OLAP

What is OLAP (Online Analytical Processing)

- **OLAP** stands for **On-Line Analytical Processing**. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.
- **OLAP** implement the multidimensional analysis of business information and support the capability for complex estimations, trend analysis, and sophisticated data modeling. It is rapidly enhancing the essential foundation for Intelligent Solutions containing Business Performance Management, Planning, Budgeting, Forecasting, Financial Documenting, Analysis, Simulation-Models, Knowledge Discovery, and Data Warehouses Reporting.

Who uses OLAP and Why?

OLAP applications are used by a variety of the functions of an organization.

Finance and accounting:

Budgeting
Activity-based costing
Financial performance analysis
And financial modeling

Sales and Marketing

Sales analysis and forecasting
Market research analysis
Promotion +analysis
Customer analysis
Market and customer segmentation

Production

Production planning
Defect analysis

OLAP cubes have two main purposes. The first is to provide business users with a data model more intuitive to them than a tabular model. This model is called a Dimensional Model.

The second purpose is to enable fast query response that is usually difficult to achieve using tabular models.

1) **Multidimensional Conceptual View:** This is the central features of an OLAP system. By needing a multidimensional view, it is possible to carry out methods like slice and dice.

2) Transparency: Make the technology, underlying information repository, computing operations, and the dissimilar nature of source data totally transparent to users. Such transparency helps to improve the efficiency and productivity of the users.

3) Accessibility: It provides access only to the data that is actually required to perform the particular analysis, present a single, coherent, and consistent view to the clients. The OLAP system must map its own logical schema to the heterogeneous physical data stores and perform any necessary transformations. The OLAP operations should be sitting between data sources (e.g., data warehouses) and an OLAP front-end.

4) Consistent Reporting Performance: To make sure that the users do not feel any significant degradation in documenting performance as the number of dimensions or the size of the database increases. That is, the performance of OLAP should not suffer as the number of dimensions is increased. Users must observe consistent run time, response time, or machine utilization every time a given query is run.

5) Client/Server Architecture: Make the server component of OLAP tools sufficiently intelligent that the various clients to be attached with a minimum of effort and integration programming. The server should be capable of mapping and consolidating data between dissimilar databases.

6) Generic Dimensionality: An OLAP method should treat each dimension as equivalent in both is structure and operational capabilities. Additional operational capabilities may be allowed to selected dimensions, but such additional tasks should be grantable to any dimension.

7) Dynamic Sparse Matrix Handling: To adapt the physical schema to the specific analytical model being created and loaded that optimizes sparse matrix handling. When encountering the sparse matrix, the system must be easy to dynamically assume the distribution of the information and adjust the storage and access to obtain and maintain a consistent level of performance.

8) Multiuser Support: OLAP tools must provide concurrent data access, data integrity, and access security.

9) Unrestricted cross-dimensional Operations: It provides the ability for the methods to identify dimensional order and necessarily functions roll-up and drill-down methods within a dimension or across the dimension.

10) Intuitive Data Manipulation: Data Manipulation fundamental the consolidation direction like as reorientation (pivoting), drill-down and roll-up, and another manipulation to be accomplished naturally and precisely via point-and-click and drag and drop methods on the cells of the scientific model. It avoids the use of a menu or multiple trips to a user interface.

11) Flexible Reporting: It implements efficiency to the business clients to organize columns, rows, and cells in a manner that facilitates simple manipulation, analysis, and synthesis of data.

12) Unlimited Dimensions and Aggregation Levels: The number of data dimensions should be unlimited. Each of these common dimensions must allow a practically unlimited number of customer-defined aggregation levels within any given consolidation path.

The main characteristics of OLAP are as follows:

1.Multidimensional conceptual view: OLAP systems let business users have a dimensional and logical view of the data in the data warehouse. It helps in carrying slice and dice operations.

2.Multi-User Support: Since the OLAP techniques are shared, the OLAP operation should provide normal database operations, containing retrieval, update, adequacy control, integrity, and security.

3.Accessibility: OLAP acts as a mediator between data warehouses and front-end. The OLAP operations should be sitting between data sources (e.g., data warehouses) and an OLAP front-end.

4.Storing OLAP results: OLAP results are kept separate from data sources.

5.Uniform documenting performance: Increasing the number of dimensions or database size should not significantly degrade the reporting performance of the OLAP system.
6.OLAP provides for distinguishing between zero values and missing values so that aggregates are computed correctly.
7.OLAP system should ignore all missing values and compute correct aggregate values.
8.OLAP facilitate interactive query and comOLAP allows users to drill down for greater details or roll up for aggregations of metrics along a single business dimension or across multiple dimension.
9.OLAP provides the ability to perform intricate calculations and comparisons.
10.OLAP presents results in a number of meaningful ways, including charts and graphs.
11.plex analysis for the users.

Benefits of OLAP

OLAP holds several benefits for businesses: -

1.OLAP helps managers in decision-making through the multidimensional record views that it is efficient in providing, thus increasing their productivity.

2.OLAP functions are self-sufficient owing to the inherent flexibility support to the organized databases.

3.It facilitates simulation of business models and problems, through extensive management of analysis-capabilities.

4.In conjunction with data warehouse, OLAP can be used to support a reduction in the application backlog, faster data retrieval, and reduction in query drag.

OLAP Operations in the Multidimensional Data Model

In the multidimensional model, the records are organized into various dimensions, and each dimension includes multiple levels of abstraction described by concept hierarchies. This organization support users with the flexibility to view data from various perspectives. A number of OLAP data cube operation exist to demonstrate these different views, allowing interactive queries and search of the record at hand. Hence, OLAP supports a user-friendly environment for interactive data analysis.

Consider the OLAP operations which are to be performed on multidimensional data. The figure shows data cubes for sales of a shop. The cube contains the dimensions, location, and time and item, where the **location** is aggregated with regard to city values, **time** is aggregated with respect to quarters, and an **item** is aggregated with respect to item types.

