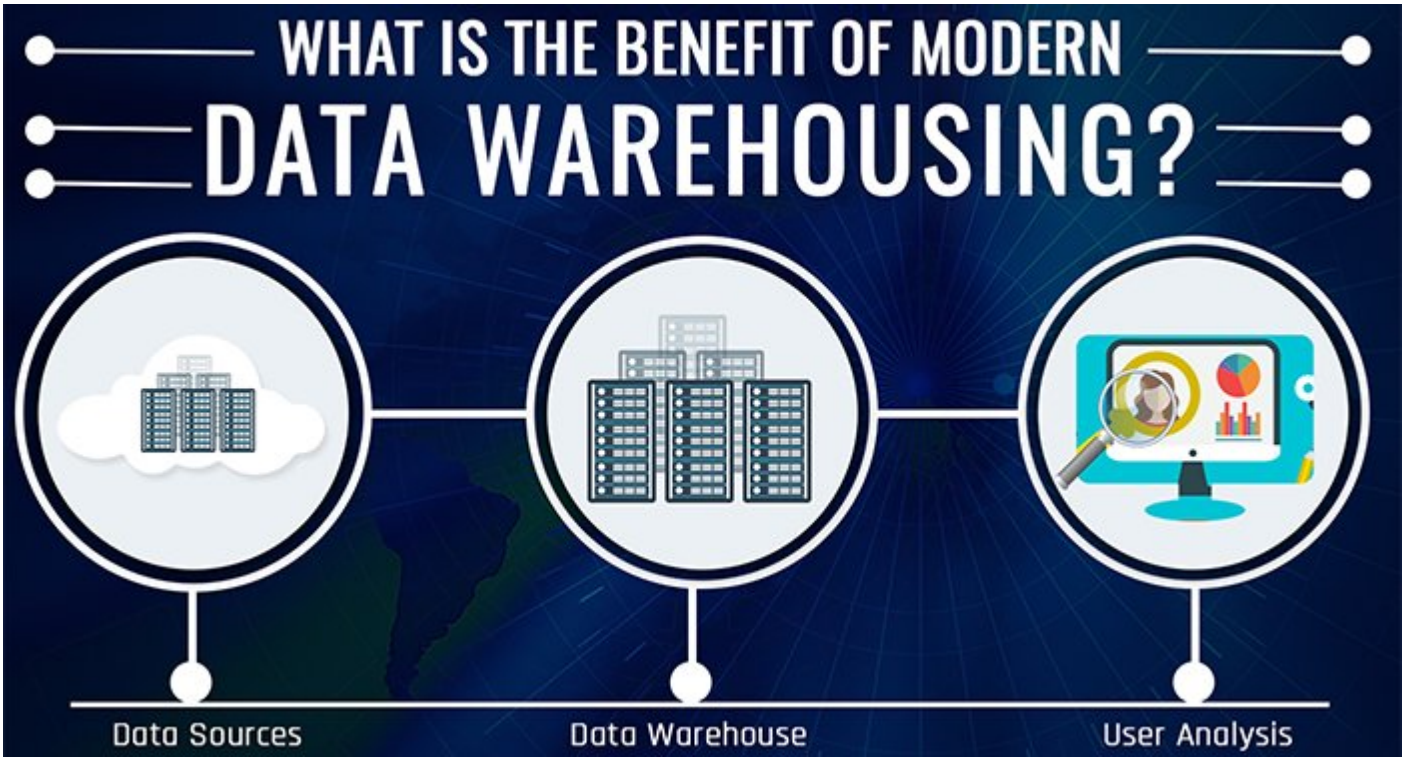
Road to Data Science in 50 Days - Day 8

Databases for Data Science



Data Warehousing

What is Data Warehousing?

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Using Data Warehouse Information

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains −

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.
- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

Data Extraction − Involves gathering data from multiple heterogeneous sources.

Data Cleaning − Involves finding and correcting the errors in data.

Data Transformation − Involves converting the data from legacy format to warehouse format.

Data Loading − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.

Refreshing − Involves updating from data sources to warehouse.

Data Warehousing Terminologies

Metadata

Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

In terms of data warehouse, we can define metadata as following −

- Metadata is a road-map to data warehouse.
- Metadata in data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Metadata Repository

Metadata repository is an integral part of a data warehouse system. It contains the following metadata −

- **Business metadata** − It contains the data ownership information, business definition, and changing policies.
- **Operational metadata** − It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** − It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
- **The algorithms for summarization** − It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

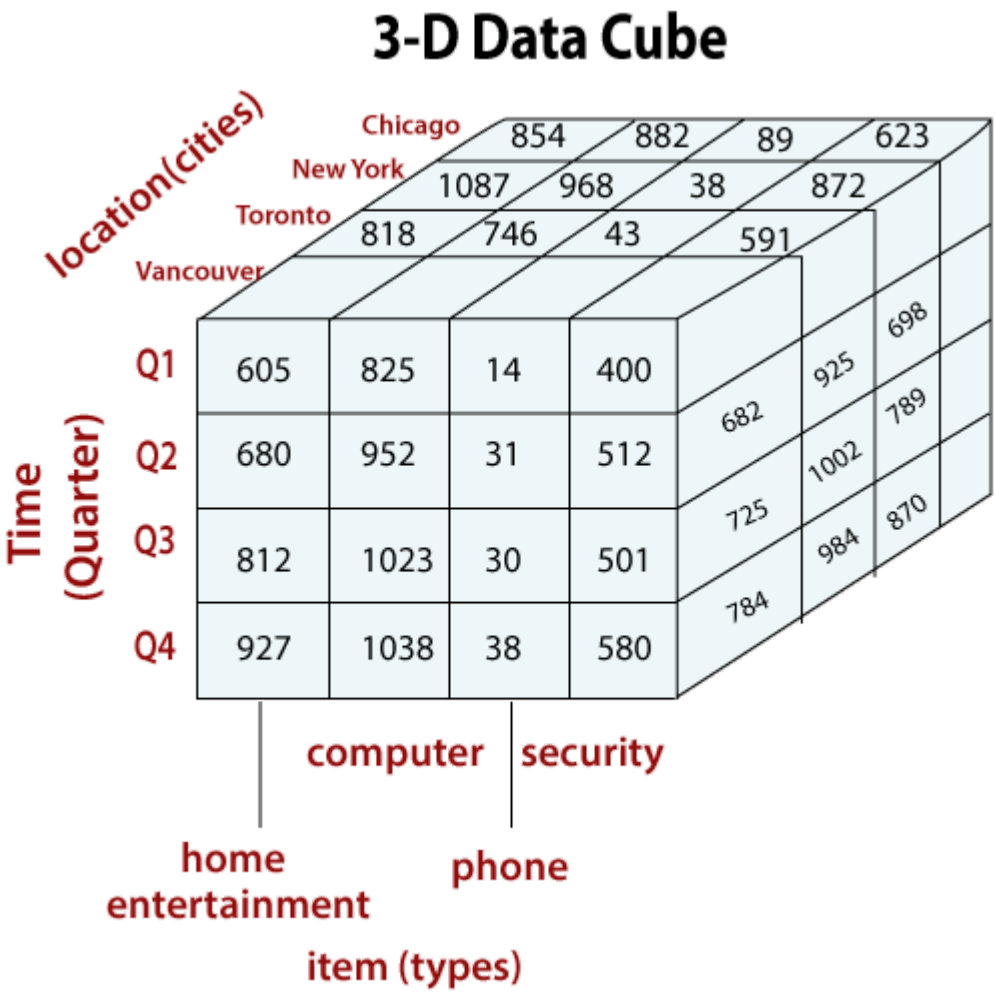
Data Cube

A data cube helps us represent data in multiple dimensions. It is defined by dimensions and facts. The dimensions are the entities with respect to which an enterprise preserves the records.

3-D view of Sales Data

location ="Chicago"					**location** ="New York"				**location** ="Toronto"			
item					item				item			
	home				home				home			
time	ent.	comp.	phone	sec.	time	comp.	phone	sec.	ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784

Conceptually, we may represent the same data in the form of 3-D data cubes, as shown in fig:



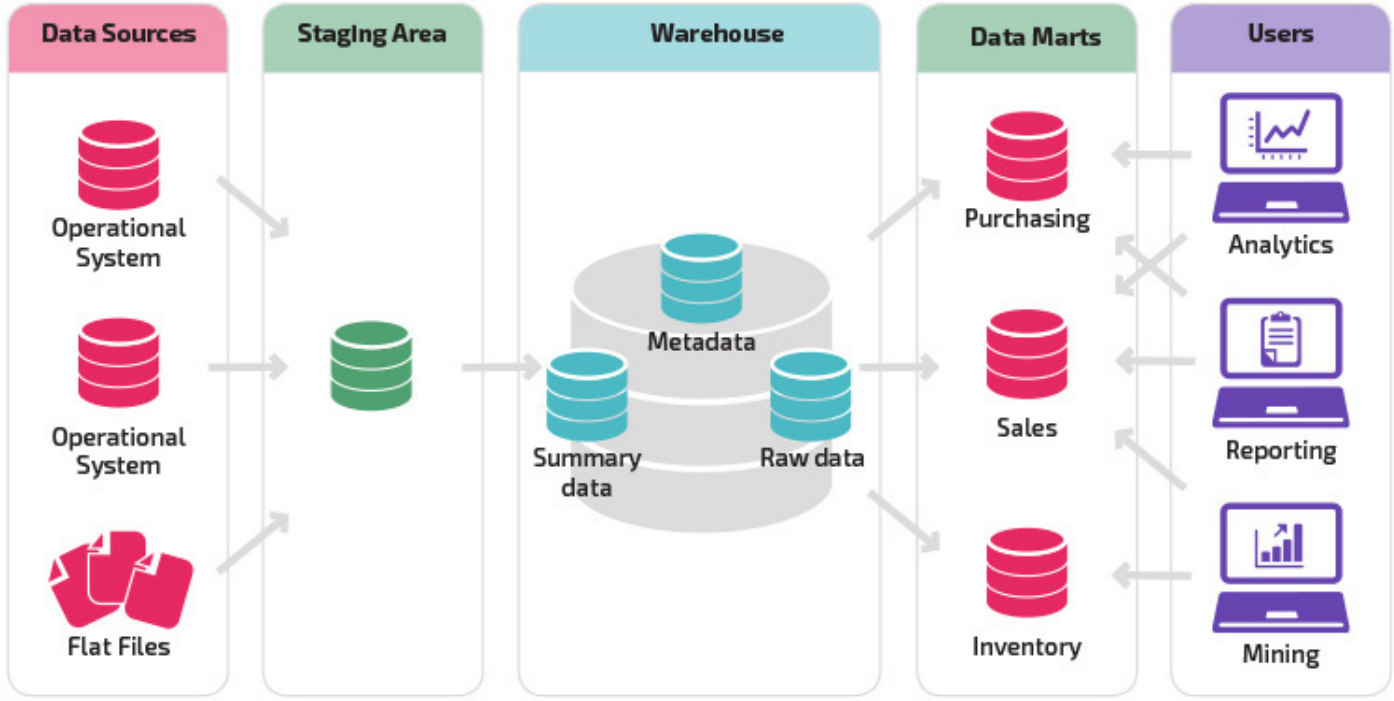
Data Mart

Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization. In other words, a data mart contains only those data that is specific to a particular group. For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.

