IDA

UNIT I

INTRODUCTION:

In the beginning times of computers and Internet, the data used was not as much of as it is today, The data then could be so easily stored and managed by all the users and business enterprises on a single computer

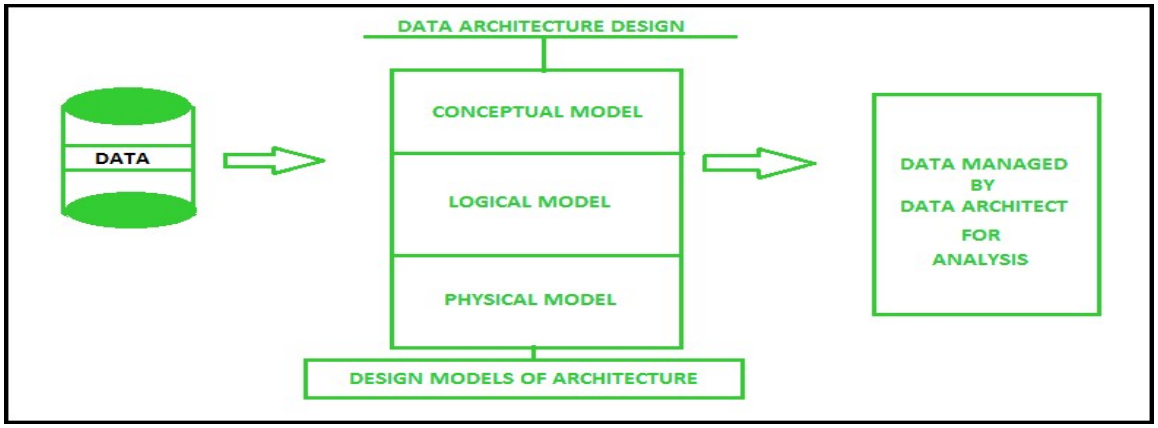
Most of the data is generated from social media sites like Facebook, Instagram, Twitter, etc, and the other sources can be e-business, e-commerce transactions, hospital, school, bank data, etc. This data is impossible to manage by traditional data storing techniques. So Big-Data came into existence for handling the data which is big and impure.

Big Data is the field of collecting the large data sets from various sources like social media, GPS, sensors etc and analyzing them systematically and extract useful patterns using some tools and techniques by enterprises. Before analyzing and determining the data, the data architecture must be designed by the architect.

Data architecture Design and Data Management :

Data architecture design is set of standards which are composed of certain policies, rules, models and standards which manages, what type of data is collected, from where it is collected, the arrangement of collected data, storing that data, utilizing and securing the data into the systems and data warehouses for further analysis.
Data is one of the essential pillars of enterprise architecture through which it succeeds in the execution of business strategy.

Data architecture design is important for creating a vision of interactions occurring between data systems
Data architecture also describes the type of data structures applied to manage data and it provides an easy way for data preprocessing. The data architecture is formed by dividing into three essential models and then are combined :



A data architect is responsible for all the design, creation, manage, deployment of data architecture and defines how data is to be stored and retrieved, other decisions are made by internal bodies.

Factors that influence Data Architecture :
Few influences that can have an effect on data architecture are business policies, business requirements, Technology used, economics, and data processing needs.

- **Business requirements –**
  These include factors such as the expansion of business, the performance of the system access, data management, transaction management, making use of raw data by converting them into image files and records, and then storing in data warehouses. Data warehouses are the main aspects of storing transactions in business.
- **Business policies –**
  The policies are rules that are useful for describing the way of processing data. These policies are made by internal organizational bodies and other government agencies.
- **Technology in use –**
  This includes using the example of previously completed data architecture design and also using existing licensed software purchases, database technology.
- **Business economics –**
  The economical factors such as business growth and loss, interest rates, loans, condition of the market, and the overall cost will also have an effect on design architecture.
- **Data processing needs –**
  These include factors such as mining of the data, large continuous transactions, database management, and other data preprocessing needs.

