Unit-1 Introduction to data Warehouse

* What is Data Warehouse?
* Data Warehouse Characteristics?
* A Multidimensional Data Model?
* Data Warehouse Architecture?
* Extraction, Transformation, Loading?
* Components or Building Blocks of Data Warehouse?
* Components?
* **OLAP operation?**

## Types of OLAP Servers?

* OLAP architecture ROLAP,MOLAP,HOLAP?

What is Data Warehouse?

According to William H. Inmon,

def

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”

What is Data Warehouse?

* Integrated: A data warehouse is constructed by integrating multiple heterogonous sources, such as relational databases, flat files, and online transaction.
* Time variant: Data are stored to provide information from an historic perspective (e.g., the past 5-10 years).
* Non-volatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.
* Subject oriented : A data warehouse is organized around major subjects such as customer, supplier, product, and sales.

OLAP vs OLTP

|  |  |  |
| --- | --- | --- |
| Sr.No. | Data Warehouse (OLAP) | Operational Database (OLTP) |
| 1 | Involves historical processing of information. | Involves day-to-day processing. |
| 2 | OLAP systems are used by knowledge workers such as executives, managers and analysts. | OLTP systems are used by clerks, DBAs, or database professionals. |
| 3 | Useful in analyzing the business. | Useful in running the business. |
| 4 | It focuses on Information out. | It focuses on Data in. |
| 5 | Based on Star Schema, Snowflake, Schema and Fact Constellation Schema. | Based on Entity Relationship Model. |
| 6 | Contains historical data. | Contains current data. |
| 7 | Provides summarized and consolidated data. | Provides primitive and highly detailed data. |
| 8 | Provides summarized and multidimensional view of data. | Provides detailed and flat relational view of data. |
| 9 | Number or users is in hundreds. | Number of users is in thousands. |
| 10 | Number of records accessed is in millions. | Number of records accessed is in tens. |
| 11 | Database size is from 100 GB to 1 TB | Database size is from 100 MB to 1 GB. |
| 12 | Highly flexible. | Provides high performance. |

**Data Warehouse Characteristics?**

Integrated: A data warehouse is constructed by integrating multiple heterogonous sources, such as relational databases, flat files, and online transaction.

Time variant: Data are stored to provide information from an historic perspective (e.g., the past 5-10 years).

Non-volatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.

Subject oriented : A data warehouse is organized around major subjects such as customer, supplier, product, and sales.

**Data Warehouse Architecture  
Data Warehouse Architecture** is complex as it’s an information system that contains historical and commutative data from multiple sources. There are 3 approaches for constructing Data Warehouse layers: Single Tier, Two tier and Three tier. This 3 tier architecture of Data Warehouse is explained as below.  
**Single-tier architecture**  
The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

**Two-tier architecture**

Two-layer architecture is one of the Data Warehouse layers which separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

**Three-Tier Data Warehouse Architecture**

This is the most widely used Architecture of Data Warehouse.

It consists of the Top, Middle and Bottom Tier.

**Bottom Tier:** The database of the Datawarehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools.

**Middle Tier:**The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database.

**Top-Tier:**The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools.

## Extraction, Transformation, Loading

## 1) Extraction

In this step of ETL architecture, 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.

**Three Data Extraction methods:**

1. Full Extraction
2. Partial Extraction- without update notification.
3. Partial Extraction- with update notification

Irrespective of the method used, extraction should not affect performance and response time of the source systems. These source systems are live production databases. Any slow down or locking could effect company’s bottom line.

## 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.

It is one of the important ETL concepts where 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**.

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.

## 3) Loading

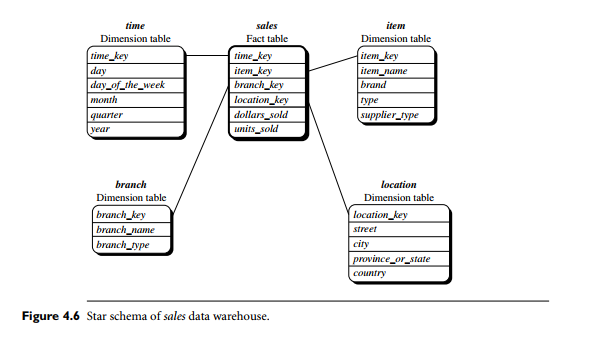
Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

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.

**Types of Loading:**

* **Initial Load** — populating all the Data Warehouse tables
* **Incremental Load**— applying ongoing changes as when needed periodically.
* **Full Refresh** —erasing the contents of one or more tables and reloading with fresh data.
* multidimensional model
* Star Schema:-

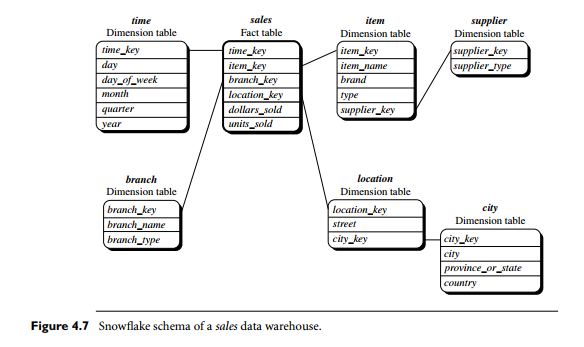
1. A set of smaller attendant tables (dimension tables), one for each dimension.
2. A large central table (fact table) containing the bulk of the data, with no redundancy



* Snowflake schema:

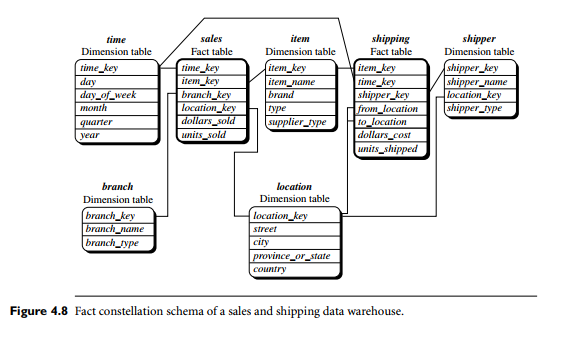
1. A snowflake schema is a variant of the star schema model,
2. Where some dimensions are normalized, thereby further splitting the data into additional tables.

3. The resulting schema similar to snowflake.

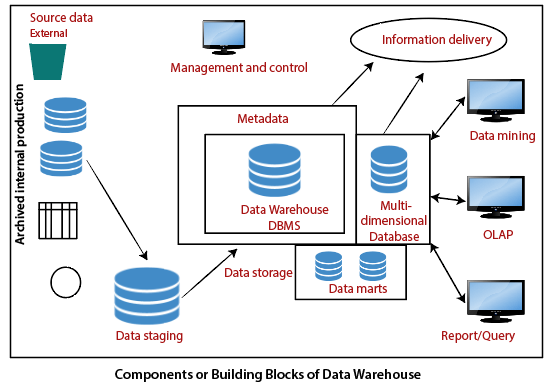


* Fact Constellation:

1. Sophisticated applications may require multiple fact tables to share dimension tables.
2. It can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.



Components or Building Blocks of Data Warehouse



## Main Components of **DWH Architecture**

A data warehouse (DWH) design consists of six main components:

1. **Data Warehouse Database**

The central component of a DW architecture is a database that stocks all enterprise data and makes it manageable for reporting. Obviously, this means you need to choose which kind of database you’ll use to store data in your warehouse.

The following are the four database types that you can use:

* **Typical relational databases** are the row-centered databases you perhaps use on an everyday basis —for example, Microsoft SQL Server, SAP, Oracle, and IBM DB2.
* **Analytics databases** are precisely developed for data storage to sustain and manage analytics, such as Teradata and Greenplum.
* **Data warehouse applications** aren’t exactly storage databases, but several dealers now offer applications that offer software for data management as well as hardware for storing data. For example, SAP Hana, Oracle Exadata, and IBM Netezza.
* **Cloud-based databases** can be hosted and retrieved on the cloud so that you don’t have to procure any hardware to set up your data warehouse—for example, Amazon Redshift, Google BigQuery, and Microsoft Azure SQL.

### 2. Extraction, Transformation, and Loading Tools (ETL)

ETL tools are central components of an enterprise data warehouse architecture. These tools help extract data from different sources, transform it into a suitable arrangement, and load it into a data warehouse.

The ETL tool you choose will determine:

* The time expended in data extraction
* Approaches to extracting data
* Kind of transformations applied and the simplicity to do so
* Business rule definition for [data validation and cleansing](https://www.astera.com/type/blog/data-cleansing-tools-what-you-need-to-know/)to improve end-product analytics
* Filling mislaid data
* Outlining information distribution from the fundamental depository to your BI applications

### 3. Metadata

Data about data is called as meta data.

In the data warehouse architecture, metadata describes the data warehouse database and offers a framework for data. It helps in constructing, preserving, handling, and making use of the data warehouse.

There are two types of metadata in data warehousing:

* **Technical Metadata** comprises of information that can be used by developers and managers when executing warehouse development and administration tasks.
* **Business Metadata** includes information that offers an easily understandable standpoint of the data stored in the warehouse.

Metadata plays an important role for businesses and the technical teams to understand the data present in the warehouse and convert it into information.

Your data warehouse isn’t a project, it’s a process. To make your implementation as effective as possible, you need to take a truly agile approach, which necessitates a [metadata-driven data warehouse architecture](https://www.astera.com/type/blog/introduction-to-metadata-architecture/).

Also, you can [test data warehouse models concurrently](https://www.astera.com/type/blog/traditional-data-warehouse-vs-mdds/) before deployment and replicate your schema in any leading database. A metadata-driven approach leads to an iterative development culture and futureproofs your data warehouse deployment, so you can update the existing infrastructure with the new requirements without disrupting your data warehouse’s integrity and usability.

Coupled with automation capabilities, a metadata-driven data warehouse architecture can [streamline design, development, and deployment](https://www.astera.com/type/blog/metadata-in-data-warehouse/), leading to a robust data warehouse implementation.

### 4. Data Warehouse Access Tools

A data warehouse uses a database or group of databases as a foundation. Data warehouse corporations generally cannot work with databases without the use of tools unless they have database administrators available. However, that is not the case with all business units. This is why they use the assistance of several no-code data warehousing tools, such as:

* **Query and reporting tools**help users produce corporate reports for analysis that can be in the form of spreadsheets, calculations, or interactive visuals.
* **Application development tools** help create tailored reports and present them in interpretations intended for reporting purposes.
* **Data mining tools for data warehousing** systematize the procedure of identifying arrays and links in huge quantities of data using cutting-edge statistical modeling methods.
* **OLAP tools** help construct a multi-dimensional data warehouse and allow the analysis of enterprise data from numerous viewpoints.

### 5. Data Mart

It defines the data flow within a data warehousing bus architecture and includes a data mart. A data mart is an access level that allows users to transfer data. It is also used for partitioning data that is produced for a particular user group.

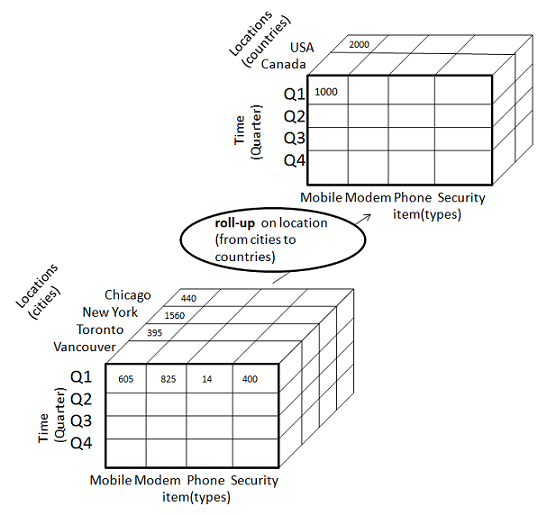
### 6. Data Warehouse Reporting Layer

The reporting layer in the data warehouse allows the end-users to access the BI interface or BI database architecture. The purpose of the reporting layer in the data warehouse is to act as a dashboard for data visualization, create reports, and take out any required information.

### Roll-up

Roll-up performs aggregation on a data cube in any of the following ways −

* By climbing up a concept hierarchy for a dimension
* By dimension reduction



The following diagram illustrates how roll-up works.

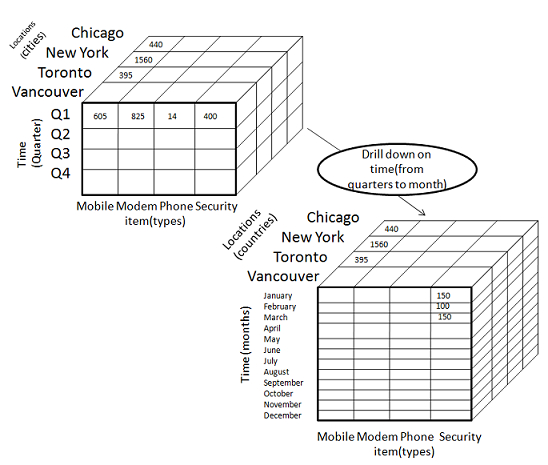
* Roll-up is performed by climbing up a concept hierarchy for the dimension location.
* Initially the concept hierarchy was "street < city < province < country".
* On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
* The data is grouped into cities rather than countries.
* When roll-up is performed, one or more dimensions from the data cube are removed.

### Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways −

* By stepping down a concept hierarchy for a dimension
* By introducing a new dimension.

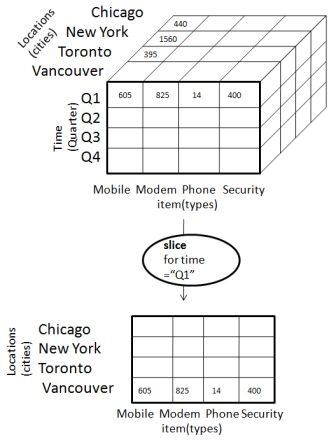
The following diagram illustrates how drill-down works

−

* Drill-down is performed by stepping down a concept hierarchy for the dimension time.
* Initially the concept hierarchy was "day < month < quarter < year."
* On drilling down, the time dimension is descended from the level of quarter to the level of month.
* When drill-down is performed, one or more dimensions from the data cube are added.
* It navigates the data from less detailed data to highly detailed data.

### Slice

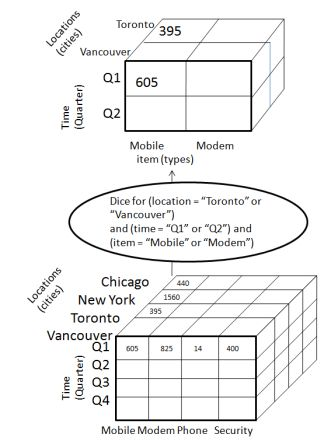
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



* Here Slice is performed for the dimension "time" using the criterion time = "Q1".
* It will form a new sub-cube by selecting one or more dimensions.

### Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

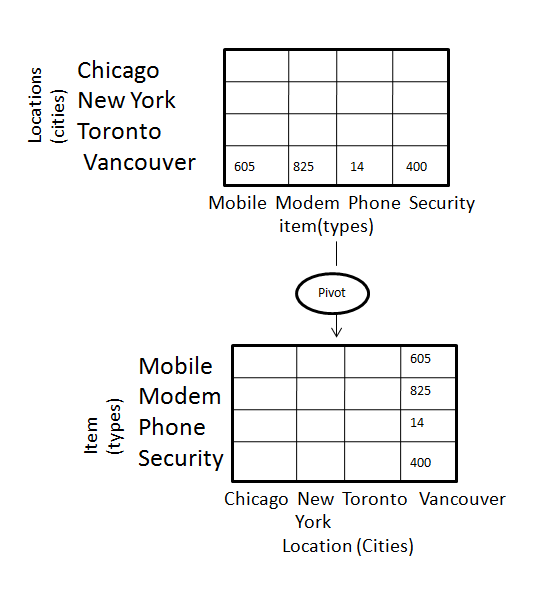


The dice operation on the cube based on the following selection criteria involves three dimensions.

* (location = "Toronto" or "Vancouver")
* (time = "Q1" or "Q2")
* (item =" Mobile" or "Modem")

### Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



* Drill-down is performed by stepping down a concept hierarchy for the dimension time.
* Initially the concept hierarchy was "day < month < quarter < year."
* On drilling down, the time dimension is descended from the level of quarter to the level of month.
* When drill-down is performed, one or more dimensions from the data cube are added.
* It navigates the data from less detailed data to highly detailed data.

## Types of OLAP Servers

We have four types of OLAP servers −

* Relational OLAP (ROLAP)
* Multidimensional OLAP (MOLAP)
* Hybrid OLAP (HOLAP)

## Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following −

* Implementation of aggregation navigation logic.
* Optimization for each DBMS back end.
* Additional tools and services.

## Multidimensional OLAP

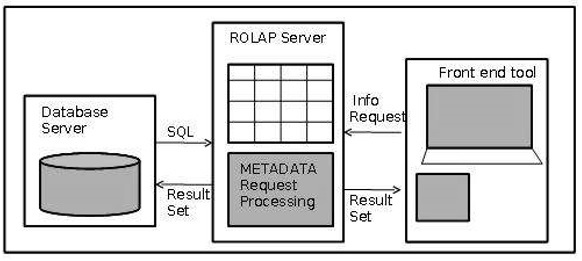
MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

## Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

## Relational OLAP Architecture

ROLAP includes the following components −

* Database server
* ROLAP server
* Front-end tool.
* 

## Advantages

* ROLAP servers can be easily used with existing RDBMS.
* Data can be stored efficiently, since no zero facts can be stored.
* ROLAP tools do not use pre-calculated data cubes.
* DSS server of micro-strategy adopts the ROLAP approach.

## Disadvantages

* Poor query performance.
* Some limitations of scalability depending on the technology architecture that is utilized.

# 

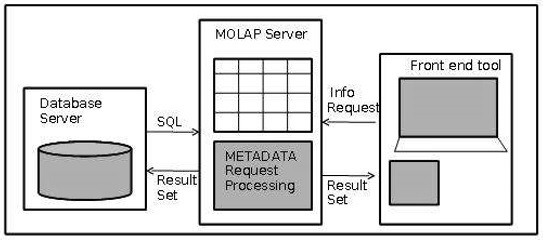
* **Multidimensional OLAP**

Multidimensional OLAP (MOLAP) uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the dataset is sparse. Therefore, many MOLAP servers use two levels of data storage representation to handle dense and sparse datasets.

* MOLAP tools process information with consistent response time regardless of level of summarizing or calculations selected.
* MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.
* MOLAP tools need fastest possible performance.
* MOLAP server adopts two level of storage representation to handle dense and sparse data sets.
* Denser sub-cubes are identified and stored as array structure.
* Sparse sub-cubes employ compression technology.

## MOLAP Architecture

MOLAP includes the following components −

* Database server.
* MOLAP server.
* Front-end tool.
* 

## Advantages

* MOLAP allows fastest indexing to the pre-computed summarized data.
* Helps the users connected to a network who need to analyze larger, less-defined data.
* Easier to use, therefore MOLAP is suitable for inexperienced users.

## Disadvantages

* MOLAP are not capable of containing detailed data.
* The storage utilization may be low if the data set is sparse.

## MOLAP vs ROLAP

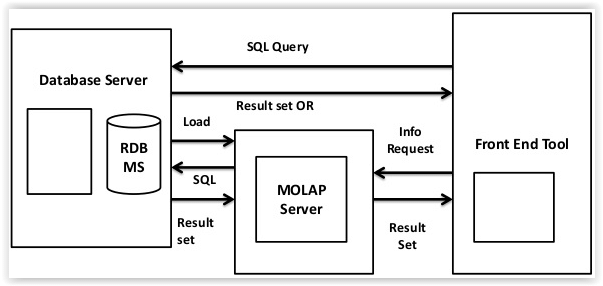
|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **MOLAP** | **ROLAP** |
| 1 | Information retrieval is fast. | Information retrieval is comparatively slow. |
| 2 | Uses sparse array to store data-sets. | Uses relational table. |
| 3 | MOLAP is best suited for inexperienced users, since it is very easy to use. | ROLAP is best suited for experienced users. |
| 4 | Maintains a separate database for data cubes. | It may not require space other than available in the Data warehouse. |
| 5 | DBMS facility is weak. | DBMS facility is strong. |

## What is HOLAP Architecture?

HOLAP represents Hybrid OLAP. It can manage a tradeoff between ROLAP’s scalability and MOLAP’s query implementation, some commercial OLAP servers are depended on the HOLAP method. In this case, the user determines which portion of the data to save in the MOLAP and which in the ROLAP. For example, generally, the low-level data are saved using a relational database, while higher-level data, including aggregations, are stored in an independent MOLAP.

HOLAP is a mixture of ROLAP (relational OLAP) and MOLAP (multidimensional OLAP) which are different implementations of OLAP. HOLAP enables storing elements of the data in a MOLAP store and another element of the data in a ROLAP store, enabling a tradeoff of the benefits of each. The degree of control that the cube designer has over this separation change from product to product.

Because hybrid OLAPs allows the use of several set of the two OLAPs, they generally save data in both a relational database and a multidimensional database. As a result, the decision to access one of the two databases is based on which is best suited for the desired processing type or software.



**Difference between ROLAP, MOLAP and HOLAP :**

| **Basis** | **ROLAP** | **MOLAP** | **HOLAP** |
| --- | --- | --- | --- |
| **Storage location for summary aggregation** | Relational Database is used as storage location for summary aggregation. | Multidimensional Database is used as storage location for summary aggregation. | Multidimensional Database is used as storage location for summary aggregation. |
| **Processing time** | Processing time of ROLAP is very slow. | Processing time of MOLAP is fast. | Processing time of HOLAP is fast. |
| **Storage space requirement** | Large storage space requirement in ROLAP as compare to MOLAP and HOLAP. | Medium storage space requirement in MOLAP as compare to ROLAP and HOLAP. | Small storage space requirement in HOLAP as compare to MOLAP and ROLAP. |
| **Storage location for detail data** | Relational database is used as storage location for detail data. | Multidimensional database is used as storage location for detail data. | Relational database is used as storage location for detail data. |
| **Latency** | Low latency in ROLAP as compare to MOLAP nad HOLAP. | High latency in MOLAP as compare to ROLAP and HOLAP. | Medium latency in HOLAP as compare to MOLAP and ROLAP. |
| **Query response time** | Slow query response time in ROLAP as compare to MOLAP and | Fast query response time in MOLAP as compare to ROLAP and HOLAP. | Medium query response time in HOLAP as compare to MOLAP and ROLAP. |