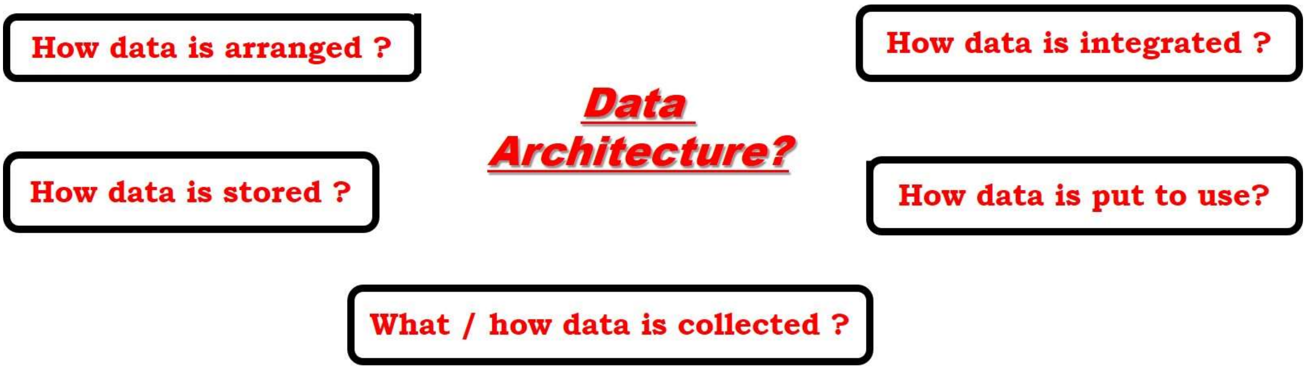
Unit I

Data Management (NOS 2101) Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/signal/GPS etc. Data Management, Data Quality (noise, outliers, missing values, duplicate data) and Data processing.

Design Data Architecture:

Data architecture is composed of *models, policies, rules or standards* that govern:



How data is arranged ?

How data is integrated ?

Data Architecture?

How data is stored ?

How data is put to use?

What / how data is collected ?

Key components of data architecture:

Data Sources: Identify and define the various sources from which data is collected, including databases, applications, APIs, and external systems.

Data Storage: Choose appropriate storage technologies (databases, data lakes, etc.) and design data structures (schemas) for efficient and organized storage.

Data Processing: Implement processes for data transformation, integration, and analysis, including batch processing and real-time processing as needed.

Data Integration: Establish methods for combining data from different sources, ensuring consistency and coherence across the organization.

Data Security and Privacy: Implement measures to secure data, control access, and ensure compliance with privacy regulations through encryption, access controls, and governance policies.

Data Quality and Governance: Define and enforce standards for data quality, metadata management, and governance practices to maintain the accuracy and reliability of data.

Scalability and Performance: Design the architecture to scale effectively as data volumes grow, optimizing performance for efficient data processing.

Monitoring and Maintenance: Implement monitoring tools to track system health, performance, and data quality. Schedule regular maintenance tasks, including backups and updates.

User Access and Analytics: Enable user access to data through well-defined APIs and integrate business intelligence tools for analytics and reporting.

Design with flexibility to adapt to evolving business needs and technological changes, adhering to industry standards for interoperability.

- *Data is one of the essential pillars of* ***enterprise architecture*** *through which it succeeds in the execution of business strategy.*
- The data architecture is formed by dividing into three essential models and then are combined :

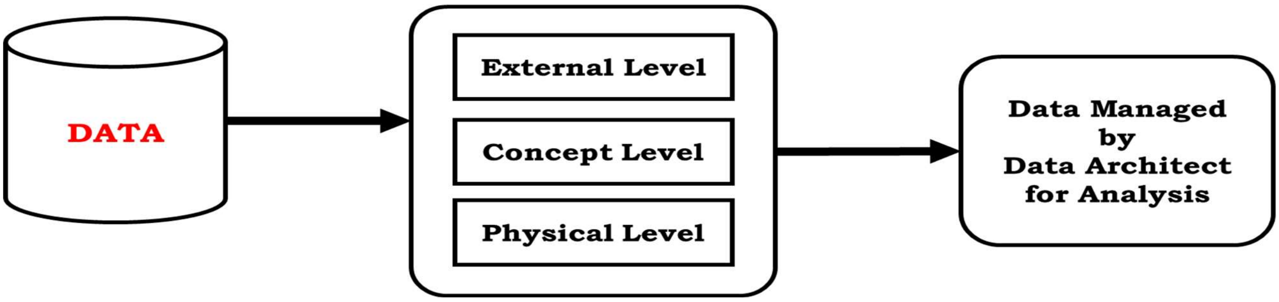


Figure: Levels of data architecture

Physical Level

- The physical level describes how data is actually stored in the database. In the lowest level, this data is stored in the external hard drives in the form of bits and at a little high level, it can be said that the data is stored in files and folders.

Conceptual Level

- It is also known as the logical level. It describes how the database appears to the users conceptually and the relationships between various data tables. The conceptual level does not care for how the data in the database is actually stored.

External Level

- It is also known as the view level. The external level only shows the relevant database content to the users in the form of views and hides the rest of the data. So different users can see the database as a different view as per their individual requirements.