Points to Remember About Data Marts

- Windows-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation cycle of a data mart is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of data marts may be complex in the long run, if their planning and design are not organization-wide.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data marts are flexible.

The following figure shows a graphical representation of data marts.



Data Mart vs. Data Warehouse

A data mart is a subset of a data warehouse oriented to a specific business line. Data marts contain repositories of summarized data collected for analysis on a specific section or unit within an organization, for example, the sales department.

A data warehouse is a large centralized repository of data that contains information from many sources within an organization. The collated data is used to guide business decisions through analysis, reporting, and data mining tools.

Data Marts Use Cases

Marketing analysis and reporting favor a data mart approach because these activities are typically performed in a specialized business unit, and do not require enterprise-wide data. A financial analyst can use a finance data mart to carry out financial reporting.

Virtual Warehouse

The view over an operational data warehouse is known as virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.

Delivery Process

IT Strategy: DWH project must contain IT strategy for procuring and retaining funding.

Business Case Analysis: After the IT strategy has been designed, the next step is the business case. It is essential to understand the level of investment that can be justified and to recognize the projected business benefits which should be derived from using the data warehouse.

Education & Prototyping: Company will experiment with the ideas of data analysis and educate themselves on the value of the data warehouse. This is valuable and should be required if this is the company first exposure to the benefits of the DS record. Prototyping method can progress the growth of education. It is better than working models. Prototyping requires business requirement, technical blueprint, and structures.

Business Requirement: It contains such as

- The logical model for data within the data warehouse.
- The source system that provides this data (mapping rules)
- The business rules to be applied to information.
- The query profiles for the immediate requirement

Technical blueprint: It arranges the architecture of the warehouse. Technical blueprint of the delivery process makes an architecture plan which satisfies long-term requirements. It lays server and data mart architecture and essential components of database design.

Building the vision: It is the phase where the first production deliverable is produced. This stage will probably create significant infrastructure elements for extracting and loading information but limit them to the extraction and load of information sources.

History Load: The next step is one where the remainder of the required history is loaded into the data warehouse. This means that the new entities would not be added to the data warehouse, but additional physical tables would probably be created to save the increased record volumes.

AD-Hoc Query: In this step, we configure an ad-hoc query tool to operate against the data warehouse.

These end-customer access tools are capable of automatically generating the database query that answers any question posed by the user.

Automation: The automation phase is where many of the operational management processes are fully automated within the DWH. These would include:

Extracting & loading the data from a variety of sources systems

Transforming the information into a form suitable for analysis

Backing up, restoring & archiving data

Generating aggregations from predefined definitions within the Data Warehouse.

Monitoring query profiles & determining the appropriate aggregates to maintain system performance.

Extending Scope: In this phase, the scope of DWH is extended to address a new set of business requirements. This involves the loading of additional data sources into the DWH i.e. the introduction of new data marts.

Requirement Evolution: This is the last step of the delivery process of a data warehouse. As we all know that requirements are not static and evolve continuously. As the business requirements will change it supports to be reflected in the system.

Concept Hierarchy

Concept hierarchy is directed acyclic graph of ideas, where a unique name identifies each of the theories.

An arc from the concept a to b denotes which is a more general concept than b. We can tag the text with ideas.

Each text report is tagged by a set of concepts which corresponds to its content.

Tagging a report with a concept implicitly entails its tagging with all the ancestors of the concept hierarchy. It is, therefore desired that a report should be tagged with the lowest concept possible.

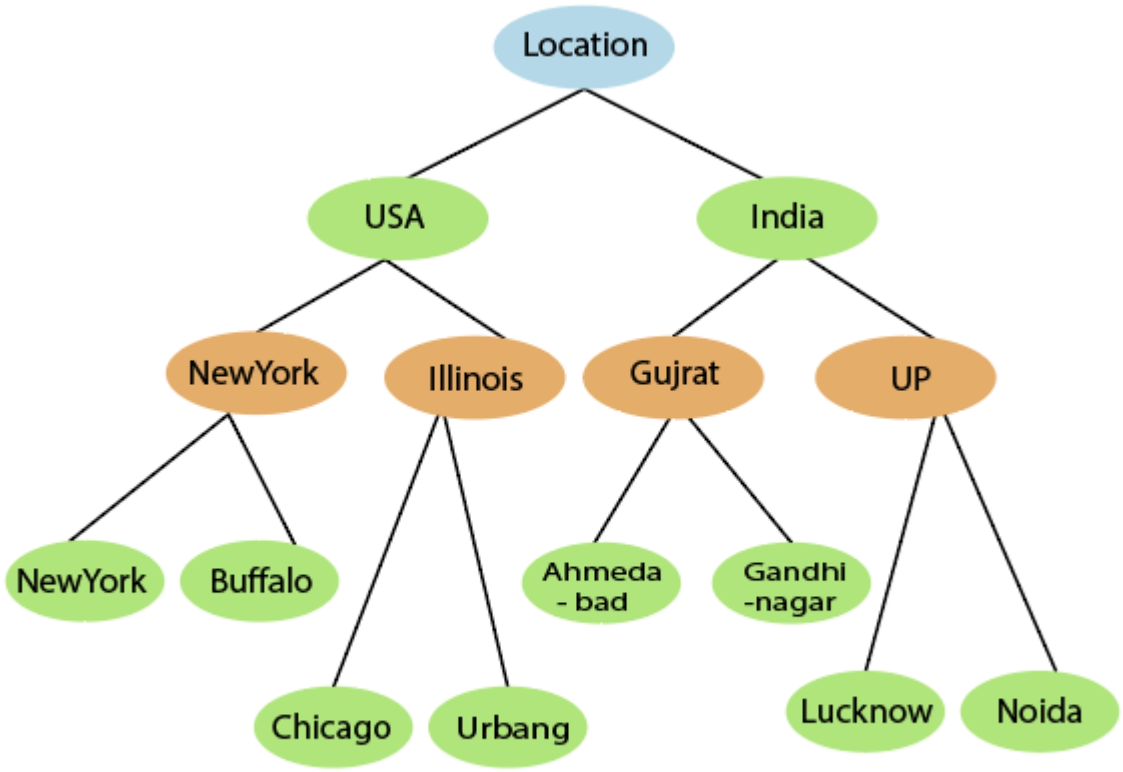
The method to automatically tag the report to the hierarchy is a top-down approach. An evaluation function determines whether a record currently tagged to a node can also be tagged to any of its child nodes.

If so, then then the tag moves down the hierarchy till it cannot be pushed any further.

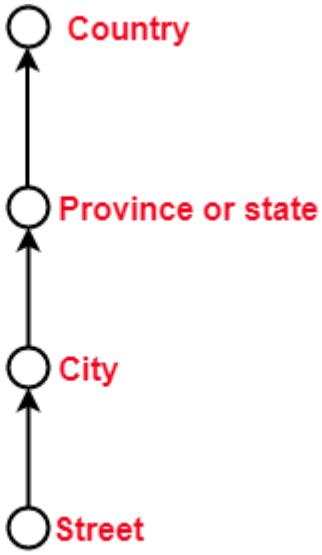
The outcome of this step is a hierarchy of report and, at each node, there is a set of the report having a common concept related to the node.

The hierarchy of reports resulting from the tagging step is useful for many texts mining process.

It is assumed that the hierarchy of concepts is called a priori. We can even have such a hierarchy of documents without a concept hierarchy, by using any hierarchical clustering algorithm, which results in such a hierarchy.



Concept hierarchy for the dimension location



Hierarchical structure for dimension location.

System Processes

We have a fixed number of operations to be applied on the operational databases and we have well-defined techniques such as use normalized data, keep table small, etc. These techniques are suitable for delivering a solution. But in case of decision-support systems, we do not know what query and operation needs to be executed in future. Therefore techniques applied on operational databases are not suitable for data warehouses.

Process Flow in Data Warehouse There are four major processes that contribute to a data warehouse −

- Extract and load the data.
- Cleaning and transforming the data.
- Backup and archive the data.
- Managing queries and directing them to the appropriate data sources.

Architecture

Three-Tier Data Warehouse Architecture Generally a data warehouses adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture.

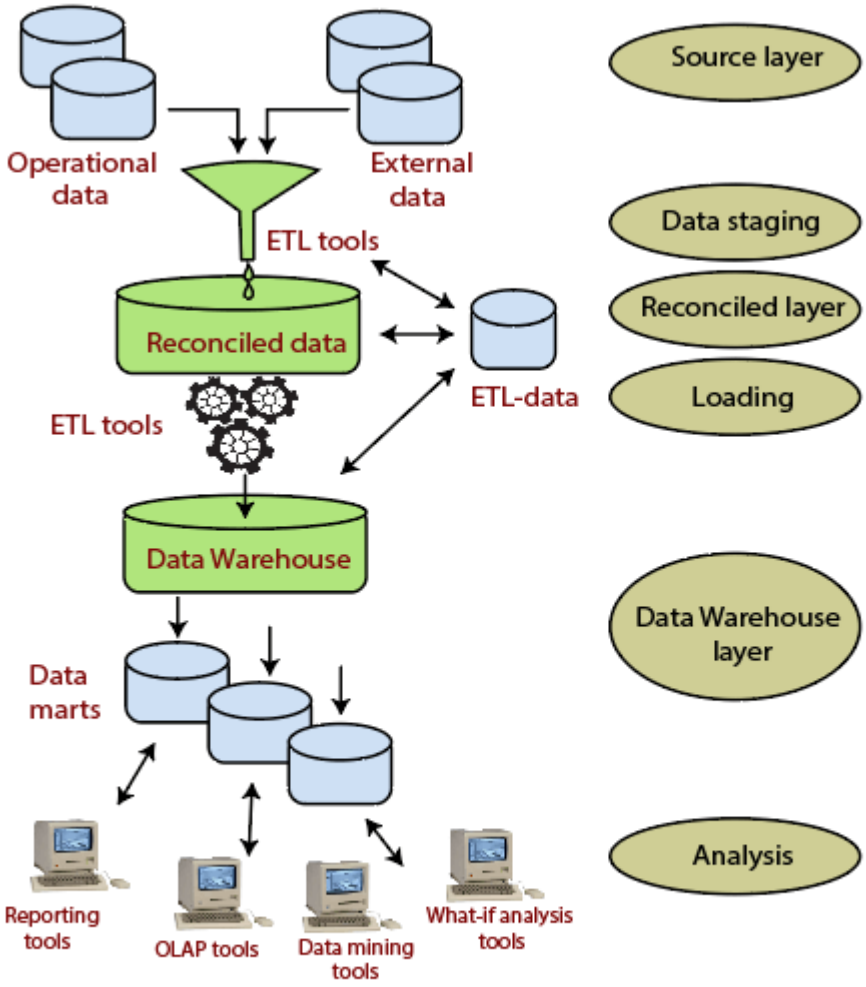
Bottom Tier − The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.

Middle Tier − In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.

By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.

By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.

Top-Tier − This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.



Three-Tier Architecture for a data warehouse system

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. OLAP enables end-clients to perform ad hoc analysis of record in multiple dimensions, providing the insight and understanding they require for better decision making.

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.

