ASSIGNMENT-1

NAME: Guvvala Swarna latha.

DATA ENGINEERING

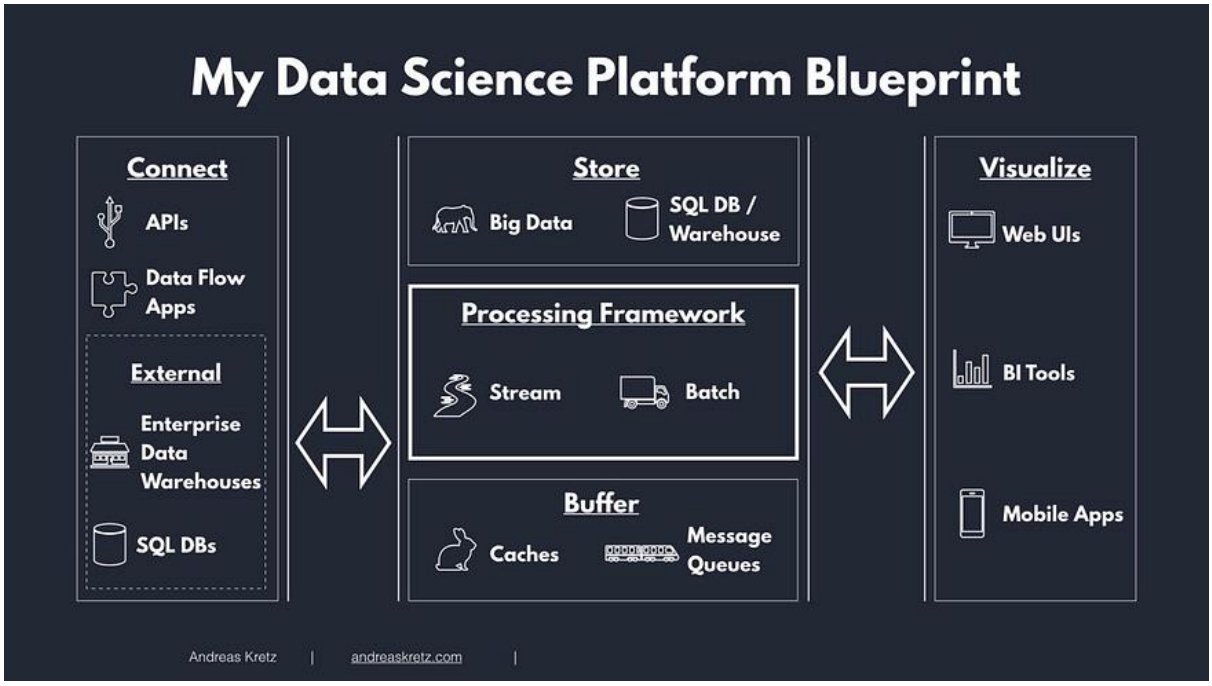
DEFINITIONS:

Data Engineer: A Data Engineer transforms data into a useful format for analysis.

Data Scientist: A Data Scientist work together with analyst and businesses to convert data insights into action.

Data Engineering: Data engineering is a processes of organizing ,managing and analysing large amounts of a data.so that it can be used by data scientists.

- ETL(Extract,Transform,Load)-Move data between systems .They extract data ,then apply rules to transform the data through steps that make it more suitable for analysis.



DATA CLASSIFICATION

Raw Data :Refers to the original, unprocessed and unstructured information collected or generated.

Eg.JSON

Processed Data:Raw data with schema applied.

Cooked Data: summarization of processed data.

BIG DATA PROPETIES

Volume: How much data you have.

Velocity: How fast data is getting to you.

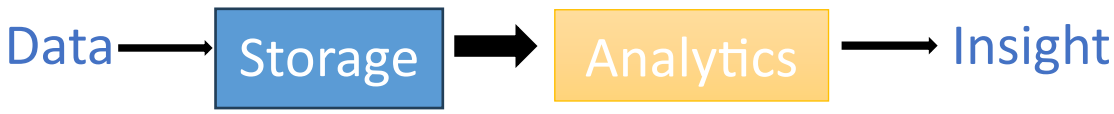
Variety: How different your data is.

Veracity: How reliable your data is.

- To provide solutions ,Data engineering comes into picture.

DATA PROCESSING METHODS

Batch processing:



Stream Processing:

Process data on the fly ,as it comes in.

STREAMING METHODS

- At Least Once
- At Most Once
- Exactly Once

PROCESSING FRAMEWORKS

MAP REDUCE:

Key -value pairing.

Organize the data into keys and values, sort by the key, combine the data with matching keys Repeat until you have the final key-value outcome.

Eg.Hadoop, Apache spark, Beam, Samza.

DATA STORAGE

Relational Database(SQL).

Document Store(NOSQL).

INTRODUCTION TO DATA WAREHOUSING

Data Warehousing: Storing and managing large volumes of structured and unstructured data for analysis and reporting purpose.

- Method of organizing and compiling data into one database.
- Collaborate data from several sources and ensure data accuracy.

