**MCA4 Big-Data-CS44**

**CS-44 Unit-1**

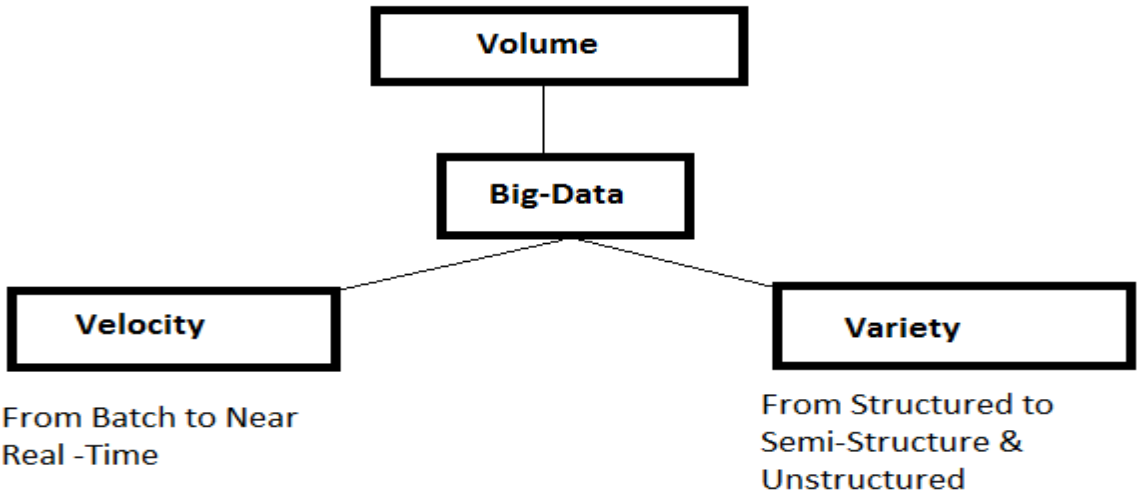
## What is Big-Data?

**Definition :-** Big data is a field of data science that treats ways to analyze, systematically extract information from or otherwise deal with data sets that are too large or complex to dealt with by traditional data-processing system. Data many cases (rows) offer greater statistical power while data with higher lead to a higher false discovery rate.

Big-data challenges include capturing data, data storage, data analysis, search, transfer, visualization, querying, updating, information privacy and data source. The term Big-data has been in use since **1990s** with some credit to **John Mashey**. It includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage and process data within a tolerable elapsed time. It also includes **structured, unstructured and semi-structured data.**

**Big-data** dependent on data size terabytes to many zetabytes of data. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from data-sets that are diverse, complex and massive scale.

**Big-Data** includes four characteristics "**Volume","Variety","Velocity" and "Veracity" known as V-4** for the large data sets. Big data used mathematical analysis, optimization,inductive statistics and concepts from non-linear system identification to infer laws from large sets of data with low information density.



# Characteristics of BIG-DATA

1. **Volume:-** 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.** The amount of data which we deal with it is very large size **Petabytes.**
2. **Variety:-** The type and nature of the data. This helps us to analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video etc.

3. **Velocity:-** The speed at which the data is generated and processed to meet the demands challenges that the path of growth and development. Big data is available in real time. Compared the 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.
4. **Veracity:-** It is extended form of Big-Data which refers to the data quality and the data value. The data quantity of captured data can vary greatly affecting the accurate analysis. Data must be processed with advanced tools (analytics and algorithms) to reveal meaningful information.

# Availability(Sources) of BIG-DATA

The Big-Data come from many sources like-

1. **Social Networking Sites: -** Facebook, Google, Linkedin all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
2. **E-Commerce sites: -** The sites like Amazon, FlipKart, Alibaba etc generates huge amount of logs from which users buys trends can be traced every day billions of data.
3. **Weather Station**:- All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
4. **Share Market: -** Stock exchange across the world generate huge amount of data through the daily trade and transition.
5. **Telecom Company :-** All the Telecom company keep customers calling record information whose data volume large and storage requires **Big-Data.**

# Big-Data Solutions:-

Solution of Big-Data requires the following working structure:-

1. **Storage: -** The huge amount of data **Hadoop** uses **HDFS (Hadoop Distributed File System)** which uses commodity hardware to form clusters and store data in a distributed fashion. It works on write once read many times principles. The Hadoop Distributed file System(HDFS) designed to run on hardware based on open standards or is called commodity hardware. This means the system is capable of running different operating systems such as windows or Linux without requiring special drivers.
2. **Processing:- MapReduce** paradigm is applied to data distributed over network to find required output. Its is a processing technique and program model for distributed computing based on **Java**. This algorithm **MapReduce** contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data where individual elements are broken down into tuples ( Key/Value Pairs).
3. **Analyze:-** To analyze a large volume of data, Big-Data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics.

