Session – Summary

Introduction to Big Data - Concept

In the previous modules, the focus was on exploring the different algorithms which can be used for analysis of data. It is equally important to understand how the size of data impacts the method of analysis and the special infrastructure required to handle it. In this lecture, starting from the core concept of data you moved towards understanding the notion of Big Data and the characteristics of Big Data. In addition, you also had a look at what Big Data Problem is and why it needs to be solved.

Concepts of Data

Data is defined as set of values that may be Qualitative or Quantitative in nature.

Qualitative data deals with forms of information that is observable but not necessarily measurable. Eg: color, texture, smell, taste, appearance, beauty, etc.

Quantitative data is related to entities that deal with numbers or that can be measured. Eg: length, height, area, volume, weight, speed, time, temperature, humidity, cost of goods, ages of set of connected people, likes and pokes etc.

The smallest unit of data measurement is byte which is made up of 8 bits. The higher units of measurement are kilobyte, megabyte, gigabyte, terabyte, petabyte and exabyte.

Fundamentals of Big Data

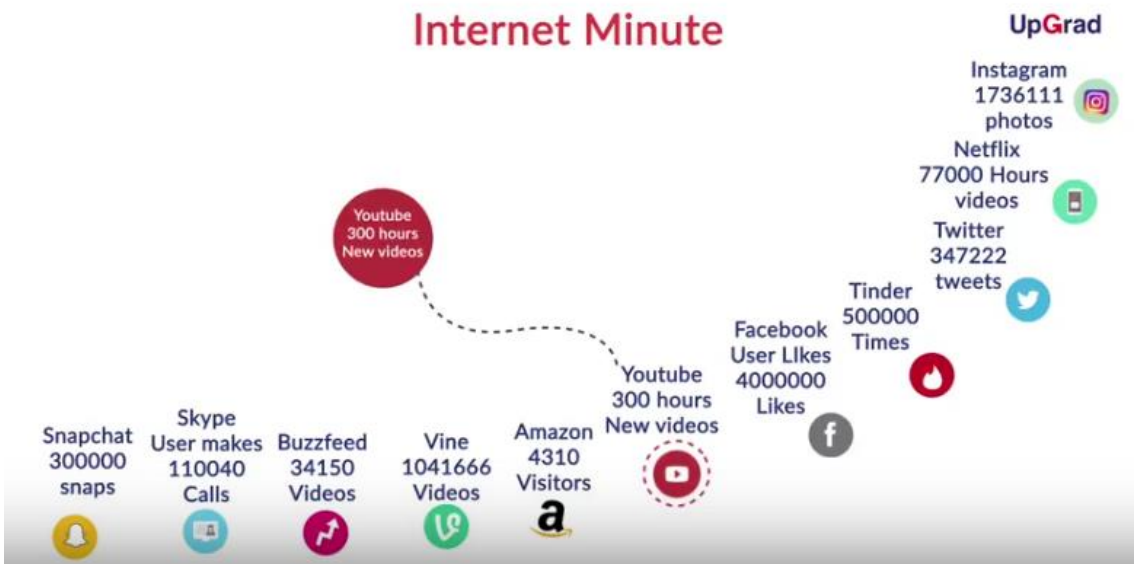

Figure 1: Data Explosion in an internet minute

With the increasing dependence on internet, every online user activity such as a search on google browser, a like on a facebook comment or sending/receiving an email leads to data generation. Refer to Figure 1 to understand the data explosion happening in an internet minute. In an Internet minute on YouTube, 300 hours of video is uploaded, 1.3 million videos are viewed. In that very minute, 2 million plus searches are made on Google, approximately 350,000 new tweets are tweeted on twitter. There are 180 million active websites in the world. You can imagine the amount of data generated every minute.

Data so large and so voluminous that it overwhelms the existing data storage and processing infrastructure, is said to be big enough to be called as - Big Data.

So how big is it really? Does it have to be Google big? Is Facebook's data big enough? Well, here are few statistics which answer the questions.

Company Name	Data Processed per day
Ebay	100PB
Google	100PB
Facebook	600 TB
Twitter	100 TB

An important point to note here is that big data doesn't refer to any specific quantity. Wherever the available infrastructure cannot handle the incoming data, it is Big Data for that set-up. The term is often used when speaking about Petabytes and Exabytes of data.