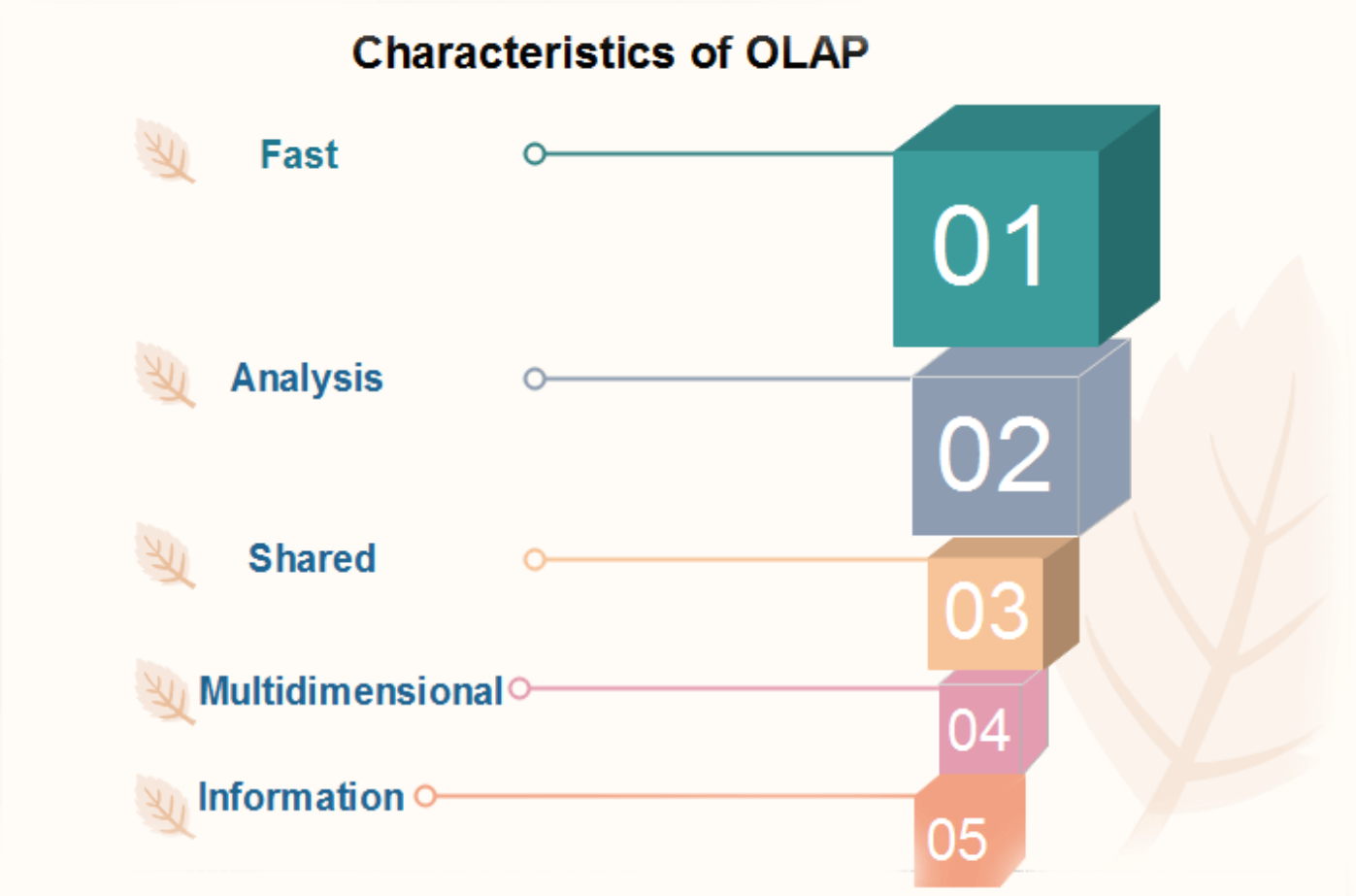
The second purpose is to enable fast query response that is usually difficult to achieve using tabular models.

How OLAP Works?

Fundamentally, OLAP has a very simple concept. It pre-calculates most of the queries that are typically very hard to execute over tabular databases, namely aggregation, joining, and grouping. These queries are calculated during a process that is usually called 'building' or 'processing' of the OLAP cube. This process happens overnight, and by the time end users get to work - data will have been updated.

Characteristics of OLAP

In the **FASMI** characteristics of OLAP methods, the term derived from the first letters of the characteristics are:



Fast

It defines which the system targeted to deliver the most feedback to the client within about five seconds, with the elementary analysis taking no more than one second and very few taking more than 20 seconds.

Analysis

It defines which the method can cope with any business logic and statistical analysis that is relevant for the function and the user, keep it easy enough for the target client. Although some preprogramming may be needed we do not think it acceptable if all application definitions have to be allow the user to define new Adhoc calculations as part of the analysis and to document on the data in any desired method, without having to program so we excludes products (like Oracle Discoverer) that do not allow the user to define new Adhoc calculation as part of the analysis and to document on the data in any desired product that do not allow adequate end user-oriented calculation flexibility.

Share

It defines which the system tools all the security requirements for understanding and, if multiple write connection is needed, concurrent update location at an appropriated level, not all functions need customer to write data back, but for the increasing number which does, the system should be able to manage multiple updates in a timely, secure manner.

Multidimensional

This is the basic requirement. OLAP system must provide a multidimensional conceptual view of the data, including full support for hierarchies, as this is certainly the most logical method to analyze business and organizations.

Information

The system should be able to hold all the data needed by the applications. Data sparsity should be handled in an efficient manner.

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 complex analysis for the users.
9. OLAP allows users to drill down for greater details or roll up for aggregations of metrics along a single business dimension or across multiple dimension.
10. OLAP provides the ability to perform intricate calculations and comparisons.
11. OLAP presents results in a number of meaningful ways, including charts and graphs.

Motivations for using OLAP

---

1) **Understanding and improving sales**: For enterprises that have much products and benefit a number of channels for selling the product, OLAP can help in finding the most suitable products and the most famous channels. In some methods, it may be feasible to find the most profitable users. For example, considering the telecommunication industry and considering only one product, communication minutes, there is a high amount of record if a company want to analyze the sales of products for every hour of the day (24 hours), difference between weekdays and weekends (2 values) and split regions to which calls are made into 50 region.

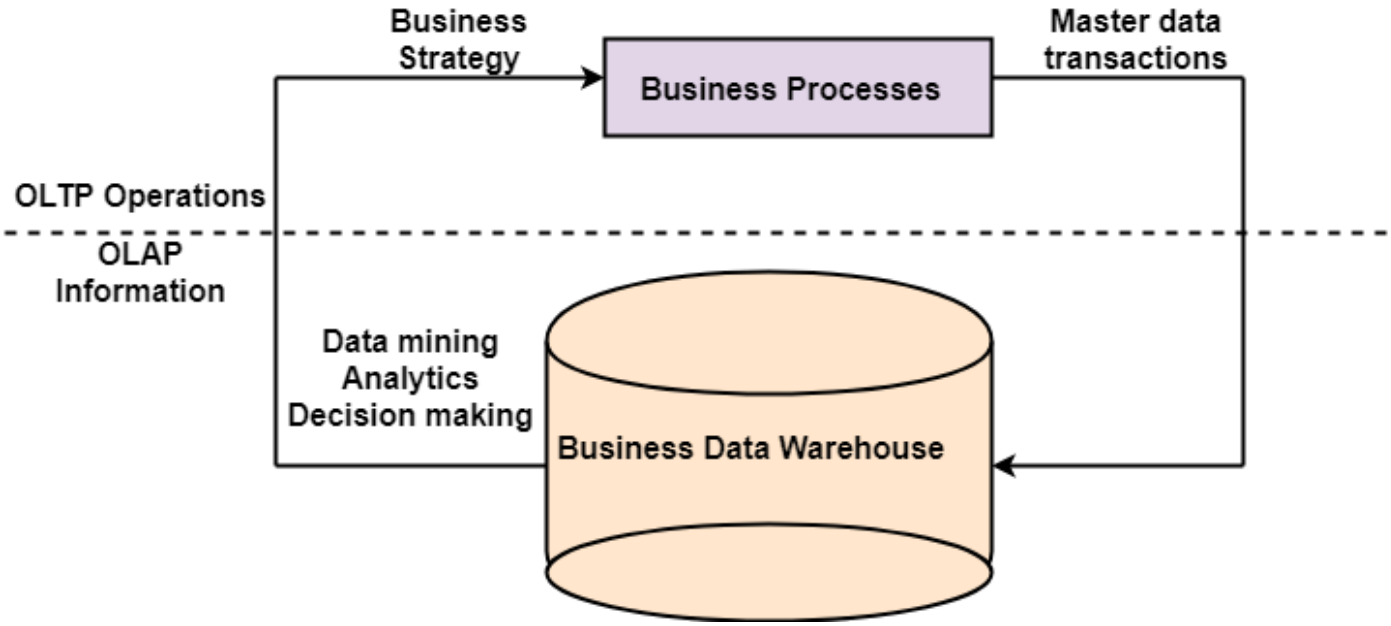
2) **Understanding and decreasing costs of doing business**: Improving sales is one method of improving a business, the other method is to analyze cost and to control them as much as suitable without affecting sales. OLAP can assist in analyzing the costs related to sales. In some methods, it may also be feasible to identify

expenditures which produce a high return on investments (ROI). For example, recruiting a top salesperson may contain high costs, but the revenue generated by the salesperson may justify the investment.

Difference between OLTP and OLAP

OLTP (On-Line Transaction Processing) is featured by a large number of short on-line transactions (INSERT, UPDATE, and DELETE). The primary significance of OLTP operations is put on very rapid query processing, maintaining record integrity in multi-access environments, and effectiveness consistent by the number of transactions per second. In the OLTP database, there is an accurate and current record, and schema used to save transactional database is the entity model (usually 3NF).

OLAP (On-line Analytical Processing) is represented by a relatively low volume of transactions. Queries are very difficult and involve aggregations. For OLAP operations, response time is an effectiveness measure. OLAP applications are generally used by Data Mining techniques. In OLAP database there is aggregated, historical information, stored in multi-dimensional schemas



Following are the difference between OLAP and OLTP system.

OLAP	OLTP
✓ Gives a multi-dimensional view of business activities.	✓ Enables a snapshot of ongoing business processes.
✓ Helps with planning, problem solving, and decision support.	✓ Useful for controlling and running fundamental business tasks.
✓ Data source is consolidated data	✓ Data source is the operational data.
✓ Includes Periodic long-running batch jobs that refresh the data.	✓ Has short and fast inserts and updates which are initiated by end users.
✓ OLAP applications are widely used by Data Mining techniques.	✓ Large number of short on-line transactions
✓ Database design is typically de-normalized and contains fewer tables.	✓ Database design in OLTP is highly normalized.
✓ Often involves complex queries along with aggregations, which in turn compels processing speed to be dependent on the amount of data involved; batch data refreshes, etc.	✓ Involves standardized and simple queries that return relatively few records hence is faster.

1) **Users**: **OLTP** systems are designed for office worker while the OLAP systems are designed for decision-makers. Therefore while an OLTP method may be accessed by hundreds or even thousands of clients in a huge enterprise, an **OLAP** system is suitable to be accessed only by a select class of manager and may be used only by dozens of users.

2) **Functions**: **OLTP** systems are mission-critical. They provide day-to-day operations of an enterprise and are largely performance and availability driven. These operations carry out simple repetitive operations. **OLAP** systems are management-critical to support the decision of enterprise support tasks using detailed investigation.

3) **Nature**: Although SQL queries return a set of data, **OLTP** methods are designed to step one record at the time, for example, a data related to the user who may be on the phone or in the store. **OLAP** system is not designed to deal with individual customer records. Instead, they include queries that deal with many data at a time and provide summary or aggregate information to a manager. OLAP applications include data stored in a data warehouses that have been extracted from many tables and possibly from more than one enterprise database.

4) **Design**: **OLTP** database operations are designed to be application-oriented while **OLAP** operations are designed to be subject-oriented. OLTP systems view the enterprise record as a collection of tables (possibly based on an entity-relationship model). OLAP operations view enterprise information as multidimensional).

5) **Data**: **OLTP** systems usually deal only with the current status of data. For example, a record about an employee who left three years ago may not be feasible on the Human Resources System. The old data may have been achieved on some type of stable storage media and may not be accessible online. On the other hand, **OLAP** systems needed historical data over several years since trends are often essential in decision making.

6) **Kind of use**: **OLTP** methods are used for reading and writing operations while **OLAP** methods usually do not update the data.

7) **View**: An **OLTP** system focuses primarily on the current data within an enterprise or department, which does not refer to historical data or data in various organizations. In contrast, an **OLAP** system spans multiple version of a database schema, due to the evolutionary process of an organization. OLAP system also deals with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, these are stored on multiple storage media.

8) **Access Patterns**: The access pattern of an **OLTP** system consist primarily of short, atomic transactions. Such a system needed concurrency control and recovery techniques. However, access to **OLAP** systems is mostly read-only operations because these data warehouses store historical information.

The biggest difference between an OLTP and OLAP system is the amount of data analyzed in a single transaction. Whereas an OLTP handles many concurrent customers and queries touching only a single data or limited collection of records at a time, an OLAP system must have the efficiency to operate on millions of data to answer a single query.

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.

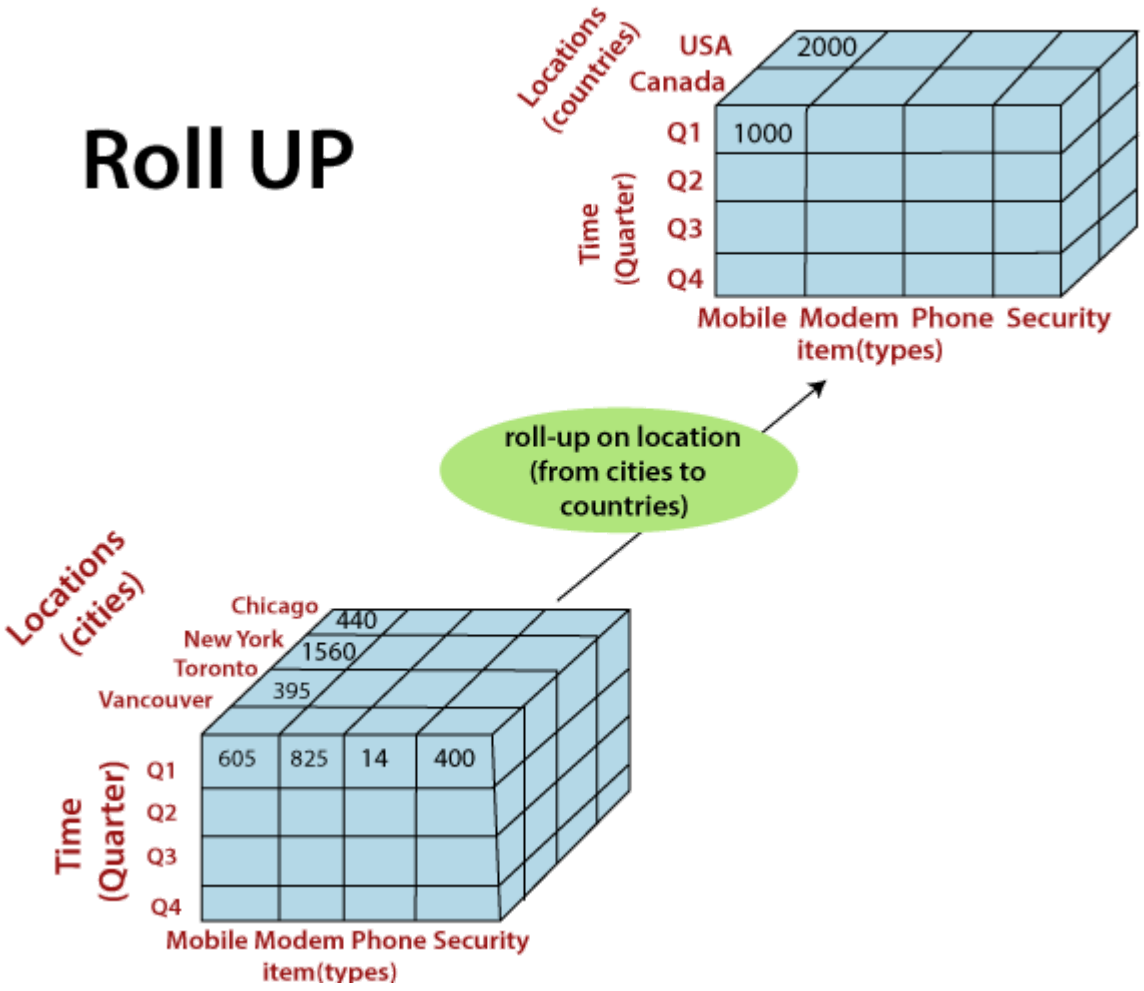
1. Roll-Up

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.



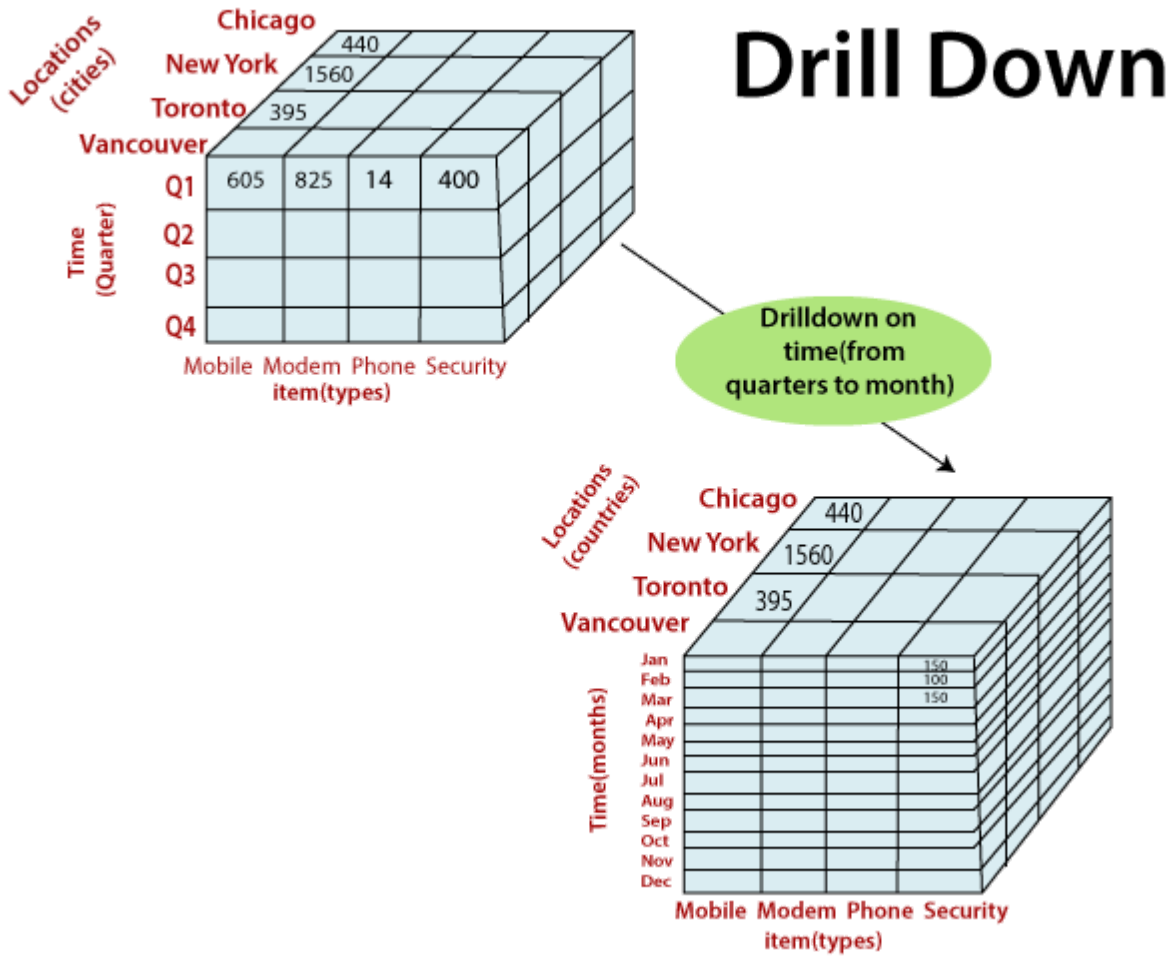
2. Drill-Down

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.

Because a drill-down adds more details to the given data, it can also be performed by adding a new dimension to a cube. For example, a drill-down on the central cubes of the figure can occur by introducing an additional dimension, such as a customer group.

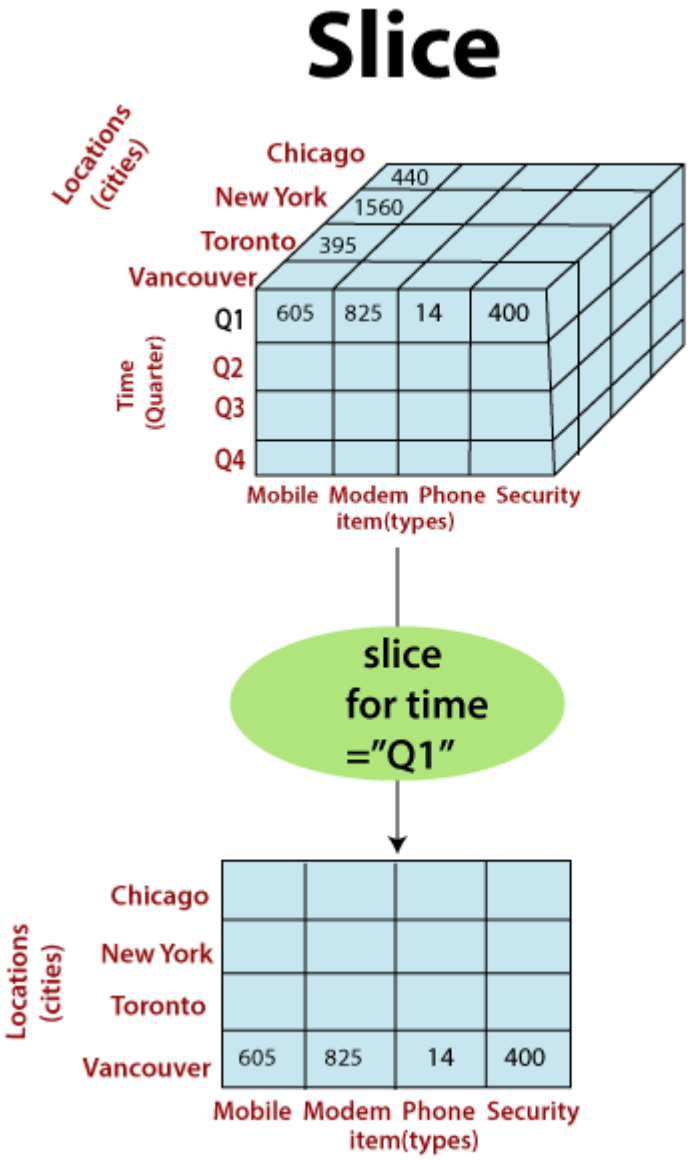
The following diagram illustrates how Drill-down works.



3. Slice

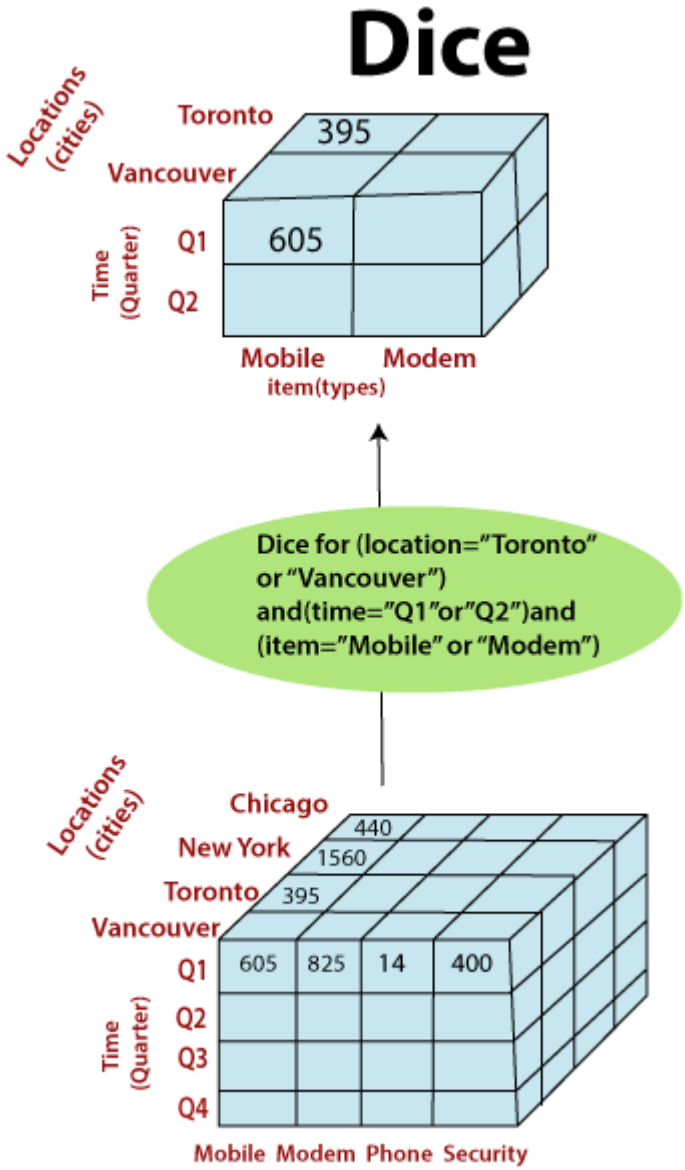
A slice is a subset of the cubes corresponding to a single value for one or more members of the dimension. For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site. So, the Slice operations perform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

The following diagram illustrates how Slice works.



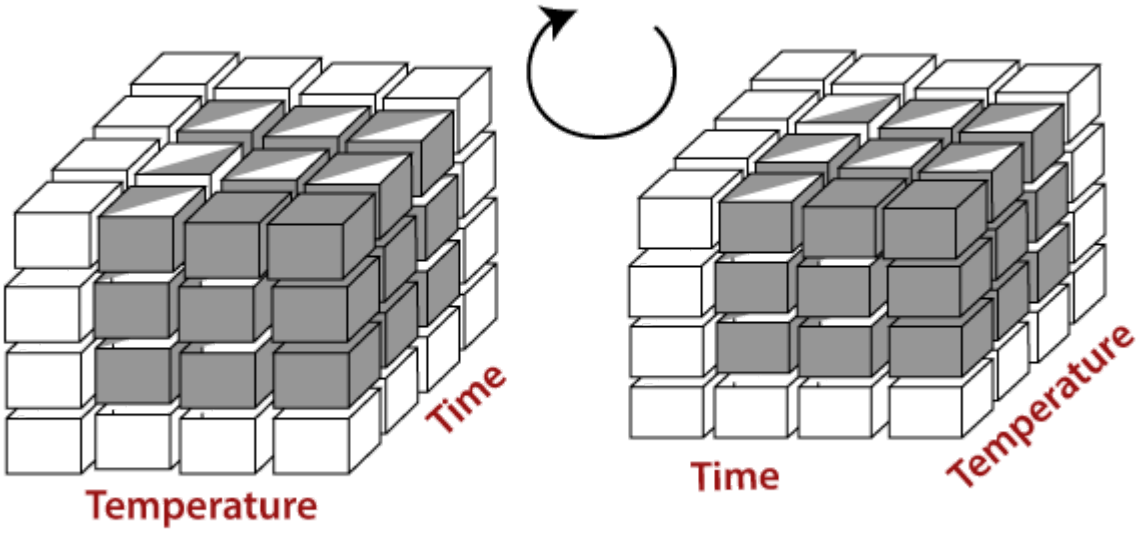
4. Dice

The dice operation describes a subcube by operating a selection on two or more dimension.



5. 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.


Time
Temperature
Time
Temperature

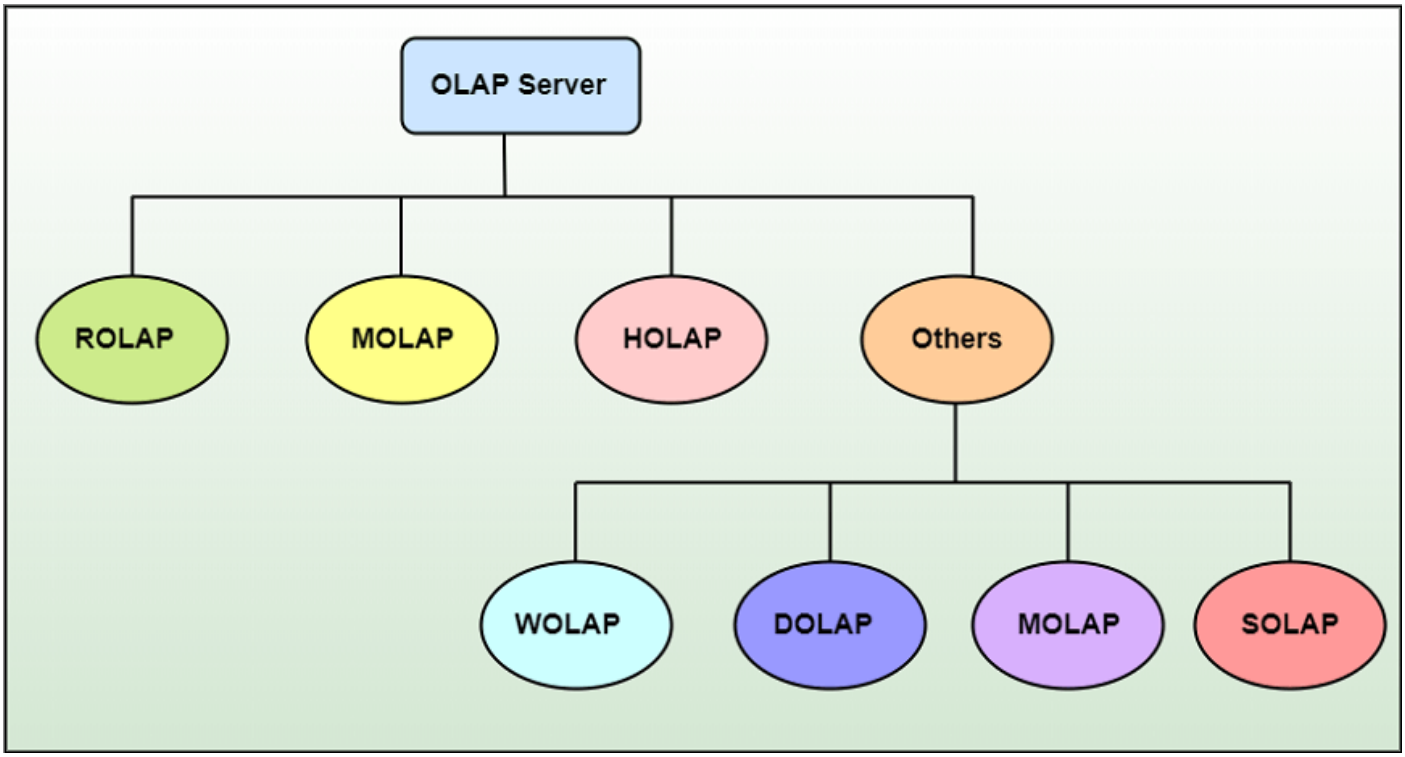
Types of OLAP

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.

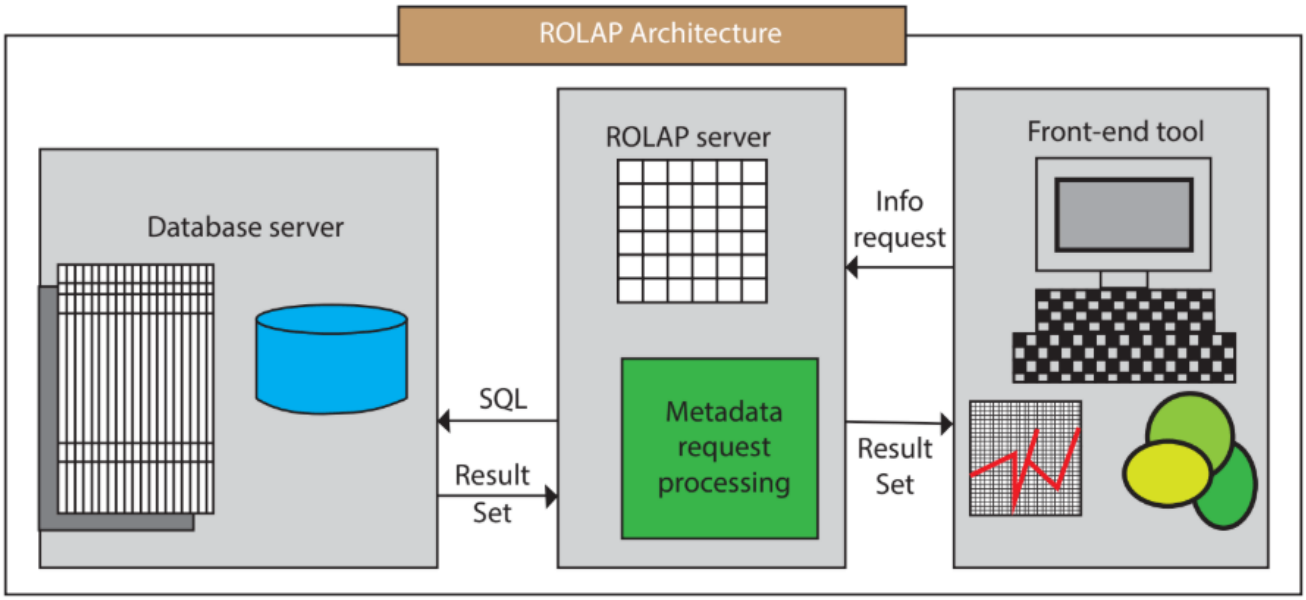
ROLAP systems work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables. This model permits the multidimensional analysis of data.

