UNIT-2

BI and DW Architectures

Business Intelligence (BI) leverages data warehousing (DW) architectures to enable informed decision-making. Common DW architectures include single-tier, two-tier, and three-tier models, each suited for different organizational needs and data volumes.

3 Types of Data Warehouse Architecture

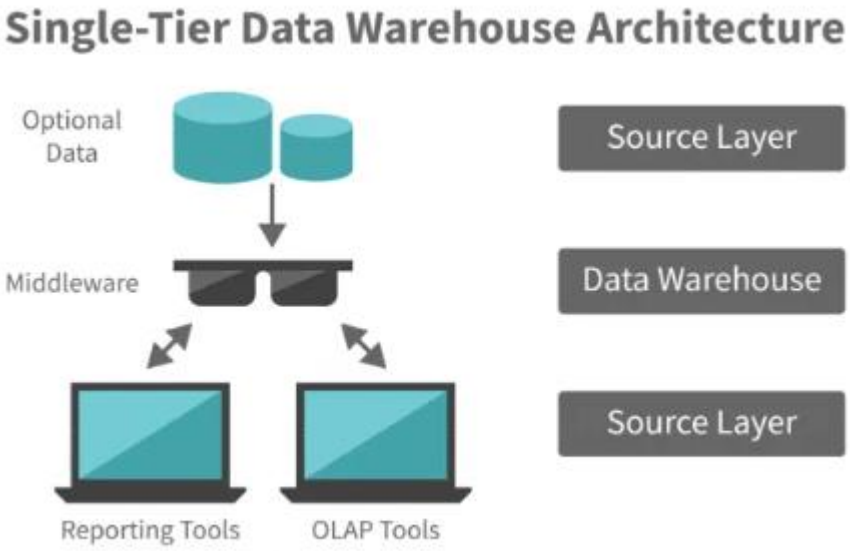
There are three common data warehouse architecture types typically used for building a data warehouse:

1. Single-Tier Architecture
2. Two-Tier Architecture
3. Three-Tier Architecture

Each type of data warehouse architecture has its own benefits and limitations. Let's explore the unique characteristics of each one of them.

1. Single-Tier Architecture

The single-tier data warehouse architecture reduces the amount of data stored in a data warehouse by building a more compact data set. Its advantage is that it helps remove data redundancies and improves the quality of your data.



However, it isn't the ideal solution for agencies that own large volumes of data and operate with multiple data streams because it's inefficient.

The single-tier architecture has three layers:

- A source layer
- A data warehouse layer

- An analysis layer

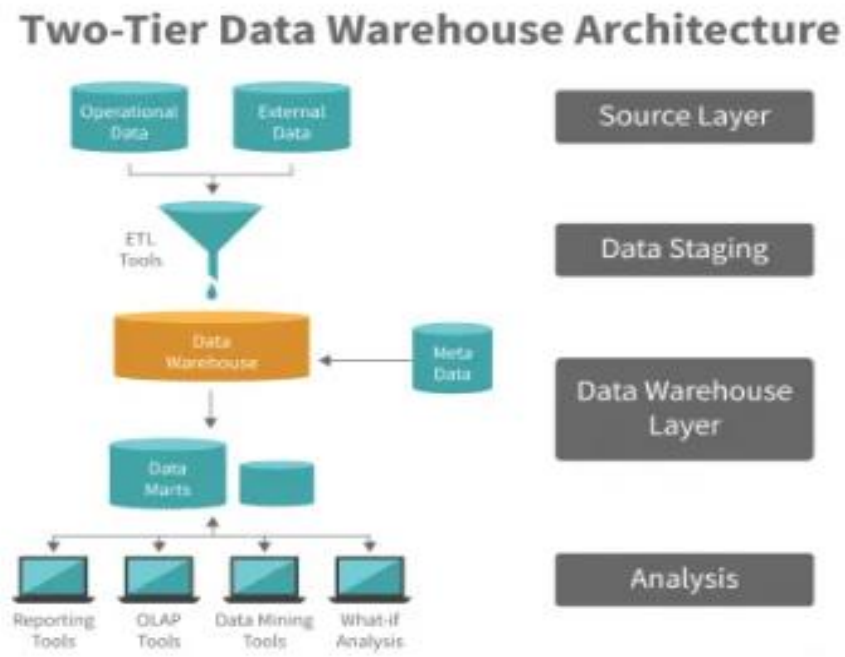
In the single-tier architecture, only the source layer is physical.The data warehouse layer is virtual and provides data in a multidimensional view, created by an intermediate processing layer.

Example: A small retail store might use a single-tier architecture to store and analyze sales data.

Drawbacks: Performance can be affected when transactional and analytical processes are not separate, and it struggles with enterprise-wide data access and scalability

2. Two-Tier Architecture

Unlike the single-tier architecture, the two-tier architecture contains a data staging area that ensures any data you load into the warehouse is cleansed and in the right format. It's found between the source layer and the data warehouse layer, as depicted in the image below.



.

Most businesses that use data marts as a server make use of the two-tier data warehouse architecture, which is also made up of two tiers:

The Data Tier

This is the layer where actual data is stored after various ETL processes have been used to load data into the data warehouse.

It's also made up of three layers:

- A source layer
- A data staging layer
- A data warehouse layer

The Client Tier

This layer is where clients can use data stored in the data warehouse to generate insights for making informed, data-driven decisions. You can modify or transform this layer based on the data trends that you discover from your analysis reports.

And it's made up of a single layer:

An analysis layer

Some disadvantages of the two-tier architecture are that it's not scalable, has network limitations, and only supports a small number of users.

3. Three-Tier Architecture

The three-tier architecture is what most organisations go for when building a data warehouse system. It solves the connectivity problems that the two-tier architecture commonly faces.

