Unit 2: Data Warehousing and on-line Analytical Processing

2.1 Data Warehouse basic concepts

A **data warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.

- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- The data are stored to provide information from a historical perspective and are typically summarized.
- A data warehouse is usually modelled by a multidimensional database structure. where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount.
- A data cube provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data

some basic concepts related to data warehouses:

- **Purpose**: The primary purpose of a data warehouse is to provide a consolidated view of an organization's data from different operational systems. It serves as a single source of truth for decision-making and analysis.
- **Data Integration:** Data warehouses integrate data from multiple sources, such as transactional databases, spreadsheets, legacy systems, and external data sources. The data is transformed, cleaned, and standardized to ensure consistency and compatibility across different sources.
- **Schema Design:** Data warehouses use a specific schema design known as a star schema or snowflake schema. In a star schema, a central fact table contains the core business metrics, and multiple dimension tables provide descriptive attributes related to the facts. Snowflake schema is an extension of the star schema where dimension tables are further normalized.
- **Historical Data:** Data warehouses store historical data, capturing snapshots of data at different points in time. This allows users to analyze trends, track performance over time, and perform historical comparisons.
- **Data Aggregation:** Aggregation is a key feature of data warehouses. It involves summarizing and consolidating data at various levels of granularity, such as by time periods, geographical regions, or product categories. Aggregated data allows for faster querying and analysis of large datasets.
- **Querying and Analysis:** Data warehouses support complex querying and analysis operations. Users can run ad-hoc queries, generate reports, perform multidimensional analysis (OLAP - Online Analytical Processing), and conduct data mining to extract insights and make informed business decisions.
- **Extract, Transform, Load (ETL):** ETL is the process of extracting data from source systems, transforming it into a suitable format, and loading it into the data warehouse. ETL tools automate this process, ensuring data consistency, quality, and timeliness.
- **Data Governance:** Data governance refers to the set of policies, standards, and procedures for managing data in the data warehouse. It includes data security, privacy, access control, data quality assurance, and compliance with regulations.
- **Data Mart:** A data mart is a subset of a data warehouse that focuses on a specific business function or department. It contains a subset of data relevant to that area, providing more specialized and targeted analysis capabilities.
- **Business Intelligence (BI):** Data warehouses serve as the foundation for business intelligence initiatives. BI tools and platforms leverage the data warehouse to provide visualizations, dashboards, and reporting capabilities, enabling users to gain insights and monitor key performance indicators.

By leveraging data warehousing concepts, organizations can centralize and organize their data, enabling efficient data analysis, reporting, and decision-making processes.

2.2 Data Cube and OLAP, Data Warehouse, Design and Usage

Data Cube and OLAP:

OLAP stands for Online Analytical Processing, which is a technology that enables multi-dimensional analysis of business data. It provides interactive access to large amounts of data and supports complex calculations and data aggregation. OLAP is used to support business intelligence and decision-making processes.

Grouping of data in a multidimensional matrix is called data cubes. In Dataware housing, we generally deal with various multidimensional data models as the data will be represented by multiple dimensions and multiple attributes. This multidimensional data is represented in the data cube as the cube represents a high-dimensional space. The Data cube pictorially shows how different attributes of data are arranged in the data model. Below is the diagram of a general data cube.

The example above is a 3D cube having attributes like branch(A,B,C,D),item type(home,entertainment,computer,phone,security), year(1997,1998,1999) .

Data cube classification:

The data cube can be classified into two categories:

- **Multidimensional data cube:** It basically helps in storing large amounts of data by making use of a multi-dimensional array. It increases its efficiency by keeping an index of each dimension. Thus, dimensional is able to retrieve data fast.
- **Relational data cube:** It basically helps in storing large amounts of data by making use of relational tables. Each relational table displays the dimensions of the data cube. It is slower compared to a Multidimensional Data Cube.

Data cube operations:

Data cube operations are used to manipulate data to meet the needs of users. These operations help to select particular data for the analysis purpose. There are mainly 5 operations listed below-

- **Roll-up:** operation and aggregate certain similar data attributes having the same dimension together. For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.
- **Drill-down:** this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis. It zooms into more detail. For example- if India is an attribute of a country column and we wish to see villages in India, then the drill-down operation splits India into states, districts, towns,

cities, villages and then displays the required information.