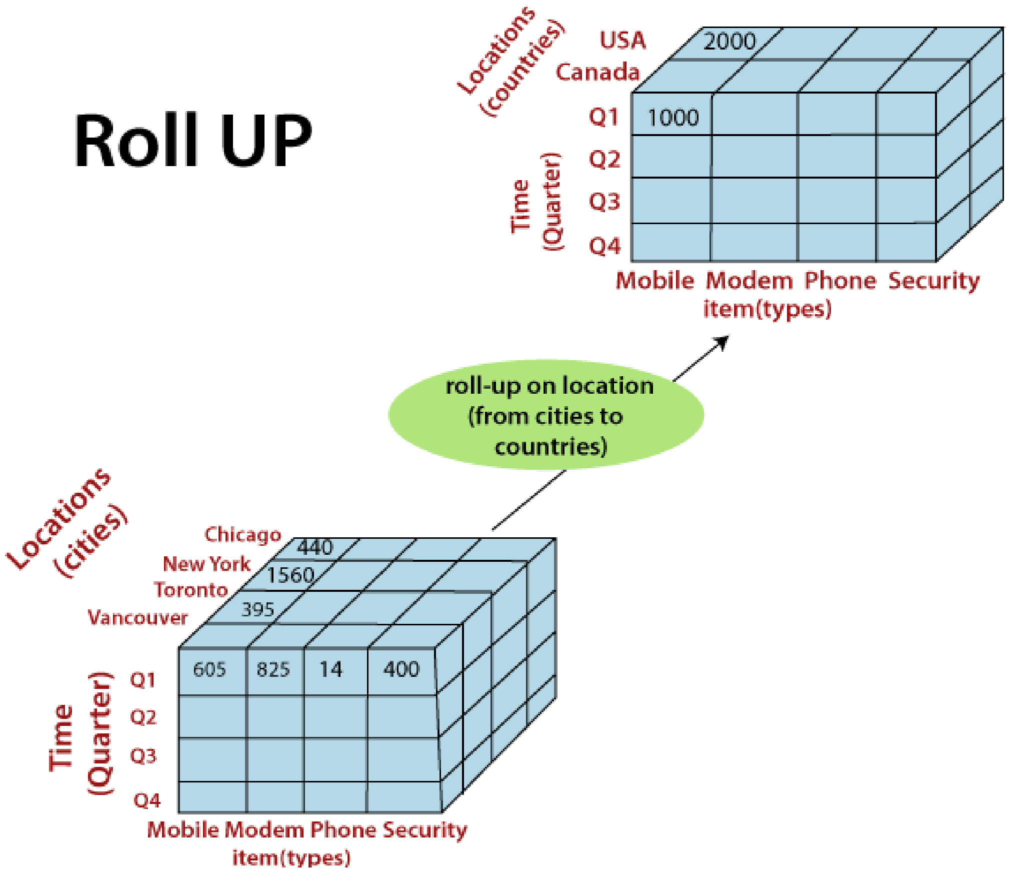
Roll-Up

The roll-up operation **(also known as drill-up or aggregation operation)** performs aggregation on a data cube, by climbing down concept hierarchies, i.e., dimension reduction. Roll-up is like **zooming-out** on the data cubes. Figure shows the result of roll-up operations performed on the dimension location. The hierarchy for the location is defined as the Order Street, city, province, or state, country. The roll-up operation aggregates the data by ascending the location hierarchy from the level of the city to the level of the country.

When a roll-up is performed by dimensions reduction, one or more dimensions are removed from the cube. For example, consider a sales data cube having two dimensions, location and time. Roll-up may be performed by removing, the time dimensions, appearing in an aggregation of the total sales by location, relatively than by location and by time.

The roll-up operation groups the information by levels of temperature.

The following diagram illustrates how roll-up works.

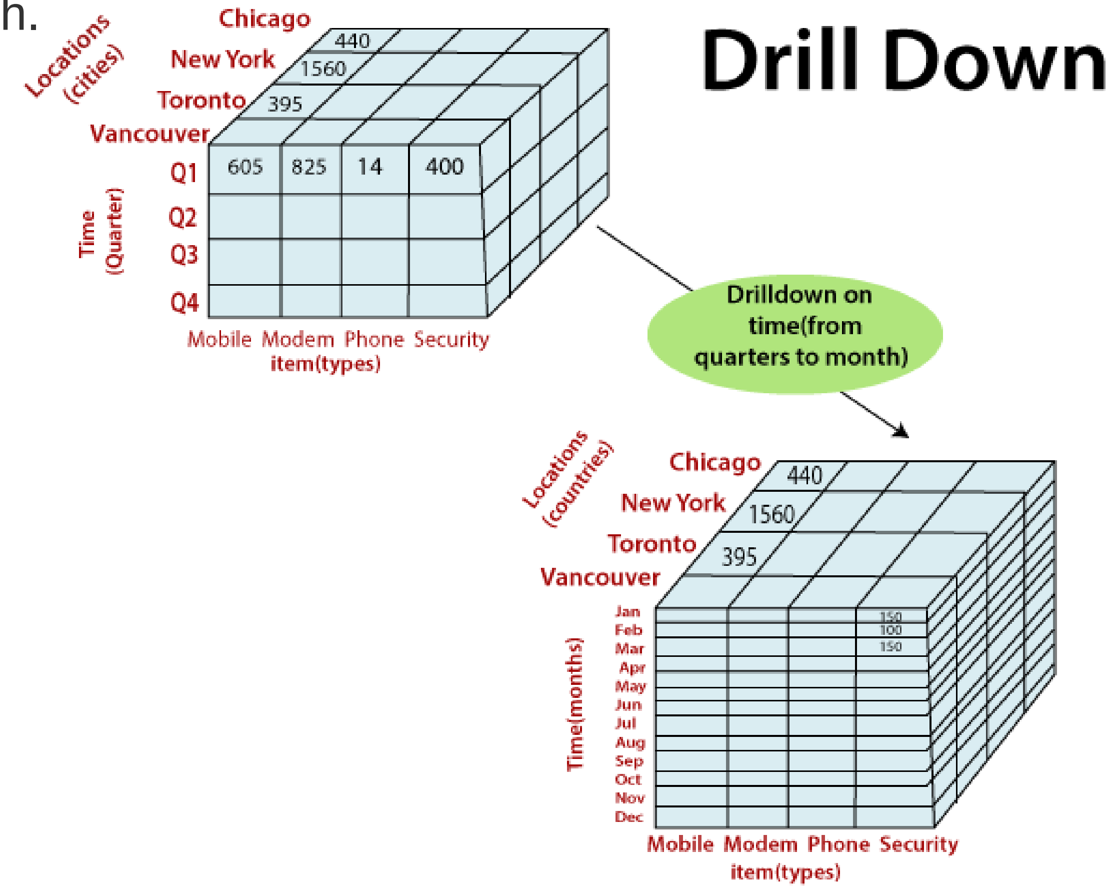


Drill-Down

The drill-down operation **(also called roll-down)** is the reverse operation of **roll-up**. Drill-down is like **zooming-in** on the data cube. It navigates from less detailed record to more detailed data. Drill-down can be performed by either **stepping down** a concept hierarchy for a dimension or adding additional dimensions.

Figure shows a drill-down operation performed on the dimension time by stepping down a concept hierarchy which is defined as day, month, quarter, and year. Drill-down appears by descending the time hierarchy from the level of the quarter to a more detailed level of the month.

The following diagram illustrates how Drill-down works.

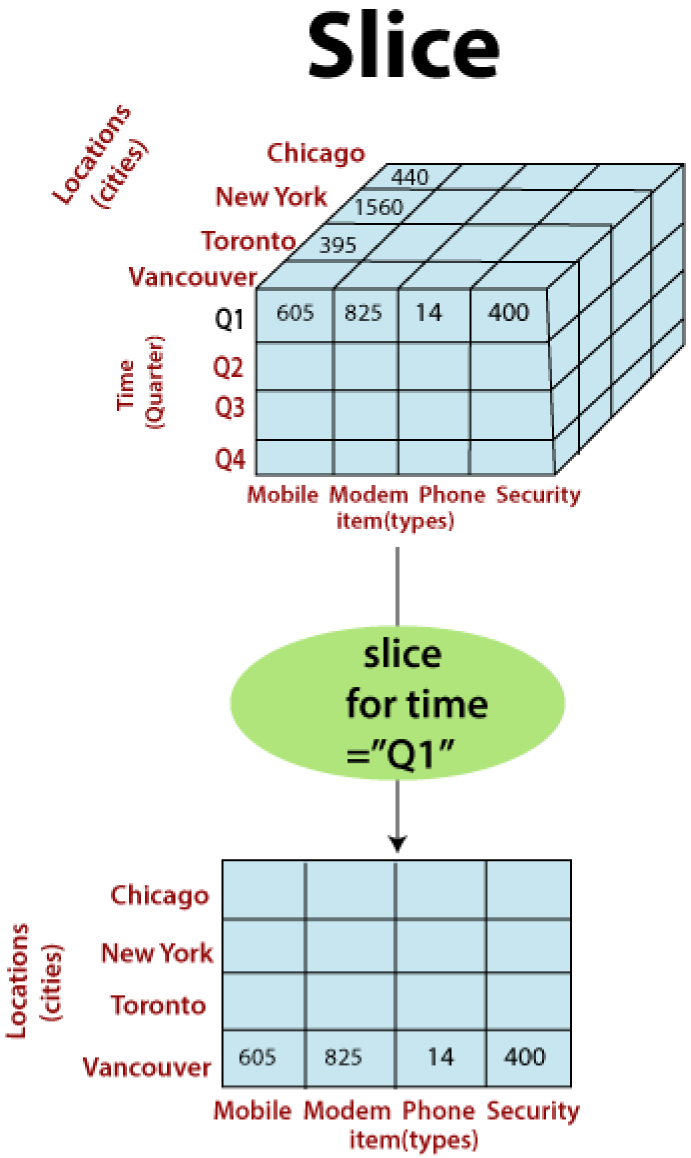


Slice

A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension. For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site. So, the Slice operations perform a selection on one dimension of the given cube, thus resulting in a subcube.