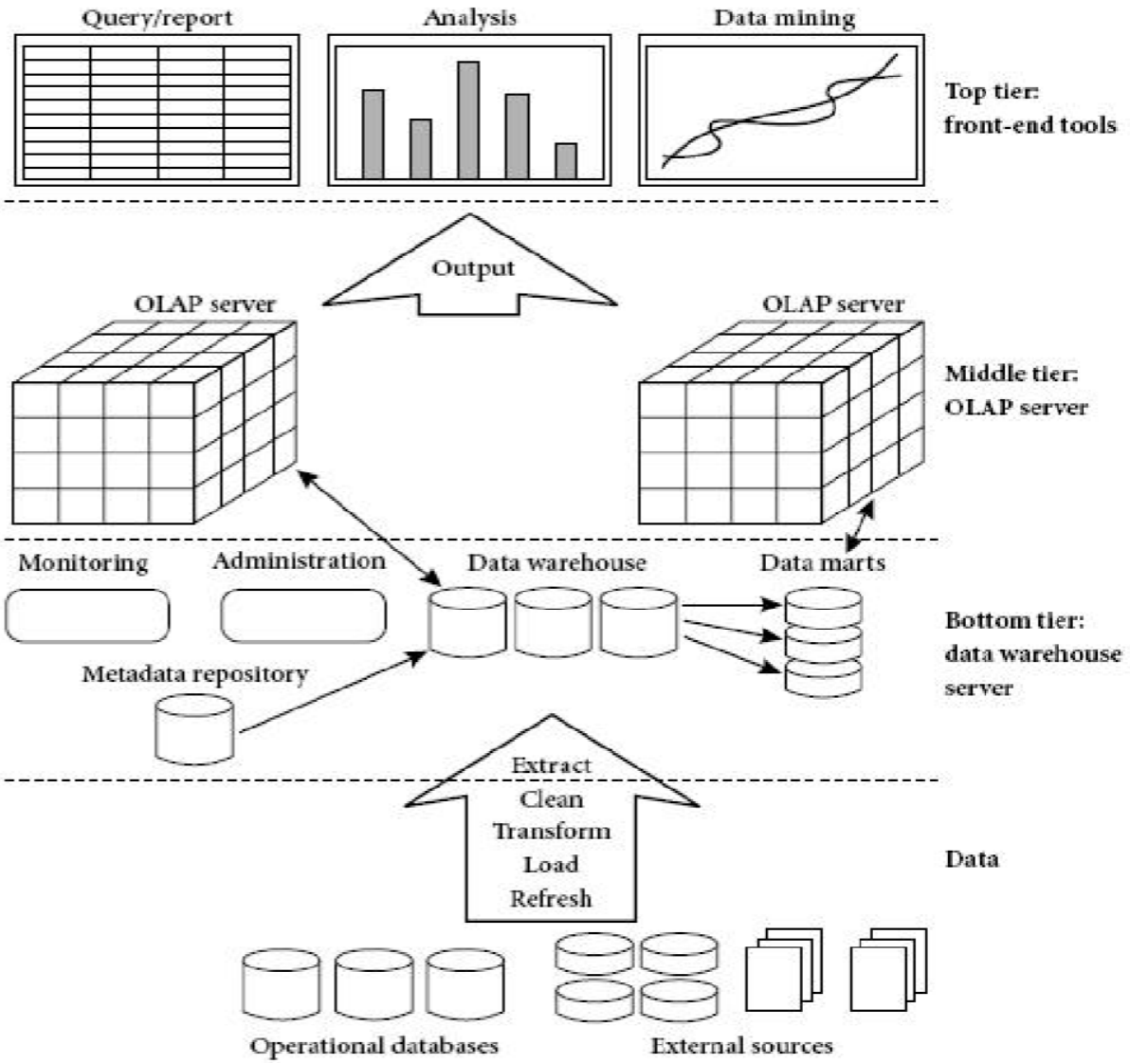
The three-tier architecture is useful for extensive, enterprise-wide systems. But its disadvantage is the additional storage space it uses through the redundant, reconciled layer.

The three-tier architecture also has three tiers:

- A bottom tier
- A top tier
- A middle tier

These three tiers are commonly called the layers of a data warehouse architecture. Let's take an in-depth look at these layers.

Three tier Data Warehouse Architecture



Layers of a Data Warehouse Architecture

1. **Bottom tier :**

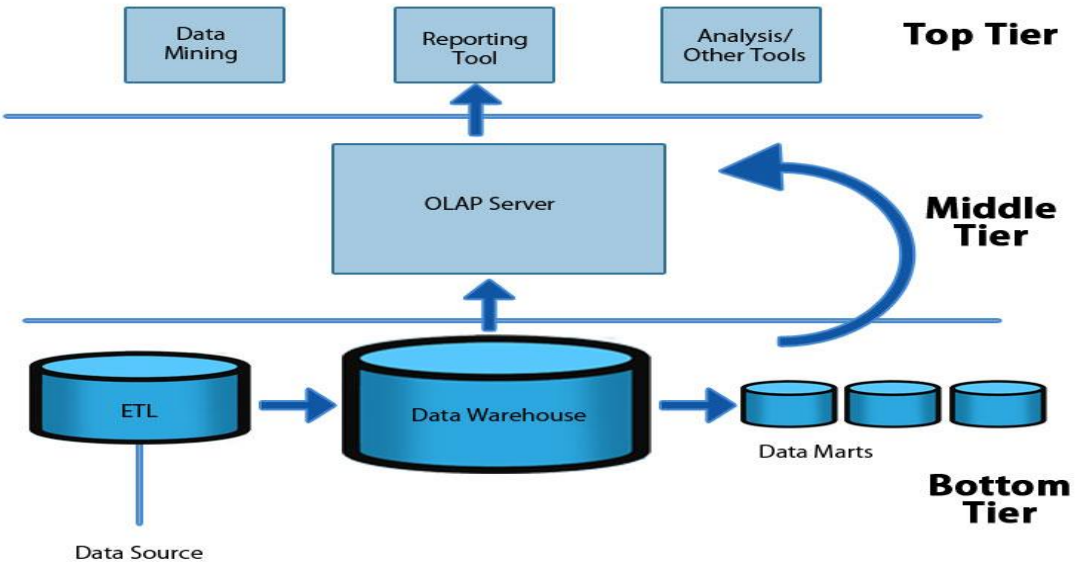
   - The bottom tier is a warehouse database server that is almost always a relational database system
   - Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources
   - These tools and utilities perform data extraction, cleaning, and transformation, as well as load and refresh functions to update the data warehouse
   - The data are extracted using application program interfaces known as gateways
   - A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server
   - Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection
   - This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

2. **Middle tier :**

- The middle tier is an OLAP server that is typically implemented using either
- A relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations or
- A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

3. **Top tier :**

- The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools



Relation between Business Intelligence (BI) and Data Warehousing:

Business Intelligence: Large business organizations usually receive large amounts of data from various sources. This data is always exploitable to obtain diverse sets of information that help in making better business decisions. These actionable insights may be descriptive, predictive, or prescriptive. BI represents the various methods and tools used for the collection, integration, analysis and visualization of business information. It could be considered synonymous with data analytics in particular to the business world.