This technique relies on manipulating the data stored in the relational database to give the presence of traditional OLAP's slicing and dicing functionality. In essence, each method of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.

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.

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.

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

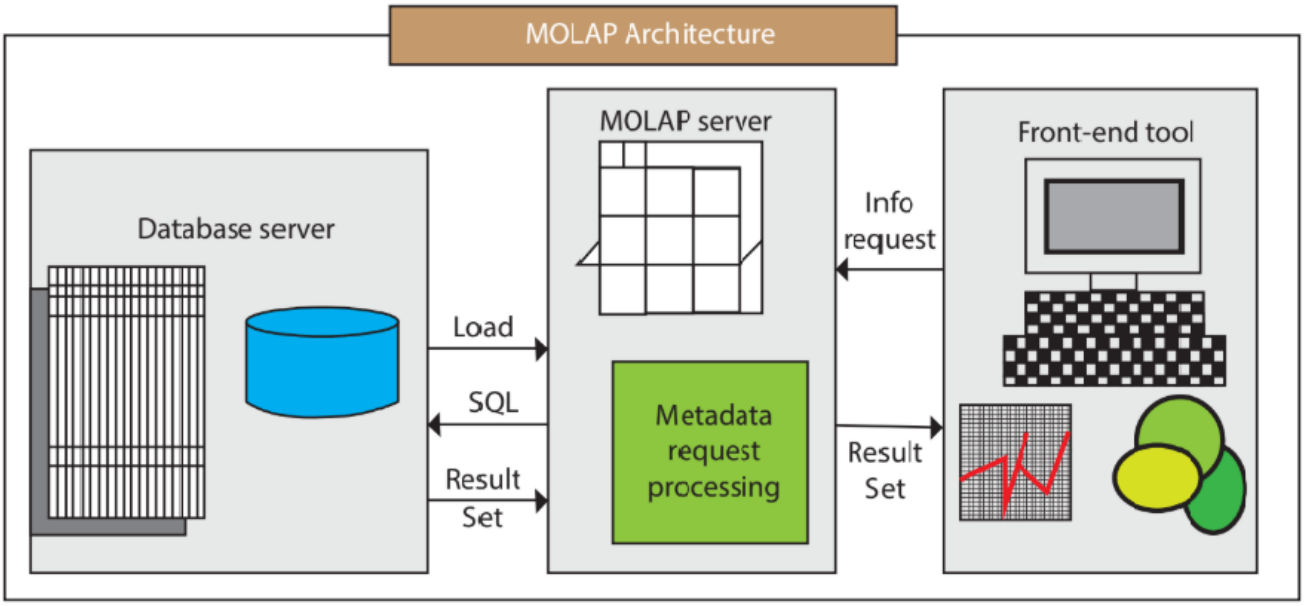
A MOLAP system is based on a native logical model that directly supports multidimensional data and operations. Data are stored physically into multidimensional arrays, and positional techniques are used to access them.

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).

Examples include Arbor Software's Essbase. Oracle's Express Server, Pilot Software's Lightship Server, Sniper's TM/1. Planning Science's Gentium and Kenan Technology's Multiway.

Some of the problems faced by clients are related to maintaining support to multiple subject areas in an RDBMS. Some vendors can solve these problems by continuing access from MOLAP tools to detailed data in and RDBMS.

This can be very useful for organizations with performance-sensitive multidimensional analysis requirements and that have built or are in the process of building a data warehouse architecture that contains multiple subject areas.

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

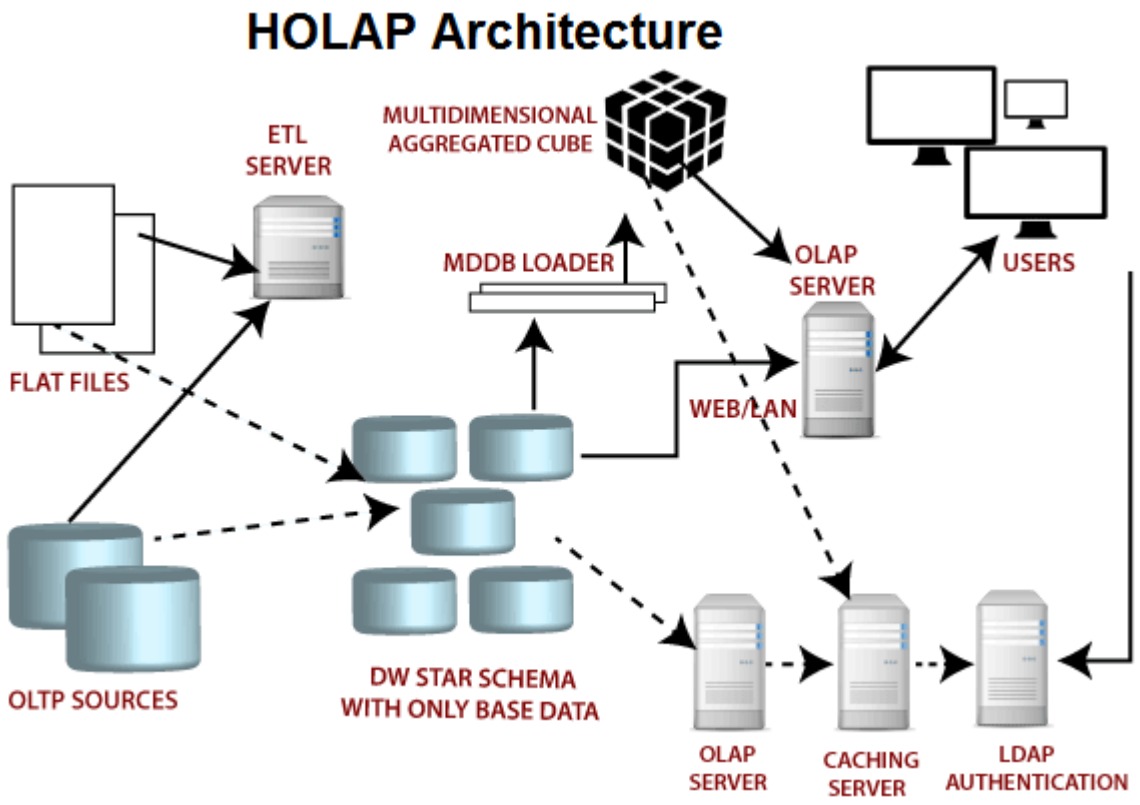
Can perform complex calculations: All evaluation have been pre-generated when the cube is created. Hence, complex calculations are not only possible, but they return quickly.

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.

Other Types

There are also less popular types of OLAP styles upon which one could stumble upon every so often. We have listed some of the less popular brands existing in the OLAP industry.

- **Web-Enabled OLAP (WOLAP) Server**

- **Desktop OLAP (DOLAP) Server**
- **Mobile OLAP (MOLAP) Server**
- **Spatial OLAP (SOLAP) Server**

Dimension Modelling

What is Dimensional Modeling?

Dimensional modeling represents data with a cube operation, making more suitable logical data representation with OLAP data management. The perception of Dimensional Modeling was developed by Ralph Kimball and is consist of "fact" and "dimension" tables.

In dimensional modeling, the transaction record is divided into either "facts," which are frequently numerical transaction data, or "dimensions," which are the reference information that gives context to the facts. For example, a sale transaction can be damage into facts such as the number of products ordered and the price paid for the products, and into dimensions such as order date, user name, product number, order ship-to, and bill-to locations, and salesman responsible for receiving the order.

Objectives of Dimensional Modeling

The purposes of dimensional modeling are:

1. To produce database architecture that is easy for end-clients to understand and write queries.
2. To maximize the efficiency of queries. It achieves these goals by minimizing the number of tables and relationships between them.

Elements of Dimensional Modeling

Fact

It is a collection of associated data items, consisting of measures and context data. It typically represents business items or business transactions.

Dimensions

It is a collection of data which describe one business dimension. Dimensions decide the contextual background for the facts, and they are the framework over which OLAP is performed.

Measure

It is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions.

Considering the relational context, there are two basic models which are used in dimensional modeling:

- Star Model
- Snowflake Model

The star model is the underlying structure for a dimensional model. It has one broad central table (fact table) and a set of smaller tables (dimensions) arranged in a radial design around the primary table. The snowflake model is the conclusion of decomposing one or more of the dimensions.

Fact Table

Fact tables are used to data facts or measures in the business. Facts are the numeric data elements that are of interest to the company.

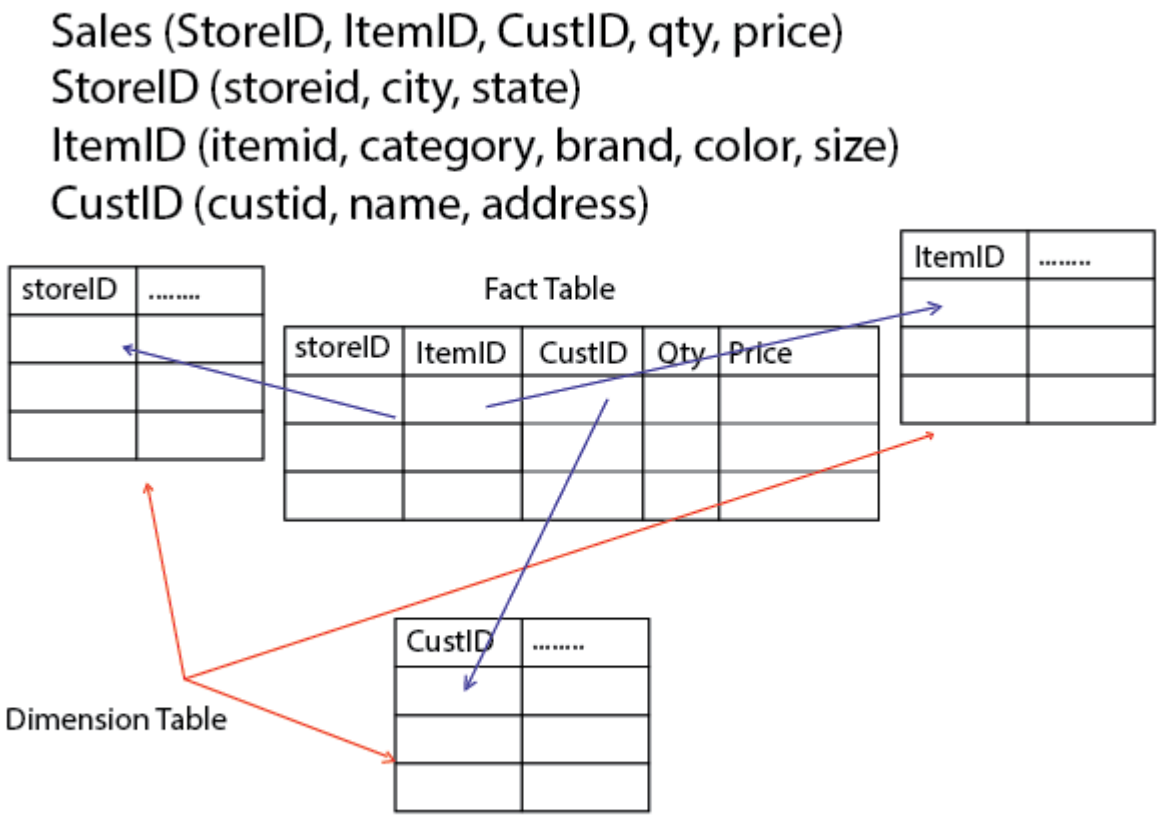
Characteristics of the Fact table

The fact table includes numerical values of what we measure. For example, a fact value of 20 might means that 20 widgets have been sold.