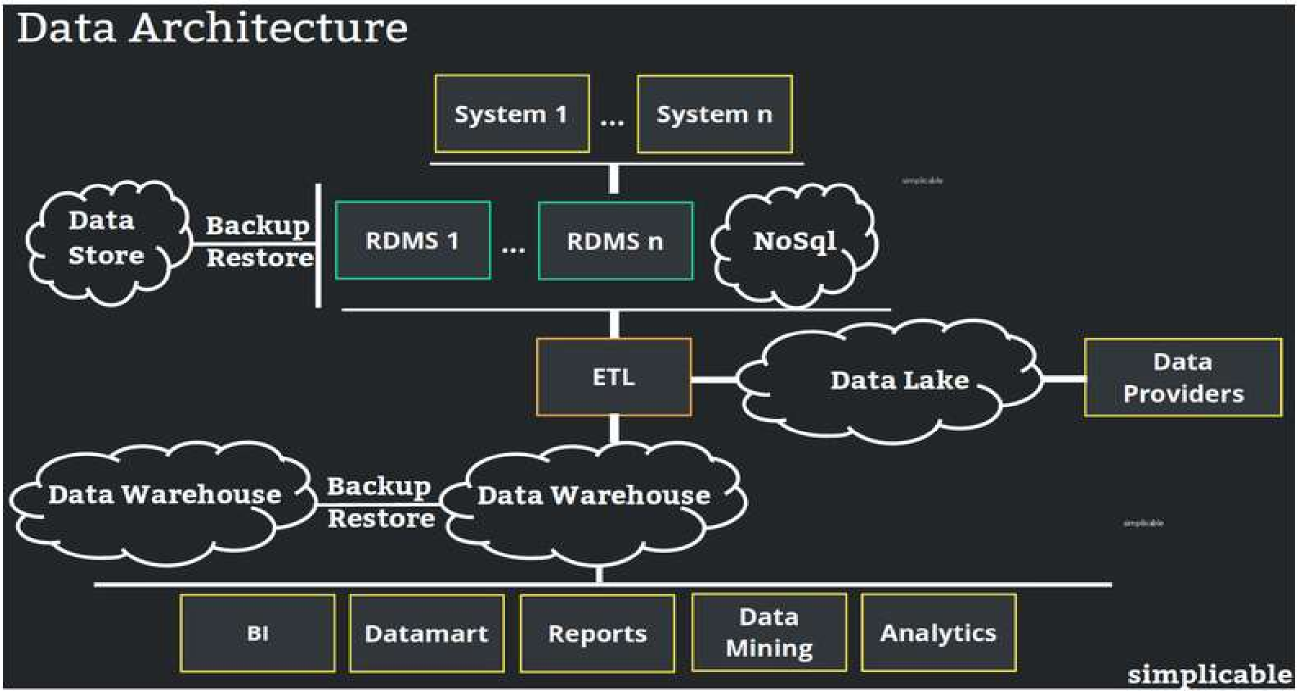


Figure: Data Architecture Example

Various constraints and influences will have an effect on data architecture design. These include

- ***Business requirements***
- ***Technology in use***
- ***Economics***
- ***Business Policies and***
- ***Data Processing needs***

Business Requirements:

These include factors such as the

- ✓ Expansion of business
- ✓ Data management
- ✓ Transaction management

- ✓ Making use of raw data by converting them into image files and records, and then storing in data warehouses.

Data Ware is common organizational requirement since it enables managerial decision making and other organizational processes.

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.

Economics:

The economic 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.

Technology in use:

This includes using the example of previously completed data architecture design and also using existing licensed software purchases, database technology.

Data Processing needs:

These include factors such as mining of the data, large continuous transactions, database management, and other data preprocessing needs.

Data management comprises all disciplines related to managing data as a valuable resource.

All disciplines includes: ***Data Mining, Data Analytics, Big Data Analytics***

or

- ➢ Data management is the process of collecting, storing, organizing, protecting, and using data. It is essential for data analytics, as it ensures that the data used for analysis is accurate, reliable, and accessible.
- ➢ 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 includes a number of key tasks, such as:
  - **Data collection and integration:** This involves identifying and collecting data from a variety of sources, such as internal databases, external databases, and web services. The data is then integrated into a central repository, such as a data warehouse or data lake.
  - **Data quality management:** This involves cleaning, transforming, and validating the data to ensure that it is accurate, consistent, and complete.
  - **Data governance:** This involves establishing policies and procedures for managing the data, such as who has access to the data and how it can be used.
  - **Data security:** This involves protecting the data from unauthorized access, use, disclosure, disruption, modification, or destruction.
  - **Data analysis:** This involves using data to identify trends, patterns, and insights that can be used to make better decisions.

Difference between Data Management and Data Architecture

- ✓ Data management and data architecture are both important components of any organization that relies on data.
- ✓ By understanding the differences between these two concepts, organizations can better manage their data and make more informed decisions.

Data Management encompasses the strategies and actions taken to collect, store, structure, and safeguard data from its creation to disposal. Its primary objective is to maintain data integrity, security, and accessibility.

- ✓ **Example:** A company uses data management to maintain customer information in a CRM system, ensuring data accuracy, regular backups, and access control for authorized personnel.

Data Architecture involves planning and designing the way data is structured, stored, and made available for use within an organization. It's like creating a detailed map that outlines how data flows and is organized in the organization's data landscape.

- ✓ **Example:** An e-commerce platform designs its data architecture to include a relational database for customer data, a NoSQL database for product catalog, and APIs to access and update this data.

Understand various sources of Data like Sensors/signal/GPS etc.

- ➢ In the process of data analysis, ***"Data collection"*** is the initial step before starting to analyze the useful information in data. The data which is to be analyzed must be collected from different valid sources.
- ➢ ***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.

Data can be generated/collected from two types of sources namely

- ✓ Primary
- ✓ Secondary

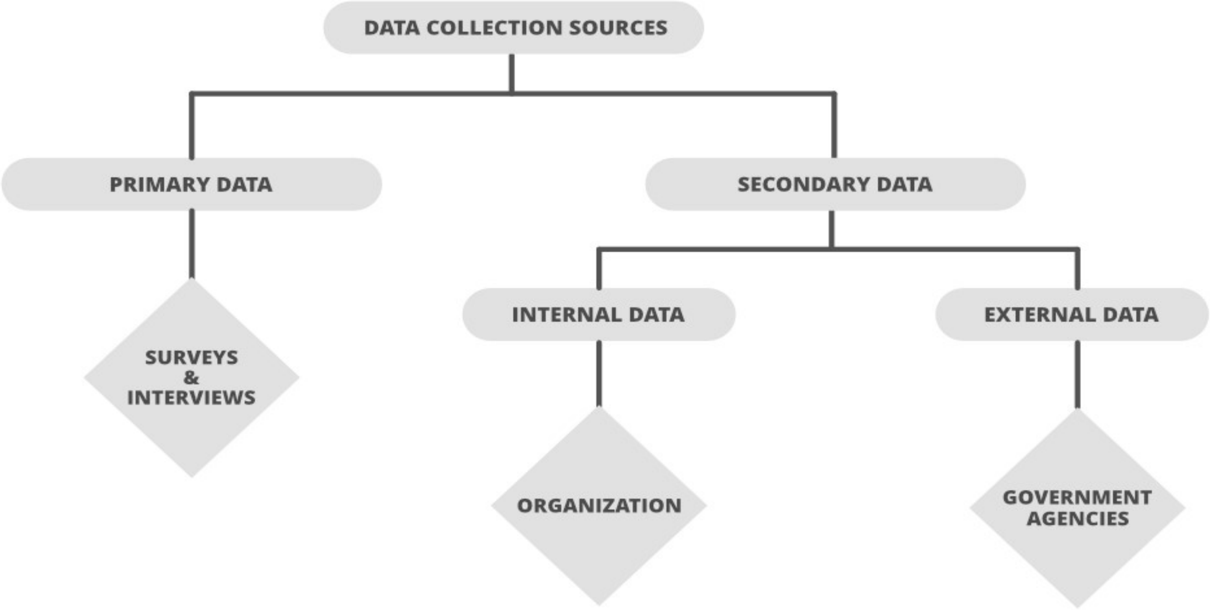


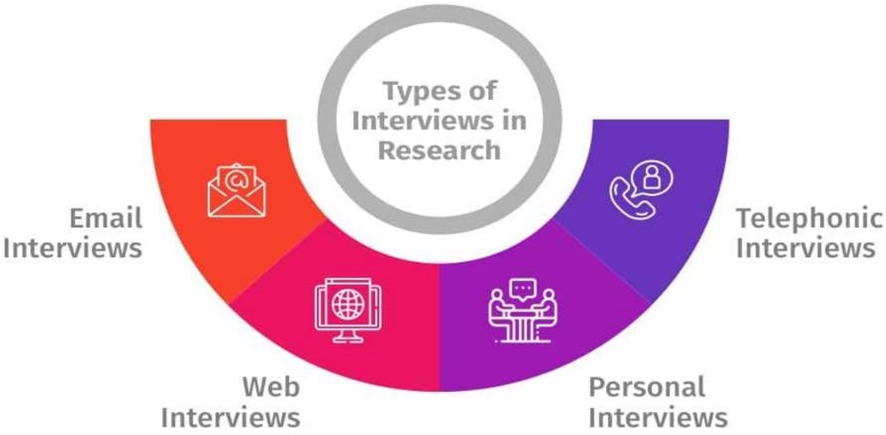
Figure: various sources of data

Primary Data:

- The primary data which is collected directly by performing techniques such as
  - ✓ ***Interview method***
  - ✓ ***Survey method***
  - ✓ ***Observational method***
  - ✓ ***Experiment method***
- 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.

Interview Method

- The interview method is a qualitative research method that involves asking questions to a participant to collect data.
- It is a systematic process of gathering information from a sample of individuals to learn about their opinions, beliefs, or experiences.
- Interviews are used in a wide variety of fields, including academia, business, and government.
- Interviews can be structured or unstructured. Structured interviews involve asking the participant a set of predetermined questions, while unstructured interviews allow the researcher to explore the participant's thoughts and experiences in more depth.



Survey Method:

- The survey method is a quantitative research method that is used to collect data from a large number of people.
- Surveys are typically conducted using questionnaires, which are self-administered instruments that ask respondents to provide information about themselves, their opinions, or their experiences.
- Surveys can be conducted online, in person, or over the phone.

Survey Method	**Description**
Mail survey	A written survey that is self-administered
Telephone survey	A survey conducted by telephone in which the questions are read to the respondents

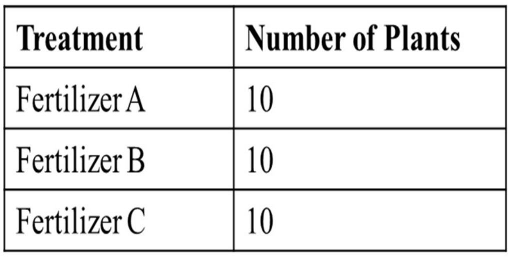
Observational Method: Observation is a research method used to systematically collect data by directly observing people, objects, or phenomena in their natural settings. It is a valuable research technique in various fields, including anthropology, psychology, sociology, education, and environmental science.

Experimental Method: The experimental method is the process of collecting data through performing experiments, the most frequently used experiment methods are CRD, RBD, LSD & FD.

