SUB-DWM DATE-25/4/2020

UNIT -3 INTRODUCTION TO DATAWAREHOUSING

Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.

The primary objective is data analysis and not data processing.

What is OLTP?

Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.

The primary objective is data processing and not data analysis

Example of OLAP

Any Datawarehouse system is an OLAP system. Uses of OLAP are as follows

- A company might compare their mobile phone sales in September with sales in October, then compare those results with another location which may be stored in a sperate database.
- Amazon analyzes purchases by its customers to come up with a personalized homepage with products which likely interest to their customer.

Example of OLTP system

An example of OLTP system is ATM center. Assume that a couple has a joint account with a bank. One day both simultaneously reach different ATM centers at precisely the same time and want to withdraw total amount present in their bank account.

However, the person that completes authentication process first will be able to get money. In this case, OLTP system makes sure that withdrawn amount will be never more than the amount present in the bank. The key to note here is that OLTP systems are optimized for **transactional superiority instead data analysis.**

Other examples of OLTP system are:

- Online banking
- Online airline ticket booking
- Sending a text message
- Order entry

- Add a book to shopping cart

Benefits of using OLAP services

- OLAP creates a single platform for all type of business analytical needs which includes planning, budgeting, forecasting, and analysis.
- The main benefit of OLAP is the consistency of information and calculations.
- Easily apply security restrictions on users and objects to comply with regulations and protect sensitive data.

Benefits of OLTP method

- It administers daily transactions of an organization.
- OLTP widens the customer base of an organization by simplifying individual processes.

Drawbacks of OLAP service

- Implementation and maintenance are dependent on IT professional because the traditional OLAP tools require a complicated modeling procedure.
- OLAP tools need cooperation between people of various departments to be effective which might always be not possible.

Drawbacks of OLTP method

- If OLTP system faces hardware failures, then online transactions get severely affected.
- OLTP systems allow multiple users to access and change the same data at the same time which many times created unprecedented situation.

Difference between OLTP and OLAP

OLTP vs OLAP

Parameters	OLTP	OLAP
Process	It is an online transactional system. It manages database modification.	OLAP is an online analysis and data retrievir process.
Characteristic	It is characterized by large numbers of short online transactions.	It is characterized by a large volume of data.

Functionality	OLTP is an online database modifying system.	OLAP is an online database que management system.
Method	OLTP uses traditional DBMS.	OLAP uses the data warehouse.
Query	Insert, Update, and Delete information from the database.	Mostly select operations
Table	Tables in OLTP database are normalized.	Tables in OLAP database are **not** normalized
Source	OLTP and its transactions are the sources of data.	Different OLTP databases become the sourc of data for OLAP.
Data Integrity	OLTP database must maintain data integrity constraint.	OLAP database does not get frequent modified. Hence, data integrity is not an issu
Response time	It's response time is in millisecond.	Response time in seconds to minutes.
Data quality	The data in the OLTP database is always detailed and organized.	The data in OLAP process might not b organized.
Usefulness	It helps to control and run fundamental business tasks.	It helps with planning, problem-solving, ar decision support.
Operation	Allow read/write operations.	Only read and rarely write.
Audience	It is a market orientated process.	It is a customer orientated process.
Query Type	Queries in this process are standardized and simple.	Complex queries involving aggregations.

Back-up	Complete backup of the data combined with incremental backups.	OLAP only need a backup from time to tim Backup is not important compared to OLTP
Design	DB design is application oriented. Example: Database design changes with industry like Retail, Airline, Banking, etc.	DB design is subject oriented. Exampl Database design changes with subjects lik sales, marketing, purchasing, etc.
User type	It is used by Data critical users like clerk, DBA & Data Base professionals.	Used by Data knowledge users like worker managers, and CEO.
Purpose	Designed for real time business operations.	Designed for analysis of business measure by category and attributes.
Performance metric	Transaction throughput is the performance metric	Query throughput is the performance metric.
Number of users	This kind of Database users allows thousands of users.	This kind of Database allows only hundreds users.
Productivity	It helps to Increase user's self-service and productivity	Help to Increase productivity of the busines analysts.
Challenge	Data Warehouses historically have been a development project which may prove costly to build.	An OLAP cube is not an open SQL server da warehouse. Therefore, technical knowledg and experience is essential to manage th OLAP server.
Process	It provides fast result for daily used data.	It ensures that response to the query quicker consistently.
Characteristic	It is easy to create and maintain.	It lets the user create a view with the help of

		spreadsheet.
Style	OLTP is designed to have fast response time, low data redundancy and is normalized.	A data warehouse is created uniquely so th it can integrate different data sources f building a consolidated database

KEY DIFFERENCE:

- Online Analytical Processing (OLAP) is a category of software tools that analyze data stored in a database whereas Online transaction processing (OLTP) supports transaction-oriented applications in a 3-tier architecture.
- OLAP creates a single platform for all type of business analysis needs which includes planning, budgeting, forecasting, and analysis while OLTP is useful to administer day to day transactions of an organization.
- OLAP is characterized by a large volume of data while OLTP is characterized by large numbers of short online transactions.

In OLAP, data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database whereas OLTP uses traditional DBMS. OLAP stands for On-Line Analytical Processing. It is used for analysis of database information from multiple database systems at one time such as sales analysis and forecasting, market research, budgeting and etc. Data Warehouse is the example of OLAP system.

OLTP stands for On-Line Transactional processing. It is used for maintaining the online transaction and record integrity in multiple access environments. OLTP is a system that manages very large number of short online transactions for example, ATM.

What is Data Warehousing?

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Using Data Warehouse Information

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be

used in any of the following domains −

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.
- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

Integrating Heterogeneous Databases

To integrate heterogeneous databases, we have two approaches −

- Query-driven Approach
- Update-driven Approach

Query-Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach was used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Process of Query-Driven Approach

- When a query is issued to a client side, a metadata dictionary translates the query into an appropriate form for individual heterogeneous sites involved.
- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

Disadvantages

- Query-driven approach needs complex integration and filtering processes.
- This approach is very inefficient.
- It is very expensive for frequent queries.
- This approach is also very expensive for queries that require aggregations.

Update-Driven Approach

This is an alternative to the traditional approach. Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In

update-driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis.

Advantages

This approach has the following advantages −

- This approach provide high performance.
- The data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.
- Query processing does not require an interface to process data at local sources.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.

Note − Data cleaning and data transformation are important steps in improving the quality of data and data mining results.

Metadata

Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

In terms of data warehouse, we can define metadata as following −

- Metadata is a road-map to data warehouse.
- Metadata in data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Metadata Repository

Metadata repository is an integral part of a data warehouse system. It contains the

following metadata −

- **Business metadata** − It contains the data ownership information, business definition, and changing policies.
- **Operational metadata** − It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** − It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
- **The algorithms for summarization** − It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Data Cube

A data cube helps us represent data in multiple dimensions. It is defined by dimensions and facts. The dimensions are the entities with respect to which an enterprise preserves the records.

Illustration of Data Cube

Suppose a company wants to keep track of sales records with the help of sales data warehouse with respect to time, item, branch, and location. These dimensions allow to keep track of monthly sales and at which branch the items were sold. There is a table associated with each dimension. This table is known as dimension table. For example, "item" dimension table may have attributes such as item_name, item_type, and item_brand.

The following table represents the 2-D view of Sales Data for a company with respect to time, item, and location dimensions.

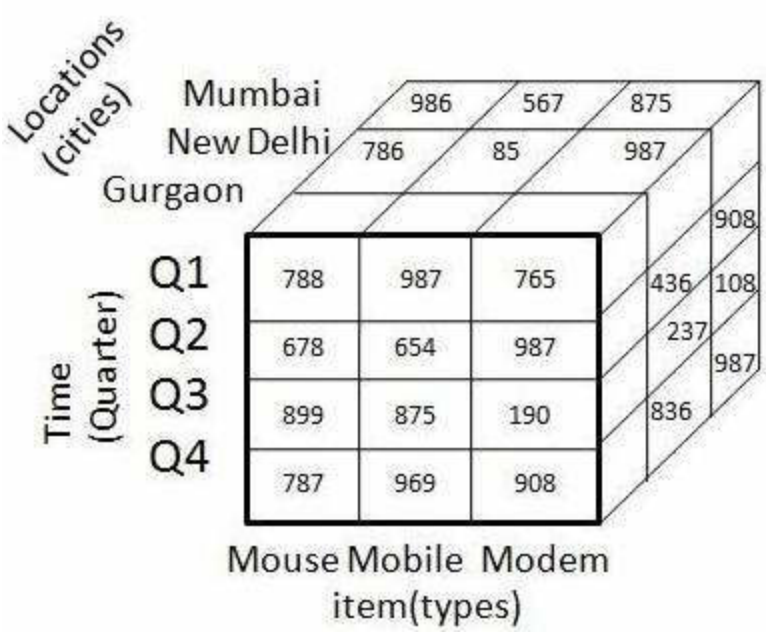
Location="New Delhi"				
Time(quarter)	Item(type)			
	Entertainment	Keyboard	Mobile	Locks
Q1	500	700	10	300
Q2	769	765	30	476
Q3	987	489	18	659
Q4	666	976	40	539

But here in this 2-D table, we have records with respect to time and item only. The

sales for New Delhi are shown with respect to time, and item dimensions according to type of items sold. If we want to view the sales data with one more dimension, say, the location dimension, then the 3-D view would be useful. The 3-D view of the sales data with respect to time, item, and location is shown in the table below –

Time	Location="Gurgaon"			Location="New Delhi"			Location="Mumbai"		
	Item			Item			Item		
	Mouse	Mobile	Modem	Mouse	Mobile	Modem	Mouse	Mobile	Modem
Q1	788	987	765	786	85	987	986	567	875
Q2	678	654	987	659	786	436	980	876	908
Q3	899	875	190	983	909	237	987	100	1089
Q4	787	969	908	537	567	836	837	926	987

The above 3-D table can be represented as 3-D data cube as shown in the following figure –



Data Mart

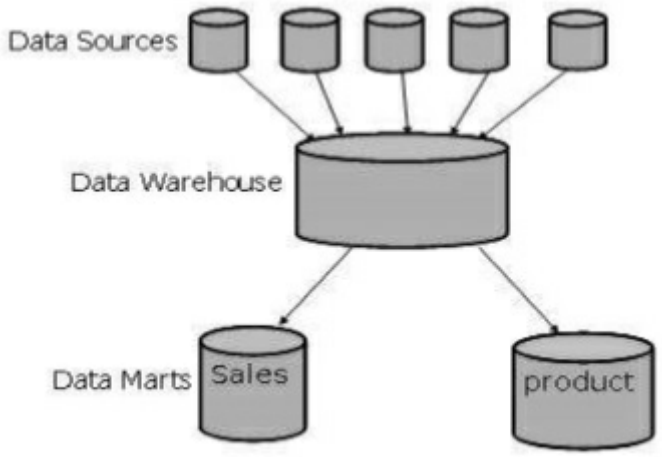
Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization. In other words, a data mart contains only those

data that is specific to a particular group. For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.

Points to Remember About Data Marts

- Windows-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation cycle of a data mart is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of data marts may be complex in the long run, if their planning and design are not organization-wide.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data marts are flexible.

The following figure shows a graphical representation of data marts.



Virtual Warehouse

The view over an operational data warehouse is known as virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database s.

The business analyst get the information from the data warehouses to measure the performance and make critical adjustments in order to win over other business holders in the market. Having a data warehouse offers the following advantages −

- Since a data warehouse can gather information quickly and efficiently, it can enhance business productivity.

- A data warehouse provides us a consistent view of customers and items, hence, it helps us manage customer relationship.
- A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period in a consistent and reliable manner.

To design an effective and efficient data warehouse, we need to understand and analyze the business needs and construct a **business analysis framework**. Each person has different views regarding the design of a data warehouse. These views are as follows −

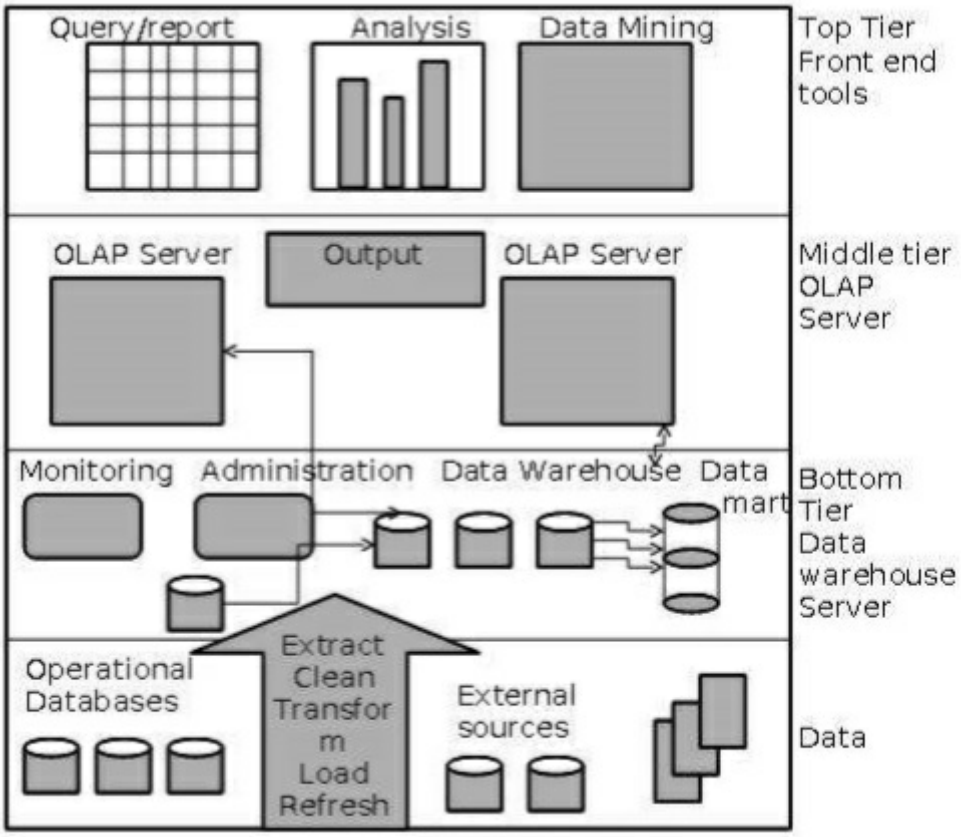
- **The top-down view** − This view allows the selection of relevant information needed for a data warehouse.
- **The data source view** − This view presents the information being captured, stored, and managed by the operational system.
- **The data warehouse view** − This view includes the fact tables and dimension tables. It represents the information stored inside the data warehouse.
- **The business query view** − It is the view of the data from the viewpoint of the end-user.

Three-Tier Data Warehouse Architecture

Generally a data warehouses adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture.

- **Bottom Tier** − The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.
- **Middle Tier** − In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.
    - By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
    - By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.
- **Top-Tier** − This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

The following diagram depicts the three-tier architecture of data warehouse −



Data Warehouse Models

From the perspective of data warehouse architecture, we have the following data warehouse models −

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

Virtual Warehouse

The view over an operational data warehouse is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.

Data Mart

Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

In other words, we can claim that data marts contain data specific to a particular group. For example, the marketing data mart may contain data related to items, customers, and sales. Data marts are confined to subjects.

Points to remember about data marts −

- Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data mart are flexible.

Enterprise Warehouse

- An enterprise warehouse collects all the information and the subjects spanning an entire organization
- It provides us enterprise-wide data integration.
- The data is integrated from operational systems and external information providers.
- This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

Load Manager

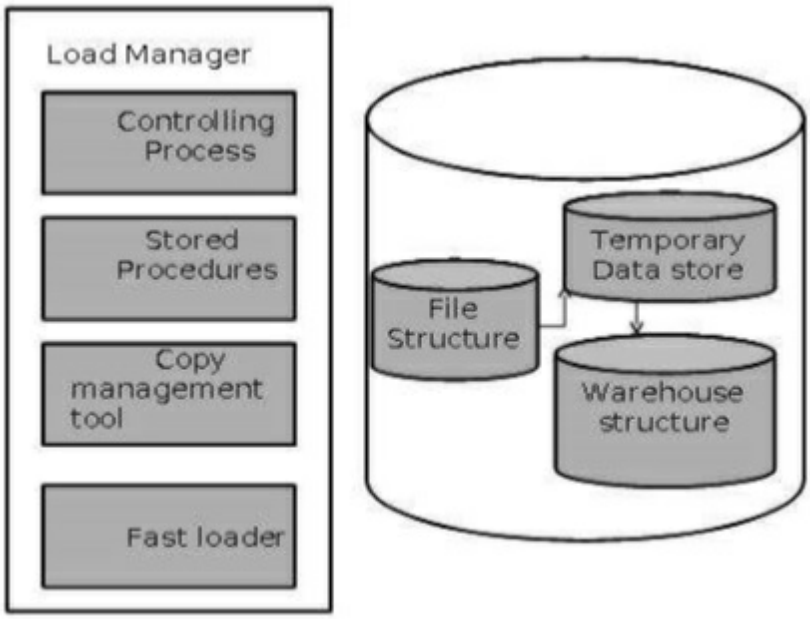
This component performs the operations required to extract and load process.

The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

Load Manager Architecture

The load manager performs the following functions −

- Extract the data from source system.
- Fast Load the extracted data into temporary data store.
- Perform simple transformations into structure similar to the one in the data warehouse.



Extract Data from Source

The data is extracted from the operational databases or the external information providers. Gateways is the application programs that are used to extract data. It is supported by underlying DBMS and allows client program to generate SQL to be executed at a server. Open Database Connection(ODBC), Java Database Connection (JDBC), are examples of gateway.

Fast Load

- In order to minimize the total load window the data need to be loaded into the warehouse in the fastest possible time.
- The transformations affects the speed of data processing.
- It is more effective to load the data into relational database prior to applying transformations and checks.
- Gateway technology proves to be not suitable, since they tend not be performant when large data volumes are involved.

Simple Transformations

While loading it may be required to perform simple transformations. After this has been completed we are in position to do the complex checks. Suppose we are loading the EPOS sales transaction we need to perform the following checks:

- Strip out all the columns that are not required within the warehouse.

- Convert all the values to required data types.

Warehouse Manager

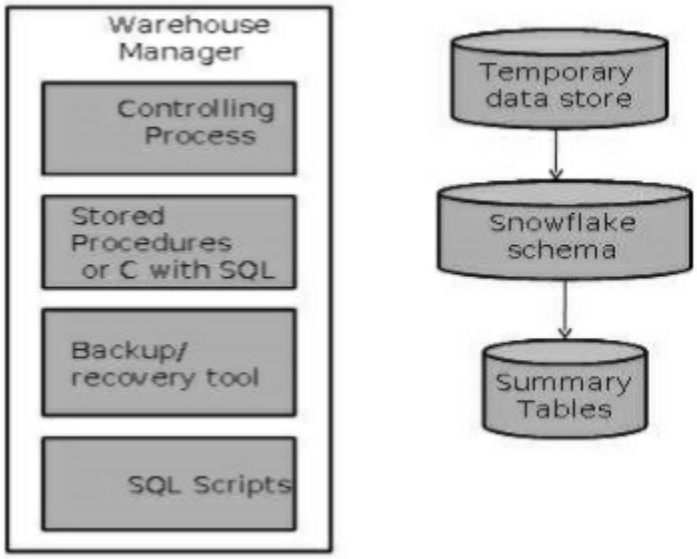
A warehouse manager is responsible for the warehouse management process. It consists of third-party system software, C programs, and shell scripts.

The size and complexity of warehouse managers varies between specific solutions.

Warehouse Manager Architecture

A warehouse manager includes the following −

- The controlling process
- Stored procedures or C with SQL
- Backup/Recovery tool
- SQL Scripts



Operations Performed by Warehouse Manager

- A warehouse manager analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.
- Generates new aggregations and updates existing aggregations. Generates normalizations.
- Transforms and merges the source data into the published data warehouse.

- Backup the data in the data warehouse.
- Archives the data that has reached the end of its captured life.

Note − A warehouse Manager also analyzes query profiles to determine index and aggregations are appropriate.