The following diagram illustrates how Slice works.

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



Locations (cities)	Mobile	Modem	Phone	Security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

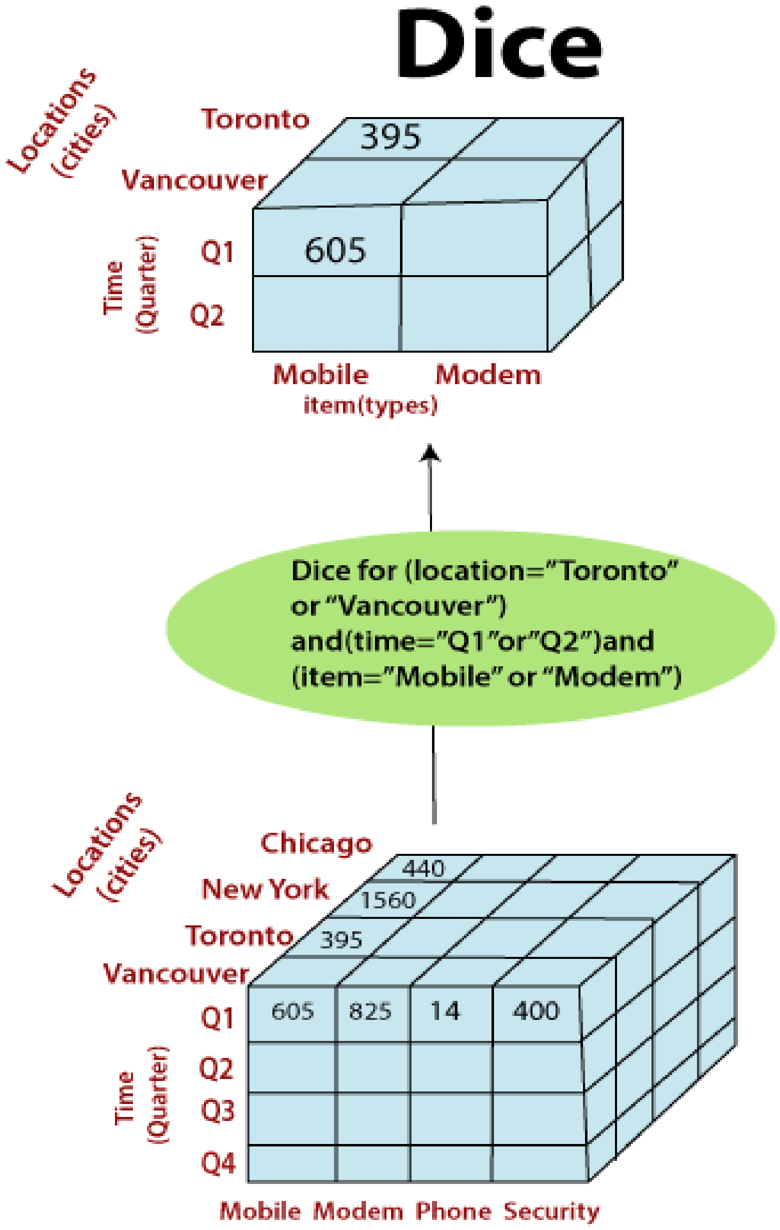
item(types)

Dice

The dice operation describes a subcube by operating a selection on two or more dimension.
Consider the following diagram, which shows the dice operations.

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

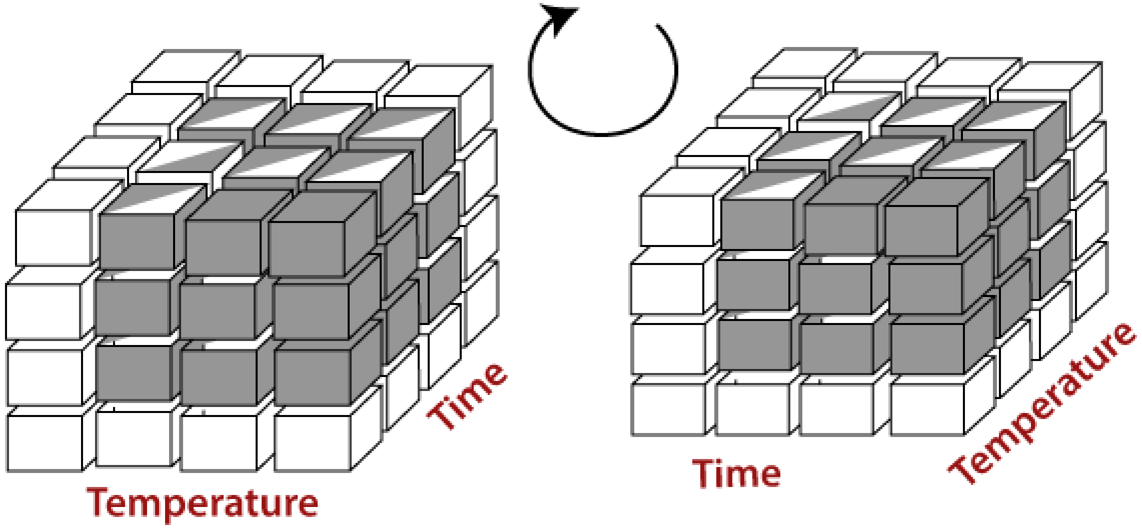
(location = "Toronto" or "Vancouver")
(time = "Q1" or "Q2")
(item =" Mobile" or "Modem")