Characteristics of Big Data

Big Data is a large volume of data (of the size of petabytes, exabytes etc.) which can be mined to carry out some insightful analysis. It may be available in any of the following three forms:

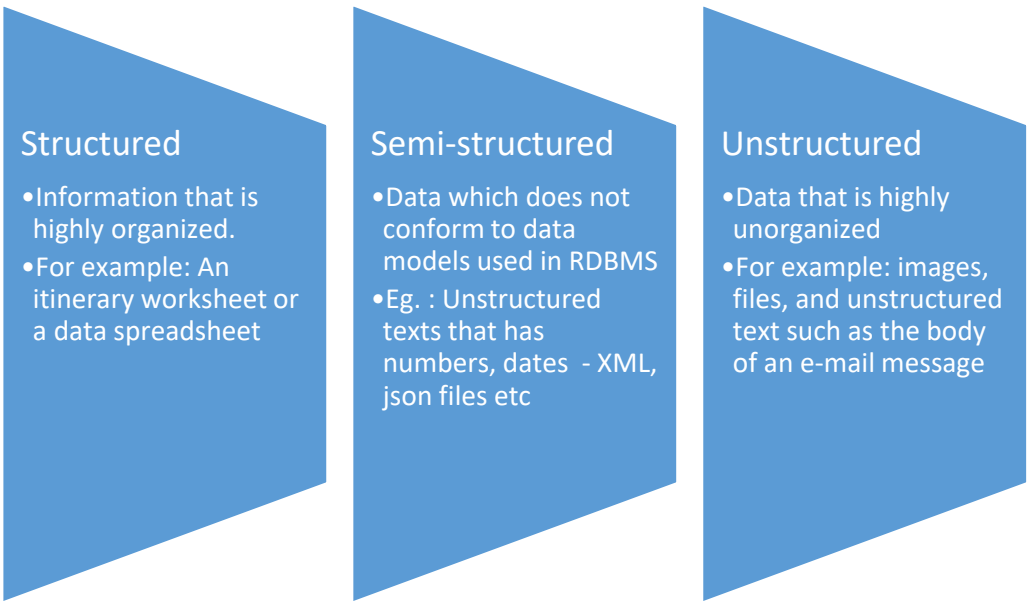

Figure 2: Different forms of data

The characteristics of the data, popularly known as The Fours V's helps to identify Big Data.

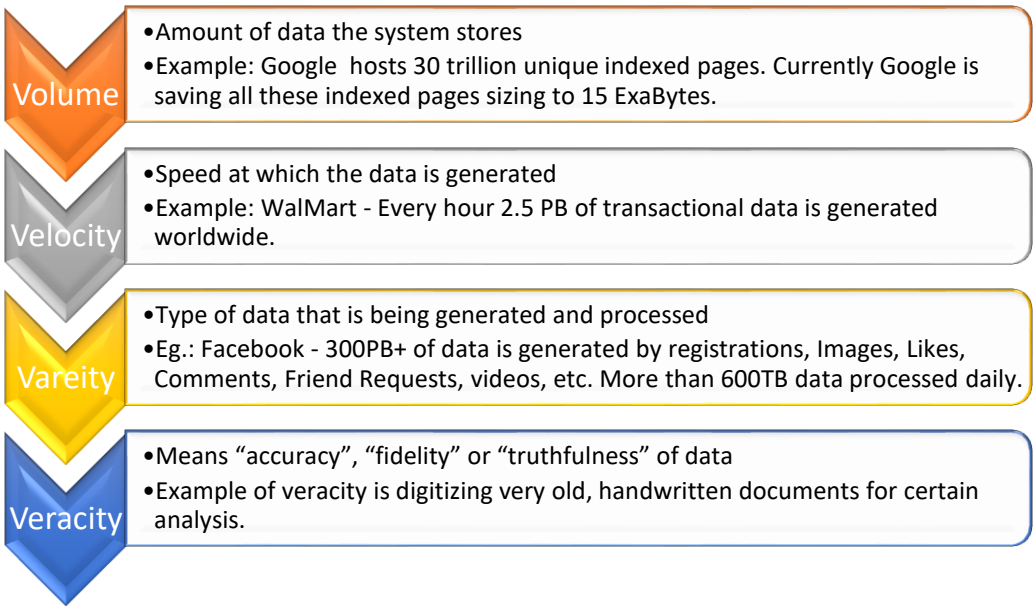

Figure 3: The Four V’s of Big Data

Big Data Problem

Big data by the virtue of its definition hints at the problem it brings. Consider the example below:

Example: An e-commerce company, say BestDeals.com has a storage capacity of 1000 TB.

