**DEFINITION OF BIG DATA**

Big Data is a term used to describe the voluminous amount of data, which could be in a structured, semi-structured or unstructured form that has the ability to be extracted into information. The **massive** amounts of data make the conventional database management systems inefficient to capture, curation, storage, search, sharing, transfer, analysis, and visualization.

EXAMPLES OF BIG DATA

Aerospace Industry- Continuous data being received from satellites, requires huge storage spaces; lot of processing and analysis of the information received in the data.

Telecom Industry – Tracking of signal strengths and various forms of delay in communication networks require lot of accumulation of data and analyzing the performance of different trans-receivers laid across the coverage area.

Cash Applications- Cash Apps like Venmo, Square Cash, Google Cash are transactional applications that requires capturing of sensitive data such as Debit card details, Capturing multiple transactional activities at a time

Email messages and Text Conversations – People exchange numerous emails and text messages. In a split of a second huge amount of data has to be captured and handled along with maintaining privacy of the data.

Youtube – Example of huge repository of data which has all categories of data Transactional (Like, Comments, Clicks), Periodic and accumulated data. People owning Youtube channels can view analysis of their views, likes, people watching their videos.

VARIATION

Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on examples like language processing, stock markets or tracking signal strength. Each of the measurements will include a small amount of error that varies from measurement to measurement. Every variable in the dataset has its own pattern of variation, which can reveal interesting information. The best way to understand that pattern is to visualise the distribution of the variable’s values.

Variation has its effect on exploratory analysis as it can reveal miraculous facts which might not be noticed otherwise unless they are represented with the correct form of ggplot for example geom\_bar (with necessary binwidth), geom\_bin2d, geom\_hex or geom\_freqpoly etc.

1. Windermere Real Estate collects GPS signals from over 100 million drivers to determine commute times throughout various times of the day in order to help market its properties. What kind of data store and analysis is necessary in this case and would the analysis require real-time response? Discuss the question first on the Discussion Forum and then formulate your own answer in 300-500 words.

GPS signals contains many variables of information such as signal strength, different types of delay, any interferences, noises in the signals, trans-receiver strength.

Collecting GPS signals is a continuous process and the data could be stored in different systems using different capturing techniques such as Excel files, XML, Text File. But these traditional data storage technologies reach their limits when used to store very large amounts of data. Since the data is in the form of Geo Data, we could look for different databases that are capable of fast update and that can foster real time analysis such as non-relational databases. In such cases we should collect data in NoSQL databases that allow data to use other data models beyond relational data model. Additionally, it has the capability of accommodating unstructured data easily.

From the six Vs that qualify the GPS signals as Big Data and reminders for dealing with humongous dataset:

Volume: The incoming data is continuous which can be in the order of Terabytes. The area covered will have 100 million drivers and the data is flowing continuously. Therefore, velocity plays a major role in this

Variability: Signals can fluctuate at the rate of milliseconds, thus for each variable we might observe huge variations in the measurements of variables that makes observations constantly changing.

Velocity: The incoming data is stored and analyzed at a very high rate. Even though its velocity is high it is important that the data must be analyzed in real-time in order to gain a strategic advantage.

Visualization: Visualization is critical as it conveys valuable information from the data captured. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets.