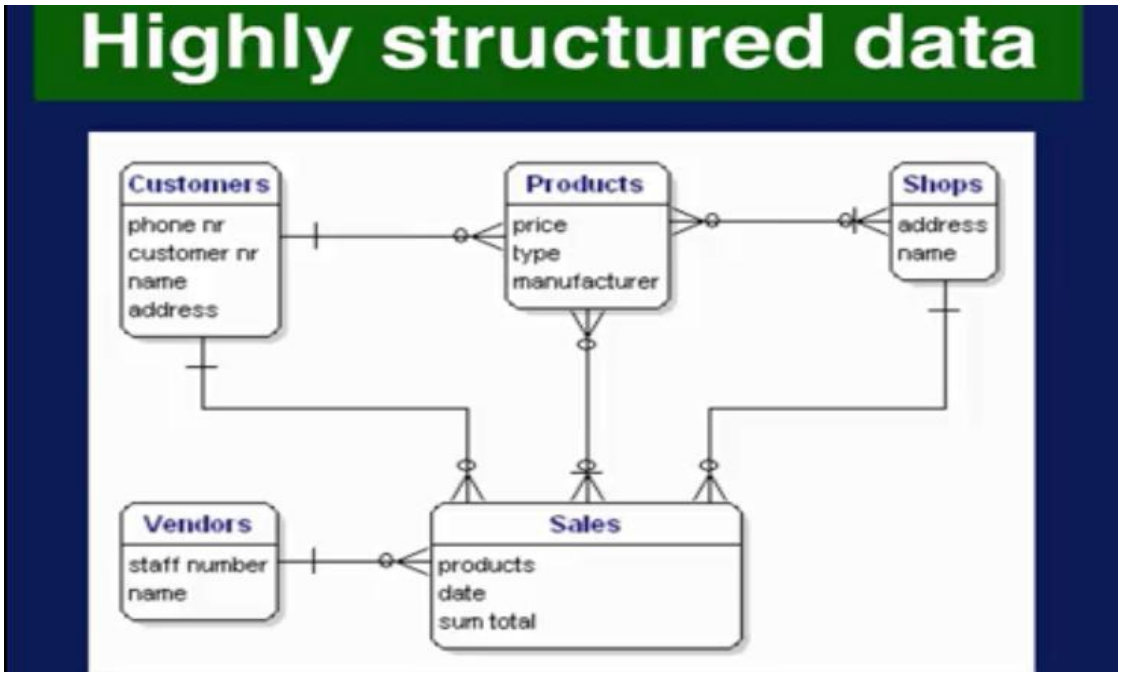

Fig. 4 Organizational Highly structured data

Organizational Data is private to that organization & so is the smallest source in global Big Data scenario. This when combined with public domain data can provide meaningful insights to take faster and better decision.

4 Conclusions

In this paper we have identified that big data can be generated at different points and sources:

- Machine - sensors and instruments
- Human - social media and e-mails
- Organization - ERP and other enterprise applications

It can be concluded as:

- Big data is here to stay and
- We need to identify new sources and points for generating Big Data

Limitations

In This paper we have only covered different ways of generating Big Data. Further scope includes identifying different ways to process, store and use this Big Data.

Application

This paper is useful for people to identify different Big Data sources across different industries.

References

[1] The McKinsey Global Institute, 2012

http://www.sdsc.edu/

SDSC *[San Diego Supercomputer System]*

An Organized Research Unit of UC San Diego. The San Diego Supercomputer Center (SDSC) is considered a leader in data-intensive computing and cyber infrastructure, providing resources, services, and expertise to the national research community including industry and academia.