Lecture Notes

Introduction to Big Data

This session helped you understand what exactly big data is and how it is different from the data that organisations used 15-20 years ago. It is always better to understand the data before performing any action on it, because this information will help you choose the appropriate processing technique, which will eventually lead to the selection of suitable big data processing tools.

In this session, starting from the core concept of data, you moved towards understanding the notion of big data and the features that come with it. Also, you took a look at what the problem with big data is and why it needs to be solved.

After completing this session, you have thoroughly understood the importance of the various nuances of big data.

What is Big Data

In the lectures, we discussed that big data shares the same definition as data, i.e. “some existing information or knowledge is represented or coded in some form suitable for better usage or processing,” with the only difference that it is enormous in size.

It’s not only about the present volume of the data, but also that the size of the dataset may not remain static. Big data has the potential to grow exponentially for an indefinite period. It can increase even to the extent where it can’t be managed or processed using traditional techniques such as RDBMSs. The graph below shows the trend of user growth on LinkedIn:

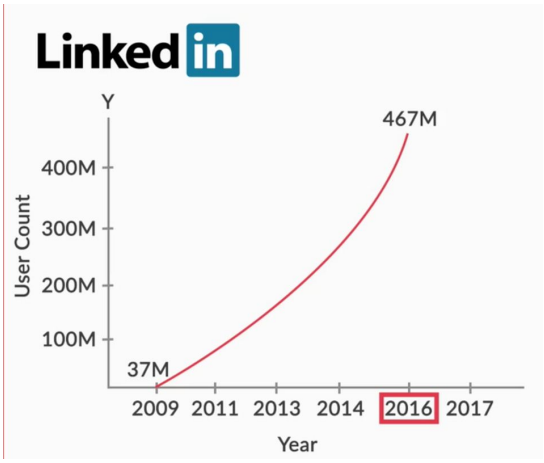

Figure 1: LinkedIn user growth trend

One of the definitions for big data, as given by our industry expert, is:

"If the data has expanded to such an extent that now a single computing system is unable to store and process it, then we can call this data big data."

If you want to quantify big data in terms of size, then you need to consider the maximum possible capacity of an HDD (hard disk drive) available today, which is 16 TB. Using this size as a scale, any data volume in terabytes or petabytes would be considered as big data because it would be difficult for a single system to accommodate a dataset in the range of TBs or PBs.

Big Data: Interesting Facts and Statistics

With the increasing dependence on the internet, every online user activity, such as a Google search, a 'like' on a Facebook post, or sending/receiving of an email, leads to data generation. Refer to Figure 1 to understand the data explosion happening in one internet minute. In an Internet minute on YouTube, 300 hours of video is uploaded, and 1.3 million videos are viewed. In that minute, more than 2 million searches are made on Google, and approximately 350,000 new tweets are tweeted on Twitter. There are 180 million active websites in the world and growing. Thus, you can imagine the cumulative growth of the amount of

data generated every minute.

