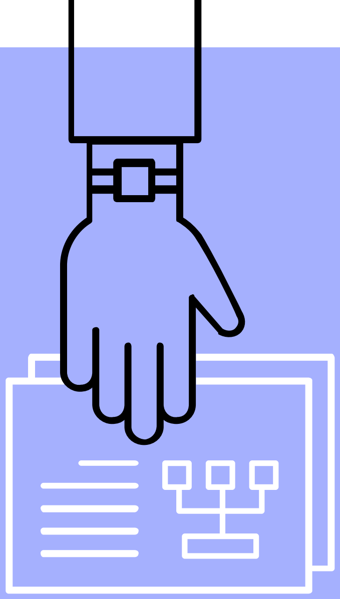
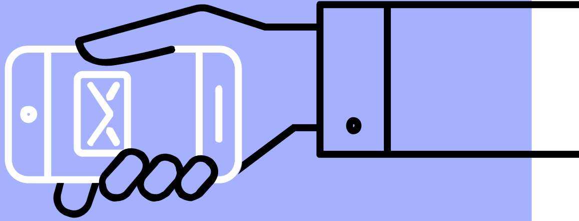
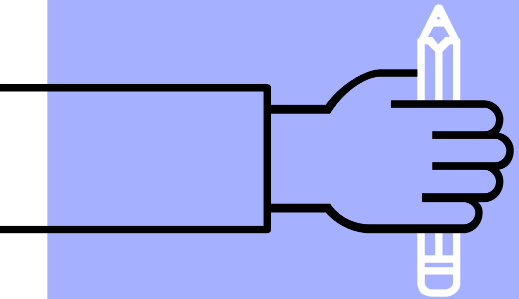
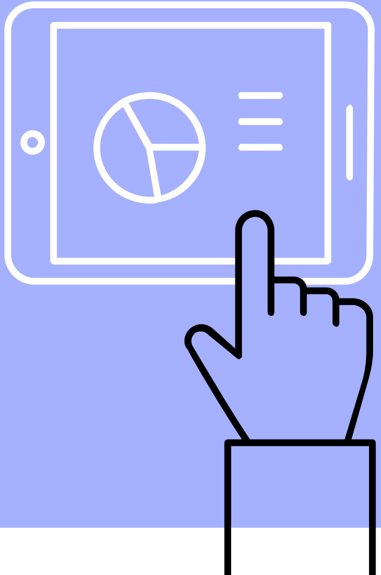




INTRODUCTION TO DATA SCIENCE

DS with Python
Lecture 4

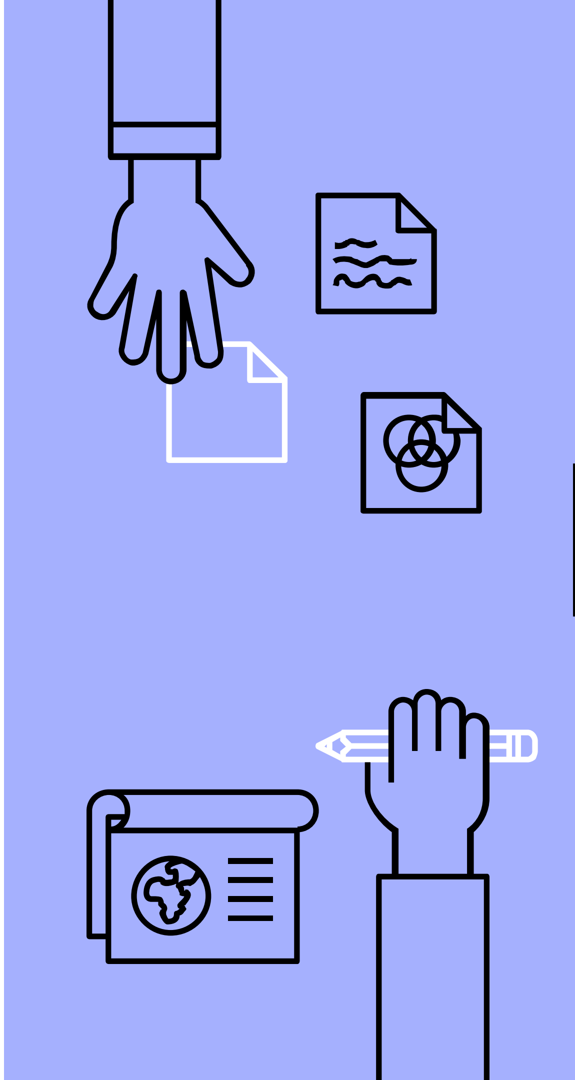




What is a Data Analysis

Analysis of Data is a process of:

- Inspecting
- Cleaning
- Transforming
- Modeling



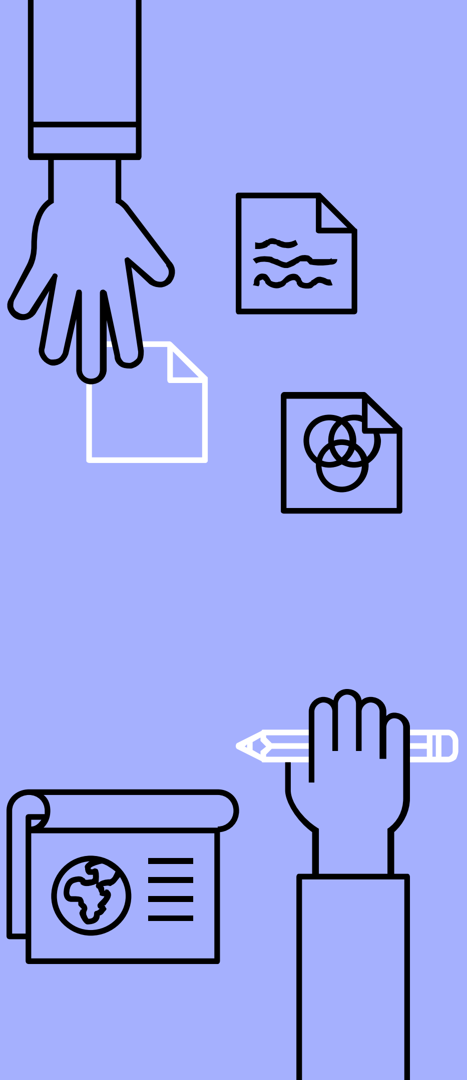
What is the Goal of Data Analysis

The goal is to --

- Highlight useful information
- Suggest Conclusions

and,

- Support Decision Making



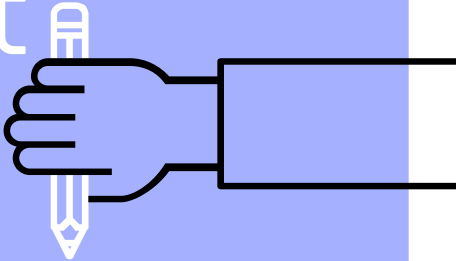
Objective!