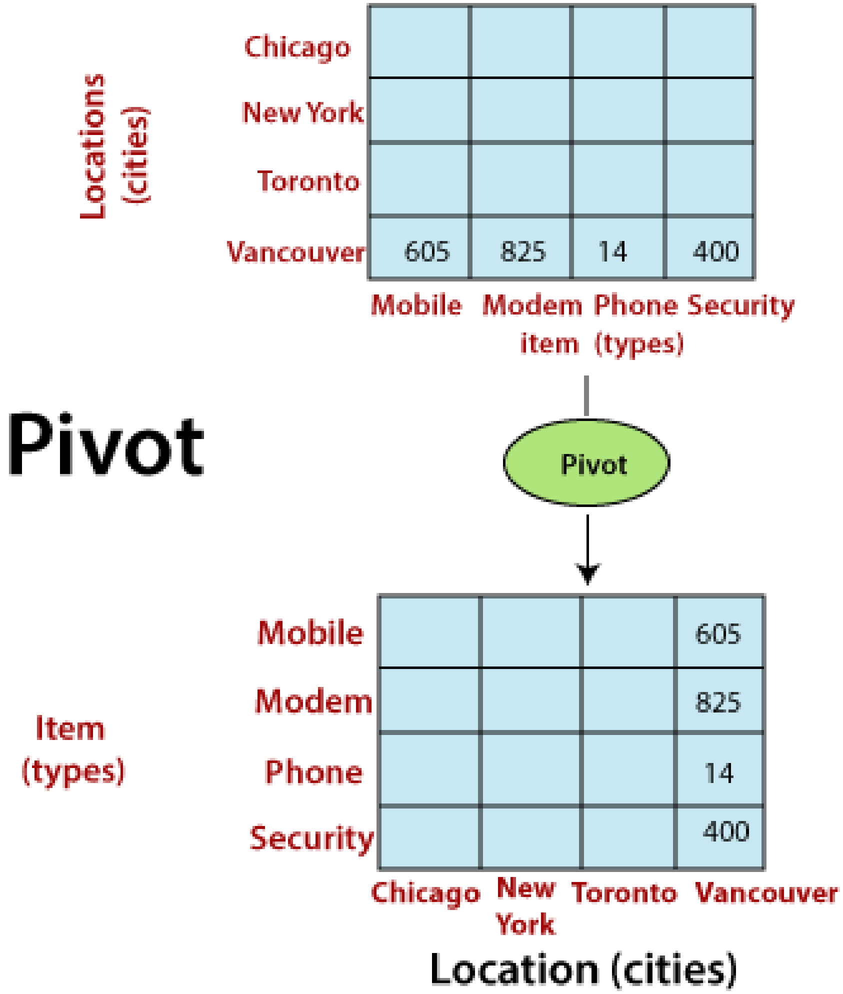


Dice for (location="Toronto" or "Vancouver") and(time="Q1"or"Q2")and (item="Mobile" or "Modem")

Pivot

The pivot operation is also called a rotation. Pivot is a visualization operations which rotates the data axes in view to provide an alternative presentation of the data. It may contain swapping the rows and columns or moving one of the row-dimensions into the column dimensions. Consider the following diagram, which shows the pivot operation.





	Mobile	Modem	Phone	Security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

	Chicago	New York	Toronto	Vancouver
Mobile				605
Modem				825
Phone				14
Security				400

Types of OLAP

There are three main types of OLAP servers are as following:

ROLAP stands for Relational OLAP, an application based on relational DBMSs.

MOLAP stands for Multidimensional OLAP, an application based on multidimensional DBMSs.

HOLAP stands for Hybrid OLAP, an application using both relational and multidimensional techniques

Relational OLAP (ROLAP) Server

These are intermediate servers which stand in between a relational back-end server and user frontend tools.

They use a relational or extended-relational DBMS to save and handle warehouse data, and OLAP middleware to provide missing pieces.

ROLAP servers contain optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.

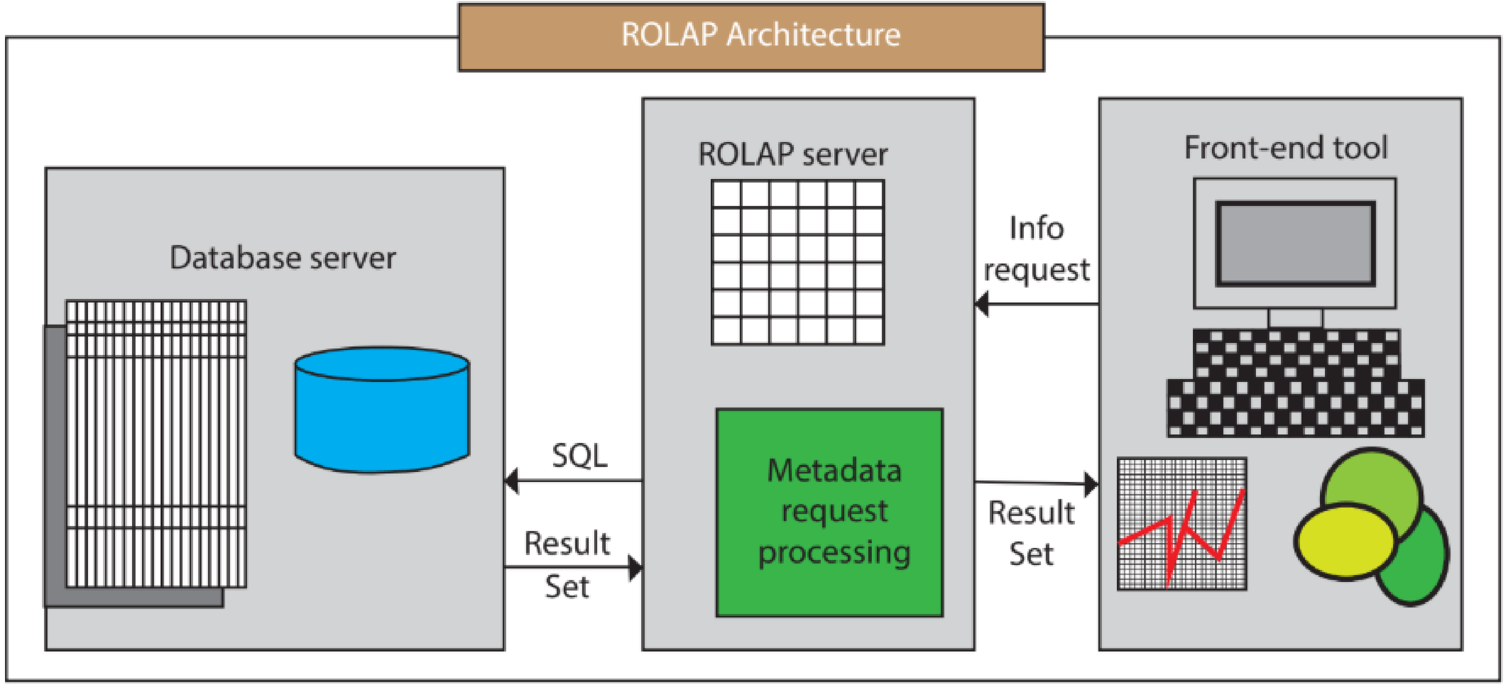
ROLAP technology tends to have higher scalability than MOLAP technology.

ROLAP systems work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables. This model permits the multidimensional analysis of data.

Relational OLAP Architecture

ROLAP Architecture includes the following components

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



Relational OLAP (ROLAP) is the latest and fastest-growing OLAP technology segment in the market. This method allows multiple multidimensional views of two-dimensional relational tables to be created, avoiding structuring record around the desired view.

Some products in this segment have supported reliable SQL engines to help the complexity of multidimensional analysis. This includes creating multiple SQL statements to handle user requests, being 'RDBMS' aware and also being capable of generating the SQL statements based on the optimizer of the DBMS engine.

Advantages

Can handle large amounts of information: The data size limitation of ROLAP technology is depends on the data size of the underlying RDBMS. So, ROLAP itself does not restrict the data amount.

<="" strong="">RDBMS already comes with a lot of features. So ROLAP technologies, (works on top of the RDBMS) can control these functionalities.

Disadvantages

Performance can be slow: Each ROLAP report is a SQL query (or multiple SQL queries) in the relational database, the query time can be prolonged if the underlying data size is large.

Limited by SQL functionalities: ROLAP technology relies on upon developing SQL statements to query the relational database, and SQL statements do not suit all needs.