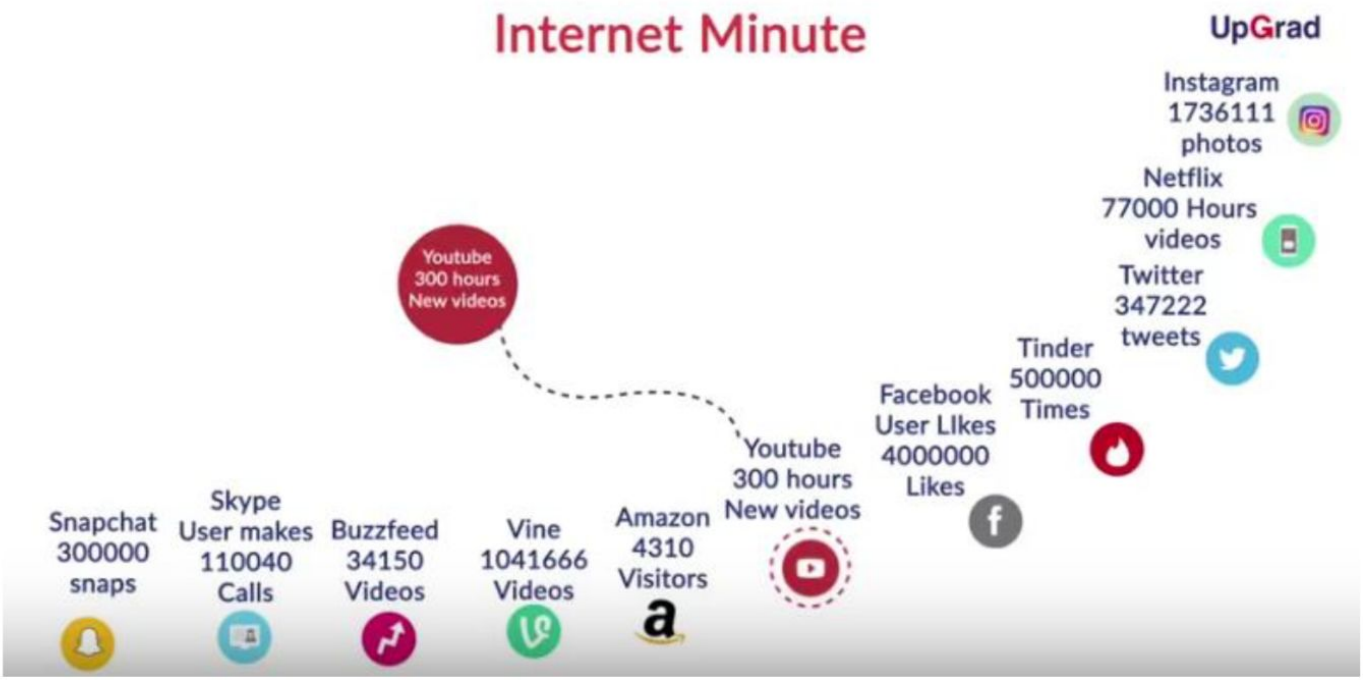

Figure 2: Data explosion in an internet minute

So how big is big data? Does it have to be as big as the data held by Google? Is Facebook's data big enough to be designated as big data? Well, here are a few statistics that'll answer the questions. Refer to figure 2 and 3 to get a fair amount of idea regarding the amount of data Facebook and Google deal with, respectively.