Data Analysis is the process of developing answers to questions through the examination and interpretation of data.

Basic Steps in Analytic Process!

1. Identification of Issues
2. Determination of Availability of suitable Data
3. Decide the methods appropriate for answering the questions of interest.
4. Apply the methods
5. Evaluate, summarize and Communicate Results.

Role of Data Analyst in Data Analysis



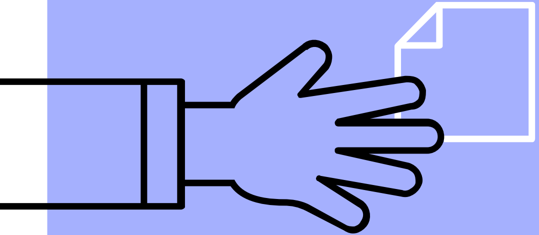
Data Analysis Sequence

Data Acquisition Pipeline

Report Structure

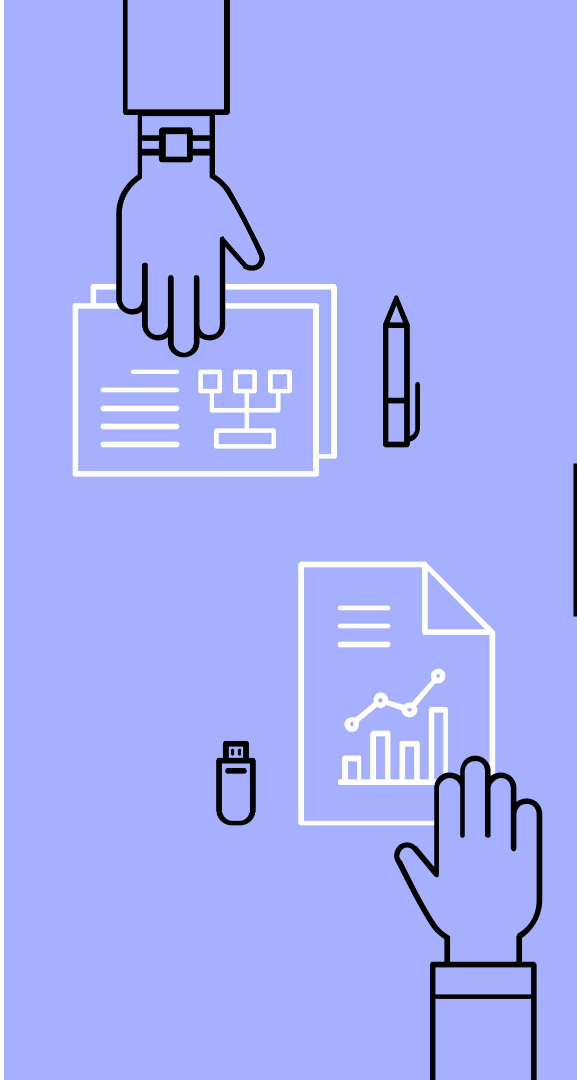
Data Cleaning

Data Pre-processing



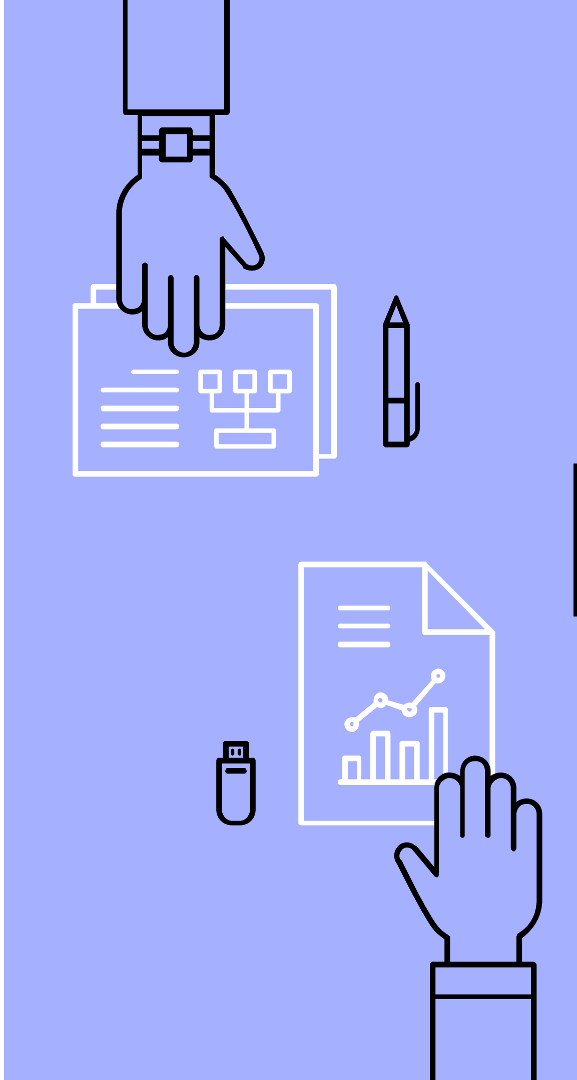
1. Gathering Information From Various Sources

- Identify the Sources that might help to better understand the situation.



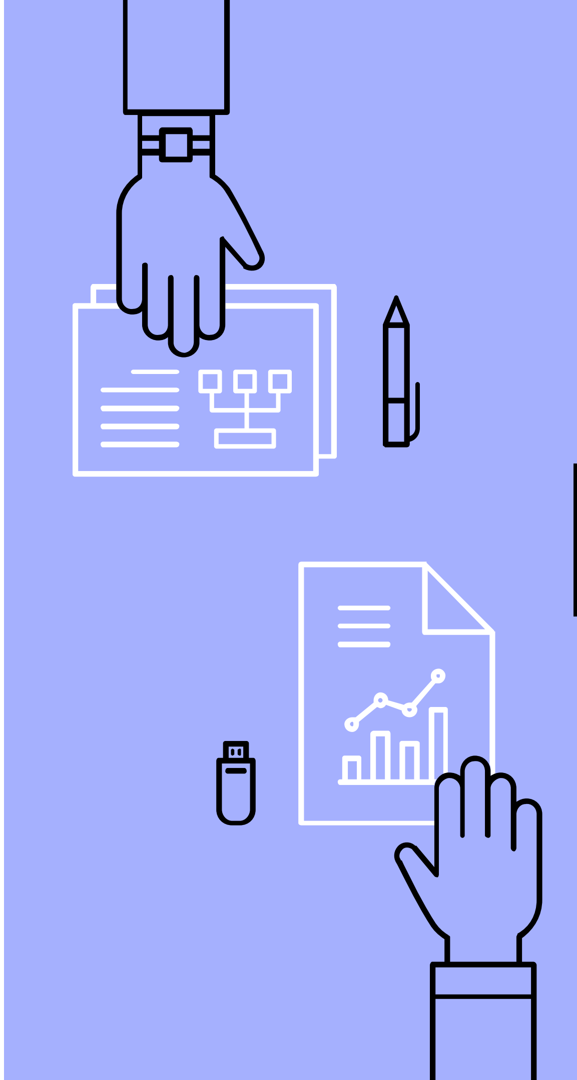
2. Identifying and Interpreting Patterns and Trends