Data Management :

- Data management is the process of managing tasks like extracting data, storing data, transferring data, processing data, and then securing data with low-cost consumption.
- Main motive of data management is to manage and safeguard the people's and organization data in an optimal way so that they can easily create, access, delete, and update the data.
- Because data management is an essential process in each and every enterprise growth, without which the policies and decisions can't be made for business advancement. The better the data management the better productivity in business.
- Large volumes of data like big data are harder to manage traditionally so there must be the utilization of optimal technologies and tools for data management such as Hadoop, Scala, Tableau, AWS, etc. Which can further used for big data analysis in achieving improvements in patterns.
- Data management can be achieved by training the employees necessarily and maintenance by DBA, data analyst, and data architects.

Understanding various Sources of Data

Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.

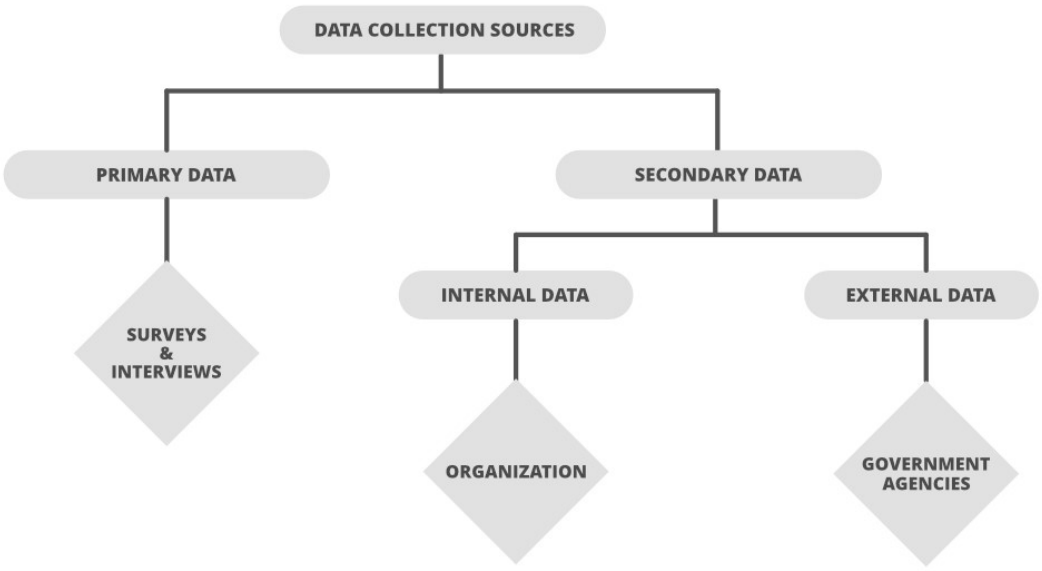
In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as “knowledge”. Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.

Data collection starts with asking some questions such as what type of data is to be collected and what the source of collection is. Most of the data collected are of two types known as “qualitative data“ which is a group of non-numerical data such as words, sentences mostly focus on behaviour and actions of the group.Another one is “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data.

The actual data is then further divided mainly into two types known as:

- **Primary data**
- **Secondary data**



I. Sources of Primary data:

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys.

The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. Interview method:
The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analysing data. Examples are online surveys or surveys through social media polls.

3.Observation method:

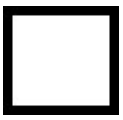
The observation method is a method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behaviour towards the products. The data obtained will be sent for processing.

4 .Experimental method:

The experimental method is the process of collecting data through performing **experiments,** research, and investigation.

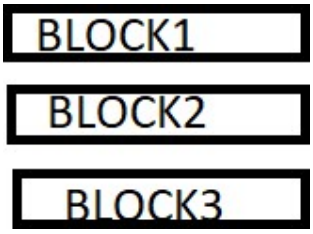
The most frequently used experiment methods are CRD, RBD, LSD, and FD.

CRD- Completely Randomized design is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.



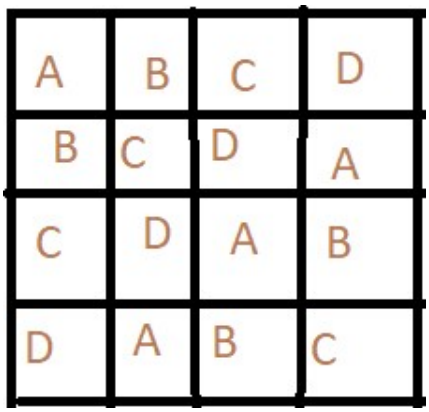
CRD: Experiment treated as a single unit

RBD- Randomized Block Design is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.