Multidimensional OLAP (MOLAP) Server

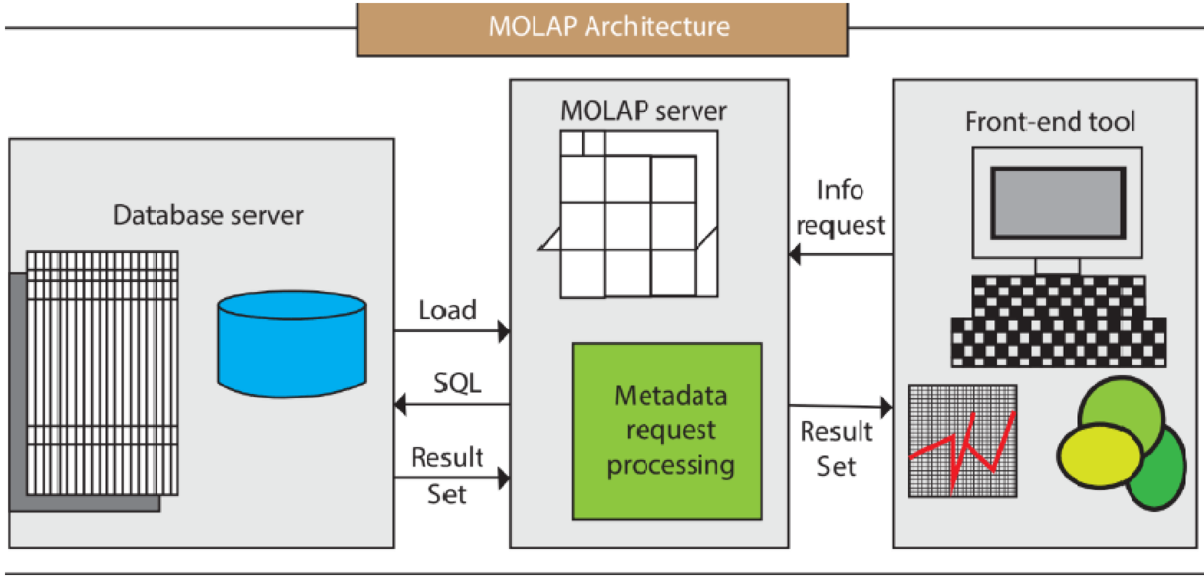
A MOLAP system is based on a native logical model that directly supports multidimensional data and operations. Data are stored physically into multidimensional arrays, and positional techniques are used to access them.

One of the significant distinctions of **MOLAP** against a **ROLAP** is that data are summarized and are stored in an optimized format in a multidimensional cube, instead of in a relational database. In MOLAP model, data are structured into proprietary formats by client's reporting requirements with the calculations pre-generated on the cubes.

MOLAP Architecture

MOLAP Architecture includes the following components

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



MOLAP structure primarily reads the precompiled data. MOLAP structure has limited capabilities to dynamically create aggregations or to evaluate results which have not been pre-calculated and stored. Applications requiring iterative and comprehensive time-series analysis of trends are well suited for MOLAP technology (e.g., financial analysis and budgeting).

Advantages

Excellent Performance: A MOLAP cube is built for fast information retrieval, and is optimal for slicing and dicing operations.

Can perform complex calculations: All evaluation have been pre-generated when the cube is created. Hence, complex calculations are not only possible, but they return quickly.

Disadvantages

Limited in the amount of information it can handle: Because all calculations are performed when the cube is built, it is not possible to contain a large amount of data in the cube itself.

Requires additional investment: Cube technology is generally proprietary and does not already exist in the organization. Therefore, to adopt MOLAP technology, chances are other investments in human and capital resources are needed.

Hybrid OLAP (HOLAP) Server

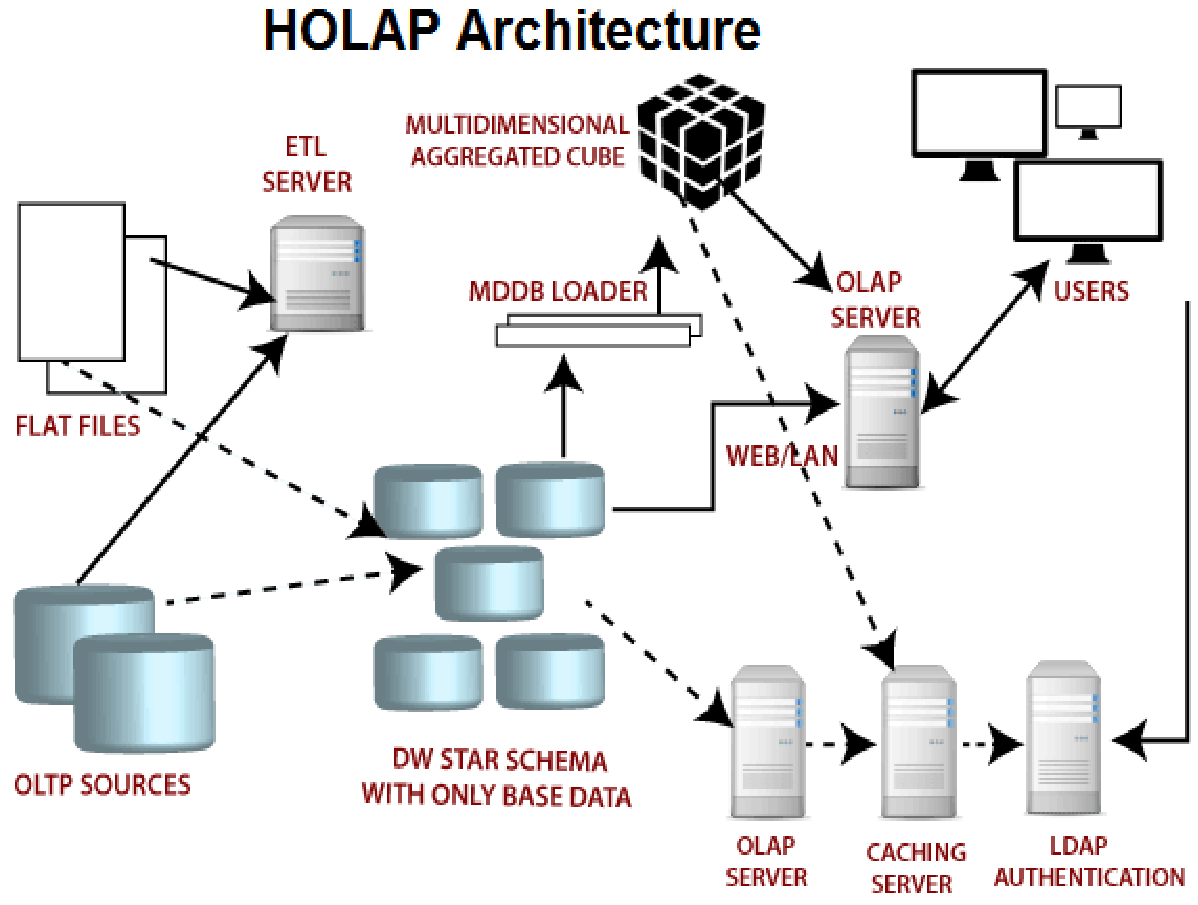
HOLAP incorporates the best features of **MOLAP** and **ROLAP** into a single architecture. HOLAP systems save more substantial quantities of detailed data in the relational tables while the aggregations are stored in the pre-calculated cubes. HOLAP also can drill through from the cube down to the relational tables for delineated data. The **Microsoft SQL Server 2000** provides a hybrid OLAP server.

Advantages of HOLAP

HOLAP provide benefits of both MOLAP and ROLAP.
It provides fast access at all levels of aggregation.
HOLAP balances the disk space requirement, as it only stores the aggregate information on the OLAP server and the detail record remains in the relational database. So no duplicate copy of the detail record is maintained.

