Big Data Analytics — UNIT-1

* **Big Data:** Bigdata is termed as collection of datasets, so large and complex that it becomes difficult to process using traditional data processing apps.

  Eg - 40 Exabytes is stored by one person using smartphone every month.

* **5Vs of BigData:**
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Value

  } 5Vs

Volume: Hospitals and clinics across world generates massive information (approx 2314 EB of data collected annually).

The amount of data generated is known as volume in bigdata.

Velocity: In the form of patient's records, & test results, all these data are generated at a very high speed which attributes to velocity.

Velocity in bigdata is defined as the speed at which data is generated, collected & analyzed.

Variety: It refers to various data types like structured (Excel records), semi-structured (Log Files), unstructured (X-Ray images).

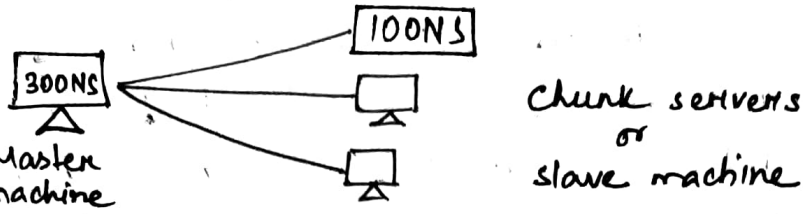
It is termed as different types of either structured or unstructured or semistructured data.

Veracity: Accuracy and trustworthiness of generated data is known as veracity, in Bigdata.

Value: It is referred as the ability to turn data into useful insights.

Analysing all the data will benefit the medical sector by enabling faster disease detection and better treatment and reduced cost.

* Usecase - Google File System:



- As a number of internet users' group throughout the last decade, Google was challenged with how to store so much user data on its ~~search queries~~ traditional servers with thousands of search queries placed every second, the retrieval process was consuming hundreds of megabytes and billions of CPU cycles. Google needed an extensive, distributed, highly fault tolerant file system to store and process the queries. In response, Google developed Google File System (GFS).

The GFS architecture consists of one master machine and multiple slave machines or chunk servers.

The master machine contains meta data and the chunk servers store data in a distributed manner. Whenever a client on an API wants to read the data, the client contacts the master which then responds to the metadata information, Client uses this metadata information to send a read/write request to the slave machines to generate a response. The files are divided into fixed sized chunks and distribute the chunk servers on slave machines.

Features of the chunk servers include:

Each piece has 64MB of data, 128MB of data from Hadoop version 2.0 & onwards. By default each piece is replicated on multiple chunk servers 3 times, if any chunk server crashes, the data file is present in other chunk servers.

1. Sampling and resembling

Sampling: Sampling is a process of selecting group of observations from the population, to study of the characteristics of the data to make conclusion about the population.

Example: Covaxin is tested over thousand of males and females before giving to all people of country.

Types of Sampling:

Sampling is classified into two major groups.

- probability Sampling
- Non-probability Sampling

probability Sampling:

In this type, data is randomly selected so that every observations of population gets the equal chance to be selected for Sampling.

probability Sampling is of 4 types

1. Simple Random Sampling
2. cluster Sampling
3. Startified Sampling
4. Systematic Sampling

Non-probability Sampling:

In this type, data is not randomly selected. It mainly depends upon how the statistician wants to select the data.

The results may or maynot be biased with the population.

Unlike probability Sampling, each observations of population doesn't get the equal chance to be selected for Sampling.

Non probability Sampling is of 4 types:

1. convenience Sampling
2. Judgmental / purposive Sampling
3. Snowball / Referral Sampling
4. Quota Sampling.

Sampling Error:

Error which occur during sampling process are known as Sampling Error.

$$\text{Sampling Error} = z \times \frac{\sigma}{\sqrt{n}}$$

Where z - Scope value based on confidence interval

σ - population standard deviation

n - Sample size

SamplingError can be reduced by

- Increasing the sample size
- classify population into different groups.

Advantage of Sampling:

- Reduce cost and Time
- Accuracy of Data
- Inferences can be applied to a large population.
- Less resource needed

Resampling:

Resampling is the method that consists of drawing repeatedly drawing samples from the population.

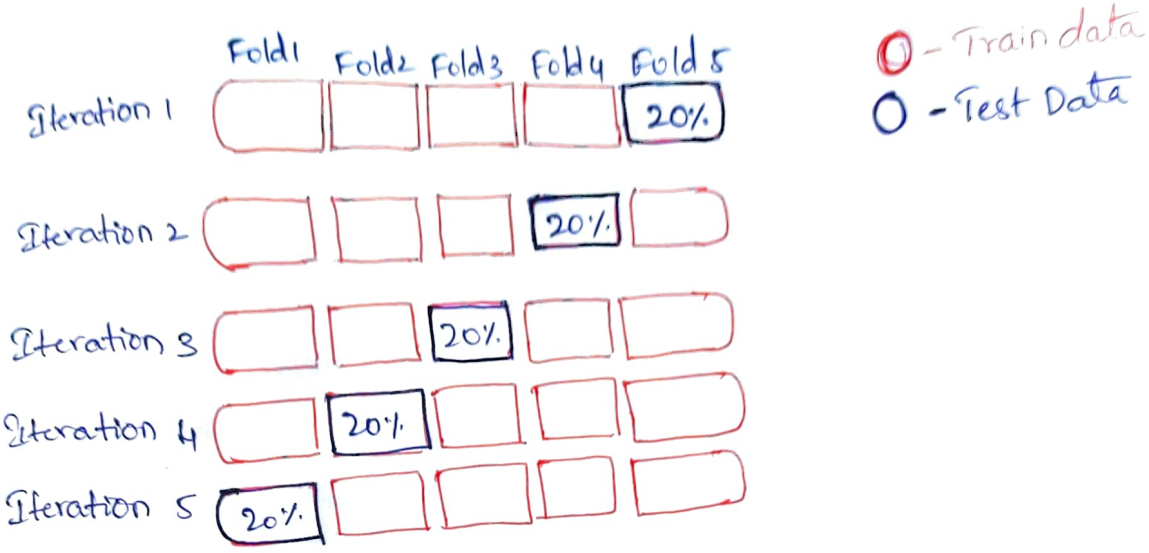
Types of Resampling:

Two common method of Resampling are:

- K-fold cross-Validation
- Bootstrapping

K-fold Cross-Validation:

- In this method population data is divided into k equal sets in which one set is considered as the test set for the experiment while all other set will be used to train the model.
- In first experiment, first set is considered as the test set and all other as trained set.
- process will be repeated k-time by choosing different sets as a test set.



Bootstrapping:

In Bootstrapping, Samples are drawn with replacement (i.e., one observation can be repeated in more than one group) and the remaining data which are not used in samples are used to test the model.