Module 1

Differences between MOLAP, ROLAP, and HOLAP

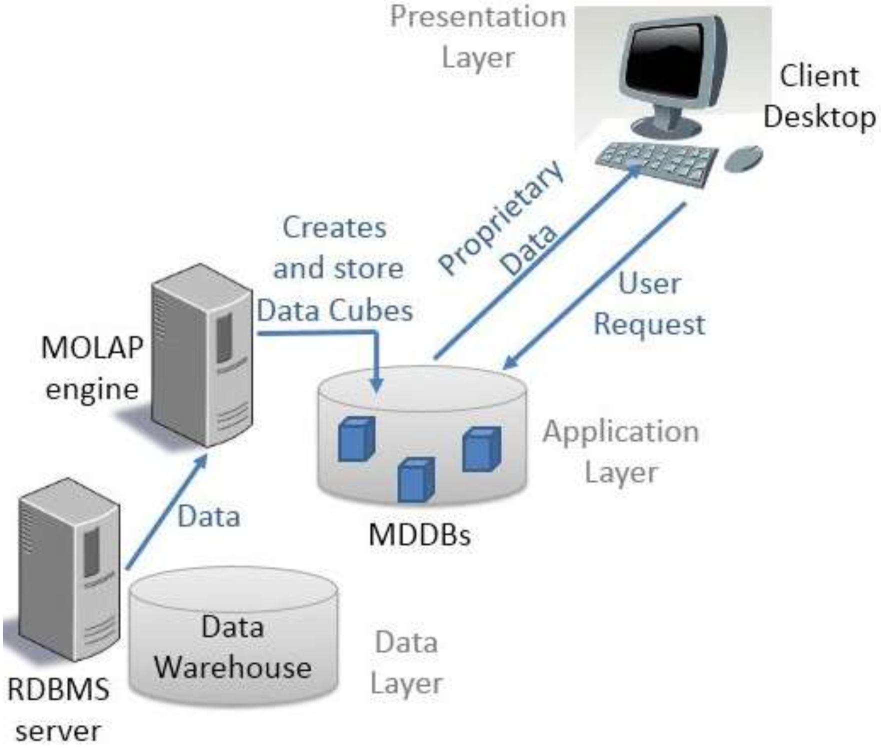
MOLAP

MOLAP is an abbreviation for Multi-dimensional Online Analytical Processing. In this type of analytical processing, multi-dimensional databases (MDDBs) are used to store data. This data is later used for analysis. MOLAP consists of data that is pre-computed and fabricated. The data cubes from MDDBs carry data that has already been calculated. This increases the speed of querying data.

The architecture of MOLAP consists of three main components:

- **Database server:** This exists in the data layer.
- **MOLAP server:** This consists of the MOLAP engine in the application layer.
- **Front-end tool:** This is usually the client desktop in the presentation layer.

The MOLAP engine in the application layer collects data from the databases in the data layer. It then loads data cubes into the multi-dimensional databases. When the user makes a query, data will move in a propriety format from the MDDBs to the client desktop in the presentation layer. This enables users to view data in multiple dimensions.



MOLAP Model

Advantages

- It performs well with operations such as slice and dice.
- Users can use it to perform complex calculations.
- It consists of pre-computed data that can be indexed fast.

Disadvantages

- It can only store a limited volume of data.
- The data used for analysis depends on certain requirements that were set (previously). This limits data analysis and navigation.