Each fact table includes the keys to associated dimension tables. These are known as foreign keys in the fact table.

Fact tables typically include a small number of columns.

When it is compared to dimension tables, fact tables have a large number of rows.

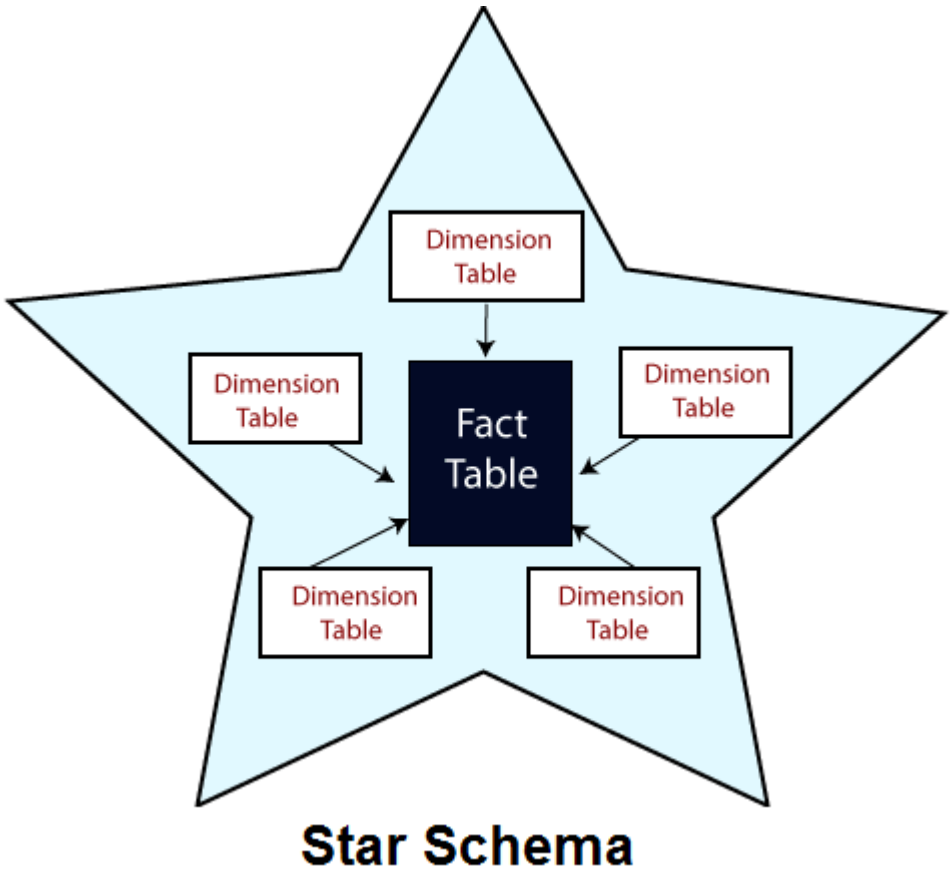


Schemas

What is 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.



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.

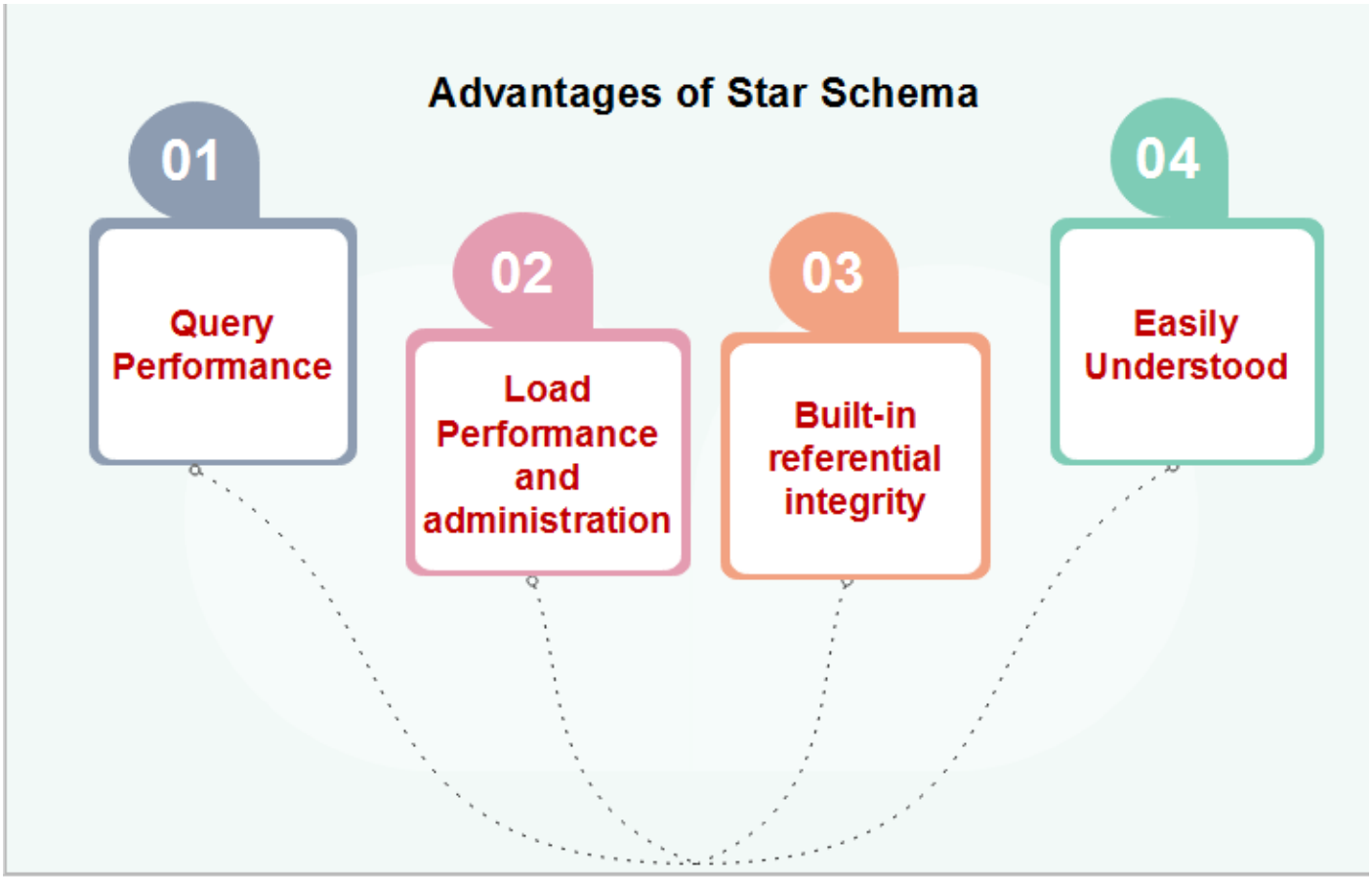
Characteristics of Star Schema

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.



- Query Performance
- Load performance and administration

- Built-in referential integrity
- Easily Understood

What is Snowflake Schema?

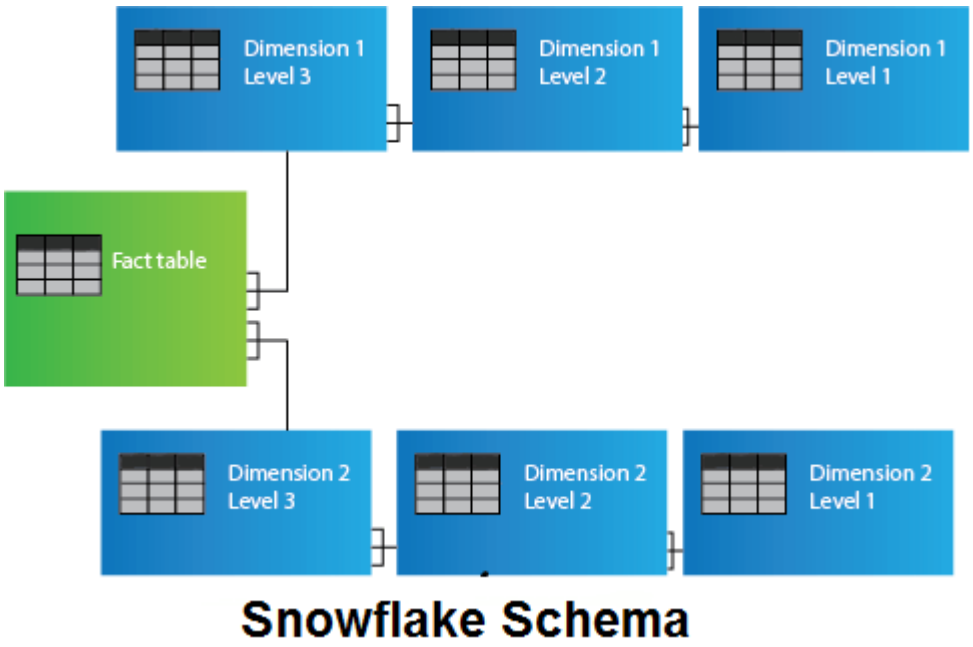
A snowflake schema is equivalent to the star schema. "A schema is known as a snowflake if one or more dimension tables do not connect directly to the fact table but must join through other dimension tables."

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.

