Data engineering:

Designing, building and scaling the systems/platforms that organize the data for analytics or pipelines for easy flow of data.

Data science :

They focus more on creating new insights from existing data.

Extract ,transform,Load(ETL):

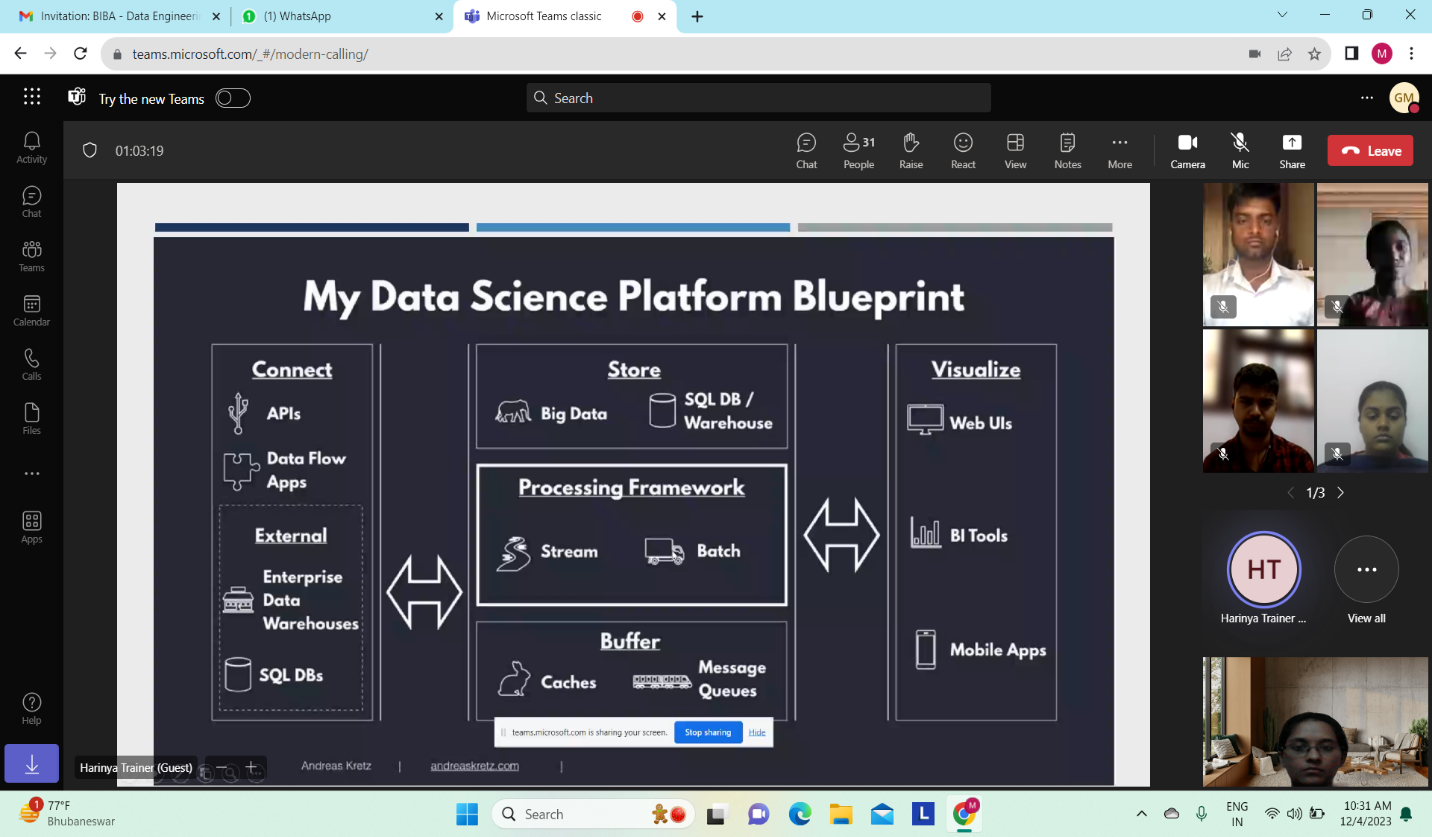
In world of data warehousing if u need to bring data from multiple sources into one centralized database ,you must first follow ETL

Extract data from its original source

Transform data by duplicating it,combining and ensuring quality

Load data into target database

Datascience platform blueprint:



Data classification:

Raw data:unprocessed data with no schema applied

Processed data:raw data with schema applied,stored in event tables .

Cooked data:

Processed data that has been summarized.