3. Assessing Data Quality and Eliminate Irrelevant Data



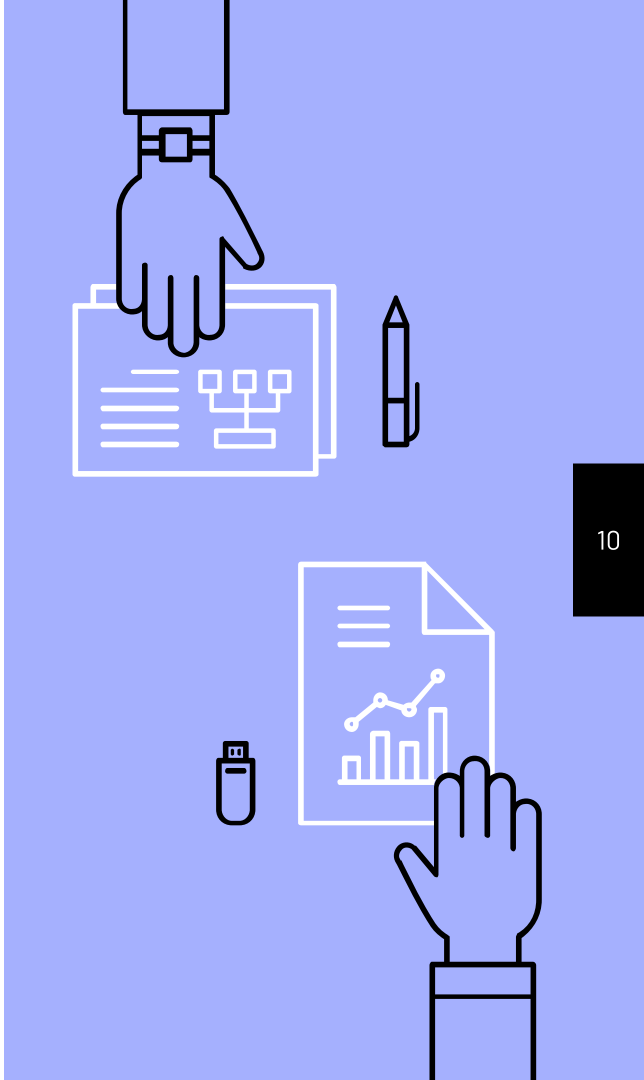
4. Documenting the types and structure of the Business Data.

5. Analyzing and Mining Business Data to identify patterns and correlations among the various data points.



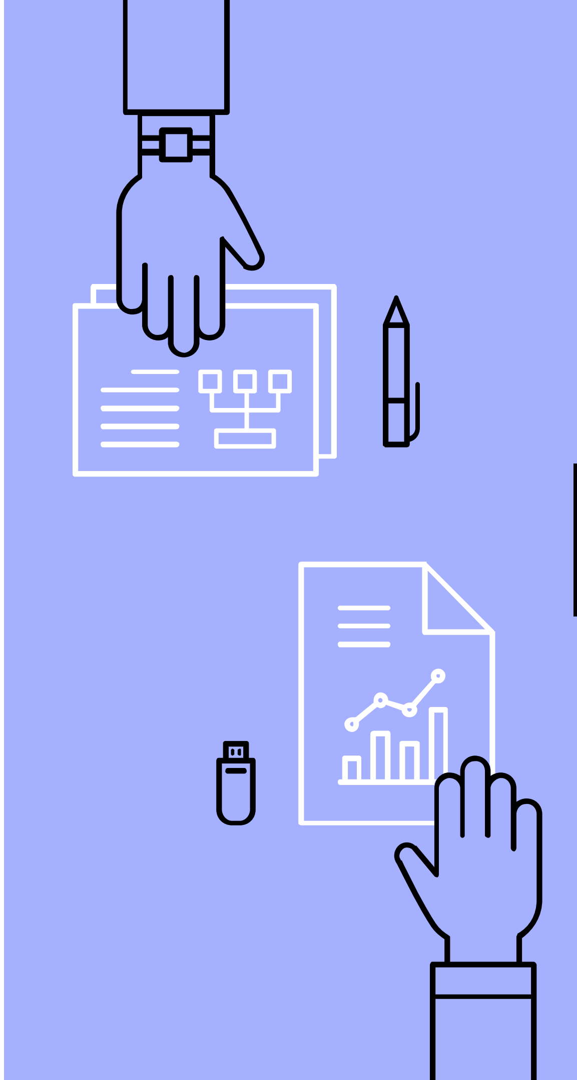
6. Perform GAP Analysis between source and Target data to avoid any loss and redundancy of data.

7. Mapping and Tracing Data from System in order to solve a given business problem.



8. Design and Create Data Reports using Reporting Tools to help business executives in their decision making.

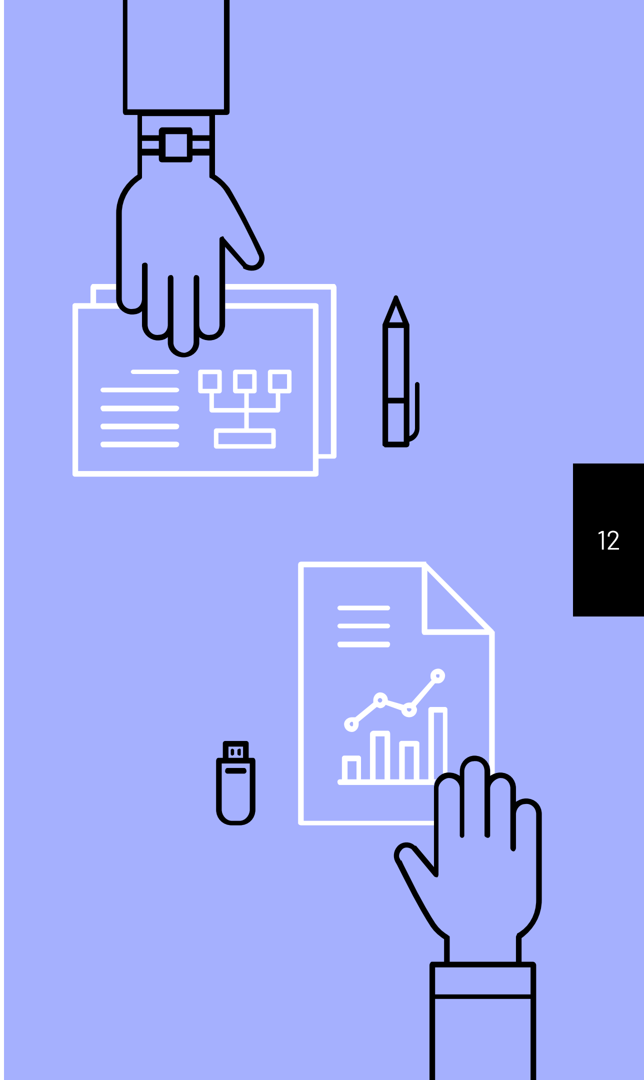
9. Performing and executing queries .