ROLAP

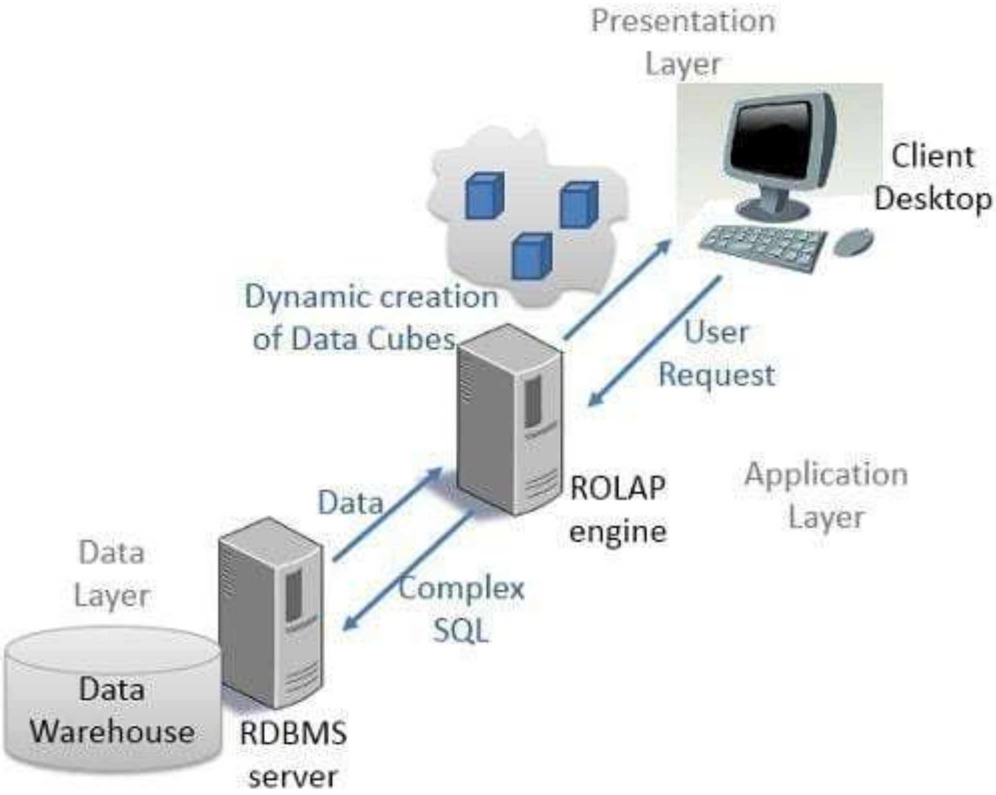
ROLAP is an abbreviation for Relational Online Analytical Processing. In this type of analytical processing, data storage is done in a relational database. In this database, the arrangement of data is made in rows and columns. Data is presented to end-users in a multi-dimensional form.

There are three main components in a ROLAP model:

1. **Database server:** This exists in the data layer. This consists of data that is loaded into the ROLAP server.
2. **ROLAP server:** This consists of the ROLAP engine that exists in the application layer.
3. **Front-end tool:** This is the client desktop that exists in the presentation layer.

Let's briefly look at how ROLAP works. When a user makes a query (complex), the ROLAP server will fetch data from the RDBMS server. The ROLAP engine will then create data cubes dynamically. The user will view data from a multi-dimensional point.

Unlike in MOLAP, where the multi-dimensional view is static, ROLAP provides a dynamic multi-dimensional view. This explains why it is slower when compared to MOLAP.



ROLAP Model

Advantages

- It can handle huge volumes of data.
- A ROLAP model can store data efficiently.
- ROLAP utilizes a relational database. This enables the model to integrate the ROLAP server with an RDBMS (relational database management system).