Big data:

Huge data

4vs :

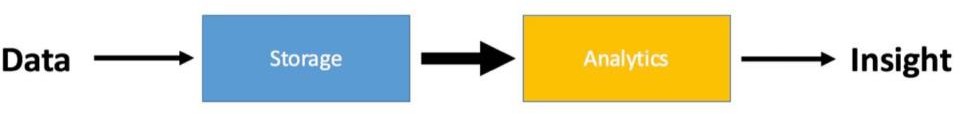
Volume:how much data u have

Velocity:how fast data is getting to u

Variety:how different your data is

Veracity:how reliable ur data is

Batch processing:



Data:collection of facts

Analytics:organizing and examining data

Insights:discovering patterns in data.

Stream processing:

Process data on the fly as it comes in.

Map reduce:

the Map Reduce in big data breaks down large data sets into smaller, easier-to-manage chunks, which are then processed concurrently.

Key – Value pairing.

Organize the data into keys and values, Sort by the key,

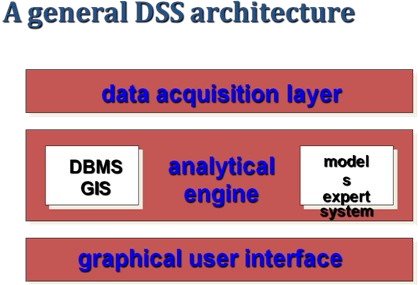
Combine the data with matching keys

Repeat until you have the final key- value outcome.

Decision support systems:

Is used to improve a company's decision-making capabilities,helps us to access and resolve everyday business questions.

It analyzes large amounts of data and presents an organization with the best possible options available.(gps)



In order to implement dss we have dss architectural styles:

OLTP (Online Transaction Processing)

-used by traditional operational systems (RDBMS).

OLAP (Online Analytical Processing)

-used by Data Warehouse.

Operational database:

It carries out regular operations which is assessed

by operational systems.

Uses OLTP architecture and OLTP Databases access the data in the form of operations like- Inserting, Deleting, and Updating data.

OLTP:

Provides end users to access large amounts of data.mainly for retrieving and transactional processing.

Ex:atm

Benefits of OLTP:

Simplicity,effiencey,data integrity,fast query processing

Downfalls: Requires instant update,Data from oltp not suitable for analysis.

Data warehouse (database used for reporting and analysis )

It is a subject oriented ,integrated,time variant,non volatile collection of data in support of management system that is designed for query and analysis rather than transaction processing .It is not used for daily operations and transaction processing but used for making decisions.

Characteristics of data warehouse:

Subject oriented :data warehouse mainly focuses on modeling and analysis of data on a specific subject for decision makers.

Integrated: A data warehouse integrates data from different sources like RDBMS, flat files etc.

Time-variant: time horizon for datawarehouse is significantly longer than operational systems.like past 5-10years data

Non-volatile: No updates are allowed.The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed. Non-Volatile defines that once entered into the warehouse, and data should not change.

The data store contains two main types of data.

-Business Data

-Business Data Model

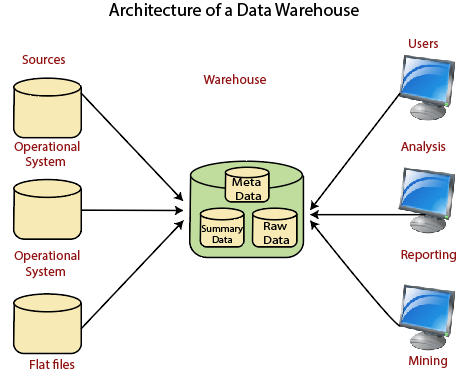
Business data are extracted from operational data+external data.

Purpose of datawarehouse:

The primary purpose of a data warehouse is to provide a central repository of information that can be quickly analyzed and queried to generate relevant insights.

Architecture of data warehouse:

Basic:



Operational systems are systems that process the day-to-day transactions of an organization.

Datawarehouse with staging area:

We must clean and process your operational information before put it into the warehouse.

Here staging takes place between daat sources and warehouse.

Staging area: is a temporary location where a some data from source systems is copied and processed before entering the warehouse.

Data warehouse with staging area and datamarts:

We may want to customize our warehouse's architecture for multiple groups within our organization.we can do that by data marts

Data marts:  A data mart is a segment of a data warehouses that can provided information for reporting and analysis on a section, unit, department or operation in the company.

Dss data vs operational data:

Operational data is just collection of data where Decision support systems allow for more informed decision-making.

The DSS data which is extracted from multiple sources differ from operational data in three main areas:time span,granularity,dimensionality.

Business data represents a snapshot of companys situation.

Operational datastore:

An operational data store takes data from multiple transactional systems for operational reporting and business reporting. They combine various real-time data sources together in their original format in a central location.

Data marts:

A data mart is a simple form of a data warehouse that is focused on a single subject or line of business, such as sales, finance, or marketing. Given their focus, data marts draw data from fewer sources than data warehouses.

Data warehouse vs data marts:

Datawarehouse is data oriented and large in scope. It collects data from various data sources.

Datamart is project oriented and smaller in scope. It generally stores data from a data warehouse.

Datawarehouse lifecycle:

