Unit III Data Provisioning and Data Visualization

3.1 Data Provisioning: Data warehouse, schemas, Data Quality, Data profiling, Data enrichment, data duplication, ETL Architecture and what is ETL, Extraction concept and Change data capture, Transformation concept, lookups, time lag, formats, consistency, Loading concept, Initial and Incremental loading, late arriving facts, What is Staging, Data marts, Cubes.

Data Provisioning Overview

Data provisioning refers to the process of preparing, transforming, and storing data so that it is readily available for analysis and decision-making. It involves gathering data from different sources, cleaning it, transforming it, and then making it accessible for business intelligence systems or analytical tools.

---

1. Data Warehouse

- **What is it?**
  A **data warehouse** is a **centralized repository** where data from various sources is stored. It is specifically designed to handle **large volumes of historical data** for analysis and reporting.
- **Key Characteristics:**
  - Optimized for **querying** and **reporting**.
  - Stores **historical data** that is processed periodically.
  - Data is **organized in schemas** like star schema or snowflake schema.

---

2. Schemas

- **What are schemas?**
  Schemas define the structure of the data in a database or data warehouse. They organize data in a way that makes it easy to store, retrieve, and analyze.
- **Types of Schemas:**

- **Star Schema**: Simple design where a **fact table** is connected to **dimension tables**.
- **Snowflake Schema**: More complex version of the star schema, with normalized **dimension tables**.

---

3. Data Quality

- **What is Data Quality?**
  Ensuring that the data is **accurate**, **consistent**, and **timely**. High-quality data is essential for accurate analytics and decision-making.
- **Key Aspects of Data Quality:**
  - **Accuracy**: Data should be correct and reliable.
  - **Consistency**: Data should be uniform and free from conflicts.
  - **Completeness**: Data should be comprehensive and not missing important values.

---

4. Data Profiling

- **What is Data Profiling?**
  It is the process of **examining** the data to **understand** its structure, relationships, and quality before transformation.
- **Key Aspects:**
  - Identifying **data anomalies**.
  - Understanding **data patterns** and **distribution**.
  - Checking for **missing or inconsistent values**.

---

5. Data Enrichment

- **What is Data Enrichment?**
  It refers to the process of **enhancing** the existing data by adding

additional information from external sources, thus improving its quality and usefulness.

- **Example:**
  Adding **geographical data** (like city or region) to customer data to enhance customer analysis.

---

6. Data Duplication

- **What is Data Duplication?**
  Data duplication occurs when the same piece of data is stored in multiple places, leading to redundancy. This can **increase storage costs** and create **inconsistencies**.
- **Key Concepts:**
  - **De-duplication** is essential in the ETL process to ensure data is unique and optimized for analysis.

---

7. ETL Architecture and ETL Concept

ETL stands for **Extract, Transform, Load**, and is the primary process for **data integration** and **data provisioning**.

- **ETL Architecture**:
  - Involves three main stages: **Extract**, **Transform**, and **Load**.
  - **Data sources** feed into the **ETL pipeline**, which processes the data and loads it into the **data warehouse** or **data marts**.

---

8. ETL Process Breakdown

✅ **Extraction Concept:**

- **What is it?**
  Extracting data means pulling data from **various source systems** (e.g., databases, APIs, flat files).

- **Change Data Capture (CDC)**:
  This technique allows capturing only the **changes** made to the source data, rather than extracting all data, which saves time and resources.

✅ **Transformation Concept:**

- **What is it?**
  Transformation involves cleaning, formatting, and converting the extracted data into the desired structure. Common transformations include:
  - **Lookups**: Matching data to reference tables.
  - **Time Lag**: Handling historical data that may not be synchronized.
  - **Consistency**: Ensuring that data conforms to predefined standards.
  - **Formats**: Changing data into appropriate formats (e.g., dates, currencies).

✅ **Loading Concept:**

- **What is it?**
  Loading is the final step where the transformed data is loaded into a **data warehouse** or **data mart** for analysis.
- **Types of Loading**:
  - **Initial Loading**: Loading the full dataset into the data warehouse.
  - **Incremental Loading**: Only new or changed data is loaded after the initial load (using CDC or timestamps).
  - **Late Arriving Facts**: Handling data that arrives late in the process, which may need to be inserted or updated in the data warehouse.

---

9. Staging

- **What is Staging?**
  Staging refers to a **temporary storage area** where data is held before it undergoes the **ETL process** (Extraction, Transformation, Loading). It allows for initial validation and cleaning before loading into the final data warehouse.

---

10. Data Marts

- **What is a Data Mart?**
  A **data mart** is a **subset of a data warehouse** designed to focus on specific business areas (e.g., sales, finance).
- **Difference from Data Warehouse**:
  While a **data warehouse** contains data from all business areas, a **data mart** contains data specific to one department or business function.

---

11. Cubes

- **What is a Cube?**
  A **data cube** is a multi-dimensional array of data used to represent **summarized data** in various dimensions (e.g., time, geography, product).
- **Key Concepts**:
  - **OLAP Cubes** (Online Analytical Processing) are used for fast querying and analysis.
  - Helps with **slice-and-dice** operations for deeper insights.