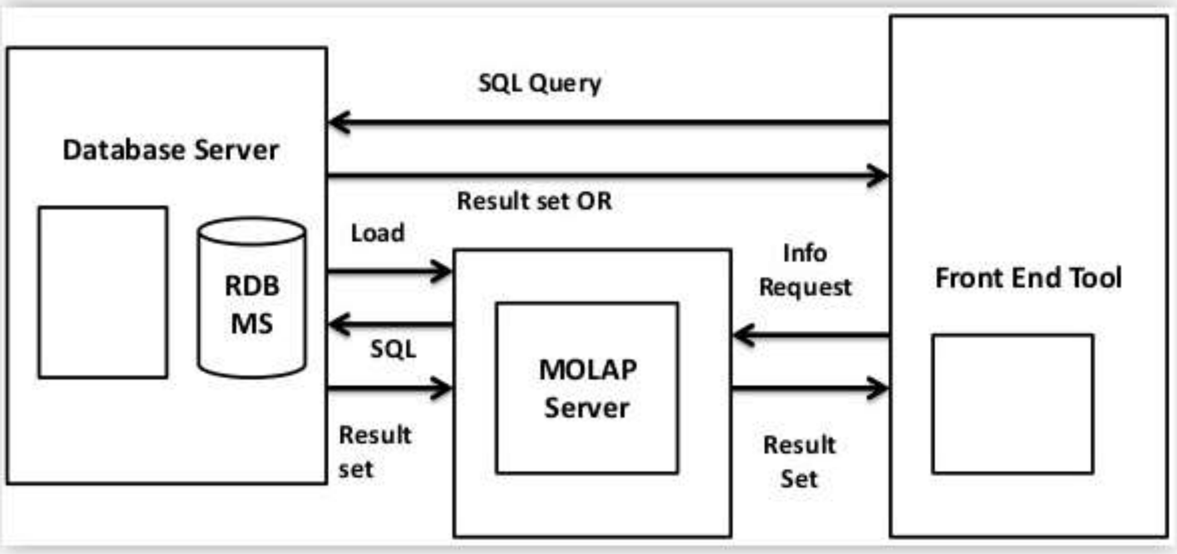
Disadvantages

- There is slow performance, especially when the volume of data is huge.
- ROLAP has certain limitations relating to SQL. For example, the SQL feature has difficulties in handling complex calculations.

HOLAP

This is an abbreviation for Hybrid Online Analytical Processing. This type of analytical processing solves the limitations of MOLAP and ROLAP and combines their attributes. Data in the database is divided into two parts: specialized storage and relational storage. Integrating these two aspects addresses issues relating to performance and scalability. HOLAP stores huge volumes of data in a relational database and keeps aggregations in a MOLAP server.

The HOLAP model consists of a server that can support ROLAP and MOLAP. It consists of a complex architecture that requires frequent maintenance. Queries made in the HOLAP model involve the multi-dimensional database and the relational database. The front-user tool presents data from the database management system (directly) or through the intermediate MOLAP.



Advantages

- It improves performance and scalability because it combines multi-dimensional and relational attributes of online analytical processing.
- It is a resourceful analytical processing tool if we expect the size of data to increase.
- Its processing ability is higher than the other two analytical processing tools.

Disadvantages

- The model uses a huge storage space because it consists of data from two databases.
- The model requires frequent updates because of its complex nature.

Summary table for MOLAP, ROLAP, and HOLAP

The following table provides a summary of the differences between MOLAP, ROLAP, and HOLAP.

Basis of Comparison	**MOLAP**	**ROLAP**	**HOLAP**
Meaning	Multi-Dimensional Online Analytical Processing	Relational Online Analytical Processing	Hybrid Online Analytical Processing
Data Storage	It stores data in a multi-dimensional database.	It stores data in a relational database.	It stores data in a relational database
Technique	It utilizes the Sparse Matrix technique.	It employs Structured Query Language (SQL).	It uses a combination of SQL and Sparse Matrix technique.
Volume of data	It can process a limited volume of data.	It processes enormous data.	It can process huge volumes of data.
Designed view	The multi-dimensional view is static.	The multi-dimensional view is dynamic.	The multi-dimensional view is dynamic.
Data arrangement	It arranges data in data cubes.	It arranges data in rows and	There is a multi-dimensional

Basis of Comparison	MOLAP	ROLAP	HOLAP
		columns (tables).	arrangement of data

What is ETL?

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.

It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

Why do you need ETL?

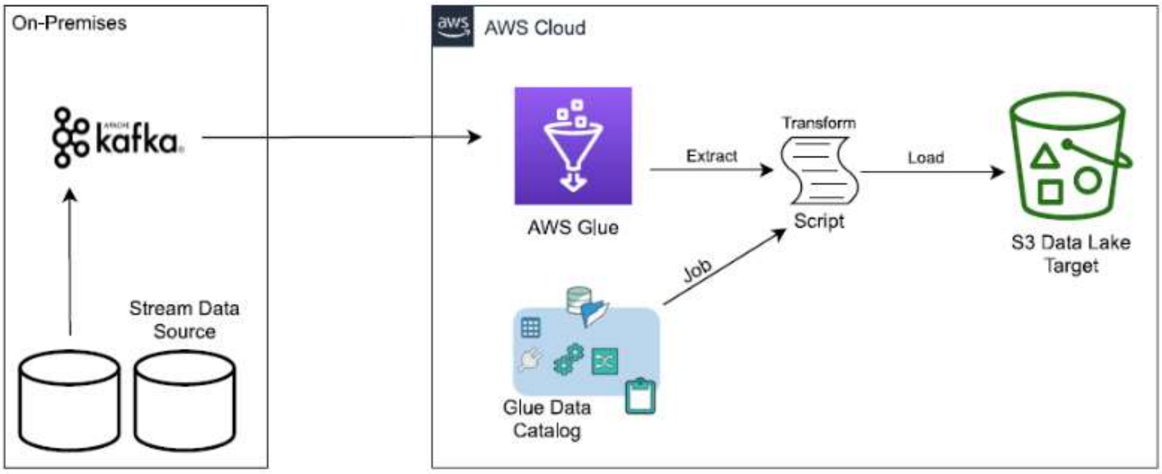
There are many reasons for adopting ETL in the organization:

- It helps companies to analyze their business data for taking critical business decisions.
- Transactional databases cannot answer complex business questions that can be answered by ETL example.
- A Data Warehouse provides a common data repository
- ETL provides a method of moving the data from various sources into a data warehouse.
- As data sources change, the Data Warehouse will automatically update.
- Well-designed and documented ETL system is almost essential to the success of a Data Warehouse project.
- Allow verification of data transformation, aggregation and calculations rules.
- ETL process allows sample data comparison between the source and the target system.
- ETL process can perform complex transformations and requires the extra area to store the data.
- ETL helps to Migrate data into a Data Warehouse. Convert to the various formats and types to adhere to one consistent system.
- ETL is a predefined process for accessing and manipulating source data into the target database.
- ETL in data warehouse offers deep historical context for the business.
- It helps to improve productivity because it codifies and reuses without a need for technical skills.

How does ETL work?

Extract, transform, and load (ETL) works by moving data from the source system to the destination system at periodic intervals. The ETL process works in three steps:

1. Extract the relevant data from the source database
2. Transform the data so that it is better suited for analytics
3. Load the data into the target database



What is data extraction?

In data extraction, extract, transform, and load (ETL) tools extract or copy raw data from multiple sources and store it in a staging area. A staging area (or landing zone) is an intermediate storage area for temporarily storing extracted data. Data staging areas are often transient, meaning their contents are erased after data extraction is complete. However, the staging area might also retain a data archive for troubleshooting purposes.

How frequently the system sends data from the data source to the target data store depends on the underlying change data capture mechanism. Data extraction commonly happens in one of the three following ways.

Update notification

In update notification, the source system notifies you when a data record changes. You can then run the extraction process for that change. Most databases and web applications provide update mechanisms to support this data integration method.

Incremental extraction

Some data sources can't provide update notifications but can identify and extract data that has been modified over a given time period. In this case, the system checks for changes at periodic intervals, such as once a week, once a month, or at the end of a campaign. You only need to extract data that has changed.

Full extraction

Some systems can't identify data changes or give notifications, so reloading all data is the only option. This extraction method requires you to keep a copy of the last extract to check which records are new. Because this approach involves high data transfer volumes, we recommend you use it only for small tables.

What is data transformation?

In data transformation, extract, transform, and load (ETL) tools transform and consolidate the raw data in the staging area to prepare it for the target data warehouse. The data transformation phase can involve the following types of data changes.

Basic data transformation

Basic transformations improve data quality by removing errors, emptying data fields, or simplifying data. Examples of these transformations follow.

Data cleansing

Data cleansing removes errors and maps source data to the target data format. For example, you can map empty data fields to the number 0, map the data value “Parent” to “P,” or map “Child” to “C.”

Data deduplication

Deduplication in data cleansing identifies and removes duplicate records.

Data format revision

Format revision converts data, such as character sets, measurement units, and date/time values, into a consistent format. For example, a food company might have different recipe databases with ingredients measured in kilograms and pounds. ETL will convert everything to pounds.

Advanced data transformation

Advanced transformations use business rules to optimize the data for easier analysis. Examples of these transformations follow.

Derivation

Derivation applies business rules to your data to calculate new values from existing values. For example, you can convert revenue to profit by subtracting expenses or calculating the total cost of a purchase by multiplying the price of each item by the number of items ordered.

Joining

In data preparation, joining links the same data from different data sources. For example, you can find the total purchase cost of one item by adding the purchase value from different vendors and storing only the final total in the target system.

Splitting

You can divide a column or data attribute into multiple columns in the target system. For example, if the data source saves the customer name as “Jane John Doe,” you can split it into a first, middle, and last name.

Summarization

Summarization improves data quality by reducing a large number of data values into a smaller dataset. For example, customer order invoice values can have many different small amounts. You can summarize the data by adding them up over a given period to build a customer lifetime value (CLV) metric.

Encryption

You can protect sensitive data to comply with data laws or data privacy by adding encryption before the data streams to the target database.

What is data loading?

In data loading, extract transform, and load (ETL) tools move the transformed data from the staging area into the target data warehouse. For most organizations that use ETL, the process is automated, well defined, continual, and batch driven. Two methods for loading data follow.

Full load

In full load, the entire data from the source is transformed and moved to the data warehouse. The full load usually takes place the first time you load data from a source system into the data warehouse.

Incremental load

In incremental load, the ETL tool loads the delta (or difference) between target and source systems at regular intervals. It stores the last extract date so that only records added after this date are loaded. There are two ways to implement incremental load.

Streaming incremental load

If you have small data volumes, you can stream continual changes over data pipelines to the target data warehouse. When the speed of data increases to millions of events per second, you can use event stream processing to monitor and process the data streams to make more-timely decisions.

Batch incremental load

If you have large data volumes, you can collect load data changes into batches periodically. During this set period of time, no actions can happen to either the source or target system as data is synchronized.

What is ELT?

Extract, load, and transform (ELT) is an extension of extract, transform, and load (ETL) that reverses the order of operations. You can load data directly into the target system before processing it. The intermediate staging area is not required because the target data warehouse has data mapping capabilities within it. ELT has become more popular with the adoption of cloud infrastructure, which gives target databases the processing power they need for transformations.

ETL compared to ELT

ELT works well for high-volume, unstructured datasets that require frequent loading. It is also ideal for big data because the planning for analytics can be done after data extraction and storage. It leaves the bulk of transformations for the analytics stage and focuses on loading minimally processed raw data into the data warehouse.

The ETL process requires more definition at the beginning. Analytics needs to be involved from the start to define target data types, structures, and relationships. Data scientists mainly use ETL to load legacy databases into the warehouse, and ELT has become the norm today.

Conclusion

OLAP is an important concept in warehousing because it enables users to query, retrieve, and analyze data. MOLAP, ROLAP, and HOLAP are the main forms of OLAP. These models can be distinguished using various aspects such as the volume of data, storage, designed view, and data arrangement.