 # Data Warehouse basic Concepts:-

A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.

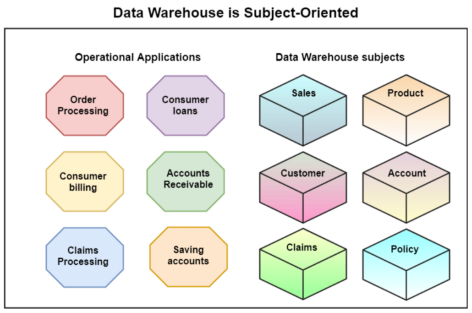
A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.

A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.

"Data Warehouse is a subject-oriented, integrated, and time-variant store of information in support of management's decisions."

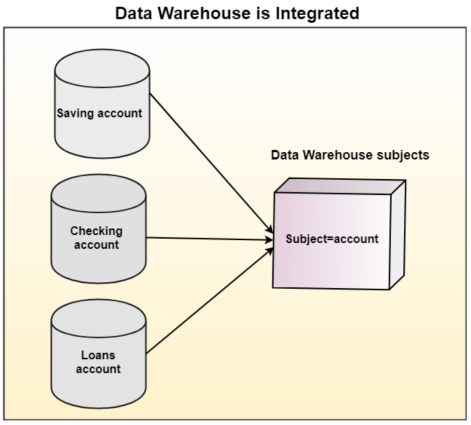
 # Characteristics of Data Warehouse:-

Subject-Oriented
A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.



Integrated

A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.

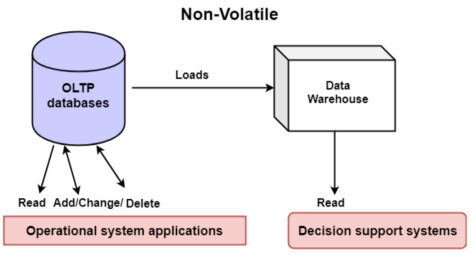


Time-Variant

Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.

Non-Volatile

The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed. It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval. Non-Volatile defines that once entered into the warehouse, and data should not change.



 # History of Data Warehouse

The idea of data warehousing came to the late 1980's when IBM researchers Barry Devlin and Paul Murphy established the "Business Data Warehouse."

In essence, the data warehousing idea was planned to support an architectural model for the flow of information from the operational system to decisional support environments. The concept attempt to address the various problems associated with the flow, mainly the high costs associated with it.

Benefits of Data Warehouse

- Understand business trends and make better forecasting decisions.
- Data Warehouses are designed to perform well enormous amounts of data.
- The structure of data warehouses is more accessible for end-users to navigate, understand, and query.
- Queries that would be complex in many normalized databases could be easier to build and maintain in data warehouses.
- Data warehousing is an efficient method to manage demand for lots of information from lots of users.
- Data warehousing provide the capabilities to analyze a large amount of historical data.

 # Data Warehouse Modelling - Data Cube and OLAP :-

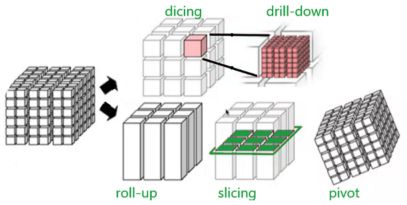
Data Cube :-

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.

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:

- 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.

What is OLAP?

Online Analytical Processing (OLAP) is a category of software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view.

Analysts frequently need to group, aggregate and join data. These OLAP operations in data mining are resource intensive. With OLAP data can be pre-calculated and pre-aggregated, making analysis faster.

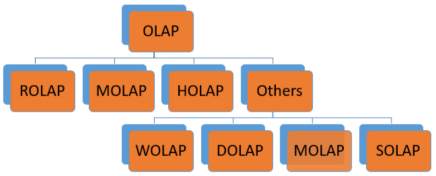
OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy. OLAP stands for Online Analytical Processing.

At the core of the OLAP concept, is an OLAP Cube. The OLAP cube is a data structure optimized for very quick data analysis.

The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the hypercube.