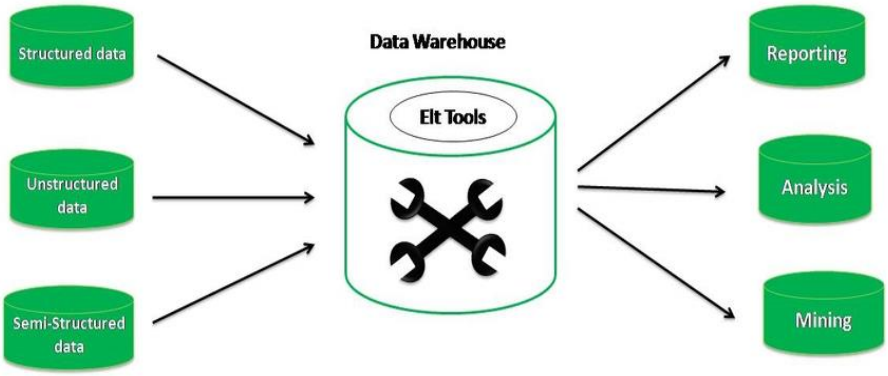
Features of Data Warehouse:

- Subject Oriented-It provides you with important data about a specific subject like suppliers, products, promotion, customers etc..
- Integrated- Different heterogeneous sources are put together to build a data warehouse such as level documents or social database.
- Time-Variant-The data collected in a data warehouse is identified with a specific period.
- Non volatile- No updates are allowed .once the data entered into the data warehouse, they are never removed.

Purpose of Data Warehouse:

The primary purpose of a data warehouse is to enable companies to access and analyze all of their data to derive the most accurate business insights and forecasting models.

Data Warehouse Architecture:



Operational Data Store(ODS):

ODS is a Subject-Oriented, Integrated, Current-Valued, Volatile and Detailed. In a Subject-Oriented ,it is organized around the significant information subject of an enterprise. In an integrated that is, it is a group of subject-oriented record from a variety of systems to provide an enterprise-wide view of the information. In a Current -Valued that is ,an ODS is up-to—date and follow the current status of the data ,an ODS does not contain historical information.In volatile that is, the data in the ODS frequently changes as new data refreshes the ODS. In detailed that is, ODS is

detailed enough to serve the need of the operational management staff in the enterprise.

Online Transaction Processing(OLTP):

OLTP is a methodology to provide end users with access to large amounts of data, it works in an intuitive and rapid manner to assist with deductions based on investigative reasoning.

OLTP refers to a class of systems that facilitate and manage transaction-oriented applications ,typically for data entry and retrieval transaction processing.

Eg. An ATM for a bank is an example of a commercial transaction processing application.

OLTP *Vs* Data Warehouse (OLAP) Applications:

The main distinction between the two systems is in their names: analytical vs. transactional. Each system is optimized for that type of processing.
OLAP is optimized for conducting complex data analysis for smarter decision-making. OLAP systems are designed for use by data scientists, business analysts and knowledge workers, and

they support business intelligence (BI), data mining and other decision support applications.

OLTP, on the other hand, is optimized for processing a massive number of transactions. OLTP systems are designed for use by frontline workers (e.g., cashiers, bank tellers, hotel desk clerks) or for customer self-service applications (e.g., online banking, e-commerce, travel reservations).

Data Marts:

A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse. For example, many companies may have a data mart that aligns with a specific department in the business, such as finance, sales, or marketing.

Data Marts *Vs* Data Warehouse:

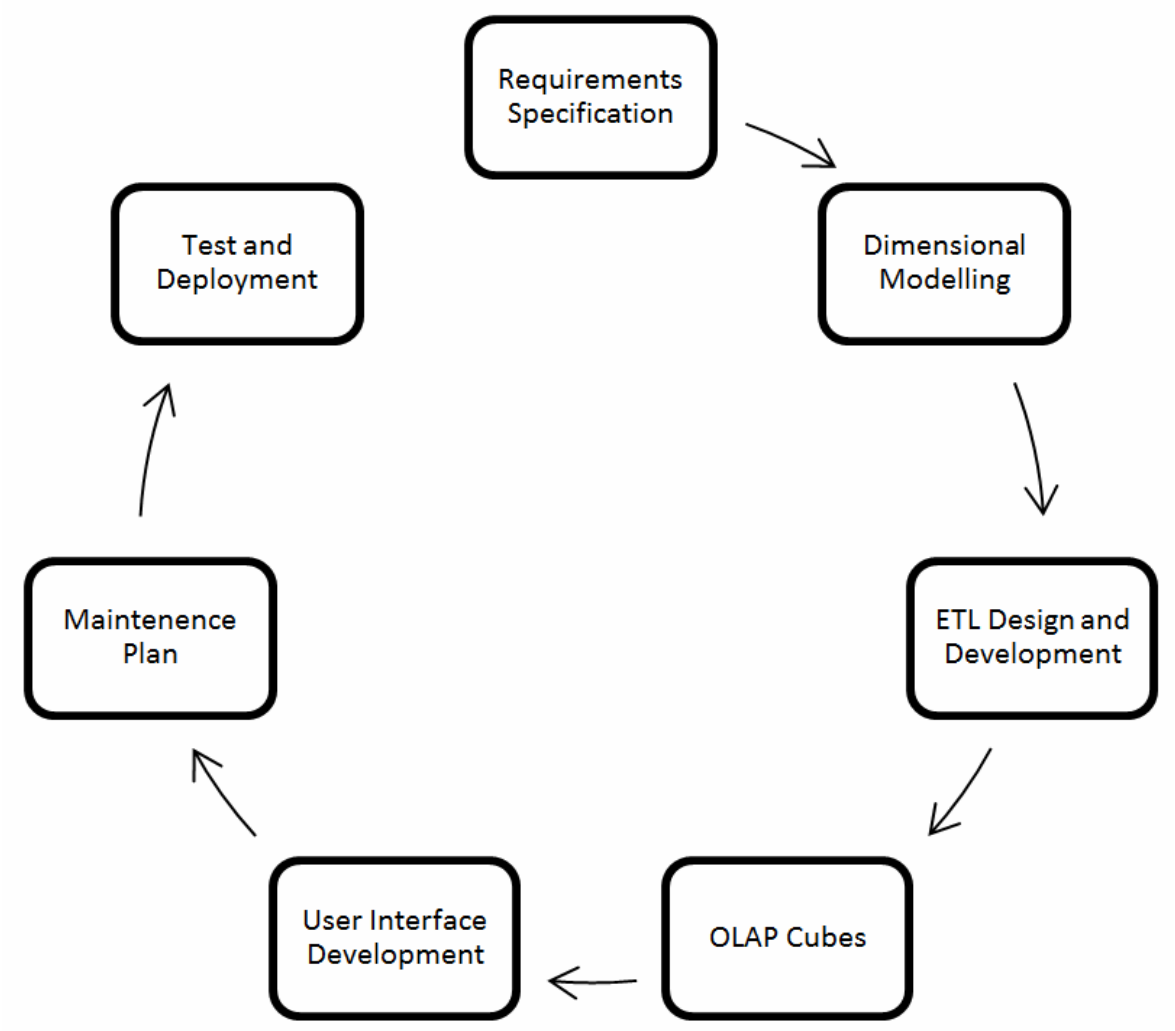
- Size: In terms of data size, data marts are generally smaller, typically encompassing less than 100 GB. In contrast, data warehouses are much larger, often exceeding 100 GB and even reaching terabyte-scale or beyond.

- Range: Data marts cater to the specific needs of a single line of business or department within the organization. On the other hand, data warehouses are designed to be enterprise-wide, spanning across multiple functional areas and serving the data requirements of the entire organization.
- Sources: Data marts draw data from a limited number of sources, while data warehouses have a more comprehensive scope, collecting data from a diverse array of sources.. Data warehouses integrates information from numerous operational systems, applications, and external feeds to offer a holistic and comprehensive view of the organization's data landscape.

Data Warehouse Life Cycle:

- Requirement Specification: It is the first step in the development of the Data Warehouse and is done by business analysts. In this step, Business Analysts prepare business requirement specification documents. After the requirements are gathered, the data modeler starts recognizing the dimensions, facts & combinations based on the requirements.

- Data Modelling: This is the second step in the development of the Data Warehouse. Data Modelling is the process of visualizing data distribution and designing databases by fulfilling the requirements to transform the data into a format that can be stored in the data warehouse. Data modelling helps to organize data, creates connections There are three data models for data warehouses:
    - Star Schema
    - Snowflake Schema
    - Galaxy Schema



- ELT Design and Development: This is the third step in the development of the Data Warehouse. ETL or Extract, Transfer, Load tool may extract data from various source systems and store it in a data lake. An ETL process can extract the data from the lake, after that transform it and load it into a data warehouse for reporting.
- OLAP Cubes: This is the fourth step in the development of the Data Warehouse. An OLAP cube, also known as a multidimensional cube or hypercube, is a data structure that allows fast analysis of data according to the multiple dimensions that define a business problem. A data warehouse would extract information from multiple data sources and formats like text files, excel sheets, multimedia files, etc. The extracted data is cleaned and transformed and is loaded into an OLAP server (or OLAP cube) where information is pre-processed in advance for further analysis.
- UI Development: This is the fifth step in the development of the Data Warehouse. The main aim of a UI is to enable a user to effectively manage a device or machine they're interacting with. There are plenty of tools in the market that helps with UI development.
- Maintenance: This is the sixth step in the development of the Data Warehouse. In this phase, we can update or make changes to the

schema and data warehouse's application domain or requirements.

- Test and Deployment: This is often the ultimate step in the Data Warehouse development cycle. Businesses and organizations test data warehouses to ensure whether the required business problems are implemented successfully or not.. The data warehouses can be deployed at their own data center or on the cloud.