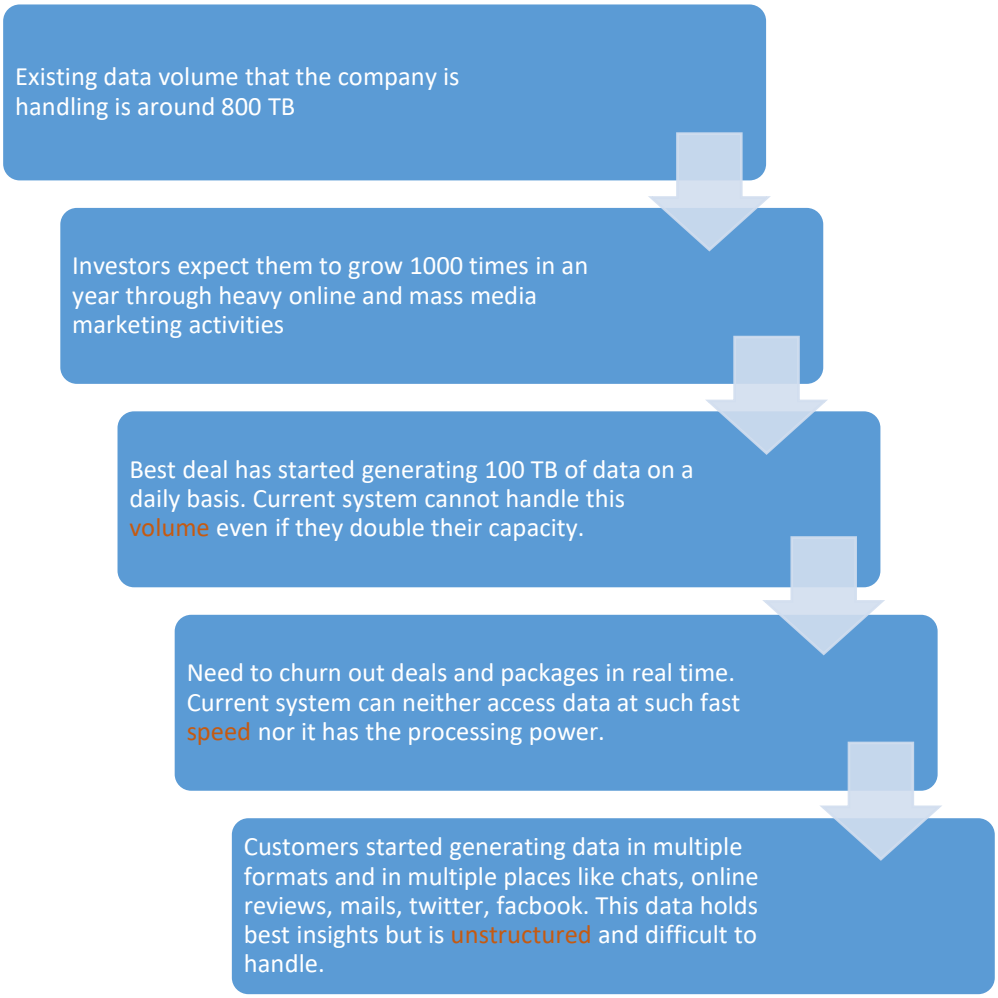

Figure 4: Flow of data generation in the e-commerce example

What are the big data problems that the above situation signifies?

The problems are -

- the volume of data
- the speed at which this volume needs to be accessed and processed
- the different formats of data types that needs to be handle at such large volumes

For years, organizations have been functioning, growing and making profit without Big Data and many may continue to do the same for several years. Then, why is it so important to tackle the 4 Vs? Why not simply ignore big data?

Consider an imaginary machine to find the answer. "Data Master Class" - a hypothetical machine that makes data storage capacity, huge, scalable, affordable and provides with unlimited processing power

Such a machine would turn the 4 Vs of big data upside down into advantages.