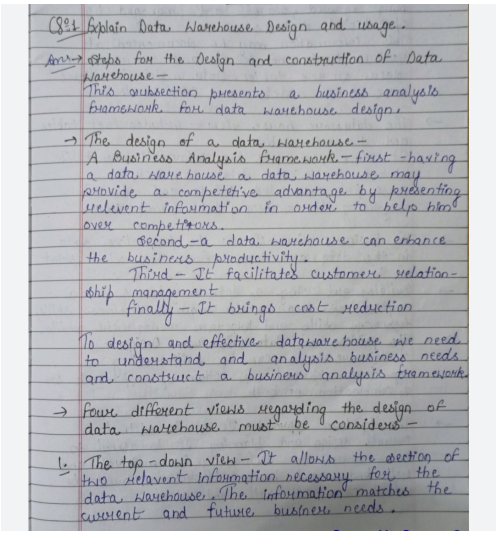
Types of OLAP systems

OLAP Hierarchical Structure



Types of OLAP Systems

Type of OLAP	Explanation
Relational OLAP(ROLAP):	ROLAP is an extended RDBMS along with multidimensional data mapping to perform the standard relational operation.
Multidimensional OLAP (MOLAP)	MOLAP Implementes operation in multidimensional data.
Web OLAP (WOLAP)	Web OLAP which is OLAP system accessible via the web browser. WOLAP is a three-tiered architecture. It consists of three components: client, middleware, and a database server.
Mobile OLAP:	Mobile OLAP helps users to access and analyze OLAP data using their mobile devices
Spatial OLAP :	SOLAP is created to facilitate management of both spatial and non-spatial data in a Geographic Information system (GIS)
Hybrid OnlineAnalytical Processing (HOLAP)	In HOLAP approach the aggregated totals are stored in a multidimensional database while the detailed data is stored in the relational database. This offers both data efficiency of the ROLAP model and the performance of the MOLAP model.
Desktop OLAP (DOLAP)	In Desktop OLAP, a user downloads a part of the data from the database locally, or on their desktop and analyze it. DOLAP is relatively cheaper to deploy as it offers very few functionalities compares to other OLAP systems.



Q.1 Explain Data Warehouse Design and usage.

Ans → Steps for the Design and construction of Data Warehouse –

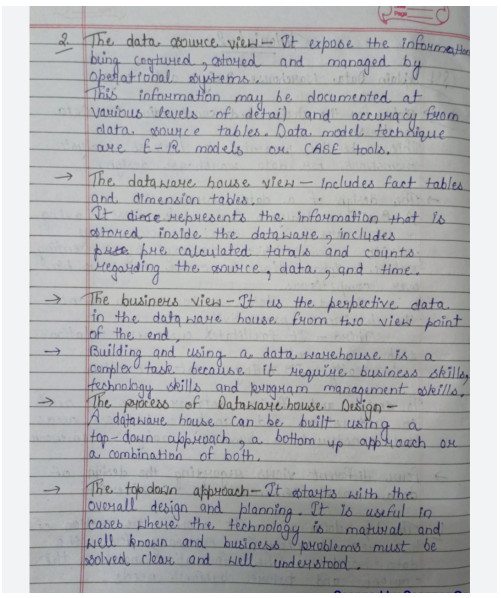
This subsection presents a business analysis framework for data warehouse design.

→ The design of a data warehouse –
A Business Analysis Framework – First - having a data warehouse a data warehouse may provide a competitive advantage by presenting relevent information in order to help him over competitors.
Second – a data warehouse can enhance the business productivity.
Third – It facilitates customer relationship management
finally – It brings cost reduction

To design and effective datawarehouse we need to understand and analysis business needs and construct a business analysis framework

→ Four different views regarding the design of data warehouse must be considers –

1. The top-down view – It allows the section of the relavent information necessary for the data warehouse. The information matches the current and future business needs.



2. The data source view – It expose the information being captured, stored and managed by operational systems.
This information may be documented at various levels of detail and accuracy from data source tables. Data model technique are E-R models or CASE tools.

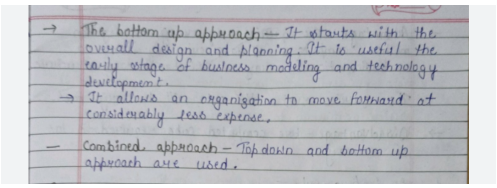
→ The datawarehouse view – Includes fact tables and dimension tables.
It also represents the information that is stored inside the datawarehouse, includes pre-calculated totals and counts regarding the source, data, and time.

→ The business view – It is the perspective data in the datawarehouse from the view point of the end.

→ Building and using a data warehouse is a complex task because it require business skills, technology skills and program management skills.

→ The process of Datawarehouse Design –
A datawarehouse can be built using a top-down approach, a bottom up approach or a combination of both.

→ The top-down approach – It starts with the overall design and planning. It is useful in cases where the technology is matured and well known and business problems must be solved clear and well understood.