Data Warehouse: Data Warehouse is a system and set of technologies at the back-end, that helps in collecting large amounts of dissimilar data from various sources and storing them for later use. Good data warehouses have business meaning backed into them facilitating future extraction and analysis. Business Intelligence is one of the applications that make use of data warehouses. Data Warehouses generally follow a multidimensional paradigm (related to OLAP) where data is held in Fact Tables (tables covering numbers such as revenue or costs) and Dimensions (things we want to view the facts by, such as region, office, or week)

Below is a table of differences between Business Intelligence and Data Warehouse:

Business Intelligence	Data Warehouse
It is a set of tools and methods to analyze data and discover, extract and formulate actionable information that would be useful for business decisions.	It is a system for storage of data from various sourc in an orderly manner as to facilitate business-mind reads and writes.
It is a Decision Support System (DSS).	It is a data storage system.
Serves at the front end.	Serves at the back end.
The aim of business intelligence is to enable users make informed, data-driven decisions.	A data warehouse's main aim is to provide the users business intelligence; a structured and comprehensi view of available data of an organization.
Collects data from the data warehouse for analysis	Collects data from various disparate sources a organizes it for efficient BI analysis.
Comprises business reports, charts, graphs, etc.	Comprises of data held in "fact tables" a "dimensions" with business meaning incorporated in them.
BI as such doesn't have much use without a da warehouse as large amounts of various and usef data is required for analysis.	BI is one of many use-cases for data warehouses, the are more applications for this system.
Handled by executives and analysts relatively high up in the hierarchy.	Handled and maintained by data engineers and syste administrators who report to/work for the executiv and analysts.
The role of Business Intelligence lies in improving t performance of business by utilizing tools a approaches that focus on counts, statistics, a visualization.	The reflection of actual database development a integration process is given by Data Warehouse and addition, Data Profiling and Company validati standards.
It deals with- • OLAP (Online Analytical Processing) • Data Visualization • Data Mining • Query/Reporting Tools	It deals with- • Acquiring/gathering of data • Metadata management • Cleaning of data • Transforming data • Data dissemination • Data recovery/backup planning

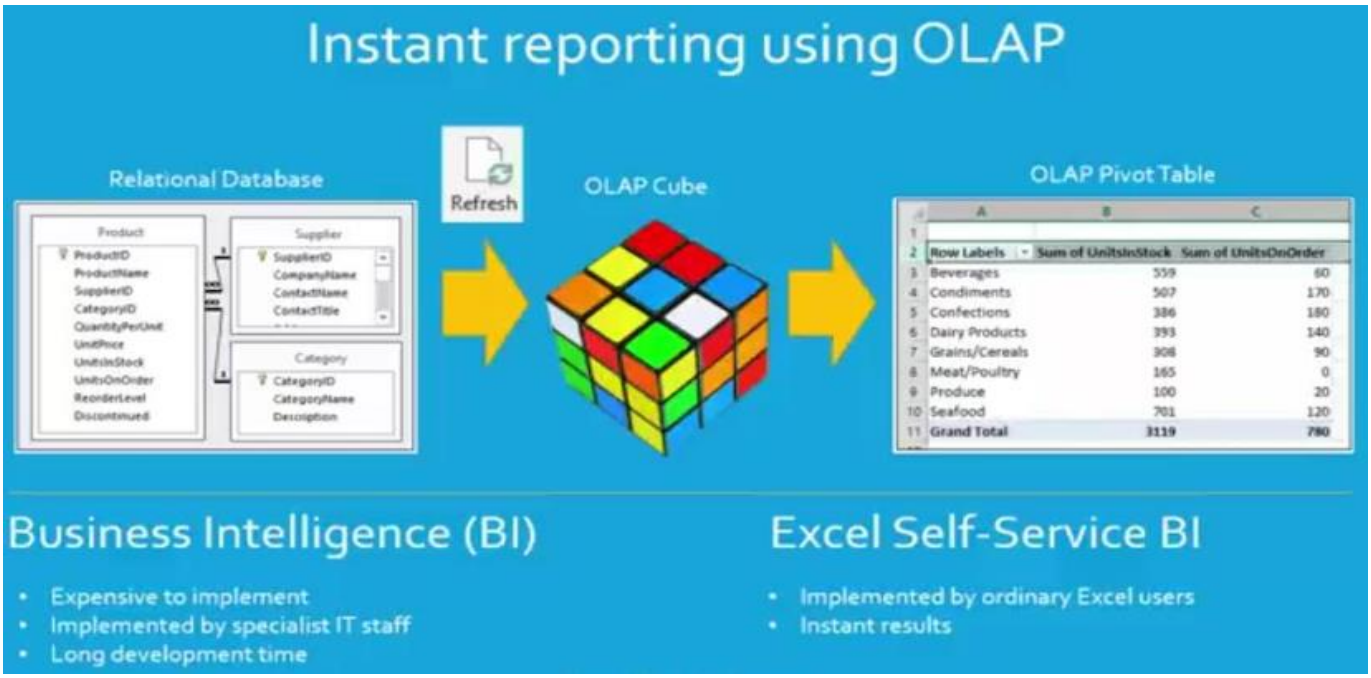
Business Intelligence	Data Warehouse
Examples of BI software: SAP, Sisense, Datapir Looker, etc.	Examples of Data warehouse software: BigQuer Snowflake, Amazon, Redshift, Panoply, etc.

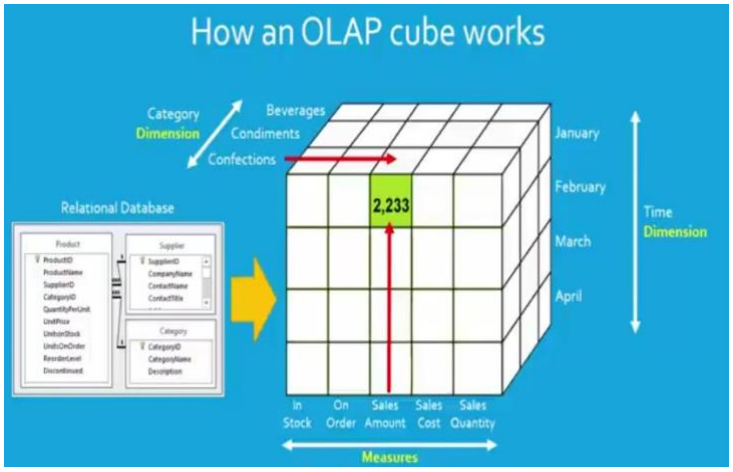
OLAP (Online Analytical Processing)

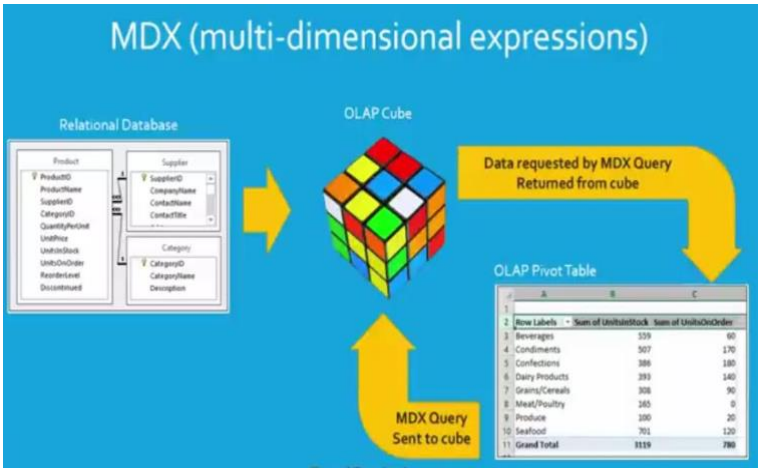
An OLAP cube is a multi-dimensional array of data. Online analytical processing is a computer-based technique of analyzing data to look for insights. The term cube here refers to a multi-dimensional dataset, which is also sometimes called a hypercube if the number of dimensions is greater than three.

OLAP stands for Online Analytical Processing, which is a technology that enables multi-dimensional analysis of business data. It provides interactive access to large amounts of data and supports complex calculations and data aggregation. OLAP is used to support business intelligence and decision-making processes.

Grouping of data in a multidimensional matrix is called data cubes. In Dataware housing, we generally deal with various multidimensional data models as the data will be represented by multiple dimensions and multiple attributes. This multidimensional data is represented in the data cube as the cube represents a high-dimensional space. The Data cube pictorially shows how different attributes of data are arranged in the data model. Below is the diagram of a general data cube.



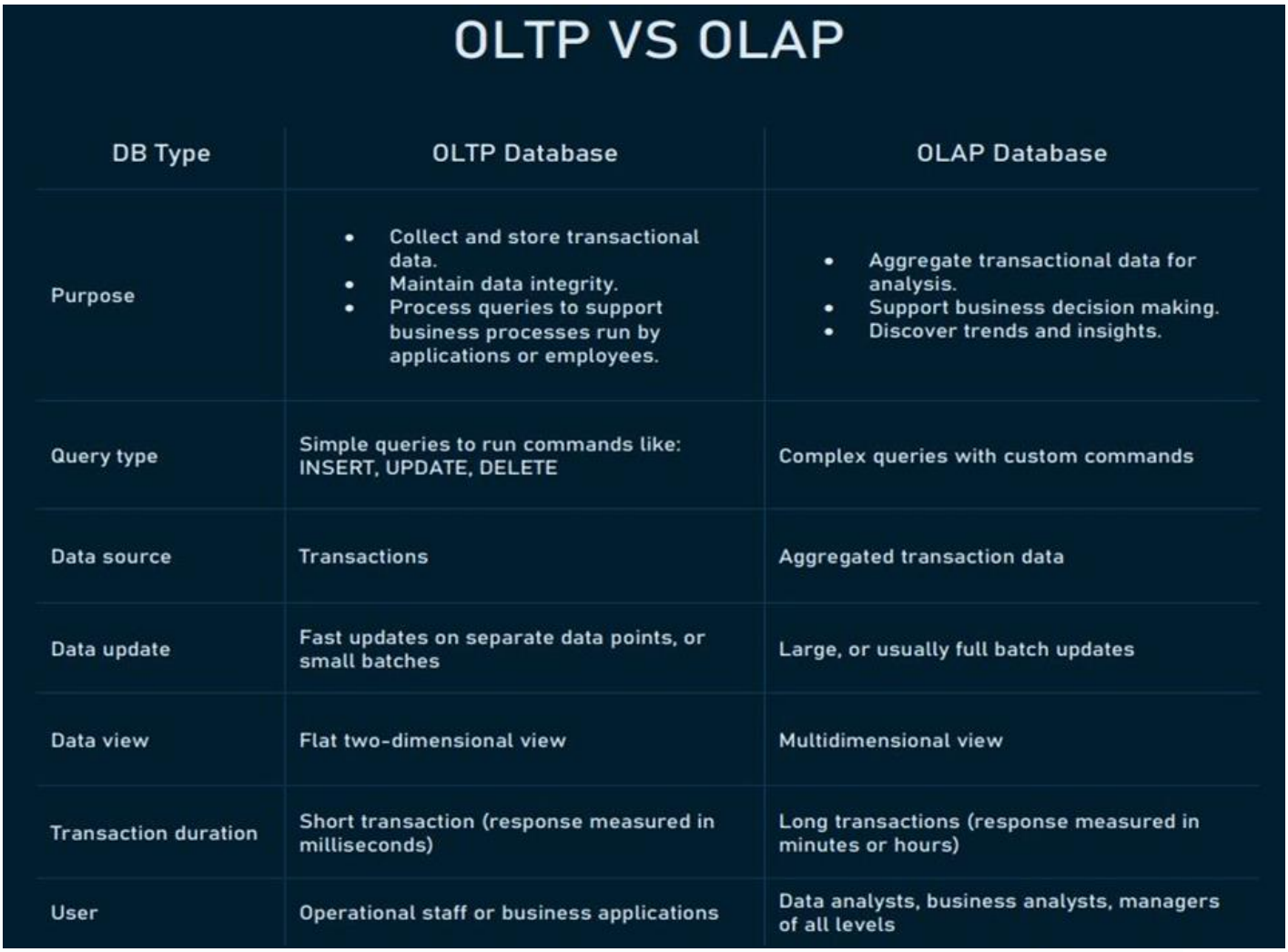




An OLAP cube allows analytics to group or slice items by different categories. They are primarily designed to run complex queries, which can't be handled by the usual OLTP databases.

OLTP vs OLAP: technology comparison

There are numerous differences between OLTP and OLAP databases in terms of purpose, information structure, and data access capabilities. The table below compares the main aspects of these two systems.



Data operations in OLTP

A transactional or OLTP database is a common storage solution we deal with to record any business information. Say, we're selling a new type of a smartphone to the customer and we want to record this transaction, including the product type, price, date, customer info, sales person name, etc. All of these items will be stored in a flat view, which allows us to quickly operate and search for the required information.The data will be saved as a set of items and values that relate to this transaction. An OLTP solution will allow a user to perform the following operations with this data:

insert,
copy,
paste,
edit/update, and
delete.

such transactions have a short response time – measured in seconds – as they are natural to OLTP. But when it comes to more complex queries that involve aggregating data from multiple tables, a transactional database will run into trouble. The more data is inquired, the more problematic and resource-intensive it is for OLTP.

Analytical requests are often much more complex than “show me total sales amount.” More often than not, we need to compare things to each other and look at the data from different dimensions. That’s where an OLAP technology kicks in.