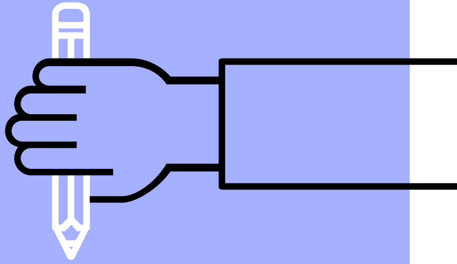


10. Performing Statistical Analysis of Business Data

11. Training the Employees in new procedures and protocols.

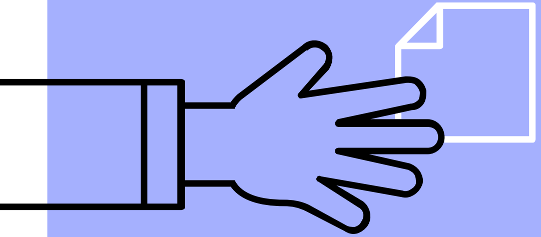
12. Undertake role of Support Analyst in case of Maintenance Projects.





Data Classification

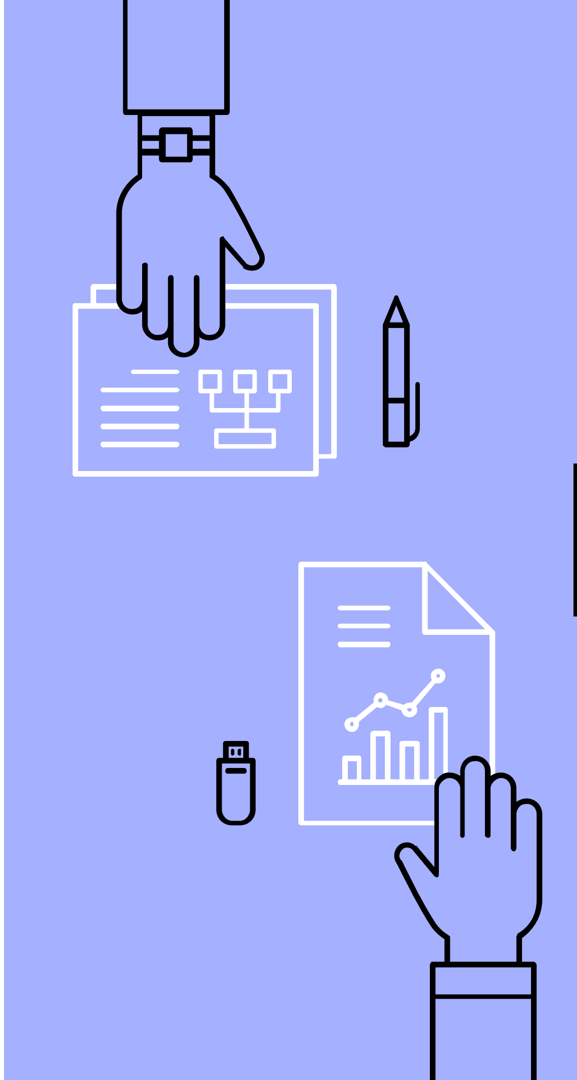
Data is classified into two major types



Structured Data

Structured Data Refers to information with a high degree of organization.

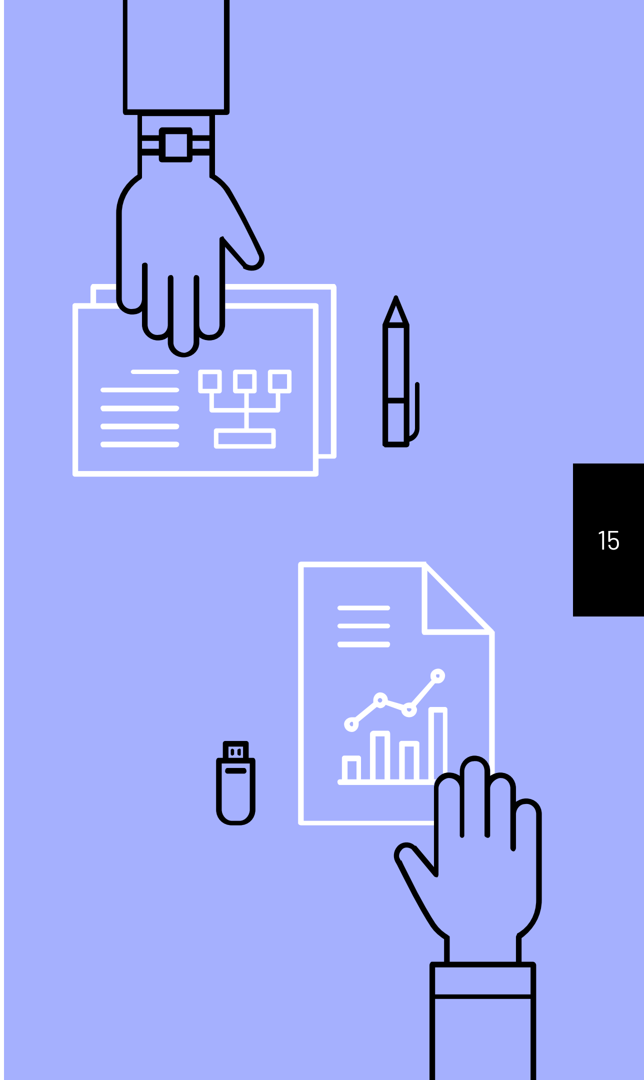
- Relational Database is readily available by simple select/insert/update operations.