- ***C R D: Completely Randomized Design***
- ***Randomized Block Design***
- ***Latin Square Design***
- ***Factorial Design***

Randomized Design-CRD: Completely Randomized Design (CRD) is a type of experimental design in which all experimental units are randomly assigned to one of several treatments. This design is useful when the experimental units are homogeneous, meaning that they are all similar in terms of their characteristics.

Example: A researcher is interested in studying the effects of three different types of fertilizer on the growth of corn plants. The researcher has 30 plants, and they randomly assign each plant to one of the three fertilizer treatments. The researcher then measures the height of each plant after 10 weeks.



Treatment	Number of Plants
Fertilizer A	10
Fertilizer B	10
Fertilizer C	10

The researcher can analyze the average plant height in each treatment group using statistics. If a significant difference is found, they can conclude that different fertilizers have varying effects on corn plant growth.

Randomized Block Design (RBD) is a type of experimental design in which experimental units are grouped into blocks, and then units within each block are randomly assigned to one of several treatments. This design is useful when the experimental units are heterogeneous, meaning that they vary in terms of their characteristics.

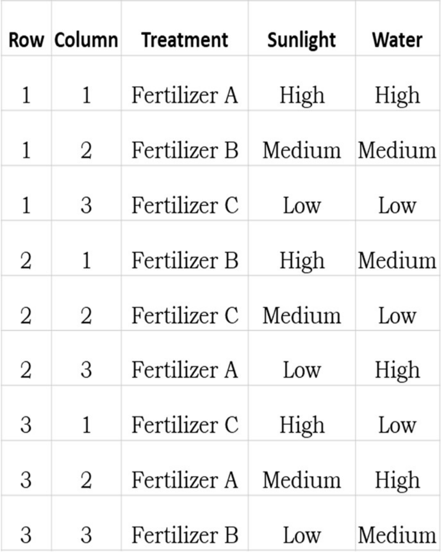
Example: To account for field heterogeneity, the researcher divides it into three homogeneous blocks and randomly assigns one of the three fertilizer treatments to each plot within each block.

The researcher measures plant height after 10 weeks, comparing average heights of plants within each fertilizer group in each block. If significant differences emerge within each block, the researcher can conclude distinct effects of the fertilizers on corn plant growth.

Block	Treatment	Plot
Block 1	Fertilizer A	1
Block 1	Fertilizer B	2
Block 1	Fertilizer C	3
Block 2	Fertilizer A	4
Block 2	Fertilizer B	5
Block 2	Fertilizer C	6
Block 3	Fertilizer A	7
Block 3	Fertilizer B	8
Block 3	Fertilizer C	9

Latin square method: A Latin Square Design (LSD) is a type of experimental design in which each treatment appears once in each row and column of a square. This design is useful when there are two extraneous variables that can affect the results of the experiment.

For example, a farmer wants to test the effects of three different fertilizers on the growth of corn plants. The farmer also knows that the amount of sunlight and the amount of water can affect the growth of corn plants. To control for these two extraneous variables, the farmer uses a Latin square design.The farmer divides the field into nine plots, arranged in a three-by-three square. Each plot is randomly assigned one of the three fertilizer treatments. The farmer then randomly assigns one of the three sunlight levels and one of the three water levels to each plot.



Row	Column	Treatment	Sunlight	Water
1	1	Fertilizer A	High	High
1	2	Fertilizer B	Medium	Medium
1	3	Fertilizer C	Low	Low
2	1	Fertilizer B	High	Medium
2	2	Fertilizer C	Medium	Low
2	3	Fertilizer A	Low	High
3	1	Fertilizer C	High	Low
3	2	Fertilizer A	Medium	High
3	3	Fertilizer B	Low	Medium

- The farmer then measures the height of each plant after 10 weeks. The farmer can then use statistical analysis to compare the average height of the plants in each treatment group, while controlling for the effects of sunlight and water.

Factorial Design: A factorial design is an experimental design in which multiple factors are varied simultaneously. This allows researchers to study the effects of each factor individually and in combination. Factorial designs are often used in experiments with two or more factors, but they can also be used with more factors. The number of possible combinations of factors in a factorial design is equal to the product of the number of levels of each factor.

For example, a 2x2 factorial design has two factors, each with two levels. This results in four possible combinations of factors.

- A researcher is interested in studying the effects of two factors on the growth of plants: fertilizer type (organic vs. inorganic) and watering frequency (daily vs. weekly). The researcher uses a 2x2 factorial design to study these two factors.
- The researcher plants 100 seeds in each of the four treatment groups:
    - ✓ Organic fertilizer, daily watering
    - ✓ Organic fertilizer, weekly watering
    - ✓ Inorganic fertilizer, daily watering
    - ✓ Inorganic fertilizer, weekly watering
- After 10 weeks, the researcher measures the height of each plant. The researcher can then use statistical analysis to compare the average height of the plants in each treatment group.

Secondary Data:

The secondary data can be obtained through

- **Internal Sources** - These are within the organization
- **External Sources** - These are outside the organization

Internal Sources: Internal sources of secondary data include:

- Organizational records, such as financial statements, sales reports, and customer feedback surveys
- Internal databases, such as employee records and product inventory databases

- Research reports and studies conducted by the organization
- Internal publications, such as newsletters and magazines

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

- External sources of secondary data include:
  - Government publications, such as census data and economic reports
  - Industry reports and studies
  - Academic journals and books
  - News articles and websites
  - Social media

Best sources for Secondary data

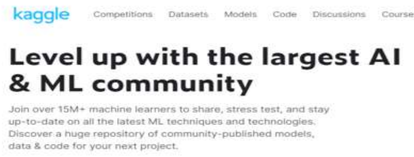
The **UCI** Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

Welcome to the UC Irvine Machine Learning Repository

We currently maintain 657 datasets as a service to the machine learning community. Here, you can donate and find datasets used by millions of people all around the world!

https://archive.ics.uci.edu/

Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC. Kaggle enables users to find and publish datasets, explore and build models in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.