RBD :Experiment divided into small units called BLOCKS.

LSD – Latin Square Design is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is a balanced arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Interchange of any rows or columns
will not cause any disturbance in arrangement .Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.



LATIN SQUARE DESIGN: Experiment divided into NxN squares

- **FD- Factorial design** is an experimental design that allows the experimenter to test two or more independent variables simultaneously. It also measures interaction effects of the variables and analyses the impacts of each of the variables.

  Factorial design can be depicted with a numbering in terms of levels of each of the

factors.

II. Source of Secondarydata:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data sources and it has two types of sources named internal sources and external sources.

1 .Internal sources:
These types of data can easily be found within the organization

- **Accounting Resources:**

  This gives so much information which can be used by market researcher

- **Internal Experts:** These are the people who are heading various departments. They can give the idea of how a particular thing is working.

- **Miscellaneous Reports:** Information we can get from operational reports.

The cost and time consumption is less in obtaining internal sources.

If the data available within organization are unsuitable and inadequate, the marketer should extend the research to the external resources.

2.External sources:
The data which can't be found at internal organizations and can be gained through external third party resources is external source data.

External data can be divided into following classes:

a.Government publications:

Government resources provide a rich pool ofdata at free of cost on internet websites.

- **Registrar general of India**-> generates demographic data like gender, age, occupation, etc.

- **Central Statistical Organization**-> publishes National Accounts Statistics such as estimates of national income for several years, growth rate, etc.

- **Ministry of Commerce and Industries**->provides information on wholesale price index related to food, power, fuel, food grains etc.

- **Planning Commission**-> provides statistics of Indian Economy.

- **RBI**-> Provides information on banking savings and investments, currency and finance report.

- **Labour Beuro**-> provides information on skilled,unskilled ,white collared jobs, etc.

- **National Sample Survey**->This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.
- **Department of Economic Affairs**-> It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.
- **State Statistical Abstract**-> This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

The cost and time consumption is more because this contains a huge amount of data.

b.Non-Government Publications- These includes publications of various industrial and trade associations, such as

- Indian Cotton Mill Association
- Various chambers of commerce
- Bombay Stock Exchange
- Various Associations of Press Media.
- Export Promotion Council.
- Confederation of Indian Industries (CII)
- Small Industries Development Board of India(SIDBI)
- Different Mills like - Woollen mills, Textile mills, etc.

c. Syndicate Services-

These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services.

So the services are designed in such a way that the information suits the subscriber.

These syndicate services provide data from both household as well as institutions.

d.International Organizations-

These includes

- The International Labour Organization (ILO)- It publishes data on the total and active population, employment, unemployment, wages and consumer prices
- The Organization for Economic Co-operation and Development (OECD) - It publishes data on foreign trade, industry, food, transport, and science and technology.
- The International Monetary Fund (IMF) - It publishes reports on national and international foreign exchange regulations.

III. Other sources:

- **SENSORS:** Sensor data is the output of a device that detects and responds to some type of input from the physical environment. The output may be used to provide information or input to another system or to guide a process. Sensors can be used to detect just about any physical element.

- **Photosensor-** detects the presence of visible light, infrared transmission (IR) and/or ultraviolet (UV) energy. Ex; card readers, remote controls.

- **Lidar-**Light Detection and Ranging is a laser-based method of detection, range-finding and mapping devices, typically uses a low-power, eye-safe pulsing laser working in conjunction with a camera.

  A Lidar instrument consists of a laser, a scanner, a specialized GPS receiver.

  Airplanes and helicopters are the most commonly used platforms for acquiring Lidar data over broad areas.

- **Charge-coupled device (CCD)–** is a light-sensitive device that stores and displays the data for an image in such a way that each pixel is converted into an electrical charge, the intensity of which is related to a color in the color spectrum.

- **Smart grid sensors**- can provide real-time data about grid conditions, detecting outages, faults, load and triggering alarms.