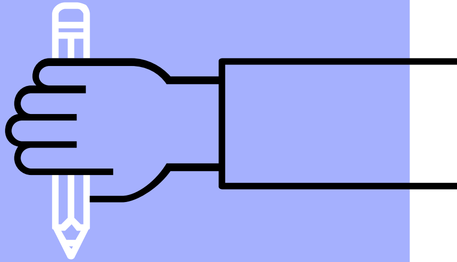


Unstructured Data

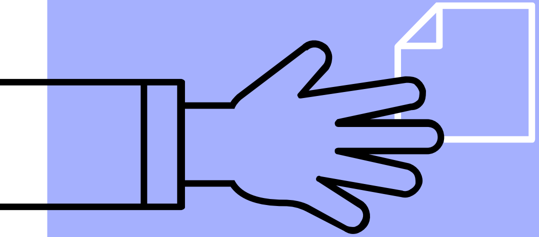
Opposite of Structured.

The lack of structure makes compilation a time and energy consuming task.





Data Processing



Let’s start with the first set of slides

Data Processing is any process that a computer program does to enter data, and summarize analyses or in some way, convert data into usable information.

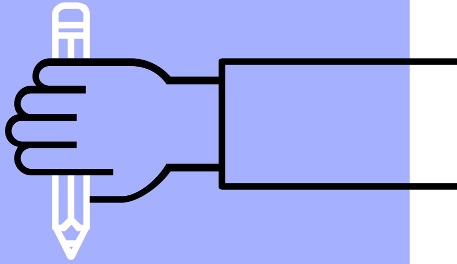
Automated Processes that run on computer involve recording, analyzing, sorting, summarizing, calculating,disseminating and storing data.

“

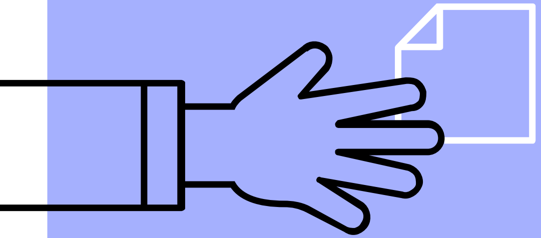
For data, In order to be processed by a computer, Data should be fed into the system and then processed.

Processing Steps

1. *Data Summarization*
2. *Data Aggregation*
3. *Data Validation*
4. *Data Tabulation*
5. *Statistical Analysis*



Storage Models



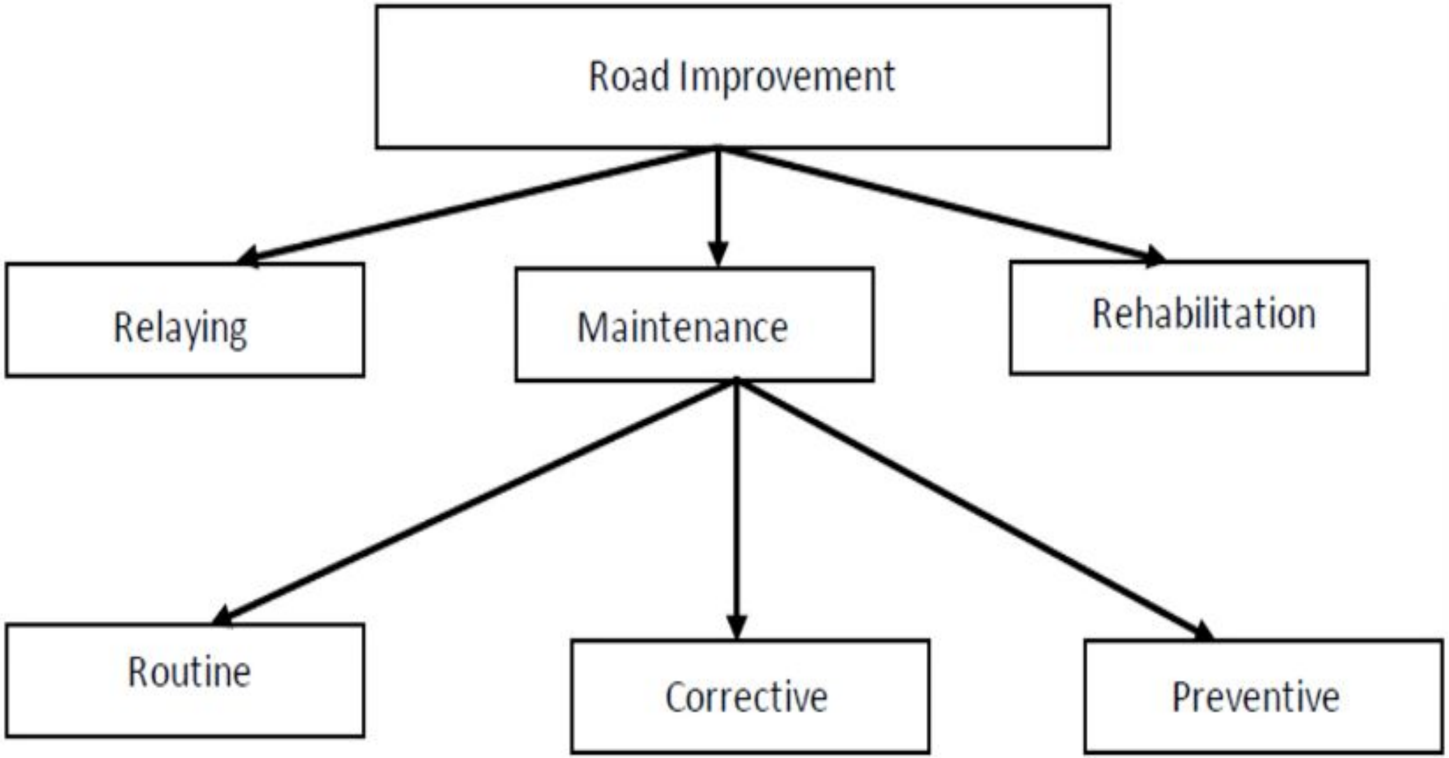
> *A database model is a type of data model that determines the logical structure of a database.*

The structure determines in which manner data can be stored, organized and manipulated.