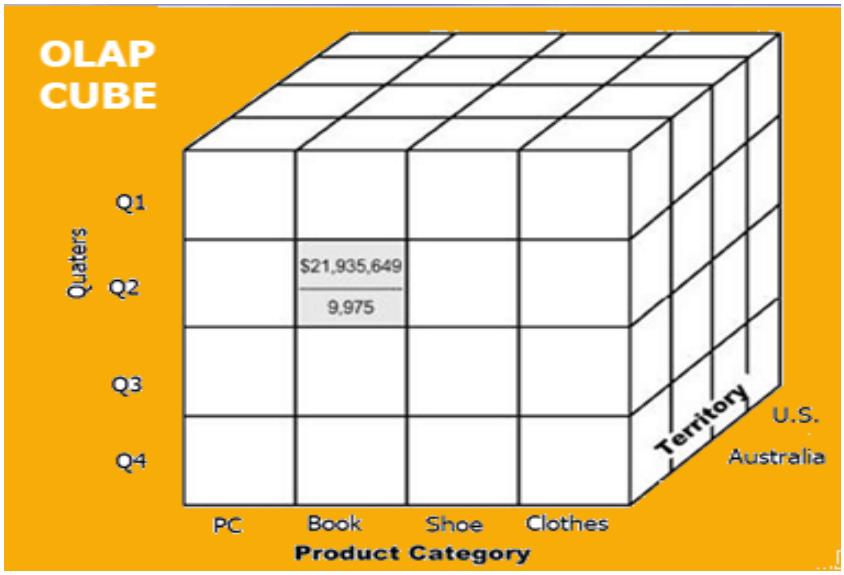
Data operations in OLAP

OLAP or Online Analytical Processing aggregates transactional data from a storage to transform it into a feasible form for analysis. As a source of data, OLAP can use some type of unified storage like a data warehouse, data lake, or data mart, or simply any place you store the historical data.

But to run complex custom queries, we must structure data properly. That’s why in most cases, there is a need for a separate OLAP database or warehouse that will model data for multidimensional analysis.

OLAP cubes

The cube may consist of several dimensions that can be used to filter the required information and form reports. OLAP systems use a specific SQL language called MDX or Multidimensional Expressions. Standard SQL queries are also supported by the most databases to perform OLAP analysis.



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

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

Usually, data operations and analysis are performed using the simple spreadsheet, where data values are arranged in row and column format. This is ideal for two-dimensional data. However, OLAP contains multidimensional data, with data usually obtained from a different and unrelated source.

Using a spreadsheet is not an optimal option. The cube can store and analyze multidimensional data in a logical and orderly manner.

operations of OLAP

Four types of analytical OLAP operations are:

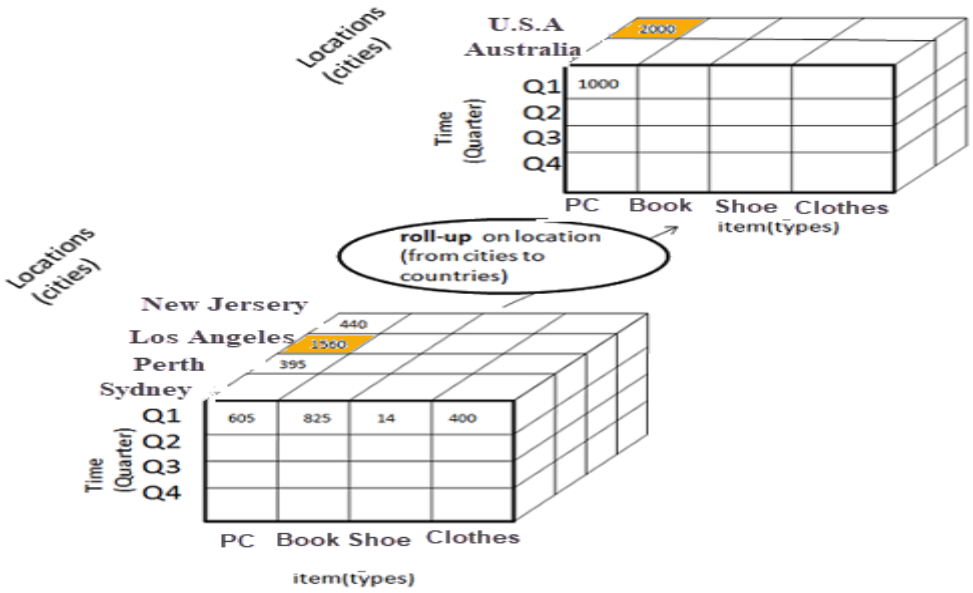
1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)

1) Roll-up:

Roll-up is also known as “consolidation” or “aggregation.” The Roll-up operation can be performed in 2 ways

1. Reducing dimensions
2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

Consider the following diagram



Roll up the location dimension

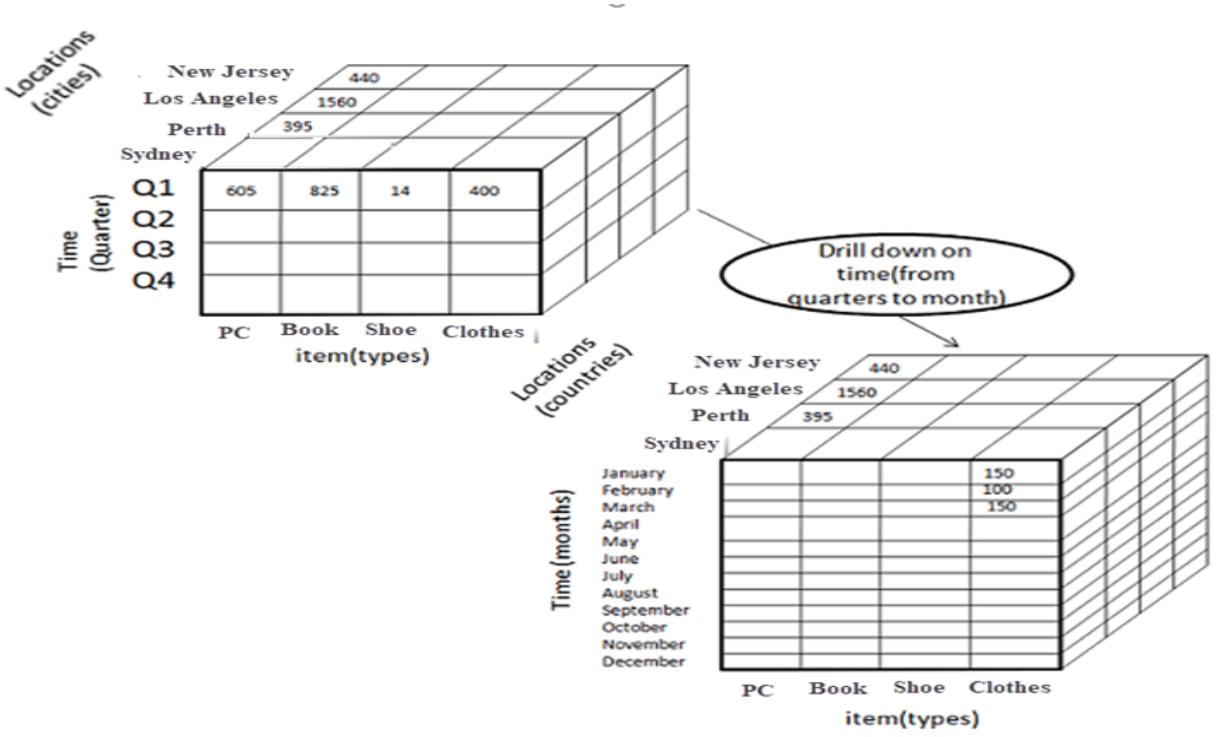
- In this example, cities New jersey and Lost Angles and rolled up into country USA
- The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up
- In this aggregation process, data is location hierarchy moves up from city to the country.
- In the roll-up process at least one or more dimensions need to be removed. In this example, Cities dimension is removed.