- Wireless Sensor Networks- combine specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Commonly monitored parameters include temperature, humidity, pressure, wind direction and speed, illumination intensity, vibration intensity, sound intensity

2.SIGNALS:

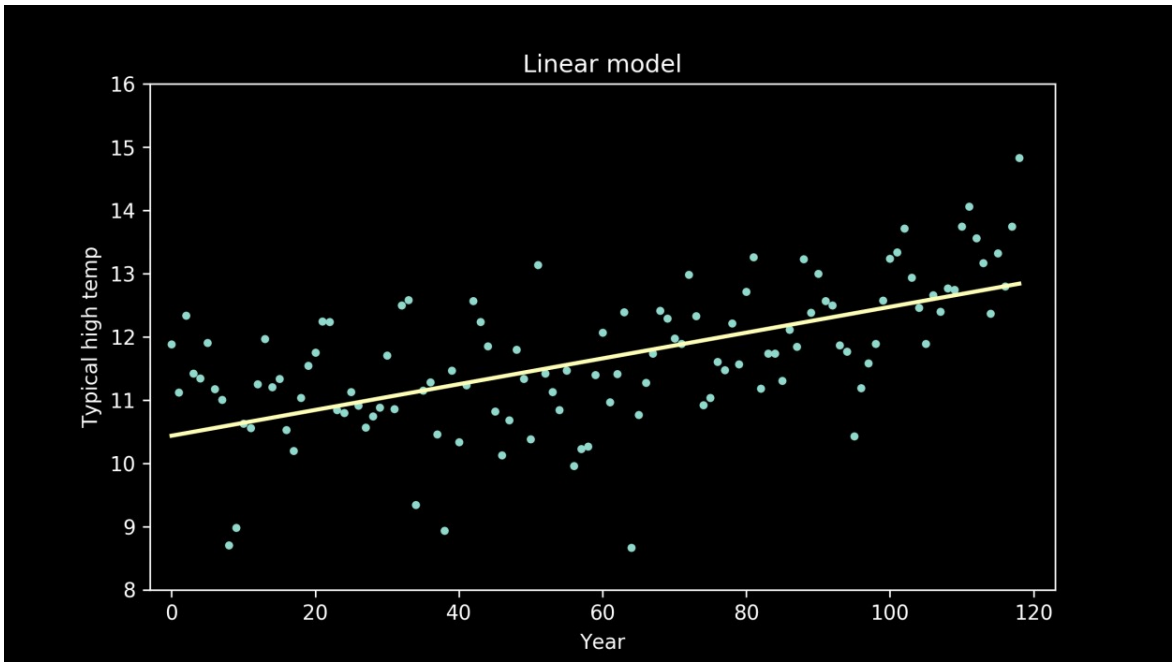
Signal is **the real pattern, the repeatable process that we hope to capture and describe**. It is the information that we care about. The signal is what lets the model generalize to new situations. The noise is everything else that gets in the way of that.

Separating signal from noise

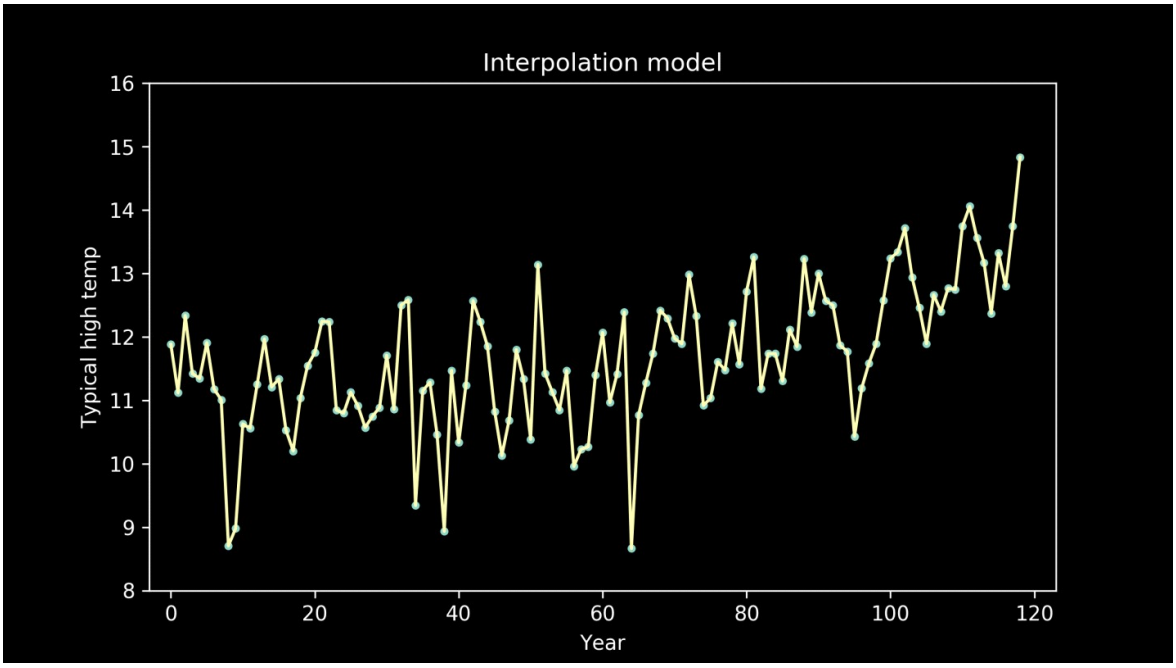
When we are building a model, we are making the assumption that our data has two parts, signal and noise. Signal is the real pattern, the repeatable process that we hope to capture and describe. It is the information that we care about. The signal is what lets the model generalize to new situations.