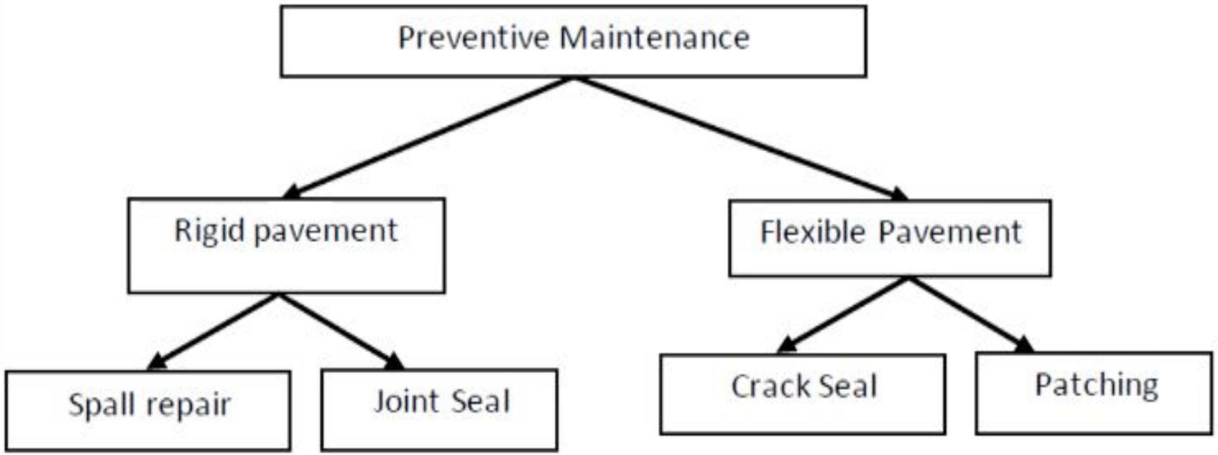
Type of Database Models

1. *Hierarchical*
2. *Network Model*
3. *Inverted File Model*
4. *Relational Model*
5. *Dimensional Model*
6. *Flat File Model*

In a hierarchical model, data is organized into a **tree-like structure, implying a single parent for each record**.



The network model expands upon the **hierarchical structure, allowing many-to-many relationships** in a tree-like structure that allows multiple parents.



Preventive Maintenance

Rigid pavement

Flexible Pavement

Spall repair

Joint Seal

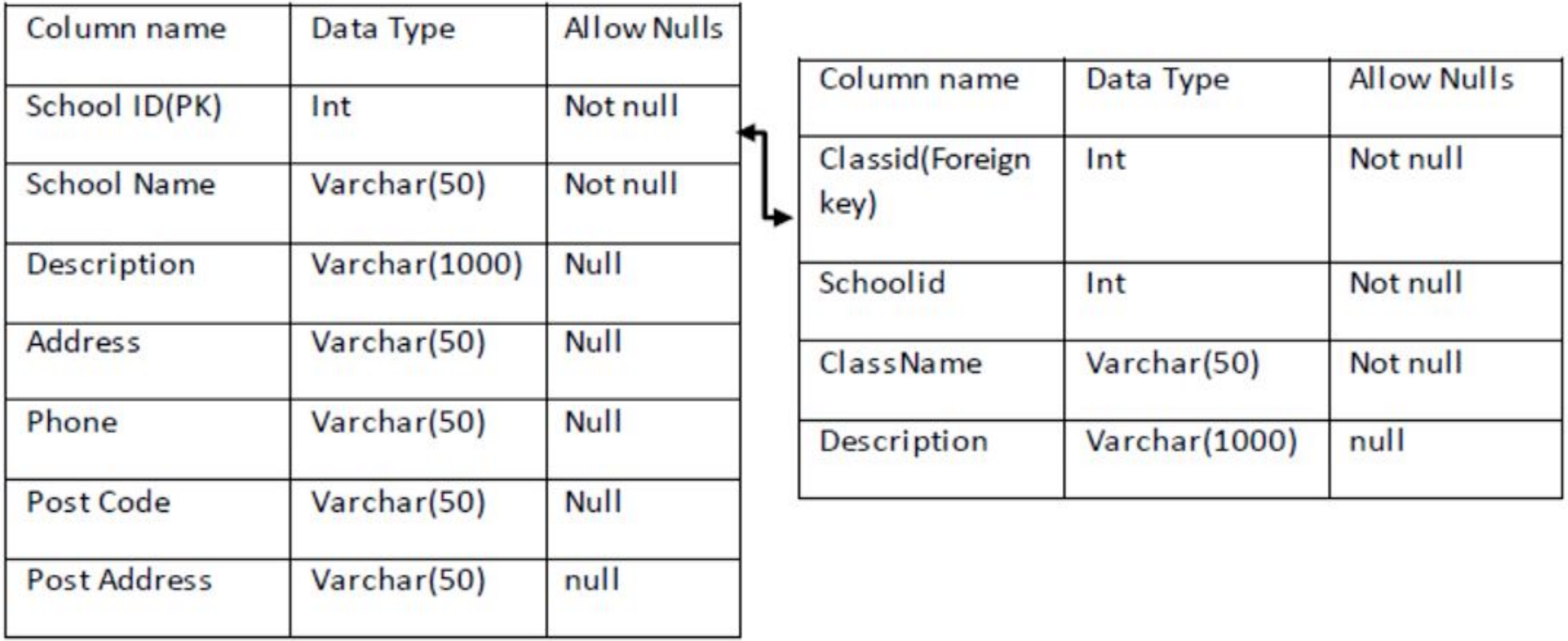
Crack Seal

Patching

❖Inverted File Model

In an inverted file or inverted index, the contents of the data are used as keys in a lookup table, and the values in the table are pointers to the location of each instance of a given content item.

- The relational model was used to make **database management systems** more independent of any particular application. It is a **mathematical model** defined in terms of **predicate logic and set theory, and systems implementing** it have been used by mainframe, midrange and microcomputer systems.



Column name	Data Type	Allow Nulls
School ID(PK)	Int	Not null
School Name	Varchar(50)	Not null
Description	Varchar(1000)	Null
Address	Varchar(50)	Null
Phone	Varchar(50)	Null
Post Code	Varchar(50)	Null
Post Address	Varchar(50)	null

Column name	Data Type	Allow Nulls
Classid(Foreign key)	Int	Not null
Schoolid	Int	Not null
ClassName	Varchar(50)	Not null
Description	Varchar(1000)	null

[redacted]