Summary

Topic	Explanation
Data Warehouse	Centralized repository for storing large datasets for analysis.

Topic	Explanation
Schemas	Define the structure of the data (e.g., Star Schema, Snowflake Schema).
Data Quality	Ensures data is accurate, complete, and consistent.
Data Profiling	Examining and understanding data quality and patterns.
Data Enrichment	Enhancing data by adding external information.
Data Duplication	Redundant data that needs to be cleaned.
ETL (Extract, Transform, Load)	The process of extracting data, transforming it, and loading it into a data warehouse.
Extraction	Pulling data from source systems.
Change Data Capture (CDC)	Capturing changes in data rather than extracting all data.
Transformation	Cleaning and converting data into a desired structure.
Loading	Loading data into the data warehouse or mart.
Staging	Temporary storage for data before ETL processing.
Data Marts	A subset of a data warehouse for specific business areas.
Cubes	Multi-dimensional data representations for quick analysis.

3.2 Data Visualization: What Is a Business Report, Components of Business Reporting Systems, Data and Information Visualization, Types of Charts and Graphs, Visual Analytics, Performance Dashboards, Business Performance Management?

Data Visualization Overview

Data visualization is the process of representing data and information in a graphical format, helping businesses understand trends, patterns, and insights. It's crucial for making data-driven decisions and presenting complex information in a clear, easy-to-understand way.

Here's a **simplified breakdown** of key topics in **data visualization**:

---

1. What is a Business Report?

- **Definition**:
  A **business report** is a structured document that presents data analysis, insights, and conclusions, usually related to specific business objectives or decisions.
- **Purpose**:
  The main purpose of a business report is to inform and guide decision-making. It typically includes:
  - **Key performance metrics** (e.g., sales figures, customer satisfaction).
  - **Trends** or patterns based on analyzed data.
  - **Recommendations** or conclusions based on the data.

---

2. Components of Business Reporting Systems

A **business reporting system** helps organizations collect, process, and present data for analysis. Key components of these systems include:

- **Data Sources**: Where the raw data comes from (e.g., databases, cloud services, spreadsheets).
- **Data Processing**: Includes cleaning, transforming, and organizing data for easy analysis.
- **Reporting Tools**: Software or platforms that create reports and visualizations (e.g., Tableau, Power BI).

- **Visualization**: Graphical representation of the processed data for better understanding.
- **Communication**: Methods for sharing reports (e.g., through dashboards, email reports).

---

3. Data and Information Visualization

- **Data Visualization**:
  The graphical representation of raw data, typically in the form of charts, graphs, or maps. It helps to reveal patterns, trends, and outliers.
- **Information Visualization**:
  A step further from data visualization, this involves creating interactive, informative displays that allow users to explore data and derive insights on their own (e.g., interactive dashboards).

Key Benefits:

- Makes **complex data** easier to interpret.
- Highlights **key trends** and insights for quicker decision-making.

---

4. Types of Charts and Graphs

Charts and graphs are commonly used to display data in a visual format. Here are some popular types:

- **Bar Chart**:
  Used to compare quantities across different categories.
- **Line Chart**:
  Ideal for showing trends over time (e.g., stock prices over a year).
- **Pie Chart**:
  Shows proportions of a whole (e.g., market share breakdown).
- **Scatter Plot**:
  Used to show relationships between two variables.

- **Histogram**:
  Displays the distribution of a dataset, often used in statistical analysis.
- **Area Chart**:
  Similar to a line chart but with the area below the line filled to emphasize volume.
- **Heatmap**:
  Uses color gradients to show data intensity, often used in geographic data or matrix-like data.

When to Use Each:

- **Bar & Line charts** for comparisons and trends.
- **Pie charts** for proportions.
- **Scatter plots** for correlations.

---

5. Visual Analytics

- **What is Visual Analytics?**
  Visual analytics is the process of using **interactive visualizations** to explore and analyze large datasets. It combines **data visualization** with **analytics tools** to help users discover insights in a more intuitive and exploratory manner.
- **Benefits**:
  - Helps uncover **hidden patterns** in data.
  - Facilitates **self-service analytics**, allowing users to explore data without needing deep technical expertise.
- **Tools for Visual Analytics**:
  Tools like **Tableau**, **Power BI**, and **QlikView** allow users to build interactive dashboards and perform ad-hoc analysis with ease.

---

6. Performance Dashboards

- **What is a Performance Dashboard?**
  A **performance dashboard** is a **visual representation** of key metrics and performance indicators (KPIs) in real-time. Dashboards give a quick overview of how well a business is performing against its goals.
- **Types of Dashboards**:
  - **Operational Dashboards**: Focus on the day-to-day operations of a business, showing real-time metrics.
  - **Strategic Dashboards**: Provide a broader overview of long-term business goals, helping to track progress.
  - **Analytical Dashboards**: Used to drill deeper into data, typically for detailed analysis.
- **Key Features**:
  - Real-time **data updates**.
  - Customizable **visualizations** to display KPIs (e.g., sales, revenue, customer satisfaction).
  - Interactive **drill-down** features to explore more detailed data.