It's easy to picture the difference between signal and noise if you imagine listening to your favorite playlist in the middle of winter while there is a heater running nearby. The music is the signal. That's the thing that you want to track and absorb. The heater fan is noise. It is additional variation piled on top of the signal. And if it gets too loud, it becomes impossible to follow the flow of the signal.

The challenge:
It is the goal of models to describe the signal, despite the noise.A perfect model describes the signal exactly, and ignores all of the noise. If a model fails to capture all of the signal, that type of error is called **bias**.
If a model captures of some of the noise, that type of error is called **variance**.
Too much bias in our model means that it will perform poorly in all situations because it hasn't captured the signal well. You may also hear this called **underfitting**.
This was the case when we fit a straight line to our temperature data. It didn't capture the underlying pattern well, and because of that, had a much higher error then the rest of our candidates.



Too much variance in our model will also cause it to fail. It won't generalize well. Instead of capturing just the pattern we care about, it will also capture a lot of extraneous noise that we don't care about. The patterns in the noise will be different from situation to situation. When we try to generalize and apply or model to a new situation, it will have extra error. This is also called **overfitting**.
The more complex our model, the greater the risk of overfitting. This was the case in the connect-the-dots interpolation model.



What is data management?

Data management is the practice of collecting, organizing, protecting, and storing an organization's data so it can be analyzed for business decisions. As organizations create and consume data at unprecedented rates, data management solutions become essential for making sense of the vast quantities of data. Today's leading data management software ensures that reliable, up-to-date data is always used to drive decisions.

Types of Data Management

Data management plays several roles in an organization's data environment, making essential functions easier and less time-intensive. These data management techniques include the following:

- **Data preparation** is used to clean and transform raw data into the right shape and format for analysis, including making corrections and combining data sets.
- **Data pipelines** enable the automated transfer of data from one system to another.
- **ETLs (Extract, Transform, Load)** are built to take the data from one system, transform it, and load it into the organization's data warehouse.
- **Data catalogs** help manage metadata to create a complete picture of the data, providing a summary of its changes, locations, and quality while also making the data easy to find.
- **Data warehouses** are places to consolidate various data sources, contend with the many data types businesses store, and provide a clear route for data analysis.

- **Data governance** defines standards, processes, and policies to maintain data security and integrity.
- **Data architecture** provides a formal approach for creating and managing data flow.
- **Data security** protects data from unauthorized access and corruption.
- **Data modeling** documents the flow of data through an application or organization.

With effective data management, people across an organization can find and access trusted data for their queries. Some benefits of an effective data management solution include:

Visibility

Data management can increase the visibility of your organization's data assets, making it easier for people to quickly and confidently find the right data for their analysis. Data visibility allows your company to be more organized and productive, allowing employees to find the data they need to better do their jobs.

Reliability

Data management helps minimize potential errors by establishing processes and policies for usage and building trust in the data being used to make decisions across your organization

Security

Data management protects your organization and its employees from data losses, thefts, and breaches with authentication and encryption tools. Strong data security ensures that vital company information is backed up and retrievable should the primary source become unavailable. Additionally, security becomes more and more important if your data contains any personally identifiable information that needs to be carefully managed to comply with consumer protection laws.