Level up with the largest AI & ML community

Join over 15M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

https://www.kaggle.com/

//Additional points for External Sources of data

Government Publications: Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data. These are:

- ***Registrar General of India :*** which generates demographic data. It includes details of gender, age, occupation etc.
- ***Central Statistical Organization:*** This organization publishes the national accounts statistics. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

- ***Director General of Commercial Intelligence:*** This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.
- ***Ministry of Commerce and Industries:*** This ministry through the office of economic advisor provides information on wholesale price index.
- ***Planning Commission:*** It provides the basic statistics of Indian Economy.
- ***Labour Bureau:*** It 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.
- ***Non-Government Publications:*** These includes publications of various industrial and trade associations, such as
    - *The Indian Cotton Mill Association*
    - *Various chambers of commerce*
    - *The Bombay Stock Exchange (it publishes a directory containing financial accounts, key profitability and other relevant matter)*
    - *Various Associations of Press Media.*
    - *Export Promotion Council.*
    - *Confederation of Indian Industries (CII)*
    - *Small Industries Development Board of India*
    - *Different Mills like - Woolen mills, Textile mills etc*

Data from Sensors/Signals/GPS:

- There are many different sources of data like sensors/signals/GPS etc. These can be used for a variety of purposes, such as to track the performance of a system, to monitor the environment, or to collect information about user behaviour.

- Sensors are devices that can detect and measure physical or chemical changes in the environment. They are used in a wide range of applications, including:
    - ✓ Automotive: Sensors are used in cars to monitor engine performance, emissions, and safety features.
    - ✓ Manufacturing: Sensors are used to monitor the quality of products and to control manufacturing processes.
    - ✓ Healthcare: Sensors are used to monitor vital signs, such as heart rate and blood pressure.
    - ✓ Environmental monitoring: Sensors are used to monitor air quality, water quality, and weather conditions.
    - ✓ Agriculture: Sensors are used to monitor soil moisture, crop growth, and livestock health.

Signals are another important source of data. Signals can be transmitted over long distances using radio waves, optical fibers, etc. Signals are used in a wide range of applications, including:

- ✓ Communications: Signals are used to transmit voice, data, and video over long distances.
- ✓ Navigation: Signals from GPS satellites are used to determine location and speed.
- ✓ Television and radio broadcasting: Signals are used to transmit television and radio programs to receivers.
- ✓ Medical imaging: Signals are used to create images of the inside of the body.

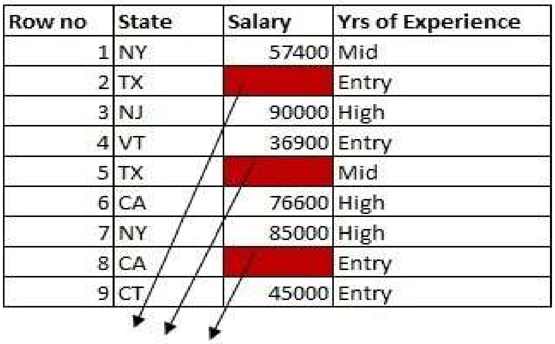
GPS (Global Positioning System) is a satellite-based system that can be used to determine location and speed. GPS is used in a wide range of applications, including:

- ✓ **Navigation:** GPS is used in cars, boats, and airplanes to help people navigate from one place to another.
- ✓ **Geospatial mapping:** GPS is used to create maps and to track the movement of objects.
- ✓ **Emergency services:** GPS is used to locate people in need of assistance.

Data Quality:

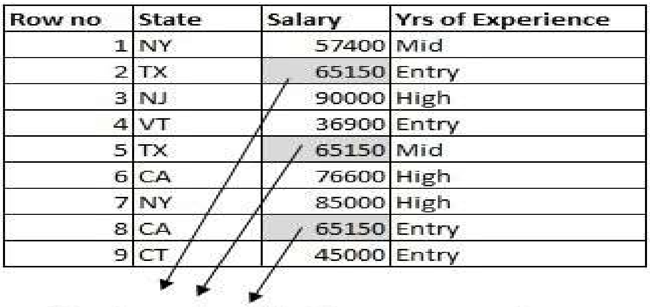
- Data quality refers to the ***level of quality of data.***
- Data quality is the measure of how well suited a data set is to serve its specific purpose.
- In order to improve the quality of data the dataset/data must be free of :
  - Missing Values
  - Noise
  - Outliers
  - Duplicate Data

Missing Values: In statistics, missing data or missing values occurs when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.



Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX	🟥	Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX	🟥	Mid
6	CA	76600	High
7	NY	85000	High
8	CA	🟥	Entry
9	CT	45000	Entry

Missing values



Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX	65150	Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX	65150	Mid
6	CA	76600	High
7	NY	85000	High
8	CA	65150	Entry
9	CT	45000	Entry

Replaced with the mean salary

Handling missing values:

- ✓ Ignore the tuples that contains missing values
- ✓ Fill in the missing value normally
- ✓ Use a global constant to fill in the missing value
- ✓ Use attribute mean to fill in the missing value
- ✓ Use the attribute mean for all samples belonging to the same class as the given tuple.
- ✓ Use most probable value to fill in the missing value.
- ✓ Use machine learning approaches to handle missing value(efficient method)

  Note: refer class notes for example.