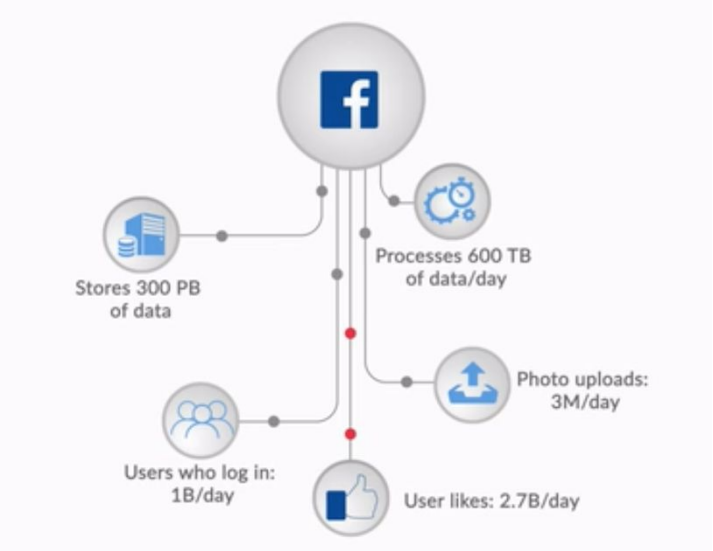

Figure 3: Facebook's Big data facts

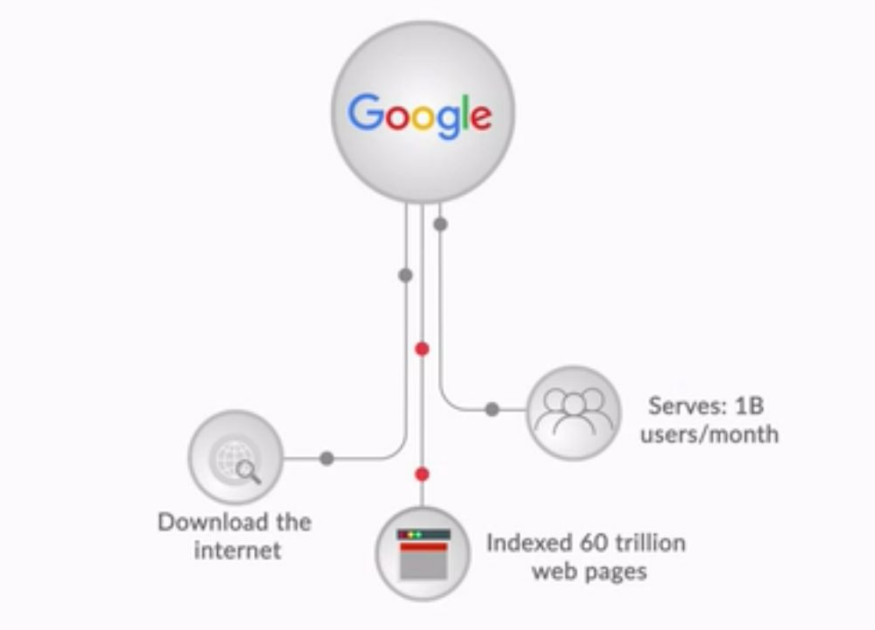

Figure 4: Google's Big data facts

To reiterating what has already been discussed, big data doesn't refer to any specific quantity. Wherever the available infrastructure can't handle the incoming data, it is designated as big data for that set-up. The term is often used when speaking of Petabytes, Exabytes, or even Terabytes of data.

Big Data: Some Basic Industry Application

As starters, just to get a feel of the potential of big data in an industrial set-up, we discussed some of the widely used applications of big data in various industries. They are —

- Application of big data in the healthcare industry
- Application of big data in the retail industry
- Application of big data in the finance industry
- Application of big data in the manufacturing industry

Let's look at each industrial set-up and its big data application in the following sections.

Big Data in the Healthcare Industry:

People visit doctors for consultation. It's a common practice that after detecting a patient's symptoms, the doctor shall suggest some tests. The test reports could either be a paper document, such as a blood test report, or it could be an image, such as a CT scan or an X-ray report. So, all these various test data is stored in a data store, and they are processed and analysed to gain valuable insights.

Figure 5: Big Data in the healthcare industry

Some of the applications of the insights or patterns derived from the analysis of a test report are —

- Real-time monitoring of patients
- Predicting outcomes or upcoming health-related hazards
- Saving cost and energy by minimising hospital visits

Big Data in the Retail Industry:

In the retail industry, the daily sales transactions are recorded and analysed. The rate of data generation is tremendous, and the volume of data is an exorbitant amount as well. This sales data is stored and studied

to discover previously unknown trends and patterns.

Figure 6: Big data in the retail industry

Some of the applications of the insights or patterns derived from the analysis of retail sales data are —

- Understanding customer preferences
- Creating a 360-degree view of a customer profile
- Buying patterns of various products

Big Data in the Finance Industry:
Some advantages of analysing and applying the insights derived from financial data are —

- Detecting and stopping fraudulent transactions
- Designing and modifying predictive models on investment strategies

Big Data in the Manufacturing Industry:
Some of the applications of manufacturing big data analysis are —

- Reading and analysing data from sensors attached to various machine parts
- Proactive maintenance of equipment
- Preventing the loss of machine hours

After learning about the benefits of these industrial applications, we discussed the various aspects of big data. They are —

- **Volume**: The size of the data has to be huge, i.e. in the range of terabytes or even more than that.

- **Rate of change**: The nature of the data has to be dynamic because of the changes in transactions. There could be multiple reasons supporting the changes in transactions, such as a change in the business logic or a change in the requirements.