- **Slicing:** this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn't need everything for analysis, rather a particular attribute. For example, country="jamaica", this will display only about jamaica and only display other countries present on the country list.

- **Dicing:** this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure). For example- the user wants to see the annual salary of Jharkhand state employees.

- **Pivot:** this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn't change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.

Advantages of data cubes:

- **Multi-dimensional analysis**: Data cubes enable multi-dimensional analysis of business data, allowing users to view data from different perspectives and levels of detail.
- **Interactivity:** Data cubes provide interactive access to large amounts of data, allowing users to easily navigate and manipulate the data to support their analysis.
- **Speed and efficiency**: Data cubes are optimized for OLAP analysis, enabling fast and efficient querying and aggregation of data.
- **Data aggregation:** Data cubes support complex calculations and data aggregation, enabling users to quickly and easily summarize large amounts of data.
- **Improved decision-making:** Data cubes provide a clear and comprehensive view of business data, enabling improved decision-making and business intelligence.
- **Accessibility**: Data cubes can be accessed from a variety of devices and platforms, making it easy for users to access and analyze business data from anywhere.
- Helps in giving a summarised view of data.
- Data cubes store large data in a simple way.
- Data cube operation provides quick and better analysis,
- Improve performance of data.

Disadvantages of data cube:

- **Complexity**: OLAP systems can be complex to set up and maintain, requiring specialized technical expertise.
- **Data size limitations**: OLAP systems can struggle with very large data sets and may require extensive data aggregation or summarization.
- **Performance issues**: OLAP systems can be slow when dealing with large amounts of data, especially when running complex queries or calculations.
- **Data integrity:** Inconsistent data definitions and data quality issues can affect the accuracy of OLAP analysis.
- **Cost:** OLAP technology can be expensive, especially for enterprise-level solutions, due to the need for specialized hardware and software.
- **Inflexibility:** OLAP systems may not easily accommodate changing business needs and may require significant effort to modify or extend.

2.3 Data Warehouse Implementation

Data Warehouse Implementation

There are various implementation in data warehouses which are as follows:

1. Requirements analysis and capacity planning: The first process in data warehousing involves defining enterprise needs, defining architectures, carrying out capacity planning, and selecting the hardware and software tools. This step will contain be consulting senior management as well as the different stakeholder.

2. Hardware integration: Once the hardware and software has been selected, they require to be put by integrating the servers, the storage methods, and the user software tools.

3. Modeling: Modelling is a significant stage that involves designing the warehouse schema and views. This may contain using a modeling tool if the data warehouses are sophisticated.

4. Physical modeling: For the data warehouses to perform efficiently, physical modeling is needed. This contains designing the physical data warehouse organization, data placement, data partitioning, deciding on access techniques, and indexing.

5. Sources: The information for the data warehouse is likely to come from several data sources. This step contains identifying and connecting the sources using the gateway, ODBC drives, or another wrapper.

6. ETL: The data from the source system will require to go through an ETL phase. The process of designing and implementing the ETL phase may contain defining a suitable ETL tool vendors and purchasing and implementing the tools. This may contains customize the tool to suit the need of the enterprises.

7. Populate the data warehouses: Once the ETL tools have been agreed upon, testing the tools will be needed, perhaps using a staging area. Once everything is working adequately, the ETL tools may be used in populating the warehouses given the schema and view definition.

8. User applications: For the data warehouses to be helpful, there must be end-user applications. This step contains designing and implementing applications required by the end-users.

9. Roll-out the warehouses and applications: Once the data warehouse has been populated and the end-client applications tested, the warehouse system and the operations may be rolled out for the user's community to use.

Implementation Guidelines

1. Build incrementally: Data warehouses must be built incrementally. Generally, it is recommended that a data marts may be created with one particular project in mind, and once it is implemented, several other sections of the enterprise may also want to implement similar systems. An enterprise data warehouses can then be implemented in an iterative manner allowing all data marts to extract information from the data warehouse.

2. Need a champion: A data warehouses project must have a champion who is active to carry out considerable researches into expected price and benefit of the project. Data warehousing projects requires inputs from many units in an enterprise and therefore needs to be driven by someone who is needed for interacting with people in the enterprises and can actively persuade colleagues.