Outliers:

- An outlier is something that behaves differently from the combination/collection of the data.
- Outlier is a point or an observation that deviates significantly from the other observations.
- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

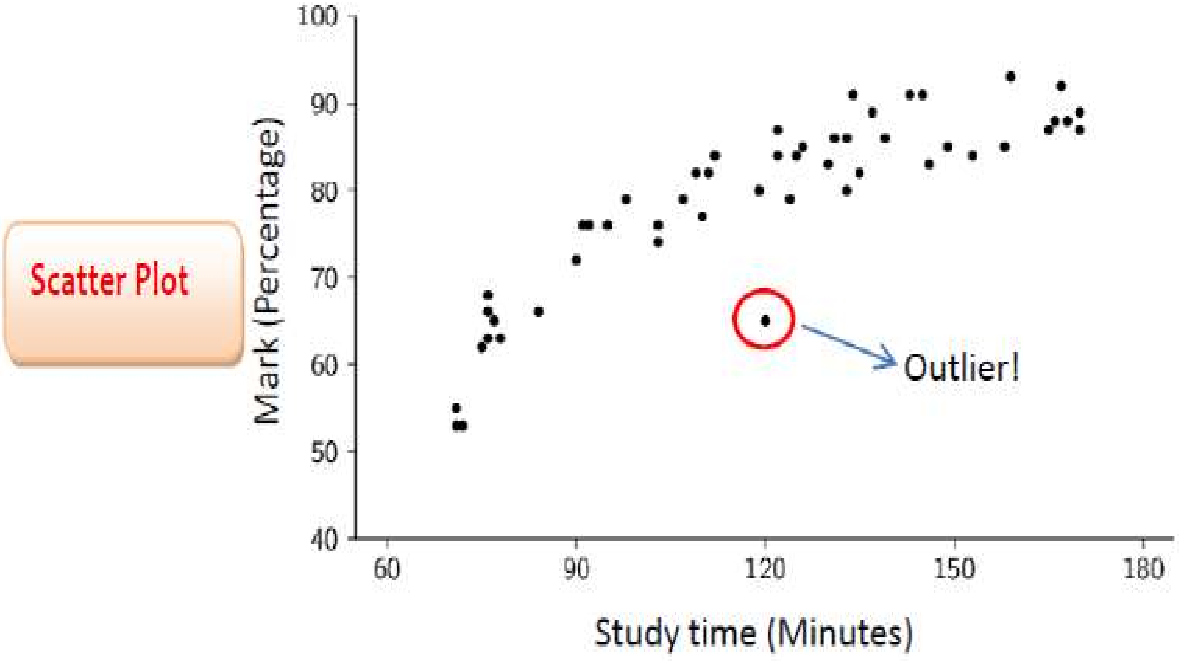


Figure: graph between mark and study time

Methods to Identify outliers

There are various ways to identify outliers in a dataset, following are some of them:

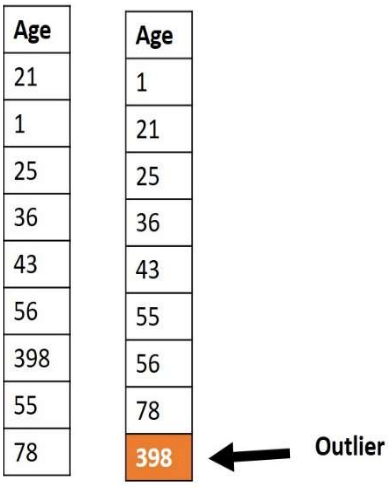
- Sorting the data
- Graphical Method
- IQR interquartile range
- Z score

Methods to handle outliers

- Deleting the values
- Changing the values
- Data transformation

- Model-Based Approaches
- Valuing the outliers

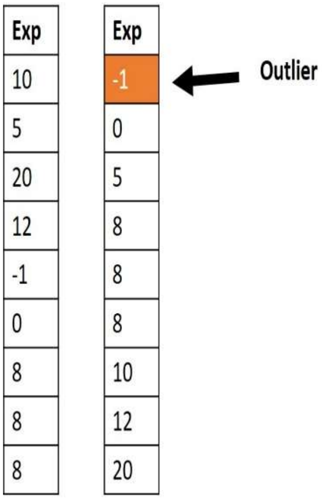
Sorting the Data Set: Sorting the dataset is the simplest and effective method to check unusual value. Let us consider an example of age dataset.



Age
21
1
25
36
43
56
398
55
78

Age
1
21
25
36
43
55
56
78
398

Outlier



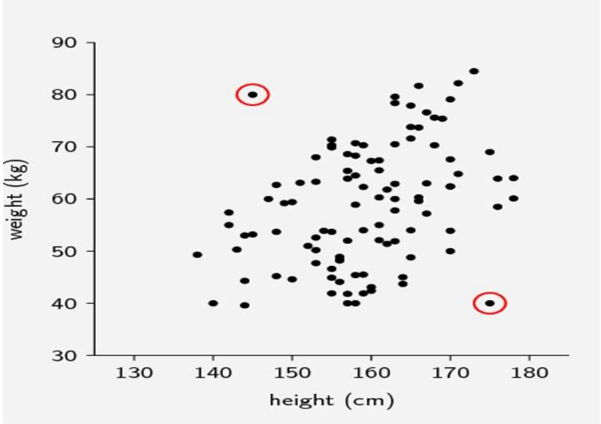
Exp
10
5
20
12
-1
0
8
8
8

Exp
-1
0
5
8
8
8
10
12
20

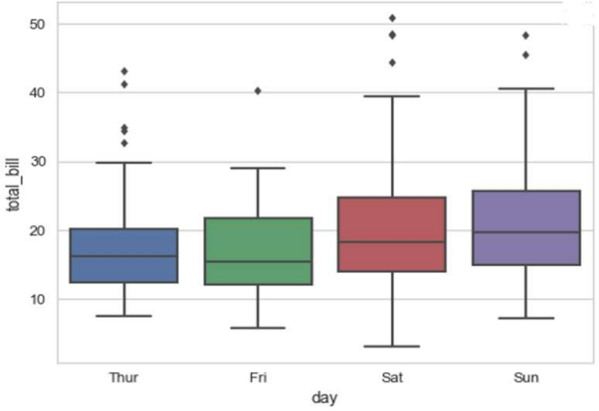
Outlier

Graphical Method

We can detect outliers with the help of graphical representation like Scatter plot and Boxplot.