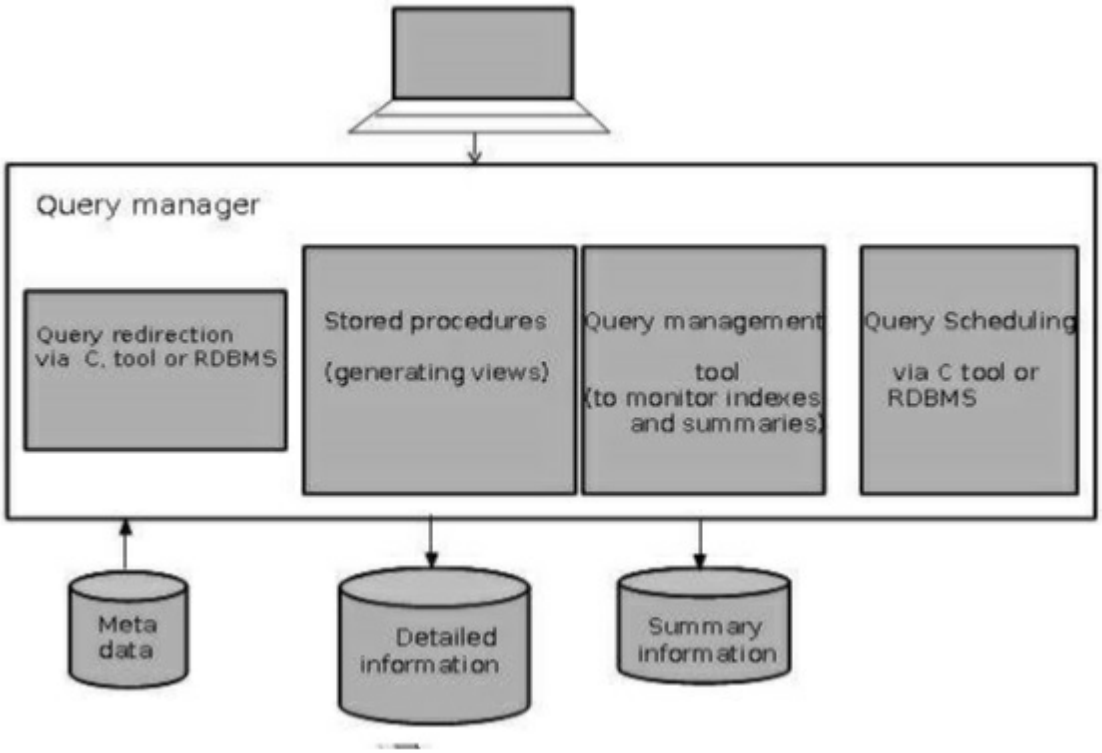
Query Manager

- Query manager is responsible for directing the queries to the suitable tables.
- By directing the queries to appropriate tables, the speed of querying and response generation can be increased.
- Query manager is responsible for scheduling the execution of the queries posed by the user.

Query Manager Architecture

The following screenshot shows the architecture of a query manager. It includes the following:

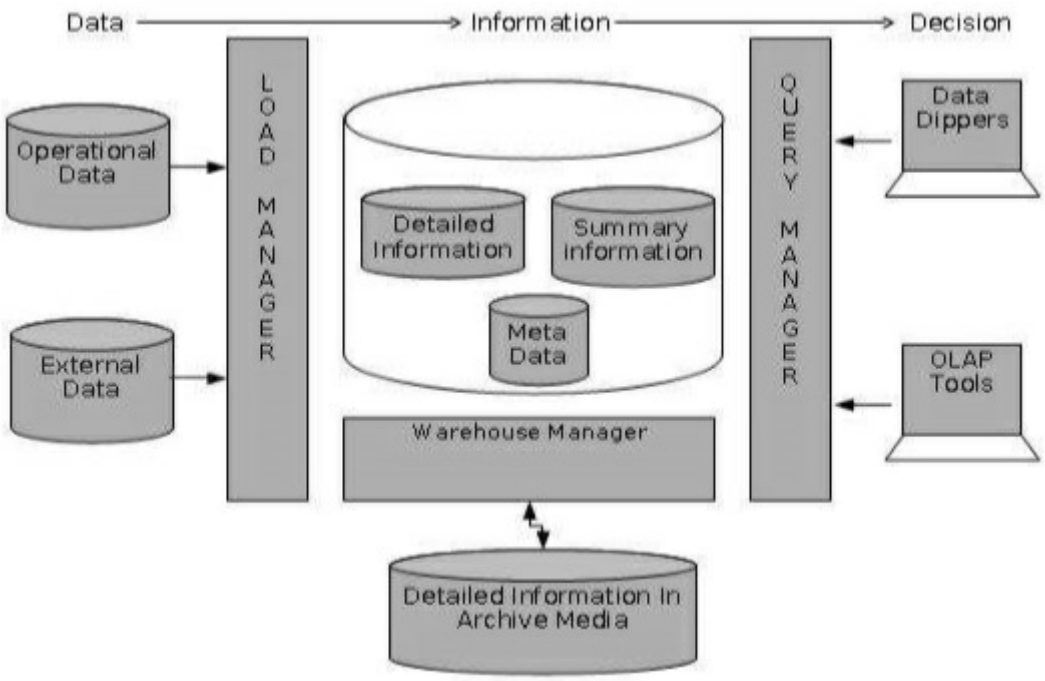
- Query redirection via C tool or RDBMS
- Stored procedures
- Query management tool
- Query scheduling via C tool or RDBMS
- Query scheduling via third-party software



Detailed Information

Detailed information is not kept online, rather it is aggregated to the next level of detail and then archived to tape. The detailed information part of data warehouse keeps the detailed information in the starflake schema. Detailed information is loaded into the data warehouse to supplement the aggregated data.

The following diagram shows a pictorial impression of where detailed information is stored and how it is used.



Note – If detailed information is held offline to minimize disk storage, we should make sure that the data has been extracted, cleaned up, and transformed into starflake schema before it is archived.

Summary Information

Summary Information is a part of data warehouse that stores predefined aggregations. These aggregations are generated by the warehouse manager. Summary Information must be treated as transient. It changes on-the-go in order to respond to the changing query profiles.

The points to note about summary information are as follows –

- Summary information speeds up the performance of common queries.
- It increases the operational cost.
- It needs to be updated whenever new data is loaded into the data warehouse.
- It may not have been backed up, since it can be generated fresh from the detailed information.

Why Do We Need a Data Mart?

Listed below are the reasons to create a data mart –

- To partition data in order to impose **access control strategies.**
- To speed up the queries by reducing the volume of data to be scanned.

- To segment data into different hardware platforms.
- To structure data in a form suitable for a user access tool.

Note − Do not data mart for any other reason since the operation cost of data marting could be very high. Before data marting, make sure that data marting strategy is appropriate for your particular solution.

