Name: Aftab Khan ID: 12985 Subject: Data Warehouse

Answer 1: Explain following terms:

a) **Data mart:** A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales or Finance or Marketing. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.

Types of Data mart:

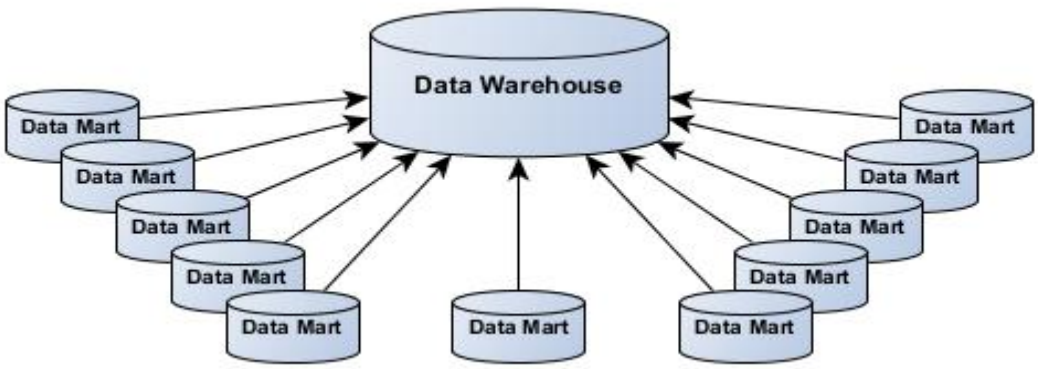
There are three main types of data marts:

Dependent: Dependent data marts are created by drawing data directly from operational, external or both sources.

Independent: Independent data mart is created without the use of a central data warehouse.

Hybrid: This type of data marts can take data from data warehouses or operational systems.

Diagram:



b) **ETL:** ETL stands for “extract, transform, and load.”

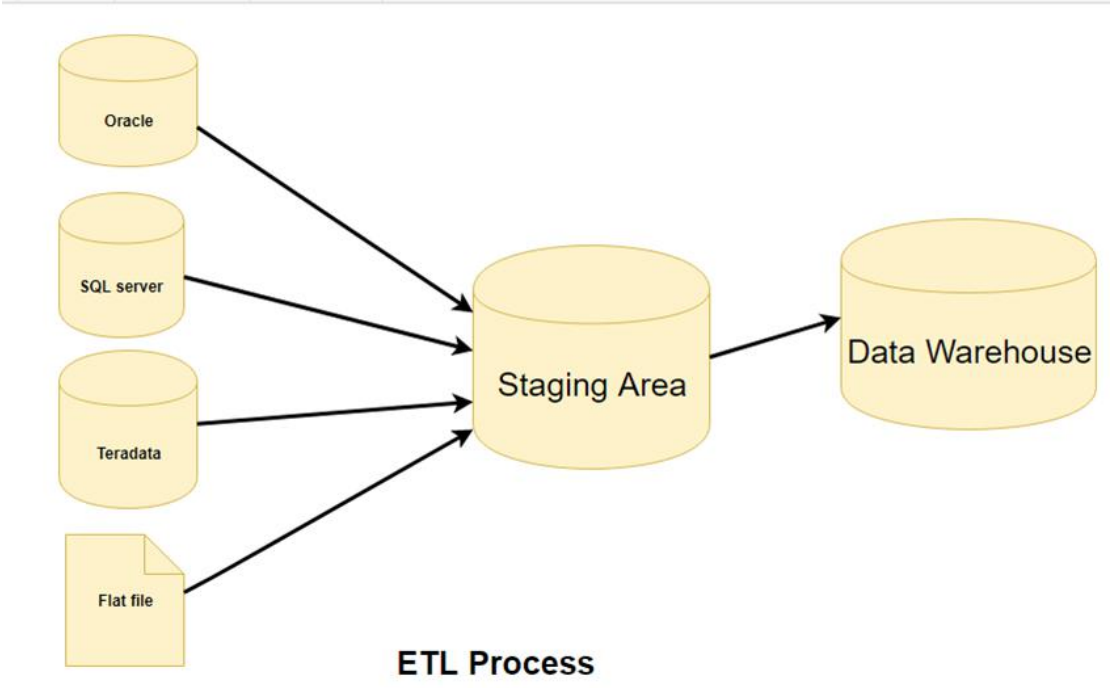
The process of ETL plays a key role in data integration strategies. ETL allows businesses to gather data from multiple sources and consolidate it into a single, centralized location. ETL also makes it possible for different types of data to work together.

Step 1) Extraction:
In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

Step 2) Transformation:
Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.
In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as **direct move or pass through data.**

Step 3) Loading:
Loading data into the target data warehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.
In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

ETL Diagram:



Answer 2:

OLTP:
An OLTP system captures and maintains transaction data in a database. Each transaction involves individual database records made up of multiple fields or columns. Examples include banking and credit card activity or retail checkout scanning.

In OLTP, the emphasis is on fast processing, because OLTP databases are read, written, and updated frequently. If a transaction fails, built-in system logic ensures data integrity.

OLAP:
OLAP applies complex queries to large amounts of historical data, aggregated from OLTP databases and other sources, for data mining, analytics, and **business intelligence** projects. In OLAP, the emphasis is on response time to these complex queries. Each query involves one or more columns of data aggregated from many rows. Examples include year-over-year financial performance or marketing lead generation trends.

OLAP databases and **data warehouses** give analysts and decision-makers the ability to use custom reporting tools to turn data into information. Query failure in OLAP does not interrupt or delay transaction processing for customers, but it can delay or impact the accuracy of business intelligence insights.

Comparison:

	OLTP	OLAP
Characteristics	Handles a large number of small transactions	Handles large volumes of data with complex queries
Query types	Simple standardized queries	Complex queries
Operations	Based on INSERT, UPDATE, DELETE commands	Based on SELECT commands to aggregate data for reporting
Response time	Milliseconds	Seconds, minutes, or hours depending on the amount of data to process
Design	Industry-specific, such as retail, manufacturing, or banking	Subject-specific, such as sales, inventory, or marketing

Source	Transactions	Aggregated data from transactions
Purpose	Control and run essential business operations in real time	Plan, solve problems, support decisions, discover hidden insights
Data updates	Short, fast updates initiated by user	Data periodically refreshed with scheduled, long-running batch jobs
Space requirements	Generally small if historical data is archived	Generally large due to aggregating large datasets
Backup and recovery	Regular backups required to ensure business continuity and meet legal and governance requirements	Lost data can be reloaded from OLTP database as needed in lieu of regular backups
Productivity	Increases productivity of end users	Increases productivity of business managers, data analysts, and executives
Data view	Lists day-to-day business transactions	Multi-dimensiona l view of enterprise data

User examples	Customer-facing personnel, clerks, online shoppers	Knowledge workers such as data analysts, business analysts, and executives
Database design	Normalized databases for efficiency	Denormalized databases for analysis

Answer 3: Difference between traditional and active data ware house.

Traditional data warehouse:
Traditional data warehouses are aimed at providing decision making support to business executives for strategic purposes

Active data warehouse:
Active Data Warehouse (ADW) is a data warehouse designed to provide real time or near-real time operational decision support

Traditional Data Warehousing Environment	Active Data Warehousing Environment
Strategic decisions only	Strategic and tactical decisions
Results sometimes hard to measure	Results measured with operations
Daily, weekly, monthly data currency is acceptable; summaries are often appropriate	Only comprehensive detailed data available within minutes is acceptable
Moderate user concurrency	High number (1,000 or more) of users accessing and querying the system simultaneously
Highly restrictive reporting used to confirm or check existing processes and patterns; often uses predeveloped summary tables or data marts	Flexible ad hoc reporting as well as machine-assisted modeling (e.g., data mining) to discover new hypotheses and relationships
Power users, knowledge workers, internal users	Operational staffs, call centers, external users

Answer 4:

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.