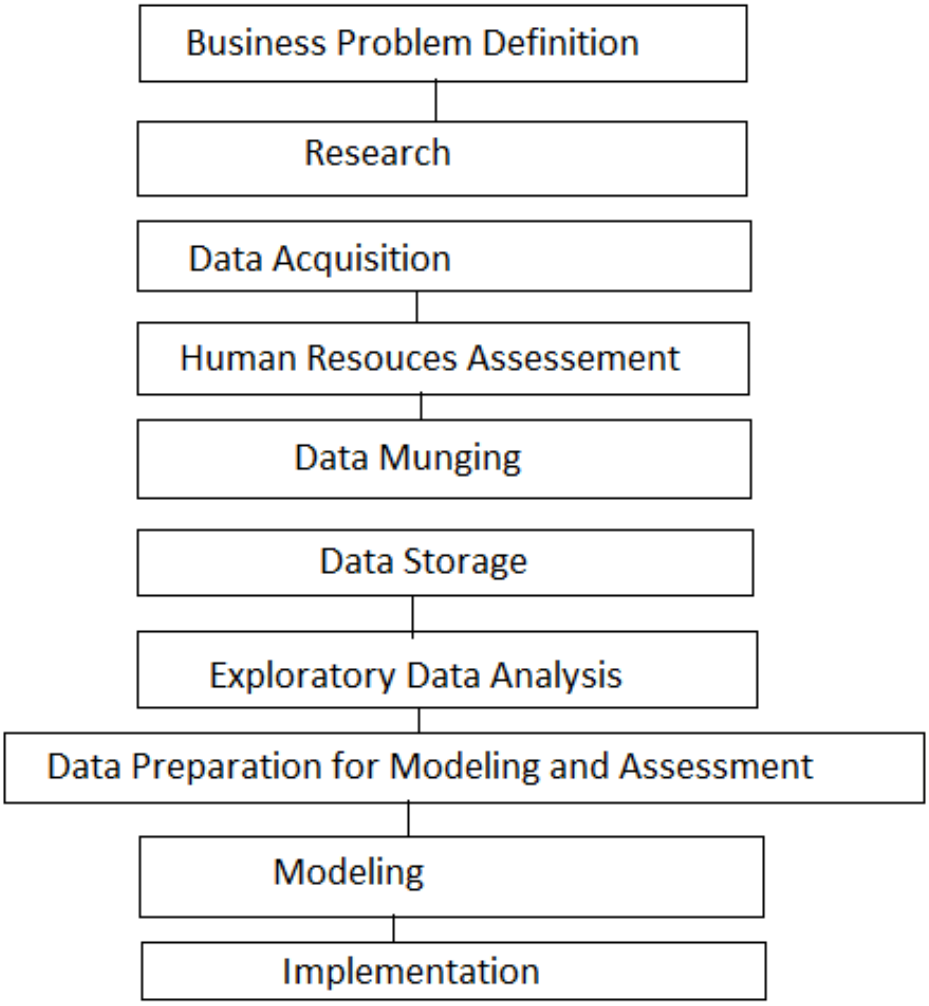
**Note:-**

**Zoho Analytics** ***is a business intelligence & analytics software that transforms large amounts of raw data into actionable reports and dashboards.***

# BIG-DATA LIFE CYCLE

In Big-Data context, the previous approaches are either incomplete or suboptimal such as SEMMA ( Sample Explore Modify Model Access) methodology disregards completely data collection and processing of different data sources. These stages normally constitute most of the work in successful big data projects. A big data analytics cycle can be described by following stage:-

1. **Business problem Definition**
2. **Research**
3. **Data Acquisition**
4. **Human Resources Assessment**
5. **Data Munging**
6. **Data Storage**
7. **Exploratory Data Analysis**
8. **Data preparation for modeling and Assessment**
9. **Modeling**
10. **Implementation**




**Life Cycle of Big-Data**

1. **Business Problem Definition :-** This is the first stage common in traditional BI and Big-Data analytics life cycle. It is non-trivial stage of a big data project to define the problem and evaluate correctly how much potential gain it may have for an organization.
2. **Research: -** This stage involves looking for solutions that are reasonable for any organization. It is adaptive and describes methodology for the future stages should be defined.
3. **Human Resources Assessment: -** This stage of big-data life cycle requires completing the project successfully. It is applied after the problem defined and method describe for optimal solution.

4. **Data Acquisition:** - This stage of Big-data life cycle is key-section. It defines which type of profiles would be needed to deliver the resultant data product. Data gathering is a non-trivial step of the process. It is normally involves gathering unstructured data from different sources.
5. **Data Munging: -** Once the data acquisition stages done, it needs to be stored in an easy-to-use format. In order to combine both the data source a decision has to be made in order to make response representations equivalents.
6. **Data Storage:-** Once the data is processed, it sometimes needs to be offer plenty of alternatives regarding this point. The most common alternative is using the Hadoop File System for storage that provides limited version of SQL known as ***Hive Query Languate***. This allows most analytics task to be done in similar ways as would be done in traditional BI data warehouses from the user perspective. Other storage options to be considered are MongoDB, Redis and Spark. This stage of the cycle is related to the human resources knowledge in terms of their ability to implement different architectures.
7. **Exploratory Data Analysis :-** Once the data has been cleaned and stored in a way that insights can be retrieved from it, the data exploration phase is mandatory. The objective of this stage is to understand the data, this is normally done with statistical techniques and also plotting the data. This is a good stage to evaluate the problem definition makes sense or is feasible.
8. **Data Preparation for Modeling and Assessment: -** This stage involves reshaping the cleaned data retrieved previously and using statistical preprocessing for missing values.
9. **Modeling: -** It is normally desired that he model would give some insight into the business. Finally the best model or combination of models is selected evaluating its performance on a left out dataset.
10. **Implementation: -**In this stage the data product developed is implemented in the data pipeline of the company. This involves setting up a validation schema while the data product is working in order to track its performance.