Cost-effective Data Marting

Follow the steps given below to make data marting cost-effective −

- Identify the Functional Splits
- Identify User Access Tool Requirements
- Identify Access Control Issues

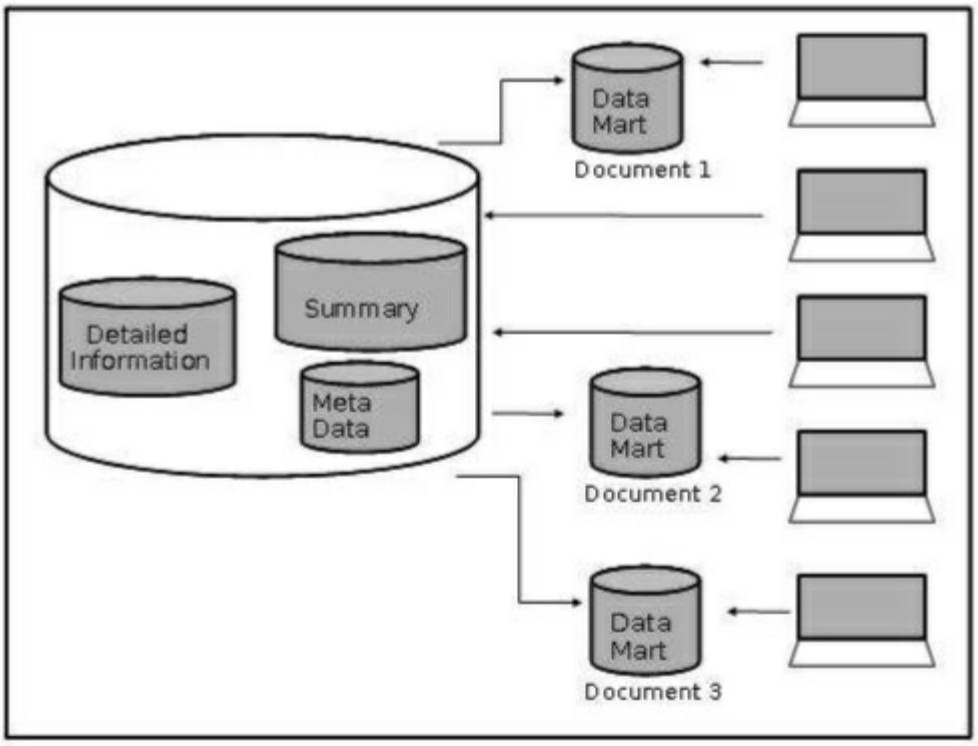
Identify the Functional Splits

In this step, we determine if the organization has natural functional splits. We look for departmental splits, and we determine whether the way in which departments use information tend to be in isolation from the rest of the organization. Let's have an example.

Consider a retail organization, where each merchant is accountable for maximizing the sales of a group of products. For this, the following are the valuable information −

- sales transaction on a daily basis
- sales forecast on a weekly basis
- stock position on a daily basis
- stock movements on a daily basis

As the merchant is not interested in the products they are not dealing with, the data marting is a subset of the data dealing which the product group of interest. The following diagram shows data marting for different users.



Given below are the issues to be taken into account while determining the functional split −

- The structure of the department may change.
- The products might switch from one department to other.
- The merchant could query the sales trend of other products to analyze what is happening to the sales.

Note − We need to determine the business benefits and technical feasibility of using a data mart.

Identify User Access Tool Requirements

We need data marts to support **user access tools** that require internal data structures. The data in such structures are outside the control of data warehouse but need to be populated and updated on a regular basis.

There are some tools that populate directly from the source system but some cannot. Therefore additional requirements outside the scope of the tool are needed to be identified for future.

Note − In order to ensure consistency of data across all access tools, the data should not be directly populated from the data warehouse, rather each tool must have its own data mart.