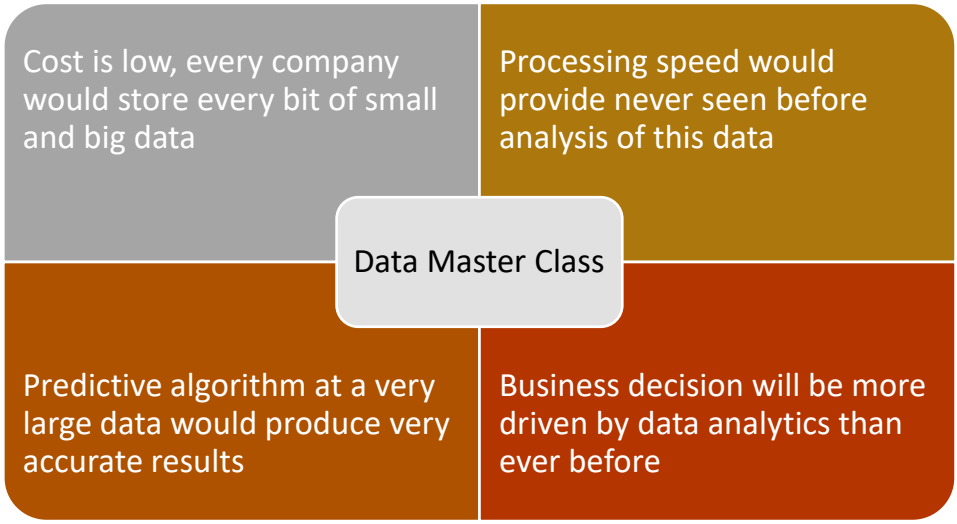

Figure 5: Hypothetical platform - Data Master Class

Every data centric sector would produce better products at lower cost. Telecommunication companies would be able to offer better plans, health care and life sciences will get results in 1/10th the cost and 1/10th the time. Fortunately, this technology is no longer hypothetical and is available in the industry. The technology is called 'Hadoop' and will be discussed in the next module.

Big Data in Industry

The session discussed how some of the bigger firms manage Big Data and what are some of the analysis being carried out by them.

Walmart

Walmart is one of the biggest retail chains in the world. It captures the data generated by every minute activity ranging from a click on its websites to its sales, customers, advertisements etc.

Figure 6: Walmart sources of data

Some of the Walmart specific Big Data solutions include:

- **Social Genome**: This software tries to analyse the various activities of its customers on social media sites to understand their behaviour and interests and recommend products accordingly.
- **Shopycat**: Recommends gifts to customers for their family and friends based on other customer's gift purchases, the budget of the customer etc.
- **Walmart Inventory Management System**: It tries to optimise the inventory of Walmart warehouses in order to ensure that the stores do not have a surplus or deficit of any product.

Vodafone

Figure 7: Vodafone Big Data Solutions

Telecommunication services providers (Telcos) are seeing a massive growth across the globe in terms of volume of data, due to a growing number of users, more affordable services and an exploding use of mobile applications by consumers to access information. This massive amount of data can be analyzed to gain a competitive edge in the market.

Vodafone, being in the telecom industry with many competitors around, need to closely analyse the churn of its customers and the reasons for the same. In addition, it also monitors the behaviour and interests of its customers so that it can recommend them appropriate talk time, netpack offers etc. All of this analysis requires Vodafone to work with the generated Big Data and handle it effectively.

Aadhaar - Unique Identification System

Figure 8: Unique Identification System

Aadhaar is an ambitious project of Indian Govt. to provide unique identification number to the entire population of India. As we know, the Indian population is more than 1.2 Billions, the scale at which the data will be collected for assigning the unique id is humongous. The challenges faced were:

- Collecting and storing this much huge data
- Retrieving and processing
- Avoiding redundant entries
- Connecting it to other Govt. welfare schemes under which people can avail the subsidies

The kind of information collected here was demographic, IRIS like eye and thumb impressions. While collecting this information, the challenges were bringing the demographic information in a structure so that it can be stored properly and processed well. India, having so many villages with same names and even districts named using same names but spelled differently, there were multiple situations where

wrong or redundant data will enter the system. The structuring of data considering all such possible situations was the biggest challenge.

The de-duplication process was then applied before generating the unique id, by applying 600 million matches per day. The process of de-duplication ensured that no single person can get two unique ids.

This processing and managing was very huge and hence the Indian Govt. opted for Big Data solutions. The further challenge faced by Indian Govt. is connecting / integrating this data with the existing subsidies, so that the intended people can avail the benefits without any trouble.