- **Variety**: Based on the form of data, it can be broadly divided into three categories:
    - **Structured**: The data is stored in a tabular format
    - **Unstructured**: Data that does not have a well-defined structure, e.g. videos and images
    - **Semi-structured**: Data that is partially structured or is a combination of both structured and unstructured data. E.g. emails. Emails are semi-structured because they have a well-defined structure that consists of a sender address, receiver addresses, subject, message body, attachments, etc. But the content mentioned in the subject or the message body is entirely unstructured.

Big Data: Major Sources

You learnt that the sources of big data can be broadly classified into three major categories:

- **People**:

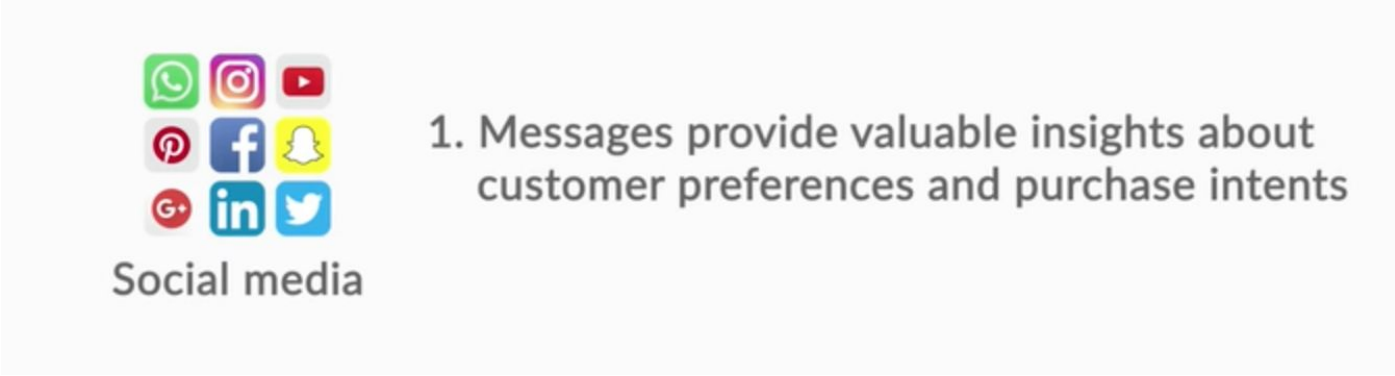

Figure 7: Social media data is an example of data generated by people

Today, people are quite active on the internet through social networking sites such as Facebook, Twitter, and Instagram. On these platforms, they share a lot of information

including what could be a valid opinion regarding a political issue or a post about their recent visit to a hill station. This is considered as data that is 'shared' by people. Even the user ratings for a movie or product can be treated as data generated by people.

- **Machine**: Data that is generated by a machine/computer in a periodic manner or at the occurrence of some event is termed as 'data generated by machines'. Some common examples of data produced by machines are —
    - Data produced by cell phone towers
    - Data produced by RFID tag scanners
    - Data produced by car sensors

Let's look at some examples of data generated by machines through images:

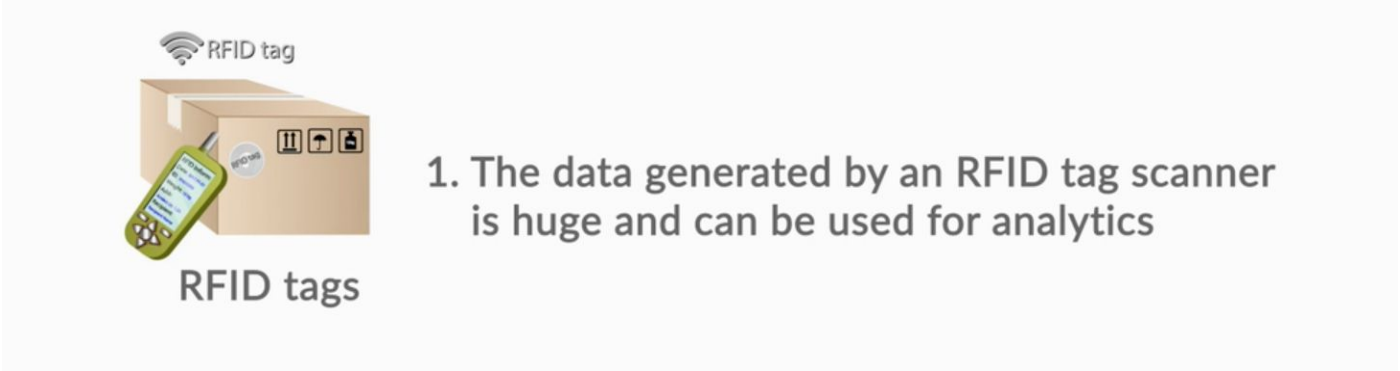

Figure 8: RFID scanner data is an example of data generated by machines

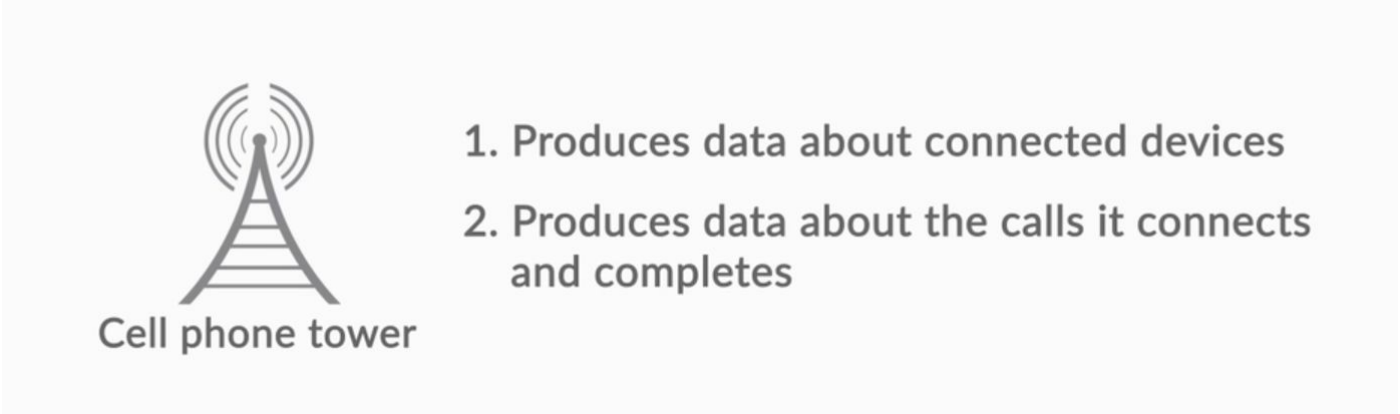

Figure 9: Cell phone tower data is an example of data generated by machines

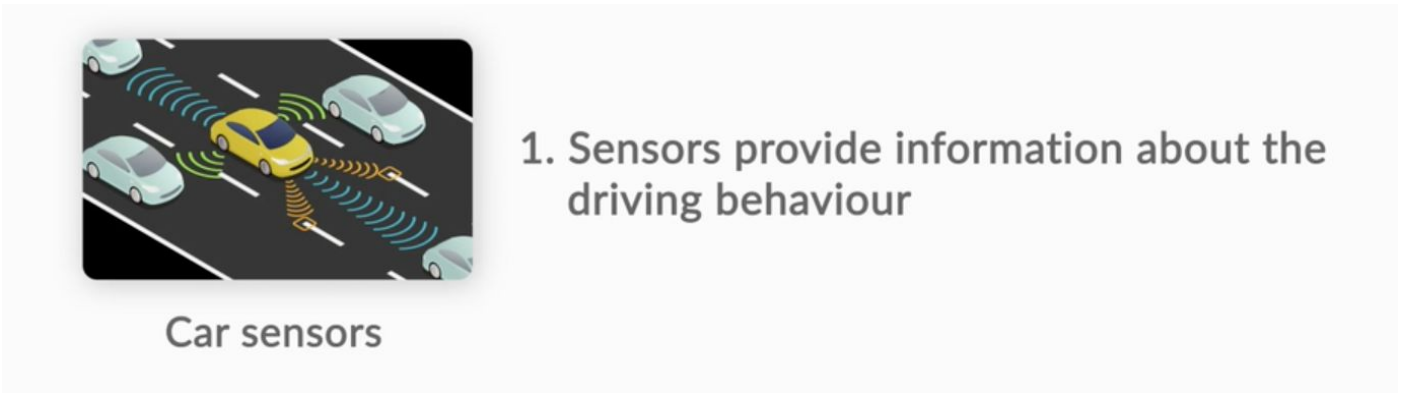

Figure 10: Car sensor data is an example of data generated by machines

- **Organisation**: This refers to the data that is generated by an organisation. This data, in most cases, has a well-defined structure. Some examples of data produced by organisations are internal sales data and customers' demographic information. This kind of data can be integrated with some external data to generate useful insights. Archived organisational data is helpful in performing comparative analyses with the latest data, as well as historical analyses, which are helpful in the prediction of future trends.

Big Data: Major Types

Data is broadly categorised into three categories. They are —

- Structured
- Semi-structured
- Unstructured

The characteristics and examples of the aforementioned data types are mentioned below:

Name	Characteristic	Example
Structured	1. Constitutes 5% of all the information being processed. 2. Has a definite schema or structure	1. Stored in DB tables
Semi-structured	1. Constitutes 5-10% of all the information being processed 2. Cannot be stored in DB tables 3. Metadata provides structure	1. XML 2. JSON 3. Data from social networking websites 4. Email and server logs
Unstructured	1. Constitutes 80% of all the information being processed 2. Does not have a definite schema or structure	**1. Human-generated** a. Social media activities b. Uploaded photos, presentations, documents, videos, etc. **2. Machine-generated** a. Satellite images b. Weather data c. CCTV footage d. RADAR and SONAR data e. Vehicular data

Vs of Big Data

Until this point, you understood the two essential characteristics of big data, i.e. it is vast and diverse. Big data is said to be diverse because it can include structured, semi-structured, and unstructured data. Apart from these two characteristics, big data can also be described with other characteristics. These are represented using the 4 Vs, often denoted as the '4 Vs of big data'. They are —