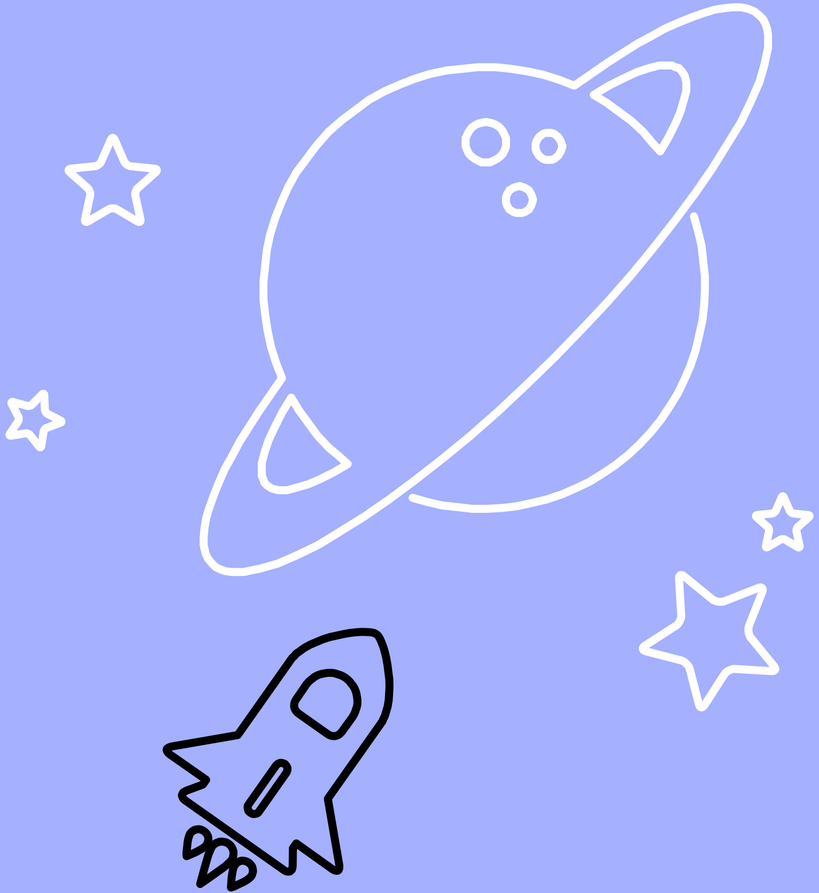
- **Dimensional Model:** The dimensional model is a **specialized adaptation of the relational model** used to represent data in **data warehouses** in a way that data can be easily summarized using **OLAP queries.**

- **Flat File model:** A Flat file can be a plain text file or a binary file. There are usually no structural relationships between the records

- The data arrangement consists of a series of columns and rows organized into a tabular format.

Example:

ID	Name	Team
1	M S DHONI	CSK
2	YUVRAJ SINGH	PW
3	ROHIT SHARMA	MI
4	RAHUK DRAVID	RR
5	VIRAT KOHLI	RCB
6	GAUTAM GAMBHIR	KKR



Normalization of Database

- Database Normalization is the process of organizing the fields and tables of a relational database to minimize redundancy and dependency.

- Normalization usually involves dividing large tables into smaller and less redundant tables and defining relationships between them.

Objectives of Normalization

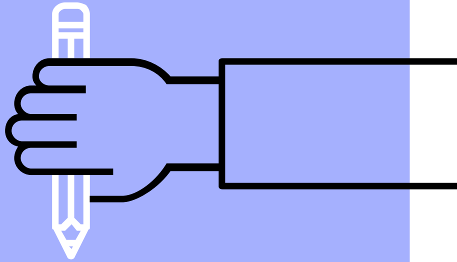
- The objective is to isolate data so that additions, deletions and modifications of a field can be made in just one table and propagated using the defined relationships.

Objectives of Normalization

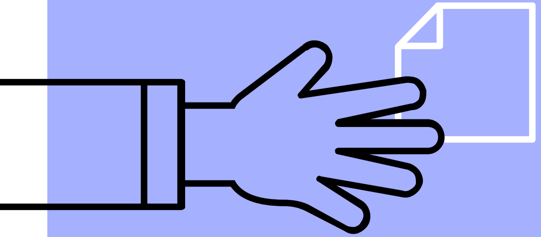
- **To free the database from modification dependencies.**
- **Minimize redesign when extending data structure.**

Denormalization

- **Denormalization is the approach to speed up read performance (data retrieval) in which the administrator selectively adds back specific instances of redundant data.**
- **When a query combines data from multiple tables into a single result table, it is called a join. Multiple joins in same query can have negative impact on performance. Denormalization can be useful for cutting down on number of joins.**



Data Models



Data Model

Data Model is a description of objects together with their properties & relationships

Objects

Suppliers

Products

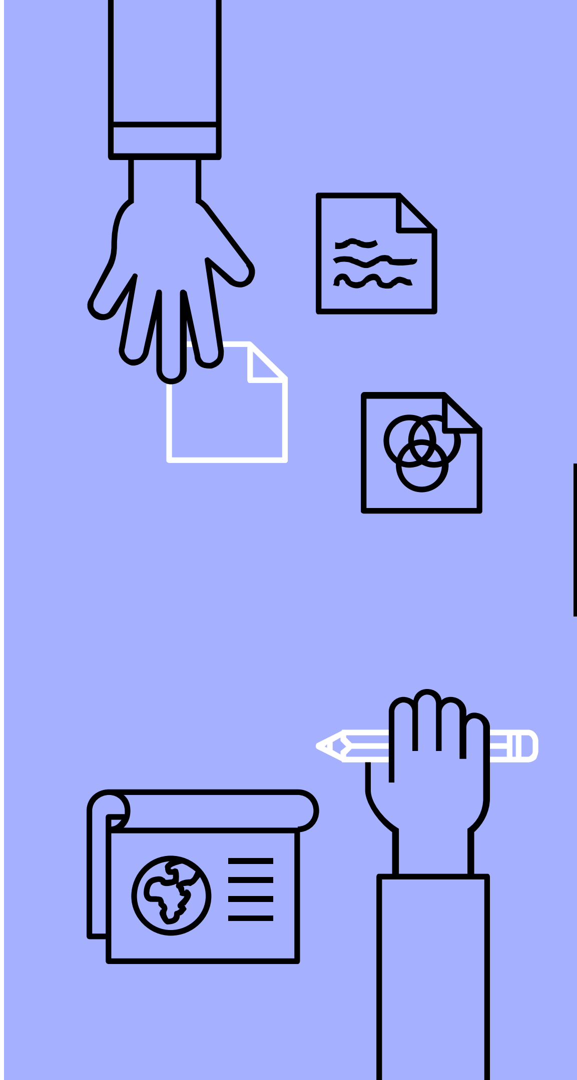
Customers

Orders

Data Dictionary

A data dictionary or metadata repository is a "centralized repository of information" about data such as meaning, relationships to other data, origin, usage and format.