2) Drill-down

In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via

- Moving down the concept hierarchy
- Increasing a dimension

Drill down allows a user to move from high-level data (e.g., annual sales) to a lower level (e.g., monthly sales). Here we use the concept of hierarchy that applies to every single dimension. So, in the "time" dimension, we can move down from yearly figures to weekly or even daily records. This depends on how you store your data and model the actual cube.



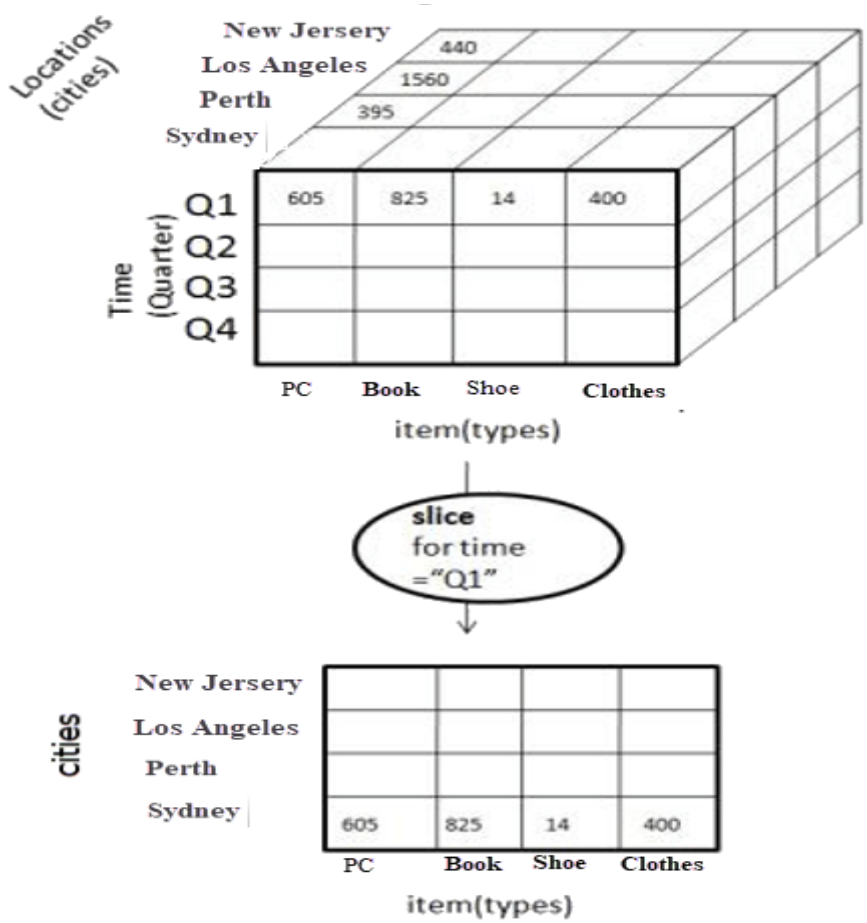
Drill-down operation in OLAP

Consider the diagram above

- Quater Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.
- In this example, dimension months are added.

3) Slice:

- Here, one dimension is selected, and a new sub-cube is created.
- Following diagram explain how slice operation performed:

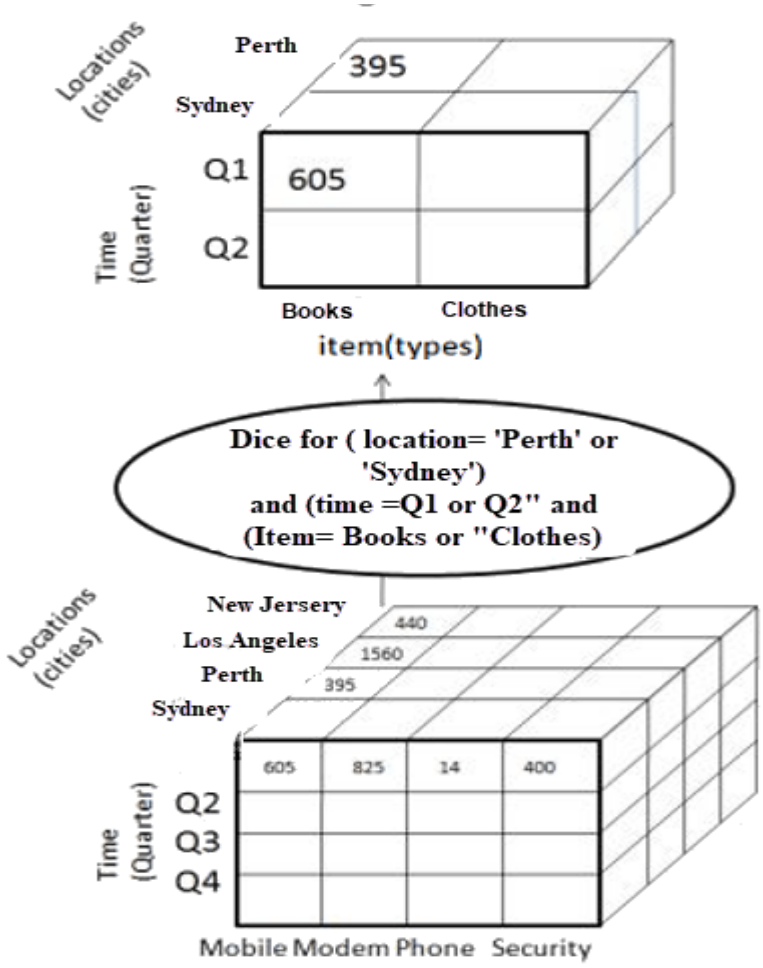


Slice operation in OLAP

- Dimension Time is Sliced with Q1 as the filter.
- A new cube is created altogether

3. Dice:

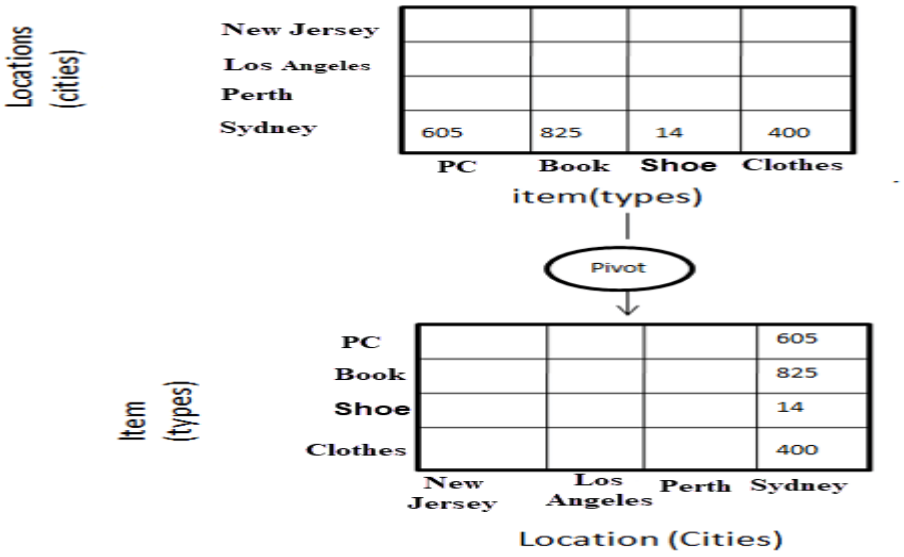
- This operation is similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube.



Dice operation in OLAP

4) Pivot

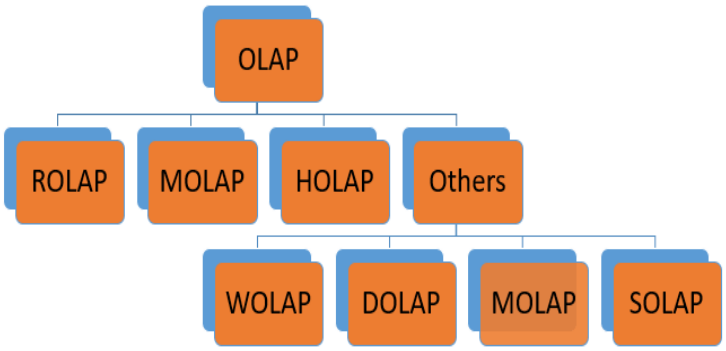
- In Pivot, you rotate the data axes to provide a substitute presentation of data.
- In the following example, the pivot is based on item types.
- Analysts can gain a new view of data by rotating the data axes of thecube.



Pivot operation in OLAP

OLAP Models

Types of OLAP Systems



Relational OLAP (ROLAP):

ROLAP works with data that exist in a relational database. Facts and dimension tables are stored as relational tables. It also allows multidimensional analysis of data and is the fastest growing OLAP.

- ROLAP works directly with relational databases. The base data and the dimensiontables are stored as relational tables and new tables are created to hold the aggregated information. It depends on a specialized schema design.
- This methodology relies on manipulating the data stored in the relational databaseto give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE"; clause in the SQL statement.
- ROLAP tools do not use pre-calculated data cubes but instead pose the query to the standard relational database and its tables in order to bring back the data required to answer the question.

- ROLAP tools feature the ability to ask any question because the methodology doesnot limit to the contents of a cube. ROLAP also has the ability to drill down to the lowest level of detail in the database.

Advantages of ROLAP model:

- **High data efficiency.** It offers high data efficiency because query performance and access language are optimized particularly for the multidimensional data analysis.
- **Scalability.** This type of OLAP system offers scalability for managing large volumes of data, and even when the data is steadily increasing.

Drawbacks of ROLAP model:

- **Demand for higher resources:** ROLAP needs high utilization of manpower, software, and hardware resources.
- **Aggregately data limitations.** ROLAP tools use SQL for all calculation of aggregate data. However, there are no set limits to the for handling computations.
- **Slow query performance.** Query performance in this model is slow when compared with MOLAP

Multidimensional OLAP (MOLAP):

MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP.

- MOLAP stores this data in an optimized multi-dimensional array storage, rather thanin a relational database. Therefore it requires the pre-computation and storage ofinformation in the cube - the operation known as processing.
- MOLAP tools generally utilize a pre-calculated data set referred to as a datacube.
- The data cube contains all the possible answers to a given range of questions.
- MOLAP tools have a very fast response time and the ability to quickly write backdata into the data set.

Hybrid OLAP (HOLAP):

Hybrid OLAP is a mixture of both ROLAP and MOLAP. It offers fast computation of MOLAP and higher scalability of ROLAP. HOLAP uses two databases.

1. Aggregated or computed data is stored in a multidimensional OLAP cube
2. Detailed information is stored in a relational database.

There is no clear agreement across the industry as to what constitutes HybridOLAP, except that a database will divide data between relational and specializedstorage.

For example, for some vendors, a HOLAP database will use relational tables to holdthe larger quantities of detailed data, and use specialized storage for at least someaspects of the smaller quantities of more-aggregate or less-detailed data.

HOLAP addresses the shortcomings of MOLAP and ROLAP by combining thecapabilities of both approaches.

HOLAP tools can utilize both pre-calculated cubes and relational data sources

Benefits of Hybrid OLAP:

- This kind of OLAP helps to economize the disk space, and it also remains compact which helps to avoid issues related to access speed and convenience.
- Hybrid HOLAP's uses cube technology which allows faster performance for all types of data.
- ROLAP are instantly updated and HOLAP users have access to this real-time instantly updated data. MOLAP brings cleaning and conversion of data thereby improving data relevance. This brings best of both worlds.

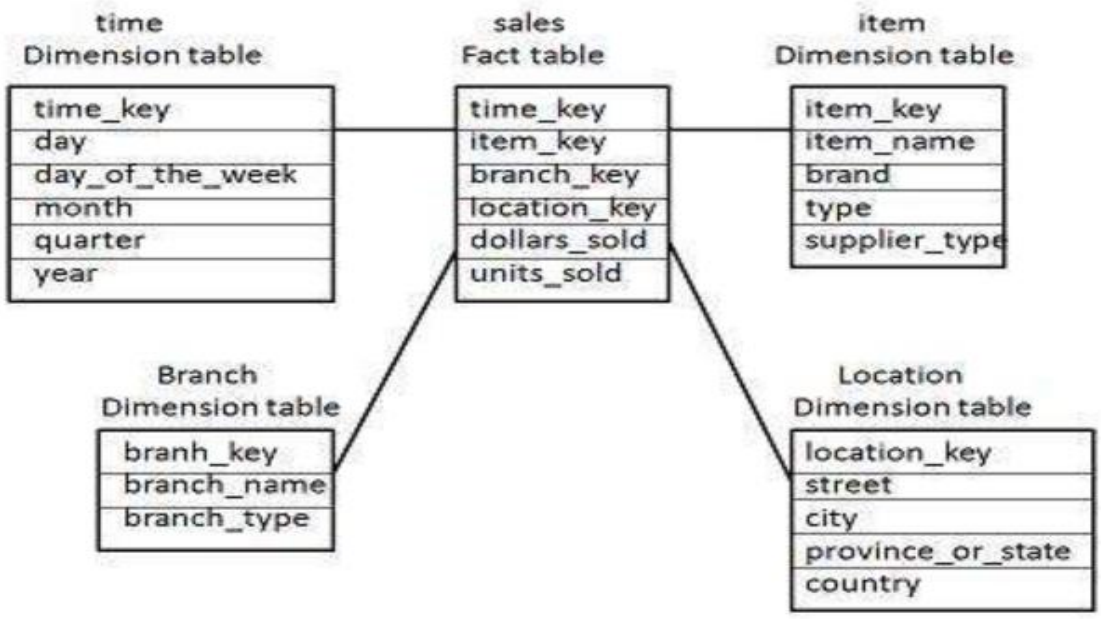
Drawbacks of Hybrid OLAP:

- Greater complexity level**:** The major drawback in HOLAP systems is that it supports both ROLAP and MOLAP tools and applications. Thus, it is very complicated.
- Potential overlaps**:** There are higher chances of overlapping especially into their functionalities.

Defining Schemas:

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location

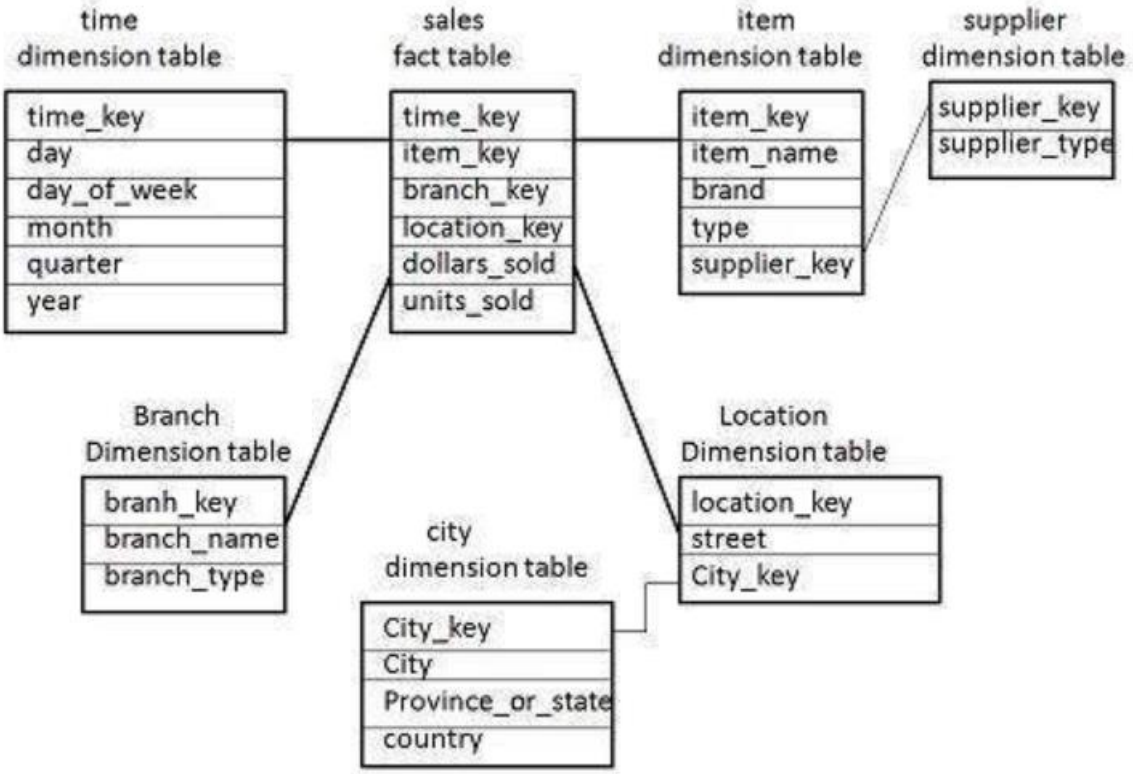


- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note − Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

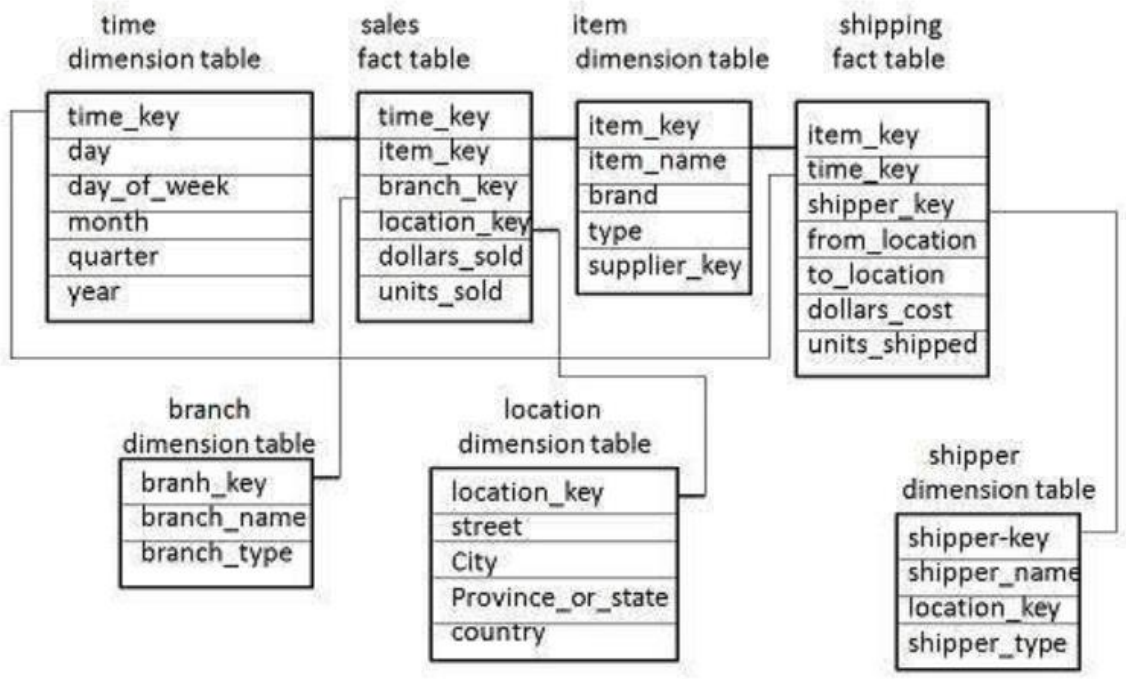


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

Note − Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



- he sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table