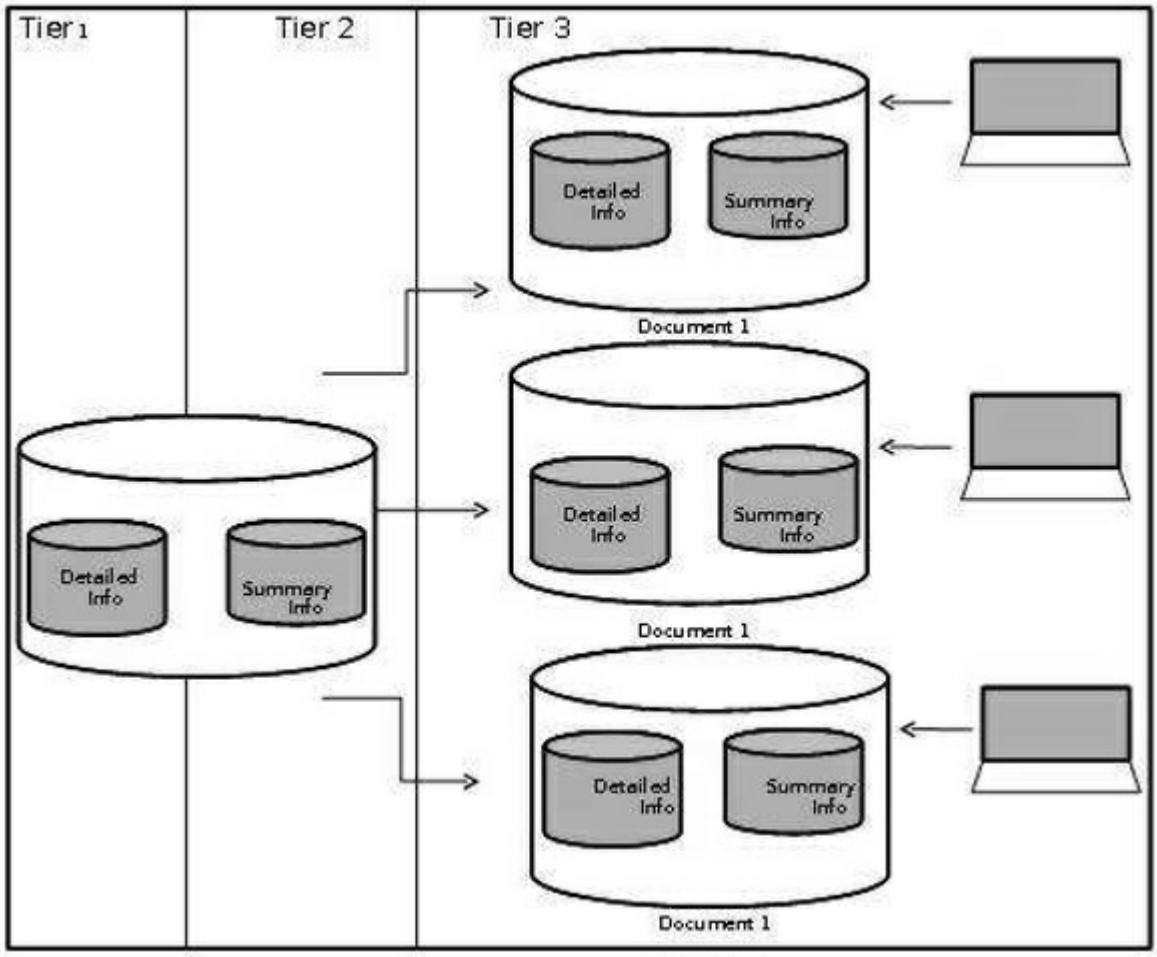
Identify Access Control Issues

There should to be privacy rules to ensure the data is accessed by authorized users only. For example a data warehouse for retail banking institution ensures that all the accounts belong to the same legal entity. Privacy laws can force you to totally prevent access to information that is not owned by the specific bank.

Data marts allow us to build a complete wall by physically separating data segments within the data warehouse. To avoid possible privacy problems, the detailed data can be removed from the data warehouse. We can create data mart for each legal entity and load it via data warehouse, with detailed account data.

Designing Data Marts

Data marts should be designed as a smaller version of starflake schema within the data warehouse and should match with the database design of the data warehouse. It helps in maintaining control over database instances.



The summaries are data marted in the same way as they would have been designed within the data warehouse. Summary tables help to utilize all dimension data in the starflake schema.

Cost of Data Marting

The cost measures for data marting are as follows −

- Hardware and Software Cost
- Network Access
- Time Window Constraints

Hardware and Software Cost

Although data marts are created on the same hardware, they require some additional hardware and software. To handle user queries, it requires additional processing power and disk storage. If detailed data and the data mart exist within the data warehouse, then we would face additional cost to store and manage replicated data.

Note − Data marting is more expensive than aggregations, therefore it should be used as an additional strategy and not as an alternative strategy.

Network Access

A data mart could be on a different location from the data warehouse, so we should ensure that the LAN or WAN has the capacity to handle the data volumes being transferred within the data mart load process.

Time Window Constraints

The extent to which a data mart loading process will eat into the available time window depends on the complexity of the transformations and the data volumes being shipped. The determination of how many data marts are possible depends on −

- Network capacity.
- Time window available
- Volume of data being transferred
- Mechanisms being used to insert data into a data mart

Short question answers

Q: Define data warehouse?

A : Data warehouse is a subject oriented, integrated, time-variant, and nonvolatile collection of data that supports management's decision-making process.

Q: What does subject-oriented data warehouse signify?

A : Subject oriented signifies that the data warehouse stores the information around a particular subject such as product, customer, sales, etc.

Q: List any five applications of data warehouse.

A : Some applications include financial services, banking services, customer goods, retail sectors, controlled manufacturing.

Q: What do OLAP and OLTP stand for?

A : OLAP is an acronym for **Online Analytical Processing** and OLTP is an acronym of Online Transactional Processing.

Q: What is the very basic difference between data warehouse and operational databases?

A : A data warehouse contains historical information that is made available for analysis of the business whereas an operational database contains current information that is required to run the business.

Q: List the Schema that a data warehouse system can implements.

A : A data Warehouse can implement star schema, snowflake schema, and fact constellation schema.

Q: What is Data Warehousing?

A : Data Warehousing is the process of constructing and using the data warehouse.

Q: List the process that are involved in Data Warehousing.

A : Data Warehousing involves data cleaning, data integration and data consolidations.

Q: List the functions of data warehouse tools and utilities.

A : The functions performed by Data warehouse tool and utilities are Data Extraction, Data Cleaning, Data Transformation, Data Loading and Refreshing.

Q: What do you mean by Data Extraction?

A : Data extraction means gathering data from multiple heterogeneous sources.

Q: Define metadata?

A : Metadata is simply defined as data about data. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

Q: What does Metadata Respiratory contain?

A : Metadata respiratory contains definition of data warehouse, business metadata, operational metadata, data for mapping from operational environment to data warehouse, and the algorithms for summarization.

Q: How does a Data Cube help?

A : Data cube helps us to represent the data in multiple dimensions. The data cube is defined by dimensions and facts.

Q: Define dimension?