Disadvantages of HOLAP

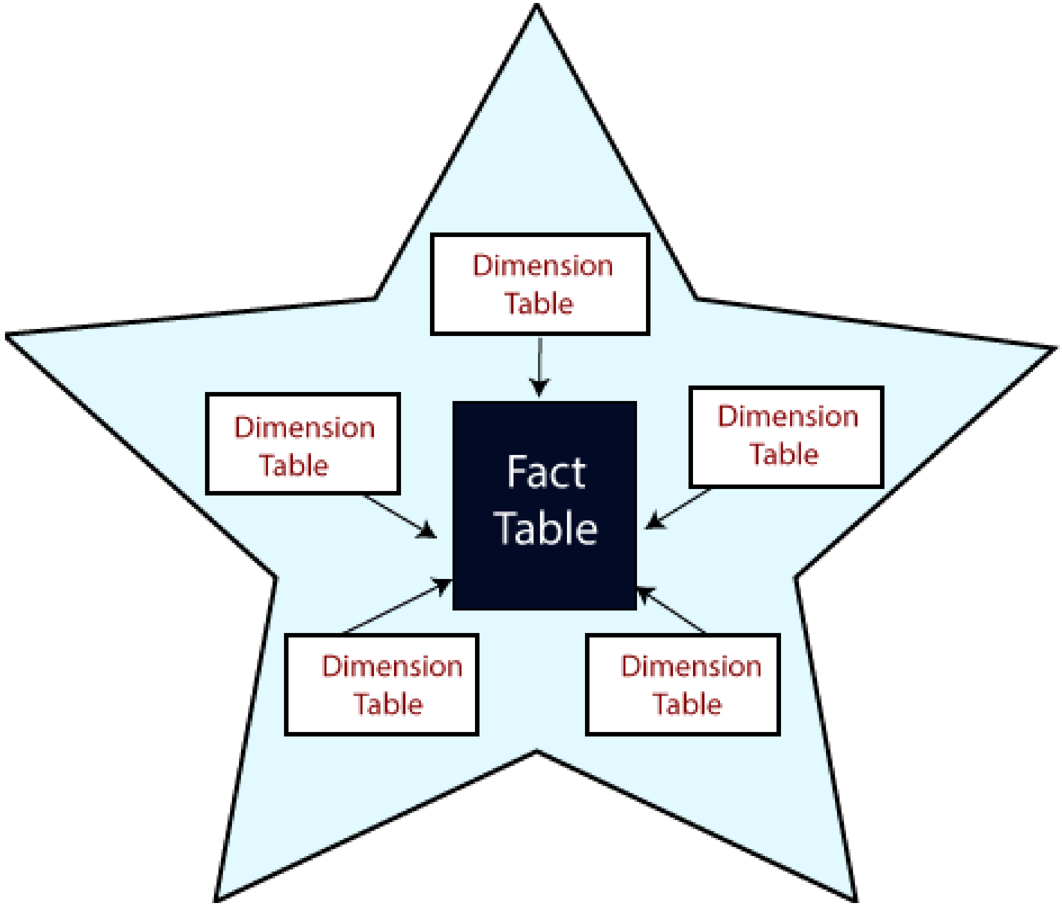
HOLAP architecture is very complicated because it supports both MOLAP and ROLAP servers.



ROLAP	MOLAP	HOLAP
ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.	HOLAP stands for Hybrid Online Analytical Processing.
The ROLAP storage mode causes the aggregation of the division to be stored in indexed views in the relational database that was specified in the partition's data source.	The MOLAP storage mode principle the aggregations of the division and a copy of its source information to be saved in a multidimensional operation in analysis services when the separation is processed.	The HOLAP storage mode connects attributes of both MOLAP and ROLAP. Like MOLAP, HOLAP causes the aggregation of the division to be stored in a multidimensional operation in an SQL Server analysis services instance.
ROLAP does not because a copy of the source information to be stored in the Analysis services data folders. Instead, when the outcome cannot be derived from the query cache, the indexed views in the record source are accessed to answer queries.	This MOLAP operation is highly optimize to maximize query performance. The storage area can be on the computer where the partition is described or on another computer running Analysis services. Because a copy of the source information resides in the multidimensional operation, queries can be resolved without accessing the partition's source record.	HOLAP does not causes a copy of the source information to be stored. For queries that access the only summary record in the aggregations of a division, HOLAP is the equivalent of MOLAP.
Query response is frequently slower with ROLAP storage than with the MOLAP or HOLAP storage mode. Processing time is also frequently slower with ROLAP.	Query response times can be reduced substantially by using aggregations. The record in the partition's MOLAP operation is only as current as of the most recent processing of the separation.	Queries that access source record for example, if we want to drill down to an atomic cube cell for which there is no aggregation information must retrieve data from the relational database and will not be as fast as they would be if the source information were stored in the MOLAP architecture.

Star Schema

- A star schema is the elementary form of a dimensional model, in which data are organized into **facts** and **dimensions**.
- A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.
- A star schema is a relational schema where a relational schema whose design represents a multidimensional data model.
- The star schema is the explicit data warehouse schema. It is known as **star schema** because the entity-relationship diagram of this schemas simulates a star, with points, diverge from a central table.
- The center of the schema consists of a large fact table, and the points of the star are the dimension tables.



Star Schema

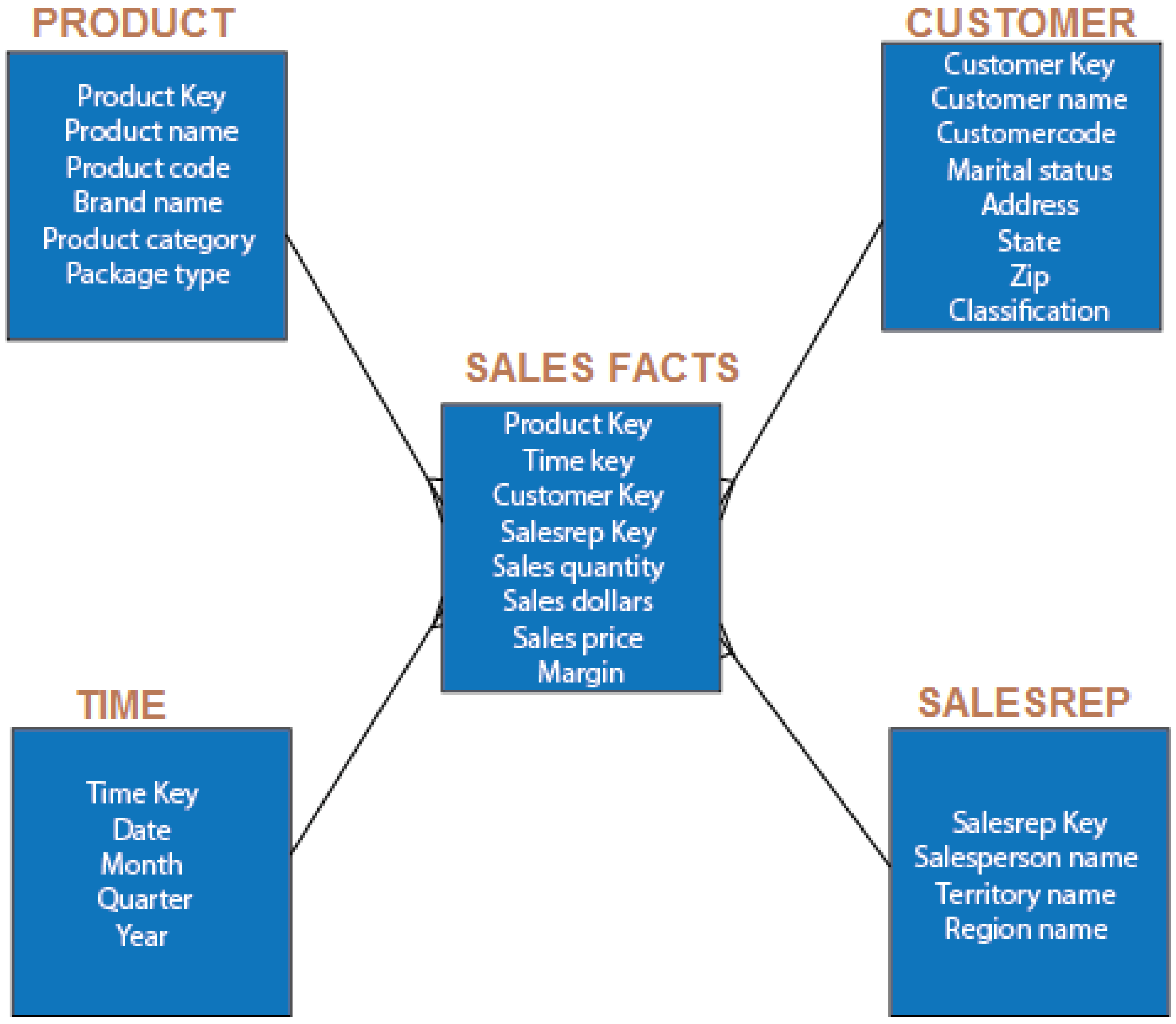
Figure shows a simple STAR schema for sales in a manufacturing company.