- Accessible
- Reliable
- Robust

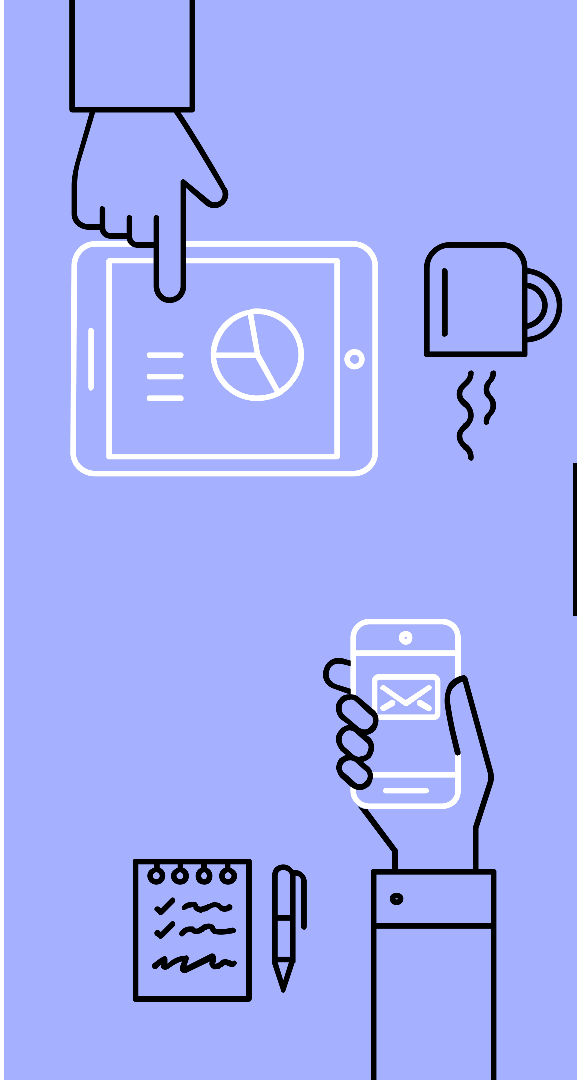


DATA MASKING

Data Masking is a method of creating a structurally similar but unauthentic version of an organization’s data that can be used for purposes such as software testing and user training.

The purpose is to protect the actual data while having a functional substitute for occasions when the real data is not required.

In data masking, the format of data remains the same, on;ly values are changed using shuffling, encryption, substitution

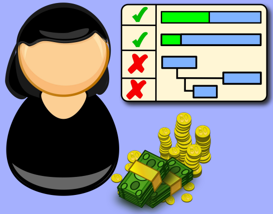


OVERVIEW

KEY ROLES OF A SUCCESSFUL ANALYST PROJECT

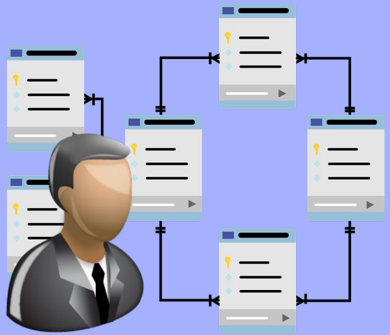
Business Users

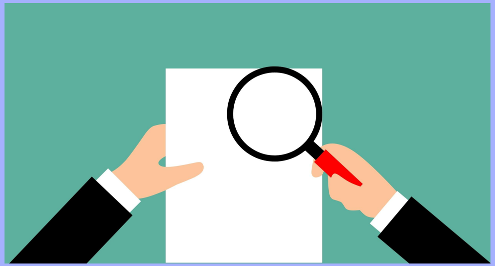
Project Sponsors


Project Manager

Business Intelligence Analyst

Database Administrator


Data Engineer

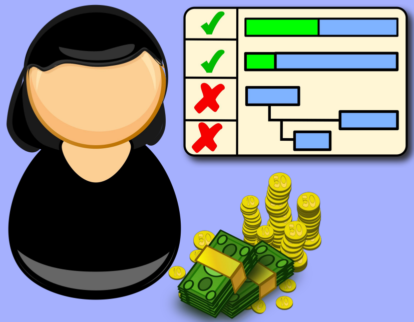

Data Scientist

Business Users

- Someone who understands the Domain Area and Benefits from the results
- Persons responsible for consulting and advising the project Team
- Person who directly deals with questions related to the context of the project, the value of the results and how the outputs will be operationalized.
- Persons who fulfill this role include - Business Analyst, Line Manager, Deep Subject Matter Expert

Project Sponsors

- Responsible for the genesis of the project
- Provides the impetus and requirements for the project and defines the core business problem.
- Generally provides the funding and gauges the degree of value from the final outputs of the working team.
- This person sets the priorities for the project and clarifies the desired outputs.



Project Manager

- Ensures that key milestones and objectives are met on time at the expected quality.

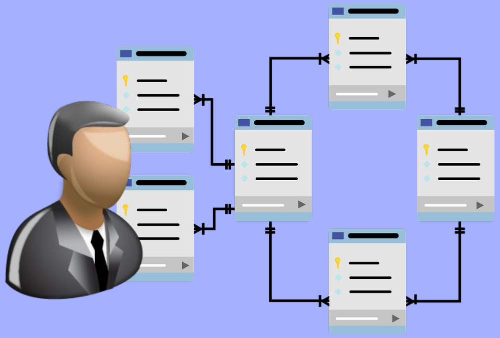
- Shares the overall responsibility for the successful planning and execution of a project.

Business Intelligence Analyst

- Provides business domain expertise based on deep understanding of the data, KPIs, Key Metrics and Business Intelligence from a reporting perspective.
- BI Analyst generally creates dashboards and reports and have knowledge of the data feeds and sources.

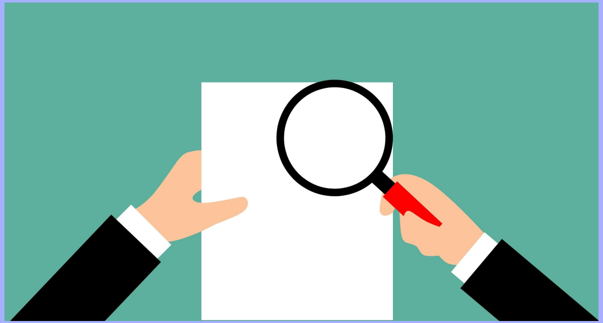
Database Administrator

- Provisions and configures the database environment to support the analytics needs of the working team.
- These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories



Data Engineer

- Leverages deep technical skills to assist with tuning SQL Queries for data management and data extraction and provides support for data ingestion into the analytic sandbox.
- The data engineer executes the actual data extractions and performs substantial data manipulations to facilitate the analytics.
- The data engineer works closely with the data scientist to help shape data in the right ways for analyses.



Data Scientist

- Provides subject matter expertise for analytical techniques, data modeling and applying valid analytical techniques to given business problems.
- Ensures overall analytics objectives are met.
- Designs and executes analytical methods and approaches with the data available to the project.

THANKS!

Any questions?

You can find me at:

ankita.sinha8118@gmail.com