- **Volume**: This represents the amount of data generated by a company/organisation. The size of big data typically ranges from petabytes to exabytes. For example, remember the amount of data generated in an internet minute, which leads to data explosion. Search engines such as Google, Yahoo, etc. deal with enormous volumes of data too.

- **Velocity**: This indicates the rate at which data is generated/consumed. Social media sites such as Twitter, Facebook, etc. create data from every activity that a user performs, leading to an enormous amount of data generated every minute.

- **Variety**: This represents the different types of data being generated. For example, the various types of data collected from Gmail may be sign-up/registration data, user login data, inbox emails, sent emails, etc.

- **Veracity**: This represents the quality and trustworthiness of data. Previously, veracity was not considered to be a characteristic of big data. But with the increasing analyses on generated data, veracity plays an important role.

Apart from the 4 Vs of big data, researchers have come up with some additional Vs. They are —

- **Variability**: This refers to the way the meaning of the data is always changing. In other words, the meaning of the data changes according to how the data is collected. One of the simplest examples of this concept would be the numerous words in the English language with multiple meanings. So, interpreting a word without considering its surrounding words will not give you its exact meaning. If you analyse the entire sentence in which the word is present, you may get the exact sense of the word.

- **Validity**: This refers to the correctness of the values present in the working dataset. The dataset is bound to have some anomalies in it, such as missing values, incorrect values, etc. It's important to get rid of these anomalies before processing this dataset. The presence of these defects will influence the output, thereby giving you erroneous results. One example of invalid data would be negative values in an age column.

- **Volatility**: It refers to determining the expiry or life expectancy of data. For better results, based on some business requirements, it's sometimes required that you archive or eliminate the old data

completely. An example of volatility would be the data corresponding to taxes such as VAT after the implementation of GST, which is archived because it will never be utilised again.