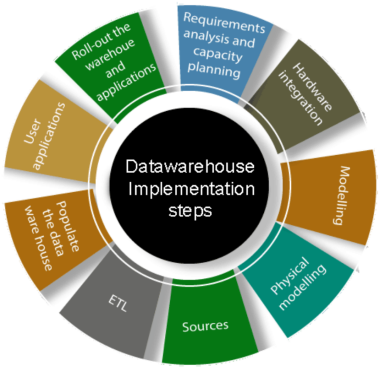
→ The bottom up approach – It starts with the small design and planning. It is useful the early stage of business modeling and technology development.

→ It allows an organization to move forward at considerably less expense.

– Combined approach – Top down and bottom up approach are used.

 # 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.

Data Generalization by Attribute-Oriented Induction :-

Data Generalization is the process of summarizing data by replacing relatively low level values with higher level concepts. It is a form of descriptive data mining.

There are two basic approaches of data generalization :

1. Data cube approach :

It is also known as OLAP approach.
It is an efficient approach as it is helpful to make the past selling graph.
In this approach, computation and results are stored in the Data cube.

2. Attribute oriented induction :

AOI stands for Attribute-Oriented Induction. The attribute-oriented induction approach to concept description was first proposed in 1989, a few years before the introduction of the

data cube approach. The data cube approach is essentially based on materialized views of the data, which typically have been pre-computed in a data warehouse.

It is an online data analysis, query oriented and generalization based approach.
In this approach, we perform generalization on basis of different values of each attributes within the relevant data set. after that same tuple are merged and their respective counts are accumulated in order to perform aggregation.

It performs off-line aggregation before an OLAP or data mining query is submitted for processing.
On the other hand, the attribute oriented induction approach, at least in its initial proposal, a relational database query – oriented, generalized based (on-line data analysis technique).
It is not limited to particular measures nor categorical data.

The generalization is implemented by attribute removal or attribute generalization.
Aggregation is implemented by combining identical generalized tuples and accumulating their specific counts. This decreases the size of the generalized data set. The resulting generalized association can be mapped into several forms for presentation to the user, including charts or rules.

Attribute oriented induction approach uses two method :
(i). Attribute removal.
(ii). Attribute generalization.

Q.2 Explain Data cube computation.

Ans → Data cube computation is needed –

→ To retrieve the information from the data cube in the most efficient way possible.

→ Queries run on the cube will be fast.

→ Cube Materialization (pre computation) –

Different Data cube Materialization include.

1. Full cube
2. Ice berg cube
3. Closed cube
4. Shell cube

→ Full cube –

- The multiway array aggregation method computes full data cube by using a multidimentional array as its basic data structure.

1. Partition array into the chunks.
2. Compute aggregate by visiting (i.e accessing the values at) cube cells.

→ Advantage – The queries way array aggregation method computes full data cube by using a multi-dimentional array as its basic data structure.

The queries run on the cube will be very fast.

→ Disadvantage – pre-computed cube required a lot of memory.

1. An Iceberg-cube –
   - Contain only those cells of the data cube that meet an aggregate condition.
   - It is called an iceberg-cube because it contains only some of the cells of the full cube, like the tip of an iceberg.
   - The purpose of the iceberg-cube is to identify and computer only those values that will most likely be required for decision support queries.
   - The aggregate condition specifies which cube values are more meaningful and shouldbe therefore be stored.
   - This is one solution to the problem of computing versus storing data cubes.

→ Advantage – pre-compute only those cells in the cube which will most likely be used for decision support queries.

→ A closed cube – A closed cube is a data cube consisting of only closed cells.

→ Shell cube – We can choose to precompute only portions or fragments of the cube. Shell

based on cuboids of interest.

→ General strategies for data cube computation –

1. Sorting, hashing and grouping – These operations facilitate aggregation, i.e., computation of the cells that share the same set of dimention values.
   These techniques can also perform –
   → Shared – sorts : sharing sorting costs across multiple cuboids.
   → Share – partitions : sharing partitioning costs across multiple cuboids.
2. Simultaneous aggregation and caching intermediate result –
   Result Reduce expensive disk I/O operations by computing higher-level group bys from computed lower-level group-bys.
   These techniques can also performs
   → Amortized – scans : Computing as many cuboids as possible at the same time to reduced disk reads.
3. Aggregation from the smallest child –
   If a parent cuboid has more than one child, it is efficient to compute it from the smallest previously computed child 'cuboid'.
4. The Apriori pruning method can be explored to compute iceberg cube efficently –
   The Apriori property, in the context of data cubes, states as follow if given cell does not satisfy minimum support, then no descendant (i.e more specialized or detailed version) of the cell will

satisfy minimum support either. ~~this~~ This property can be used to substaintially reduce the computation of iceberg cubes.