**Name:** V Venkata Sri Prasad

**Batch:** Data Engineering

**Date:** 17/01/2024 – (Day 1)

**Topic: Explain Data Warehousing?**

**Introduction to Data Warehousing:**

Data warehousing is a comprehensive approach to collecting, storing, and managing large volumes of data from various sources within an organization. The primary goal is to provide a centralized repository for data analysis and reporting, enabling decision-makers to gain valuable insights and make informed strategic decisions.

**Purpose of Data Warehouse:**

1. **Centralize Data:** Aggregate data from different sources into a single, unified repository.
2. **Facilitate Analysis:** Provide a platform for complex querying and analytical processing.
3. **Support Decision Making:** Enable decision-makers to access timely and relevant information.
4. **Store Historical Data:** Maintain historical records for trend analysis and pattern recognition.
5. **Improve Data Quality:** Implement mechanisms to enhance and ensure the quality of stored data.

**Data Warehouse Architecture:**

**Bottom Tier: Data Warehouse Server**

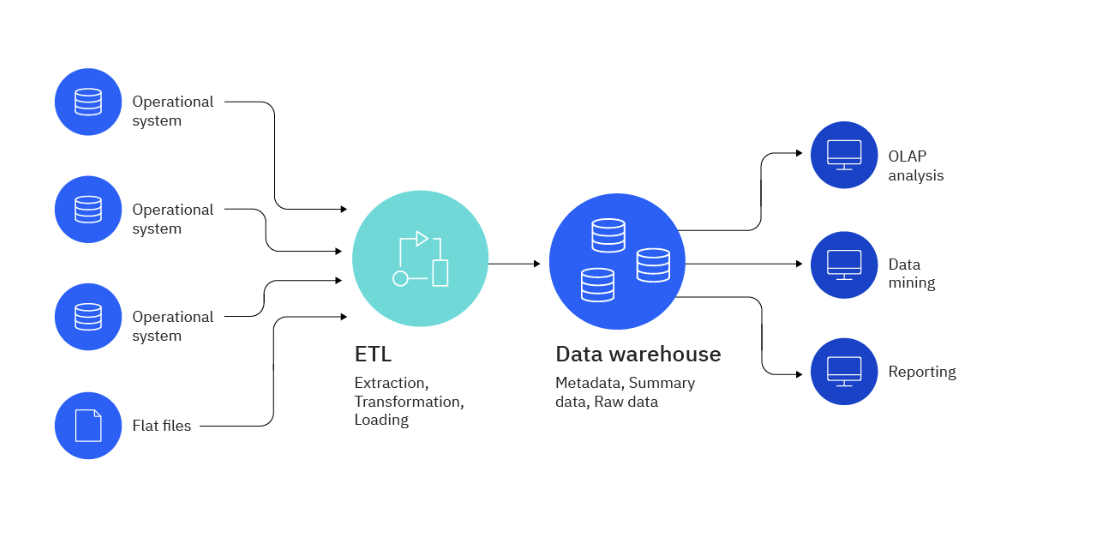
This tier features a data warehouse server, typically a relational database system. Its role is to collect, cleanse, and transform data from multiple sources using the Extract, Transform, and Load (ETL) or Extract, Load, and Transform (ELT) processes.

**Middle Tier: OLAP Server**

In the middle tier, an OLAP (online analytical processing) server takes center stage, ensuring rapid query speeds. Three OLAP models (ROLAP, MOLAP, HOLAP) are employed based on the type of database system in use, providing flexibility for efficient data processing.

**Top Tier: User Interface/Reporting Tool**

At the top tier, a user-friendly interface or reporting tool empowers end users to engage in ad-hoc data analysis on their business data. This interface serves as a control panel, allowing users to navigate and interpret insights from the data warehouse effortlessly.



**Operational Data Store (ODS):**

An Operational Data Store is an interim database that stores detailed and current data from operational systems. It acts as a staging area for data before it is loaded into the data warehouse. ODS facilitates near real-time reporting and supports operational decision-making.

**OLTP vs. Warehouse Applications:**

**OLTP (Online Transaction Processing):**

1. Handles day-to-day transactions.
2. Optimized for fast data retrieval and updates.
3. Supports operational applications like order processing.

**Data Warehouse Applications:**

1. Focuses on analytical processing.
2. Optimized for complex queries and reporting.
3. Supports decision support and business intelligence.

**Data Marts:**

Data marts are smaller, specialized subsets of a data warehouse, focusing on specific business functions or departments. They are designed to address the needs of a particular group within the organization.

**Data Marts vs. Data Warehouses:**

**Data Warehouses:**

1. Comprehensive, organization-wide data repository.
2. Centralized and integrated data.
3. Supports enterprise-level analysis.

**Data Marts:**

1. Specialized, department-specific data subset.
2. May contain a subset of data from the data warehouse.
3. Tailored for specific business units or users.

**Data warehouse development life cycle**

**1. Requirement Specification:**

Business analysts document business needs, gathering over 50% of requirements from clients in a 3-4 month process. This phase lays the overall blueprint for the data warehouse.

**2. Data Modeling:**

Data modeling organizes and designs databases to transform data into a storable format, using techniques like Star Schema, Snowflake Schema, and Galaxy Schema. It's a complex phase crucial for warehouse development.

**3. ELT Design and Development:**

ETL tools like SAS Data Management or IBM Information Server are employed to extract, transform, and load data into the warehouse, ensuring optimal speeds and consistent data pipelines.

**4. OLAP Cubes:**

OLAP cubes, multidimensional structures, process data from multiple sources for advanced analysis. Traditional in the past, they face challenges with real-time analysis, making them less favorable today.

**5. UI Development:**

A user interface (UI) is created using tools like Tableau or PowerBI, enabling user-computer interaction for efficient data analysis and report generation, shifting focus from backend processes.

**6. Maintenance:**

Maintenance involves updating schema, adapting to changing application domains or requirements, and tracking modifications. Operations include insertion, dimension changes, and category adjustments.

**7. Test and Deployment:**

Testing involves scrutinizing large volumes of data for completeness, transformation, integrity, etc. Post-testing, the data warehouse is deployed, allowing users immediate access for analysis and benefit realization.