The Difference Between Veracity and Validity:

To find out how veracity is different from validity, we discussed a few examples from our day-to-day lives.

Veracity refers to the truthfulness of data. In other words, it determines whether the given data is trustworthy or not. For instance, a piece of data will suffer from veracity issues if its source is not credible. Suppose that a student is working on a case study on the 'Pollution of Rivers in India'. To create this, he/she must collect water samples from various rivers and must get them examined at a lab to determine the degree of pollution in each stream. Here, if the student has collected a single sample from every river, then you can say that the data collected suffers from veracity issues. This is because, ideally, the entire course of the river is not polluted. A river is the least polluted near its origin, and the most polluted towards the end of its path where it meets the sea. Let's assume that, in this example, the student has collected a single sample from each river at its origin. If the data is analysed, the lab readings will show that the rivers are not polluted. However, the rivers may be highly polluted after they've flown through highly populated regions. So, ideally, the data would not suffer from veracity issues if, for each river, multiple samples were collected at various points of its entire course.

Validity refers to the correctness of the data. While performing lab experiments, if properly calibrated standard instruments are not used, then the readings will be inaccurate. It can then be established that these readings are not valid.

Digitisation: One of the Major Cause for Big Data Generation

In this lecture, you learnt how digitisation has led to the exponential growth of big data. One of the first examples of digitisation was the advent of emails. With emails, message-sharing became convenient and time-saving. Hence, we still use emails for both personal and professional communications. Consequently, we've never looked back. Almost all the services around us have become digital, from booking tickets to transferring money to another bank account. Let's see how the digital data generated by a cell phone is utilised in the flow diagram below.

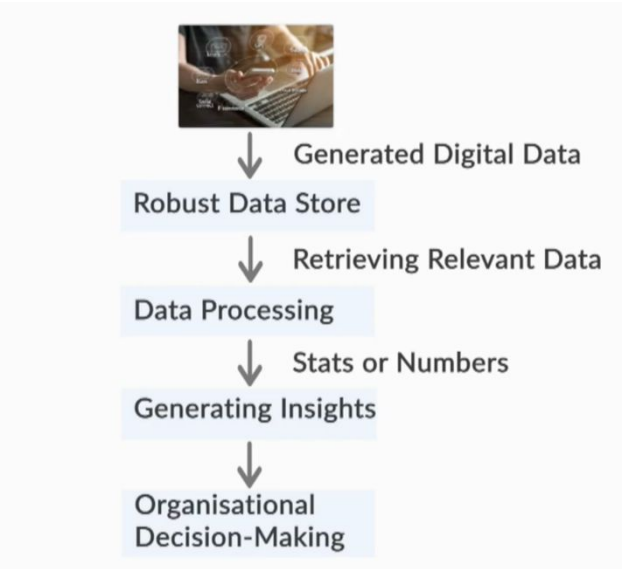

Figure 11: Storing and processing of digital data generated by cell phone

As already discussed, today, accessing the internet has become easily affordable, and internet usage has grown widespread. Each online activity, such as sending emails, booking movie tickets, uploading blogs, posting reviews on an e-commerce portal, etc. generates a massive volume of data. This growing use of the internet facilitates the generation of digital data and gives easy access to the generated data. Thus, we have entered a world that has become data-driven. Organisations and various industries leverage the power of this huge digital data reservoir to take important decisions. Additionally, the increasing usage of Internet of Things (IoT) devices has accelerated the generation of digital data.

Some of the case studies that make use of this digital data are mentioned below:

Car Digitisation:
It is assumed that all the moving parts of a car are equipped with a sensor. So, whenever the vehicle is in motion, these sensors kick in and start generating data. For example, the steering wheel, pistons in the engine, brake plates, and other sensor-fitted parts can generate data.

The data collected from these sensors gives the following information:

- The car's style of driving
- Engine temperature
- Oil temperature
- The duration the car was idle
- The car's start time

- Pedal positions

From the aforementioned information, the derived insights are —

- **Driving styles of consumers**: If the drivers are driving rash, then they can be educated about driving the right way
- Predicting the warranty costs
- Real-time health monitoring of the car

Analysing car digitisation problems in relation to the discussed Vs:

- **Volume**: Sensor data from cars across the globe
- **Velocity**: Gathering data in real time and then processing it
- **Variety**: Data is semi-structured

Sentiment Analysis Using Social Media Data:

Traditionally, after approximately six months of launching a new product, the organisation starts performing offline market surveys to collect the users' feedback.

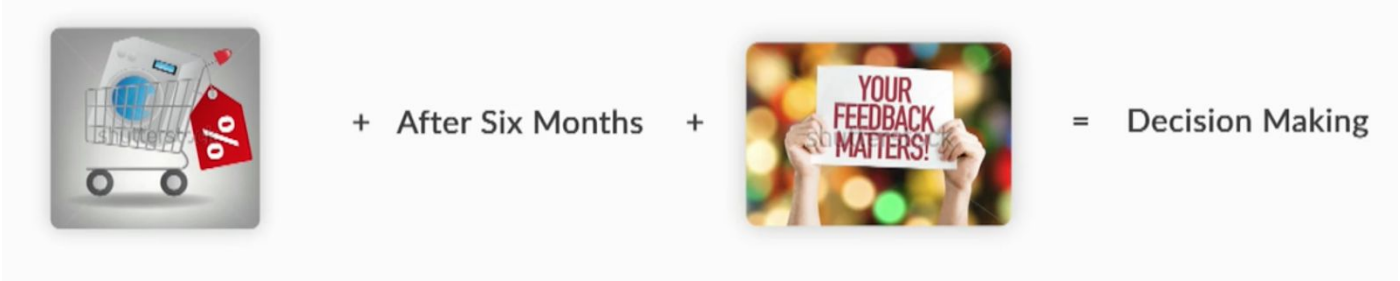

Figure 12: The older method of determining customer feedback

Today, because of the advent of social media and e-commerce websites, organisations have access to a myriad of unstructured data, which includes customer reviews, customer feedback, ratings, etc. By analysing these datasets, organisations are able to get instantaneous feedback regarding their products. Here's an image depicting how decision-making takes place in an organization:

Figure 13: A new method of determining customer feedback

Analysing the sentiment analysis problem in relation to the discussed Vs:

- **Volume**: A huge amount of social data
- **Velocity**: The data gathered is in real time and then processed
- **Variety**: The data is unstructured

Conventional Data Processing Systems and Big Data

Previously, the most commonly used data-storing and data-processing systems were RDBMSs (relational database management systems). An RDBMS uses tables to store data in a row-column format. These tables have a well-defined schema/metadata, and the data that is stored in each table must comply with the underlying schema. In most cases, operational (transactional) data is stored and analysed using RDBMSs. Storing and processing of big data is not done using traditional systems anymore.

Traditional systems fail to manage big data because of its huge size and diversity. Some big data sets are generated rapidly, because of which storing and processing such high-speed data is beyond the scope of traditional systems. Traditional systems perform well when the data is free of noise and bias. Also, sometimes the data at hand may not be credible. The lack of a mechanism to detect credibility renders traditional systems useless for the handling of big data.

Some of the typical reasons why traditional systems fail in handling big data are —

- **Types of data**: Traditional databases only support structured data. Today, big data also includes semi-structured and unstructured data, which cannot be processed using traditional systems.
- **Volume of data**: Traditional systems cannot store and process massive volumes of data efficiently. Big data processing systems store data in distributed file systems, which lead to efficient storage and processing of data.
- **Scaling**: Big data processing systems follow the scale-out architecture and distributed computing for data processing. Thus, the load of computation is shared among multiple systems. This is not possible in the cases of traditional systems because they run on single servers.
- **Data schema**: Traditional databases follow a strict schema for the data. Big data is first stored in a raw format, and then a schema is applied on it while reading it.