3. Senior management support: A data warehouses project must be fully supported by senior management. Given the resource-intensive feature of such project and the time they can take to implement, a warehouse project signal for a sustained commitment from senior management.

4. Ensure quality: The only record that has been cleaned and is of a quality that is implicit by the organizations should be loaded in the data warehouses.

5. Corporate strategy: A data warehouse project must be suitable for corporate strategies and business goals. The purpose of the project must be defined before the beginning of the projects.

6. Business plan: The financial costs (hardware, software, and peopleware), expected advantage, and a project plan for a data warehouses project must be clearly outlined and understood by all stakeholders. Without such understanding, rumors about expenditure and benefits can become the only sources of data, subversion the projects.

7. Training: Data warehouses projects must not overlook data warehouses training requirements. For a data warehouses project to be successful, the customers must be trained to use the warehouses and to understand its capabilities.

8. Adaptability: The project should build in flexibility so that changes may be made to the data warehouses if and when required. Like any system, a data warehouse will require to change, as the needs of an enterprise change.

9. Joint management: The project must be handled by both IT and business professionals in the enterprise. To ensure that proper communication with the stakeholder and which the project is the target for assisting the enterprise's business, the business professional must be involved in the project along with technical professionals.

2.4 Data Generalization by Attribute-Oriented Induction, Data Cube Computation

Data Generalization by Attribute-Oriented Induction

Data generalization is a process in data mining that involves summarizing and aggregating data at higher levels of abstraction. Attribute-Oriented Induction (AOI) is a technique used for data generalization. It aims to find general patterns and rules by analyzing data based on specific attributes.

Here's an overview of the **process of data generalization by Attribute-Oriented Induction:**

1. **Attribute Selection:** Identify the relevant attributes that will be used for data generalization. These attributes should capture the key characteristics of the data and the desired level of abstraction.
2. **Attribute Hierarchy Construction:** Create an attribute hierarchy for each selected attribute. An attribute hierarchy represents the different levels of abstraction for an attribute. For example, a time attribute hierarchy can have levels such as year, quarter, month, and day.
3. **Attribute-Oriented Induction:** Apply the Attribute-Oriented Induction algorithm to generate generalized patterns or rules from the data. This algorithm considers the attribute hierarchy and identifies patterns that hold at different levels of abstraction.
4. **Generalization:** Use the attribute hierarchy to generalize the data. This involves summarizing or aggregating the data at higher levels of abstraction. For example, sales data for individual products can be generalized to sales data for product categories or product groups.
5. **Rule Extraction:** Extract generalized rules from the generalized data. These rules capture the patterns and relationships found in the data at different levels of abstraction. These rules can provide insights and help in decision-making.

Data Cube Computation:

Data cube computation is a technique used in OLAP (Online Analytical Processing) to generate multidimensional views of data. It involves aggregating data along multiple dimensions to create a data cube, which provides a comprehensive and summarized representation of the data.

Here's an overview of the process of data cube computation:

1. **Dimension Selection:** Identify the dimensions based on which the data cube will be constructed. Dimensions represent the different attributes or characteristics of the data. For example, in a sales data cube, dimensions can include time, product, location, and customer.
2. **Measure Selection:** Select the measures that will be used for aggregation in the data cube. Measures represent the numerical values or metrics that are being analyzed, such as sales revenue, units sold, and profit.
3. **Cube Construction:** Generate the data cube by aggregating the data along the selected dimensions and measures. This involves creating all possible combinations of dimension values and calculating the aggregated values for each combination.
4. **Aggregation:** Perform aggregation operations to summarize the data at different levels of granularity within the data cube. Aggregation can be done by summing, averaging, counting, or applying other aggregation functions to the measures.
5. **Drill-Down and Roll-Up:** The data cube allows for drill-down and roll-up operations, which involve navigating through different levels of data granularity. Drill-down involves moving from higher-level aggregated data to more detailed data, while roll-up involves summarizing detailed data to higher levels of aggregation.
6. **Data Exploration and Analysis:** Use the data cube to explore and analyze the data from various perspectives. Users can perform ad-hoc queries, slice and dice the data, and generate reports and visualizations to gain insights and make informed decisions.