Big Data

**Big data** is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with.

Eg. IoT sensor data, large medical data, aadhar data, etc

**Characteristics:**

**Volume - Quantity of data**

The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not.

**Variety - Nature of data**

The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

**Velocity - Speed of data generation**

In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data are produced more continually. Two kinds of velocity related to Big Data are the frequency of generation and the frequency of handling, recording, and publishing.

**Veracity - Quality of data**

It is the extended definition for big data, which refers to the data quality and the data value.The data quality of captured data can vary greatly, affecting the accurate analysis.

**Architecture of Big data analytics:**

5C architecture:

connection, conversion, cyber, cognition, and configuration

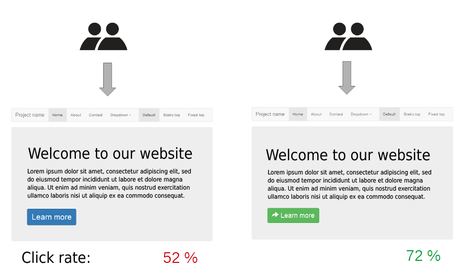
6C architecture:

* Connection (sensor and networks)
* Cloud (computing and data on demand)
* Cyber (model and memory)
* Content/context (meaning and correlation)
* Community (sharing and collaboration)
* Customization (personalization and value)

**Techniques for analysing:**

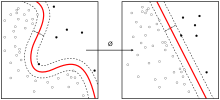
**A/B testing** (**bucket tests** or **split-run testing**): It A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.is a randomized experiment with two

variants, A and B.

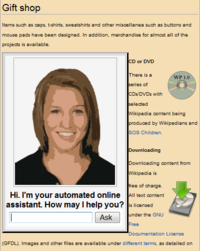


Example of A/B testing on a website. By randomly serving visitors two versions of a website that differ only in the design of a single button element, the relative efficacy of the two designs can be measured.

**Machine Learning:** Machine learning (ML) is the study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task.



**Natural Language Processing:** is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

****

Big data, especially multidimensional data can be represented using data cubes, or mathematically TENSORS.

**Softwares:**

Apache Hadoop - Open source Massive data computations

Apache Spark - Open source distributed general-purpose cluster-computing framework

MapReduce - process large data sets

Apache HBase - storing large data sets

Oracle NoSQL database - storing big data

**YARN:** a cluster management technology and one of the key features in second-generation Hadoop.  
**MapReduce:** a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.  
**Spark:** an open source, parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.  
**HBase:** a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).  
**Hive:** an open source data warehouse system for querying and analyzing large data sets stored in Hadoop files.  
**Kafka:** a distributed publish/subscribe messaging system designed to replace traditional message brokers.  
**Pig:** an open source technology that offers a high-level mechanism for the parallel programming of MapReduce jobs executed on Hadoop clusters.

**Sampling Big Data:**

Sampling refers to selecting a few data points from all the data points to represent the properties of the entire population.

Analysis:

**Descriptive** - Gives what happened in the past using data and reports of past events but cannot give any insight to future.

**Diagnostic** - Seeks to find why certain past events occured by diagnosing reports and past events

**Predictive** - Goes through past data to predict what might happen in the future

**Prescriptive** - Automated Predictive approach using A/B testing

**Criticisms:**

1. Sometimes we make strong mathematical assumptions in analysing big data that may be completely wrong.
2. Big data is usually a buzz word and is used very vaguely.
3. In most cases analysis is done on smaller sets and they give more insight than analysing big data.