---

7. Business Performance Management (BPM)

- **What is BPM?**
  Business Performance Management refers to the practices and tools used to **measure and monitor** a company's performance against its strategic objectives. It involves analyzing data from various sources and using the insights to improve business operations and decision-making.
- **Key Components**:
  - **KPIs (Key Performance Indicators)**: Specific metrics used to measure business success.
  - **Benchmarking**: Comparing performance against industry standards or competitors.

    - **Strategic Goals**: Setting clear business goals and aligning them with performance measures.
    - **Feedback Loops**: Regular monitoring and adjustment based on performance insights.

- **Tools for BPM**:
  Performance dashboards and **data visualization tools** are commonly used to track performance metrics and provide insights into how to improve operations

3.3 BI Tools: Tableau, power BI, Dundas BI, Oracle BI,bMs excel

Business Intelligence (BI) Tools Overview

Business Intelligence (BI) tools help businesses collect, process, and analyze data to support decision-making. These tools offer data visualization, reporting, and dashboards that allow users to interact with data and gain insights easily.

Here's an overview of some of the most widely used **BI tools**:

---

1. Tableau

- **Overview**:
  Tableau is one of the most popular **data visualization** and **BI tools** that helps organizations understand and analyze their data through interactive, visual reports and dashboards.
- **Key Features**:
    - **Data Connection**: Can connect to a wide range of data sources like spreadsheets, SQL databases, cloud databases, and more.
    - **Drag-and-Drop Interface**: Offers an easy-to-use drag-and-drop interface for creating dashboards and visualizations.
    - **Real-Time Data Analysis**: Supports real-time analytics for up-to-date insights.
    - **Interactive Dashboards**: Allows users to filter, drill down, and explore data interactively.

- **Best For**:
  Tableau is ideal for users who need to create sophisticated visualizations and interactive dashboards quickly and easily.

---

2. Power BI (Microsoft Power BI)

- **Overview**:
  Power BI is a cloud-based **BI platform** from Microsoft. It offers powerful data visualization tools, reporting features, and a seamless integration with other Microsoft products, such as Excel and Azure.
- **Key Features**:
  - **Integration with Microsoft Tools**: Deep integration with Microsoft Excel, SQL Server, and other Microsoft products.
  - **Customizable Dashboards**: Users can create interactive and shareable dashboards with a variety of visuals.
  - **Natural Language Query**: Users can ask questions about the data in natural language, and Power BI will provide answers in the form of visuals.
  - **Cost-Effective**: Power BI is available at a competitive price, with a free version and affordable premium options.
- **Best For**:
  Power BI is great for organizations that already use Microsoft tools and are looking for an easy-to-use, cost-effective BI solution.

---

3. Dundas BI

- **Overview**:
  Dundas BI is an advanced **BI and analytics platform** that enables data exploration, interactive dashboards, and reporting with flexible customization options.
- **Key Features**:

- **Advanced Data Visualization**: Supports interactive dashboards and visualizations with complex customizations.
- **Data Integration**: Can connect to numerous data sources, including cloud storage, SQL databases, and flat files.
- **Self-Service BI**: Allows business users to create their own dashboards and reports with minimal IT involvement.
- **Real-Time Analytics**: Provides real-time data processing and analytics.

- **Best For**:
  Dundas BI is well-suited for organizations that require advanced analytics, custom reporting, and self-service BI capabilities.

---

4. Oracle BI (Oracle Business Intelligence)

- **Overview**:
  Oracle BI is a **comprehensive BI suite** that provides analytics, reporting, and data visualization tools. It's typically used by larger organizations that require enterprise-grade BI solutions.
- **Key Features**:
  - **Advanced Analytics**: Supports both **descriptive** and **predictive** analytics, including integration with Oracle's AI and machine learning tools.
  - **Custom Reports**: Offers highly customizable reports and dashboards to meet the needs of specific business users.
  - **Data Integration**: Seamless integration with Oracle's suite of database and cloud tools, along with other third-party data sources.
  - **Scalability**: Ideal for large enterprises that need to scale their BI solutions across many departments.

- **Best For**:
  Oracle BI is best suited for large enterprises with complex data needs and existing infrastructure in Oracle systems.

---

5. Microsoft Excel (BI Features)

- **Overview**:
  While **Excel** is not a traditional BI tool, its powerful **BI features** and **add-ins** (like Power Pivot, Power Query, and Power BI integration) make it a versatile tool for data analysis and reporting.
- **Key Features**:
  - **Power Pivot**: Allows users to perform data modeling, create complex calculations, and handle large datasets.
  - **Power Query**: Offers data transformation and cleansing features to shape data before analysis.
  - **Charts and Graphs**: Excel includes a wide range of charts and graphs for visualizing data.
  - **Pivot Tables**: Great for summarizing and analyzing large datasets interactively.
  - **Integration with Power BI**: You can publish Excel reports directly to Power BI for advanced visualization and sharing.
- **Best For**:
  Excel is ideal for smaller-scale BI needs, data analysis, and reporting. It's especially useful for users familiar with Excel's interface and looking for easy-to-use BI tools for personal or small business use.