The sales fact table include quantity, price, and other relevant metrics. SALESREP, CUSTOMER, PRODUCT, and TIME are the dimension tables.

The STAR schema for sales, as shown above, contains only five tables, whereas the normalized version now extends to eleven tables.

We will notice that in the snowflake schema, the attributes with low cardinality in each original dimension tables are removed to form separate tables.

These new tables are connected back to the original dimension table through artificial keys.



STAR Schema

Fact Tables

- A table in a star schema which contains facts and connected to dimensions. A fact table has two types of columns: those that include fact and those that are foreign keys to the dimension table.
- The primary key of the fact tables is generally a composite key that is made up of all of its foreign keys.
- A fact table might involve either detail level fact or fact that have been aggregated (fact tables that include aggregated fact are often instead called summary tables).
- A fact table generally contains facts with the same level of aggregation.

Dimension Tables

- A dimension is an architecture usually composed of one or more hierarchies that categorize data.
- If a dimension has not got hierarchies and levels, it is called a **flat dimension** or **list**.
- The primary keys of each of the dimensions table are part of the composite primary keys of the fact table.
- Dimensional attributes help to define the dimensional value.
- They are generally descriptive, textual values.
- Dimensional tables are usually small in size than fact table.
- Fact tables store data about sales while dimension tables data about the geographic region (markets, cities), clients, products, times, channels.

Characteristics of Star Schema

- The star schema is intensely suitable for data warehouse database design because of the following features:
- It creates a DE-normalized database that can quickly provide query responses.
- It provides a flexible design that can be changed easily or added to throughout the development cycle, and as the database grows.
- It provides a parallel in design to how end-users typically think of and use the data.
- It reduces the complexity of metadata for both developers and end-users.

Advantages of Star Schema

- Star Schemas are easy for end-users and application to understand and navigate.
- With a well-designed schema, the customer can instantly analyze large, multidimensional data sets.

The main advantage of star schemas in a decision-support environment are:

<u>Query Performance</u>

- A star schema database has a limited number of table and clear join paths, the query run faster than they do against OLTP systems. Small single-table queries, frequently of a dimension table, are almost instantaneous.
- Large join queries that contain multiple tables takes only seconds or minutes to run.
- In a star schema database design, the dimension is connected only through the central fact table.
- When the two-dimension table is used in a query, only one join path, intersecting the fact tables, exist between those two tables.
- This design feature enforces authentic and consistent query results.

Load performance and administration

Structural simplicity also decreases the time required to load large batches of record into a star schema database.

By describing facts and dimensions and separating them into the various table, the impact of a load structure is reduced.

Dimension table can be populated once and occasionally refreshed.

We can add new facts regularly and selectively by appending records to a fact table.

Built-in referential integrity

A star schema has referential integrity built-in when information is loaded.

Referential integrity is enforced because each data in dimensional tables has a unique primary key, and all keys in the fact table are legitimate foreign keys drawn from the dimension table.

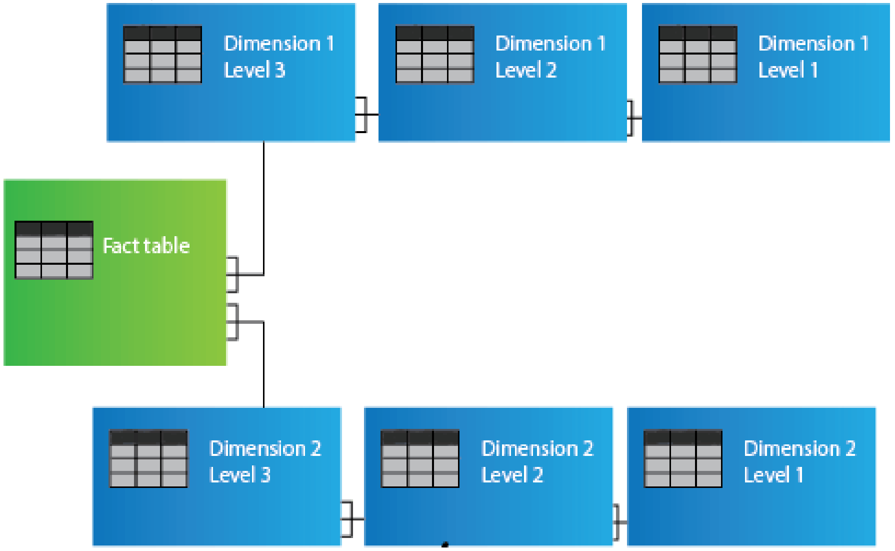
A record in the fact table which is not related correctly to a dimension cannot be given the correct key value to be retrieved.

Easily Understood

- A star schema is simple to understand and navigate, with dimensions joined only through the fact table.
- These joins are more significant to the end-user because they represent the fundamental relationship between parts of the underlying business.
- Customer can also browse dimension table attributes before constructing a query.