Scalability

Data management allows organizations to effectively scale data and usage occasions with repeatable processes to keep data and metadata up to date. When processes are easy to repeat, your organization can avoid the unnecessary costs of duplication, such as employees conducting the same research over and over again or re-running costly queries unnecessarily.

Data Quality

To maintain the data we need to maintain good quality of data there by increasing the profits of a company

For example if a company is maintaining all the requriments according to the need of customers then we can call the data as good quality of data

Similarly if the company is not able to provide needs of a customer then we can say that company is maintaining the poor quality of data.

Poor data quality negatively affects many data processing efforts.

Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data
- Wrong data

Noisy data

For objects, noise is considered an extraneous object.

For attributes, noise refers to modification of original values.

Examples:

1. Distortion of a person's voice when talking on a poor phone

2. When you are on a call when other person besides you speaks loudly then we are calling it as a noise because the original data that is a person is speaking is lost and the other person voice comes to picture which is noise

For any type of data the SNR value i,e signals to noise ratio should be low or shouldnot be there.

Origins of noise

- **outliers** -- values seemingly out of the normal range of data
- **duplicate records** -- good database design should minimize this (use DISTINCT on SQL retrievals)
- **incorrect attribute values** -- again good db design and integrity constraints should minimize this
- **numeric only**, deal with rogue strings or characters where numbers should be.
- **null handling** for attributes (nulls=missing values)

Missing Data Handling

Missing data can be any of the following :

- malfunctioning equipment
- changes in experimental design
- collation of different data sources
- measurement not possible.
- People may wish to not supply information.
- Information is not applicable (childen don't have annual income)

What can we do if we are having missing values

- **Discard** records with missing values:means just discard those values

- **Ordinal-continuous** data, could **replace with attribute means Substitute** with a value from a similar instance
- **Ignore** missing values, i.e., just proceed and let the tools deals with them
- **Treat** missing values **as equals** (all share the same missing value code ex if we give 20 all missing values are replaced with 20)
- **Treat** missing values **as unequal values** ex if we give 20 for one missing value others should be given with other values like 21,31,etc I,e no same values)

BUT...Missing (null) values may have significance in themselves (e.g. missing test in a medical examination, deathdate missing means still alive!)

Missing completely at random (MCAR)

- Missingness of a value is independent of attributes
- Fill in values based on the attribute as suggested above (e.g. attribute mean)
- Analysis may be unbiased overall

Missing at Random (MAR)

- Missingness is related to other variables
- Fill in values based other values (e.g., from similar instances)
- Almost always produces a bias in the analysis

Missing Not at Random (MNAR)

- Missingness is related to unobserved measurements
- Informative or non-ignorable missingness

Not possible to know the situation from the data. You need to know the context,application field, data collection process, etc.

DATA PREPROCESSING

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data.The quality of the data should be checked before applying machine learning or datamining algorithms.

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, datacleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled invarious ways.

Some of them are:

Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiplevalues are missing within a tuple.

Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing valuesmanually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled infollowing ways :

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided intosegments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

Regression:

Here data can be made smooth by fitting it to a regression function.The regressionused may be linear (having one independent variable) or multiple (having multiple independent variables).

Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or itwill fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable formining process. This involves following ways:

Normalization:

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to1.0)

Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to helpthe mining process.

Discretization:

This is done to replace the raw values of numeric attribute by interval levels or

conceptual levels.

Concept Hierarchy Generation:

Here attributes are converted from lower level to higher level in hierarchy. ForExample-
The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order toget rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

Attribute Subset Selection:

The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

Numerosity Reduction:

This enable to store the model of data instead of whole data, for example: RegressionModels.