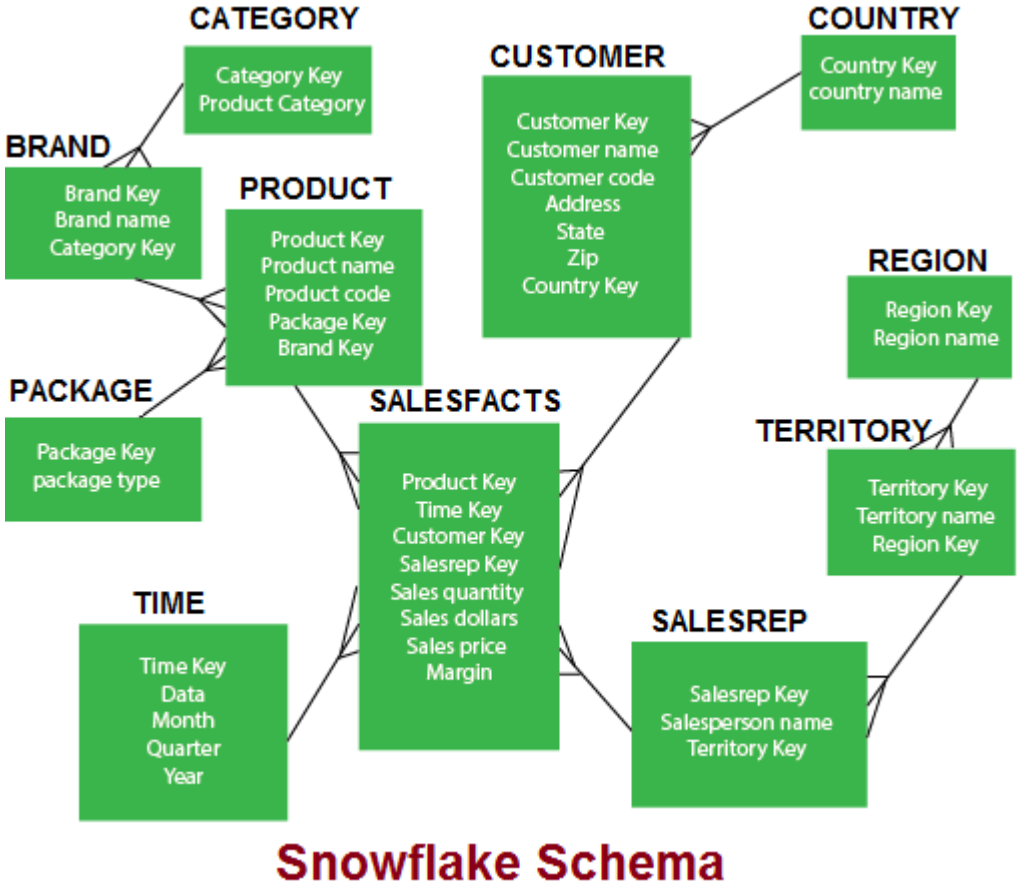
Snowflaking is used to develop the performance of specific queries. The schema is diagramed with each fact surrounded by its associated dimensions, and those dimensions are related to other dimensions, branching out into a snowflake pattern.

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 relationship. 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



Snowflake Schema

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.

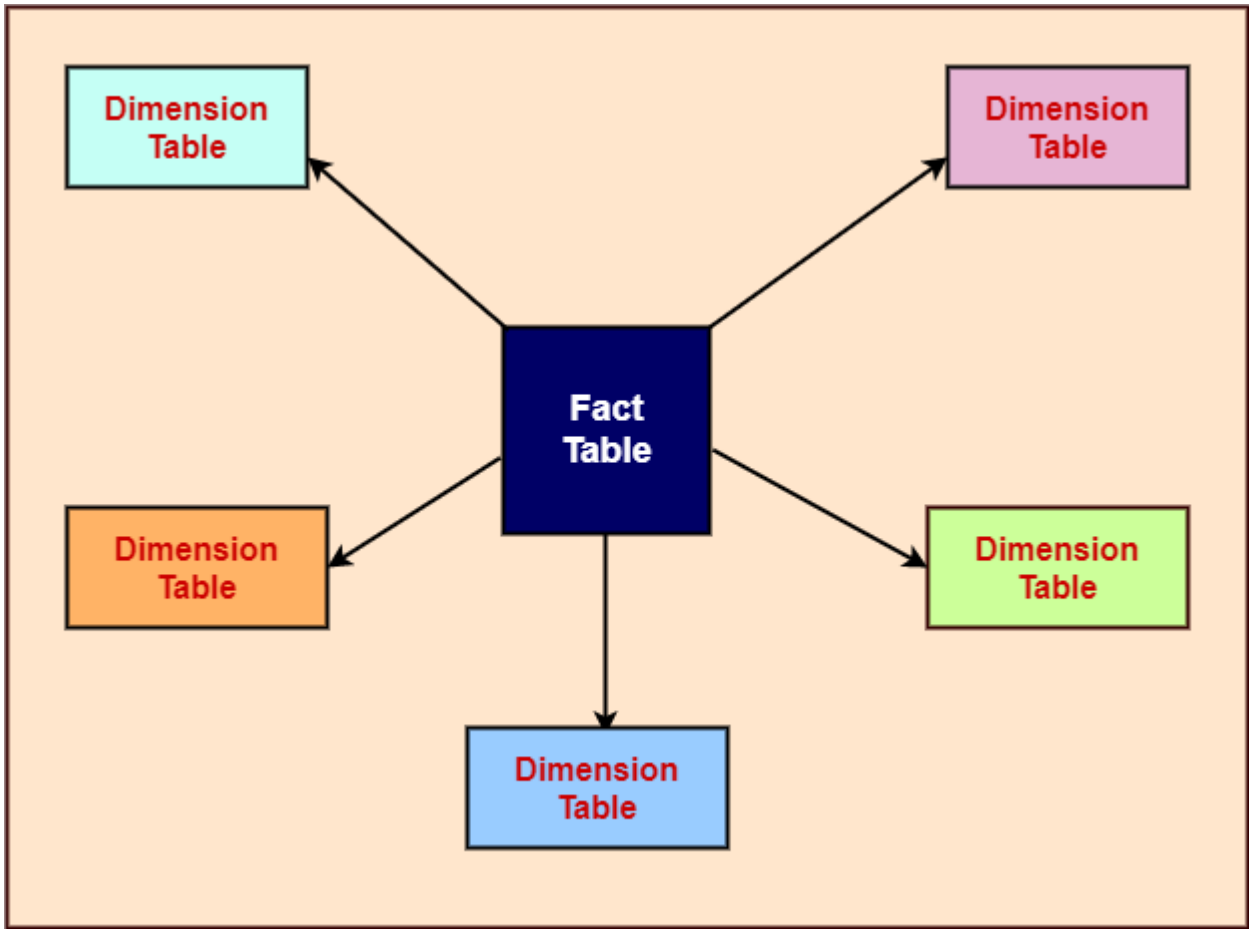
Disadvantage of Snowflake Schema

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.

Difference between Star and Snowflake Schemas

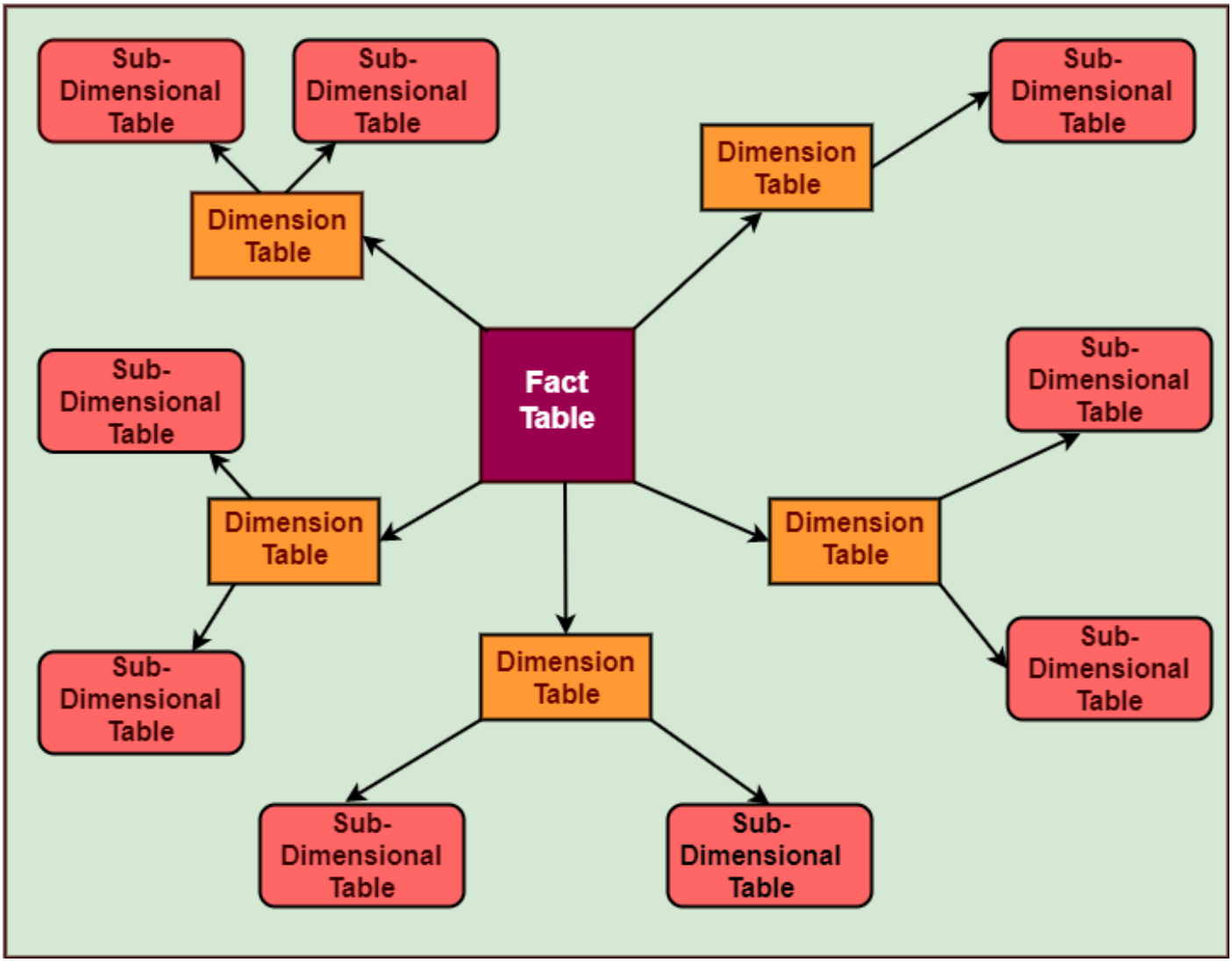
Star Schema

- In a star schema, the fact table will be at the center and is connected to the dimension tables.
- The tables are completely in a denormalized structure.
- SQL queries performance is good as there is less number of joins involved.
- Data redundancy is high and occupies more disk space.



Snowflake Schema

- A snowflake schema is an extension of star schema where the dimension tables are connected to one or more dimensions.
- The tables are partially denormalized in structure.
- The performance of SQL queries is a bit less when compared to star schema as more number of joins are involved.
- Data redundancy is low and occupies less disk space when compared to star schema.



References and Links to Follow

Data Warehouse tutorials by TutorialsPoint: https://www.tutorialspoint.com/dwh/index.htm (https://www.tutorialspoint.com/dwh/index.htm)

Data Warehouse tutorials by JavaPoint: https://www.javatpoint.com/data-warehouse (https://www.javatpoint.com/data-warehouse)

Video Tutorials:

Data Warehousing Tutorials for Beginners by Edureka: https://www.youtube.com/watch?v=J326LIUrZM8 (https://www.youtube.com/watch?v=J326LIUrZM8)

Data Warehousing Concepts by Edureka: https://www.youtube.com/watch?v=CHYPF7jxlik (https://www.youtube.com/watch?v=CHYPF7jxlik)