A : The dimensions are the entities with respect to which an enterprise keeps the records.

Q: Explain data mart.

A : Data mart contains the subset of organization-wide data. This subset of data is valuable to specific groups of an organization. In other words, we can say that a data mart contains data specific to a particular group.

Q: What is Virtual Warehouse?

A : The view over an operational data warehouse is known as virtual warehouse.

Q: List the phases involved in the data warehouse delivery process.

A : The stages are IT strategy, Education, Business Case Analysis, technical Blueprint, Build the version, History Load, Ad hoc query, Requirement Evolution, Automation, and Extending Scope.

Q: Define load manager.

A : A load manager performs the operations required to extract and load the process. The size and complexity of load manager varies between specific solutions from data warehouse to data warehouse.

Q: Define the functions of a load manager.

A : A load manager extracts data from the source system. Fast load the extracted data into temporary data store. Perform simple transformations into structure similar to the one in the data warehouse.

Q: Define a warehouse manager.

A : Warehouse manager is responsible for the warehouse management process. The warehouse manager consist of third party system software, C programs and shell scripts. The size and complexity of warehouse manager varies between specific solutions.

Q: Define the functions of a warehouse manager.

A : The warehouse manager performs consistency and referential integrity checks, creates the indexes, business views, partition views against the base data, transforms and merge the source data into the temporary store into the published data warehouse, backs up the data in the data warehouse, and archives the data that has reached the end of its captured life.

Q: What is Summary Information?

A : Summary Information is the area in data warehouse where the predefined aggregations are kept.

Q: What does the Query Manager responsible for?

A : Query Manager is responsible for directing the queries to the suitable tables.

Q: List the types of OLAP server

A : There are four types of OLAP servers, namely Relational OLAP, Multidimensional OLAP, Hybrid OLAP, and Specialized SQL Servers.

Q: Which one is faster, Multidimensional OLAP or Relational OLAP?

A : Multidimensional OLAP is faster than Relational OLAP.

Q: List the functions performed by OLAP.

A : OLAP performs functions such as roll-up, drill-down, slice, dice, and pivot.

Q: How many dimensions are selected in Slice operation?

A : Only one dimension is selected for the slice operation.

Q: How many dimensions are selected in dice operation?

A : For dice operation two or more dimensions are selected for a given cube.

Q: How many fact tables are there in a star schema?

A : There is only one fact table in a star Schema.

Q: What is Normalization?

A : Normalization splits up the data into additional tables.

Q: Out of star schema and snowflake schema, whose dimension table is normalized?

A : Snowflake schema uses the concept of normalization.

Q: What is the benefit of normalization?

A : Normalization helps in reducing data redundancy.

Q: Which language is used for defining Schema Definition?

A : Data Mining Query Language (DMQL) is used for Schema Definition.

Q: What language is the base of DMQL?

A : DMQL is based on Structured Query Language (SQL).

Q: What are the reasons for partitioning?

A : Partitioning is done for various reasons such as easy management, to assist backup recovery, to enhance performance.

Q: What kind of costs are involved in Data Marting?

A : Data Marting involves hardware & software cost, network access cost, and time cost.

Data Warehousing Interview Questions and Answers

Q1). How will you define the concept of Data Warehousing?

A data warehouse is the data repository that is used for the decision support system. A data warehouse is made up of a wide variety of data that has a high level of business conditions at a particular point of time. In simple words, this is a repository of integrated information that is available for queries and analysis.

Q2). Define the concept of Business Intelligence.

Business Intelligence is also named Decision Support Systems that refers to technologies, applications, and practices for the collection, integration, and analysis of business-related information or data.

Q3). What are the dimension tables?

A dimension table contains attributes of measurements stored in fact tables. This table is made up of hierarchies, nodes, and categories that can be used to traverse in nodes.

Q4). Define the meaning of Fact table.

A Fact Table contains the measurements of business processes and it will contain the foreign keys for dimension tables.

Q5). Define the meaning of Data mining.

Data mining is a process for analyzing data from different perspectives or dimensions and summarizing the same into meaningful content. Data can be retrieved or queried from the database in their own format.

Q6). Explain the different stages of data warehousing.

- Offline Operational Database
- Offline Data Warehouse
- Real-Time Datawarehouse
- Integrated Datawarehouse

Q7). How will you define the OLAP?

OLAP or Online Analytical Processing is set to be a system that collects, processes, and manages multi-dimensional data for analysis purposes.

Q8). How can you define the OLTP concept?

Read: How to work with Deep Learning on Keras?
OLTP or online transaction processing is an application that modifies the data whenever received from a large number of users.

Q9). How can you differentiate OLTP and OLAP from each other? ***OLTP vs OLAP***

OLTP	OLAP

Data is extracted from a single original source.	Data is extracted from multiple sources.
Simple queries are made by users.	Complex queries are generated by the system.
Normalized small database.	De-normalized large database.
There are fundamental business tasks.	There are multi-dimensional business tasks.

Q10). What is the full form and meaning of ODS?

ODS or Operational Data Source is a repository of real-time operational data instead of long-term trend data.

Q11). How will you differentiate the View and Materialized View?

A View is a virtual table that takes outputs from queries and can be used in place of tables. A materialized view is indirect access to table data by storing the result of a query into a separate schema.

Q12). How will you define the ETL?

ETL means Extract, Transform, and Load. This is a software application that can read the data from a particular data source and extracts the needed subset of data. In the next step, it will transform the data using lookup tables or rules and convert it to the desired state. In the end, load function is used to load the resulting data from the target database.