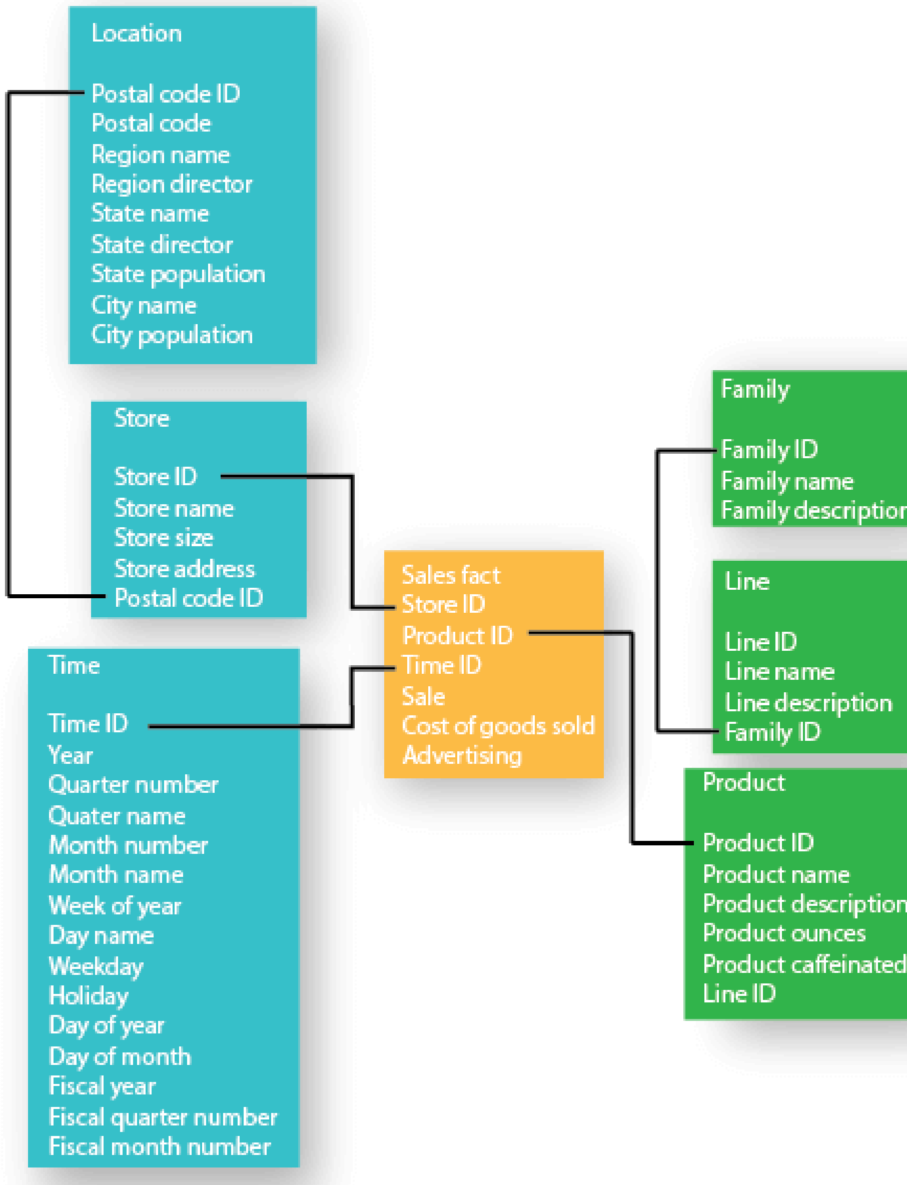
Snowflake Schema

- A snowflake schema is equivalent to the star schema. "A schema is known as a snowflake if one or more dimension tables do not connect directly to the fact table but must join through other dimension tables."
- The snowflake schema is an expansion of the star schema where each point of the star explodes into more points. It is called snowflake schema because the diagram of snowflake schema resembles a snowflake.
- **Snowflaking** is a method of normalizing the dimension tables in a STAR schemas.
- When we normalize all the dimension tables entirely, the resultant structure resembles a snowflake with the fact table in the middle. Snowflaking is used to develop the performance of specific queries.
- The schema is diagramed with each fact surrounded by its associated dimensions, and those dimensions are related to other dimensions, branching out into a snowflake pattern.
- The snowflake schema consists of one fact table which is linked to many dimension tables,
- which can be linked to other dimension tables through a many-to-one relationshi
- Tables in a snowflake schema are generally normalized to the third normal form.
- Each dimension table performs exactly one level in a hierarchy.
- The following diagram shows a snowflake schema with two dimensions, each having three levels. A snowflake schemas can have any number of dimension, and each dimension can have any number of levels.



Snowflake Schema

- **Example:** Figure shows a snowflake schema with a Sales fact table, with Store, Location, Time, Product, Line, and Family dimension tables. The Market dimension has two dimension tables with Store as the primary dimension table, and Location as the outrigger dimension table. The product dimension has three dimension tables with Product as the primary dimension table, and the Line and Family table are the outrigger dimension tables.
- A star schema store all attributes for a dimension into one denormalized table. This needed more disk space than a more normalized snowflake schema. Snowflaking normalizes the dimension by moving attributes with low cardinality into separate dimension tables that relate to the core dimension table by using foreign keys. Snowflaking for the sole purpose of minimizing disk space is not recommended, because it can adversely impact query performance.
- In snowflake, schema tables are normalized to delete redundancy. In snowflake dimension tables are damaged into multiple dimension tables.



Location

Postal code ID
Postal code
Region name
Region director
State name
State director
State population
City name
City population

Store

Store ID
Store name
Store size
Store address
Postal code ID

Time

Time ID
Year
Quarter number
Quater name
Month number
Month name
Week of year
Day name
Weekday
Holiday
Day of year
Day of month
Fiscal year
Fiscal quarter number
Fiscal month number

Sales fact
Store ID
Product ID
Time ID
Sale
Cost of goods sold
Advertising

Family

Family ID
Family name
Family description

Line

Line ID
Line name
Line description
Family ID

Product

Product ID
Product name
Product description
Product ounces
Product caffeinated
Line ID

Advantage of Snowflake Schema

1.The primary advantage of the snowflake schema is the development in query performance due to minimized disk storage requirements and joining smaller lookup tables.

2.It provides greater scalability in the interrelationship between dimension levels and components.

3.No redundancy, so it is easier to maintain.

Disadvantage of Snowflake Schema

1.The primary disadvantage of the snowflake schema is the additional maintenance efforts required due to the increasing number of lookup tables. It is also known as a multi fact star schema.

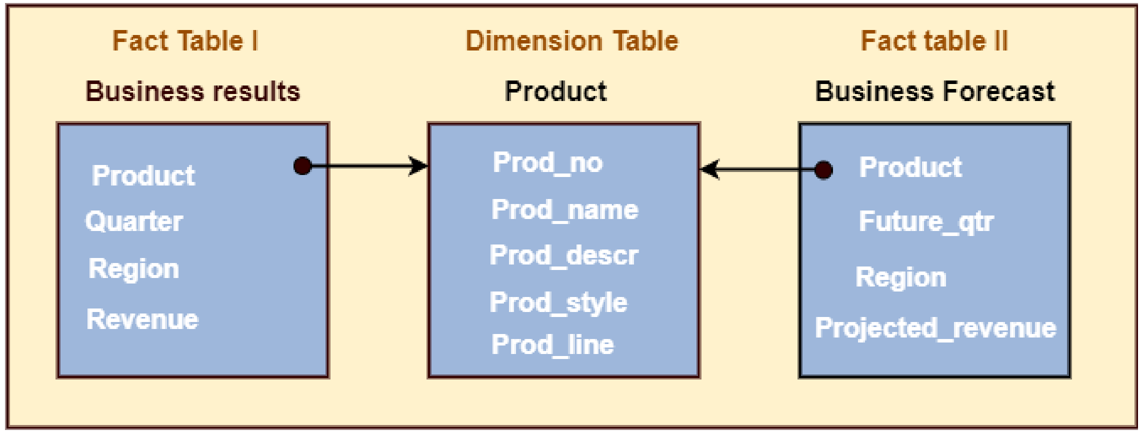
2.There are more complex queries and hence, difficult to understand.

3.More tables more join so more query execution time.

Fact Constellation Schema

A Fact constellation means two or more fact tables sharing one or more dimensions. It is also called **Galaxy schema**.

Fact Constellation Schema describes a logical structure of data warehouse or data mart. Fact Constellation Schema can design with a collection of de-normalized FACT, Shared, and Conformed Dimension tables.



FACT Constellation Schema

Fact Constellation Schema is a sophisticated database design that is difficult to summarize information.

Fact Constellation Schema can implement between aggregate Fact tables or decompose a complex Fact table into independent simplex Fact tables.