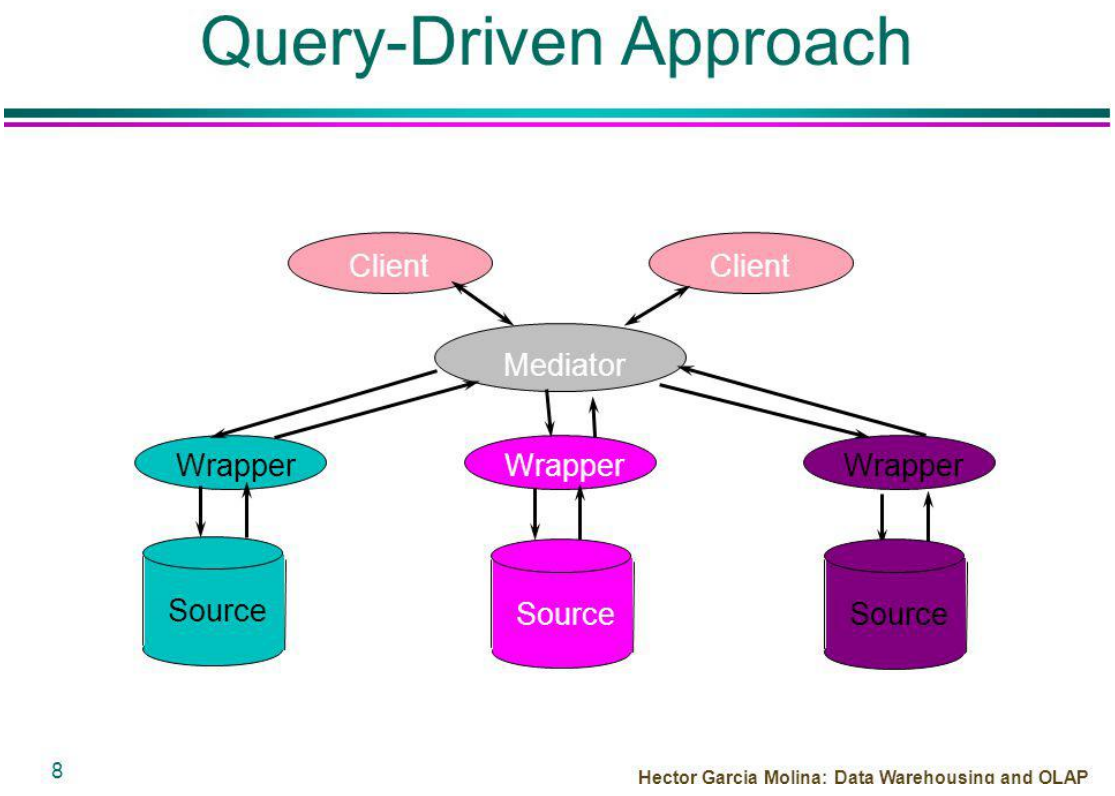
Disadvantages:
Query-driven approach needs complex integration and filtering processes.

This approach is very inefficient.

It is very expensive for frequent queries.

This approach is also very expensive for queries that require aggregations.

Diagram:



Answer 5:

limitations of data warehouse:

Maintenance costs outweigh the benefits:
Data warehouses for a huge IT project would involve high maintenance systems which may affect the revenue for medium scale organizations. The cost to benefit ratio is on the lower side as it not only involves systems with equipped technology but also longer hours as an investment from the IT department. This would restrict the organization's growth especially when it's a business which is adapting to its market conditions.

Data Ownership:
An important concern of Data warehouses is the security of data. Primarily, Data warehouses are marked for software applications for service. This restricts your data security as the

data which has been implemented locally might be sensitive only for a certain department.
Leaking of data within the same organization could lead to hiatus and cause problems for the executives. You can avoid this by ensuring that the individuals entrusted with the analysis are trusted employees of the company with no departmental lineage as it could lead to reluctance because of data censorship.

Data Rigidity:
The type of data imported into a data warehouse is often static data sets which have the least flexibility to generate specific solutions. For the data to be used, it has to be transformed and cleansed which could take several days or weeks.
Moreover, warehouses are subjected to ad hoc queries which are extremely difficult as they have least processing speed and query speed. Even though the queries are restricted to the data marts used during consolidation and integration, most of them are ad hoc queries.

Underestimation of ETL processing time:
Often organizations do not estimate the time required for the ETL process and find their work interrupted leading to backlogs. A significant portion of the time required for the entire process of data warehouse development is for extraction, cleaning, and loading of consolidated data into the warehouse. Even with tools to make the process faster, efficient transformation takes up to several days or weeks.

Hidden problems of the Source:
Hidden problems of the source arise when an organization finds themselves with problems related to the original source systems which were involved in the importing of data into the warehouse after several years of operation.
Practically, a human error while entering data like property details, like for example, leaving certain fields incomplete or improperly filled could be considered as void property data.
Also See: Data Warehouse Applications

Inability to capture required data:
There is always the probability that the data which was required for analysis by the organization was not integrated into the warehouse leading to loss of information. Consider the example of property registration, apart from the regular details, the date of registration plays an important role in statistical analysis at the end of the month. However, such data could be overlooked and not imported.

Increased demands of the users:
After success with the initial few queries, users of the facility may ask more complicated queries which would increase the workload on the system and server. With awareness of the features of the data warehouse, there might also be an increase in the number of queries posed by the staff which also increase the server load.
However, if your organization's systems and servers are equipped with high-end hardware, this wouldn't be a problem.

Long-duration project:
A comprehensive warehouse project might take up to three years to complete. Not all organizations are able to dedicate themselves entirely and are hence more reluctant in investing in a data warehouse. If the organization has to offer very low historical data, a majority of the data mart features will not be utilized as much and will only be a limitation.

Complications:
The integration feature is one of the most important aspects of a data warehouse. Which is why it is recommended that an organization pays special attention to the disparate and equally compelling data warehousing tools and their results to arrive at a proper business conclusion and make their decision. This could be a challenging task if the organization's management is not dedicated and lack experience. So it was all
about Disadvantages and Limitations of a Data Warehouse.