a. Scatter Plot



b. Box Plot

Outlier handling should be customized based on context. While removing outliers is tempting, it's not always optimal. Methods like replacing with mean/median or dropping should be discussed with stakeholders. Deleting outliers may introduce bias.

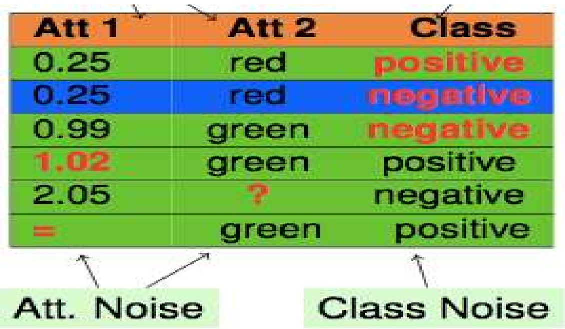
For large datasets, treating outliers separately during modelling can be effective but time-consuming and costly. The approach depends on dataset characteristics and analysis goals.

Duplicate Data:

- Duplication of data is called data redundancy. Data redundancy occurs when the same piece of data is stored in two or more separate places and is a common occurrence.
- The ***Data Deduplication*** is a technique for eliminating duplicate copies of repeating data.

Noise Data:

- Noisy data is a meaningless data.
- The term often been used as a synonym for corrupt data.
- Noise data is a data that cannot be understood and interpreted correctly by machines such as unstructured text.



Att 1	Att 2	Class
0.25	red	positive
0.25	red	negative
0.99	green	negative
1.02	green	positive
2.05	?	negative
=	green	positive

Att. Noise
Class Noise

- It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  - **Binning Method**
  - **Regression**
  - **Clustering**

Binning Method:

- This method works on sorted data in order to smooth it.
- The whole data is divided into segments 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.

Sorted data for price (in dollars) : 4,815,21,2124,25,28,34

Partition into equal-frequency bins:
Bin1: 4,8,15
Bin2: 21,21,24
Bin3: 25,28,34

Smoothing by bin means
Bin1: 9,9,9
Bin2: 22,22,22
Bin3: 29,29,29

Smoothing by bin boundaries
Bin1: 4,4,9
Bin2: 21,21,24
Bin3: 25,25,34

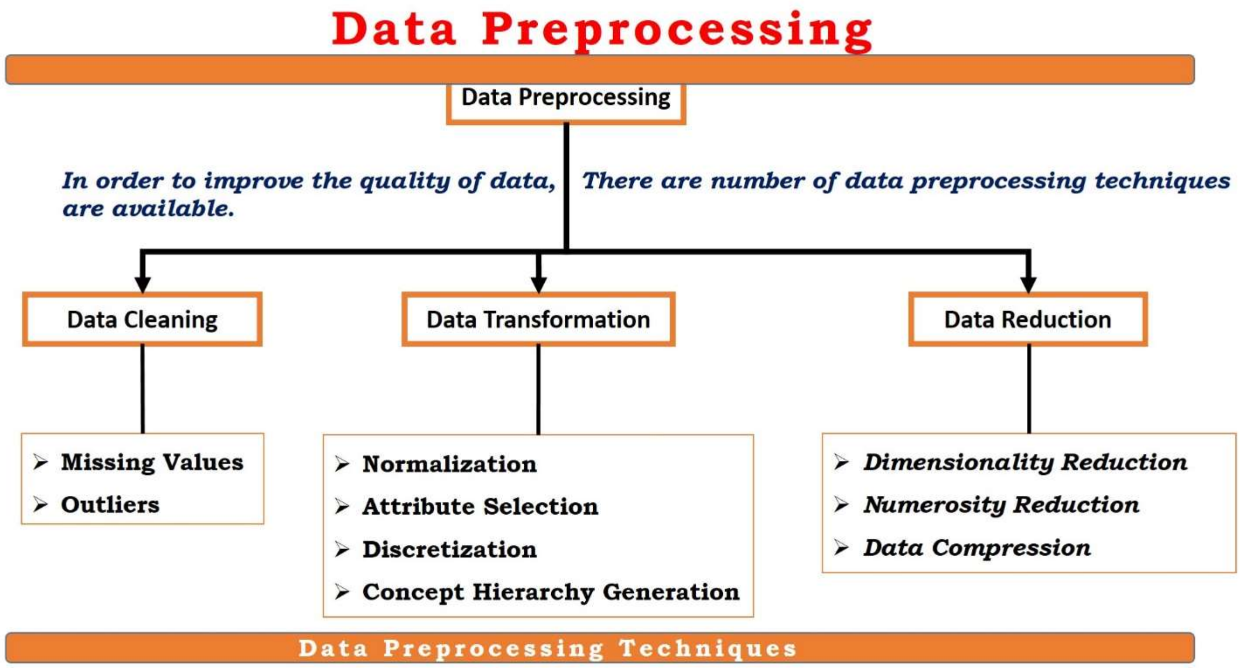
figure: Binning Methods

Regression: Here data can be made smooth by fitting it to a regression function. The regression used may be

- Linear (having one independent variable)
- Multiple (having multiple independent variables).