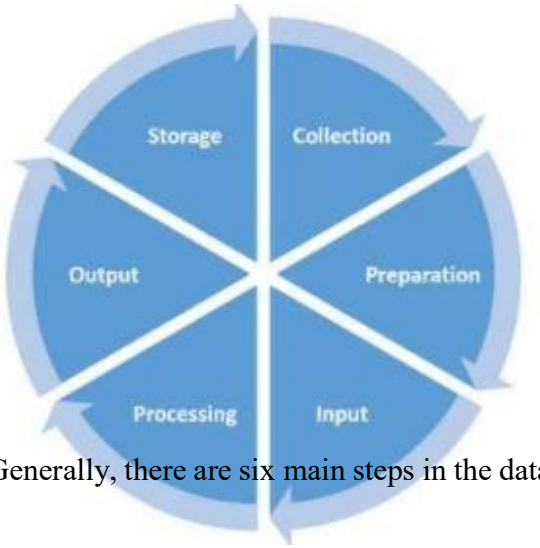
Dimensionality Reduction:

This reduce the size of data by encoding mechanisms.It can be lossy or lossless. Ifafter reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

What Is Data Processing?

Data in its raw form is not useful to any organization. Data processing is the method of collecting raw data and translating it into usable information. It is usually performed in a step-by-step process by a team of data scientists and data engineers inan organization. The raw data is collected, filtered, sorted, processed, analyzed, stored, and then presented in a readable format.

Data Processing Cycle



Generally, there are six main steps in the data processing cycle:Step 1:

Collection

The collection of raw data is the first step of the data processing cycle. The type of raw data collected has a huge impact on the output produced. Hence, raw data shouldbe gathered from defined and accurate sources so that the subsequent findings are valid and usable. Raw data can include monetary figures, website cookies, profit/lossstatements of a company, user behavior, etc.

Step 2: Preparation

Data preparation or data cleaning is the process of sorting and filtering the raw data toremove unnecessary and inaccurate data. Raw data is checked for errors, duplication, miscalculations or missing data, and transformed into a suitable form for further analysis and processing. This is done to ensure that only the highest quality data is fedinto the processing unit.

Step 3: Input

In this step, the raw data is converted into machine readable form and fed into the

processing unit. This can be in the form of data entry through a keyboard, scanner orany other input source.

Step 4: Data Processing

In this step, the raw data is subjected to various data processing methods using machine learning and artificial intelligence algorithms to generate a desirable output.This step may vary slightly from process to process depending on the source of data being processed (data lakes, online databases, connected devices, etc.) and the intended use of the output.

Step 5: Output

The data is finally transmitted and displayed to the user in a readable form like graphs, tables, vector files, audio, video, documents, etc. This output can be stored and furtherprocessed in the next data processing cycle.

Step 6: Storage

The last step of the data processing cycle is storage, where data and metadata are stored for further use. This allows for quick access and retrieval of information whenever needed, and also allows it to be used as input in the next data processingcycle directly.

Types of Data Processing

There are different types of data processing based on the source of data and the steps taken by the processing unit to generate an output. There is no one-size-fits-all methodthat can be used for processing raw data.

Batch Processing

Data is collected and processed in batches. Used for large amounts of data.Eg: payroll system

Real-time Processing

Data is processed within seconds when the input is given. Used for small amounts ofdata.

Eg: withdrawing money from ATM

Online Processing

Data is automatically fed into the CPU as soon as it becomes available. Used forcontinuous processing of data.

Eg: barcode scanning

Multiprocessing

Data is broken down into frames and processed using two or more CPUs within asingle computer system. Also known as parallel processing.

Eg: weather forecasting

Time-sharing

Allocates computer resources and data in time slots to several users simultaneously.Data

Processing Methods

There are three main data processing methods - manual, mechanical and electronic.Manual Data

Processing

In this data processing method, data is processed manually. The entire process of datacollection, filtering, sorting, calculation and other logical operations are all done withhuman intervention without the use of any other electronic device or automation software. It is a low-cost method and requires little to no tools, but produces high errors, high labor costs and lots of time.

Mechanical Data Processing

Data is processed mechanically through the use of devices and machines. These can include simple devices such as calculators, typewriters, printing press, etc. Simple data processing operations can be achieved with this method. It has much lesser errorsthan manual data processing, but the increase of data has made this method more complex and difficult.

Examples of Data Processing

Data processing occurs in our daily lives whether we may be aware of it or not. Hereare some real-life examples of data processing:

A stock trading software that converts millions of stock data into a simple graph

An e-commerce company uses the search history of customers to recommend similarproducts

A digital marketing company uses demographic data of people to strategize location-specific campaigns

A self-driving car uses real-time data from sensors to detect if there are pedestriansand other cars on the road