Clustering: This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Data Pre-processing:



Data Cleaning:

- Real world data may be incomplete noisy and inconsistent.
- Data cleaning routines
  - ✓ attempt to fill in missing values
  - ✓ smooth out noise data
  - ✓ Identify outliers and correct inconsistencies in the data.

Data Transformation:

- This step is taken in order to transform the data in appropriate forms for analysis.
- **Normalization:** It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
- **Attribute Selection:** In this strategy, new attributes are constructed from the given set of attributes to help the analysis process.
- **Discretization:** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels. Ex: age was replaced by intervals(0-10, 11-20 etc.) or youth, adult, senior.
- **Concept Hierarchy Generation:** Here attributes are converted from lower level to higher level in hierarchy. Ex: The attribute "city" can be converted to "country".

Data Reduction Techniques:

Data Reduction Techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume.

It includes:

- ***Dimensionality Reduction***
- ***Numerosity Reduction***
- ***Data Compression***

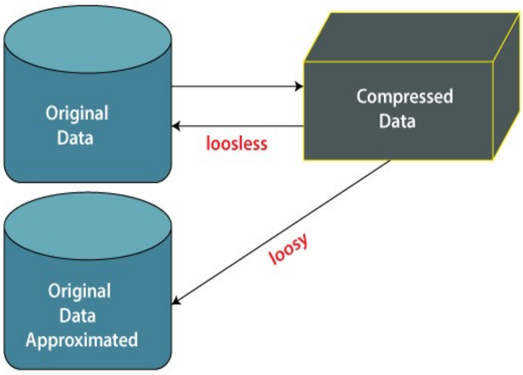
Dimensionality Reduction: It is the process of reducing the number of attributes under consideration. Attribute subset selection is a method of dimensional reduction in which irrelevant or redundant are detected and removed.

Numerosity Reduction: The numerosity reduction technique replace the original data by alternative small forms of data representation. These techniques may be parametric or non-parametric.

- **Parametric methods** a model to used to estimate the data, so that data parameters need to be stored, instead of the actual data.
- **Non Parametric methods** for storing reduced representations of the data includes histograms, sampling and data cube aggregation

Data Compression: Transformations are applied so as to obtain a reduced or compressed representation of original data.

- If original data can be reconstructed from the compressed data without any information loss, the data reduction is called lossless.
- If we reconstruct only an approximation of original data, then the data reduction is called lossy.



Data Processing:

- Data processing is a method of collecting raw data and converting it into useful information.
- Data Processing is performed in a predetermined procedure by a team of data scientists and data engineers in an organization.

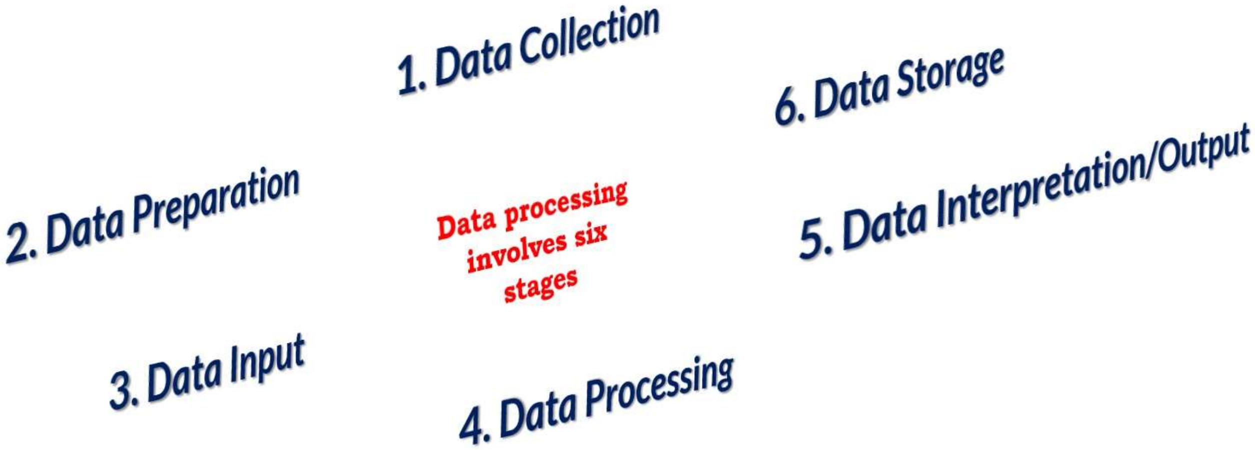


Figure: Data Processing Phases

Data Collection: The primary stage of data processing is to collect data. Data is acquired from sources like data lakes and data warehouses. The collected data must be of high quality.

Data Preparation: Also called "pre-processing", this stage is where the collected data is cleansed by checking for errors and arranged for the following data processing stage. Elimination of useless data and to generate quality data.

Data Input: The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

Processing: The processing of input data is accomplished by machine learning algorithms. Their process is variable depending on the data which is processed (connected devices, social networks, data lakes, etc.) and the intended use (medical diagnosis, ascertain customer wants, examining advertising patterns, etc.).

Data Interpretation: The non-data scientists find this data very helpful. The data is converted into graphs, tables, images and plain text. Members of a company can start analyzing this data and applying it into their projects/business.

Data Storage: The last step of 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 processing cycle directly.