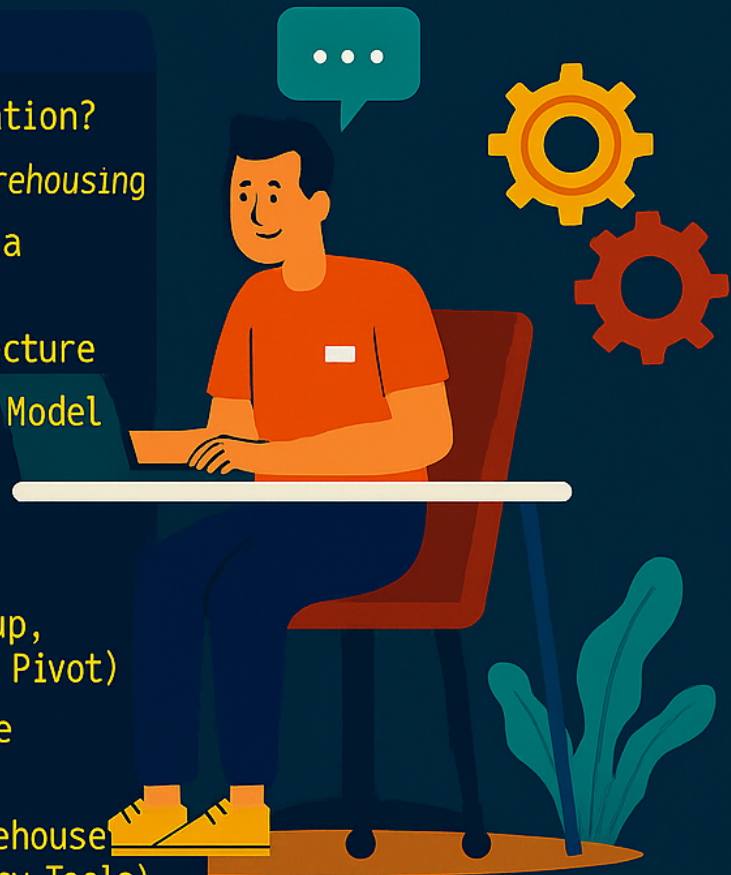


DATA WAREHOUSE FUNDAMENTALS

CORE CONCEPTS EXPLAINED

- What is Data & Information?
- Introduction to Data Warehousing
- Characteristics of Data Warehouse
- Data Warehouse Architecture
- Multi-dimensional Data Model (Facts & Dimensions)
- Star, Snowflake & Fact Constellation Schemas
- OLAP Operations (Roll-up, Drill-down, Slice, Dice, Pivot)
- Types of Data Warehouse (EDW, ODS, Data Mart)
- Components of Data Warehouse (DB, ETL, Metadata, Query Tools)



BY ARVIND KUMAR

DATAWAREHOUSE INTRODUCTION

What is Data and Information?

Data is an individual unit that contains raw materials which do not carry any specific meaning.

Information is a group of data that collectively carries a logical meaning.

Data doesn't depend on information.

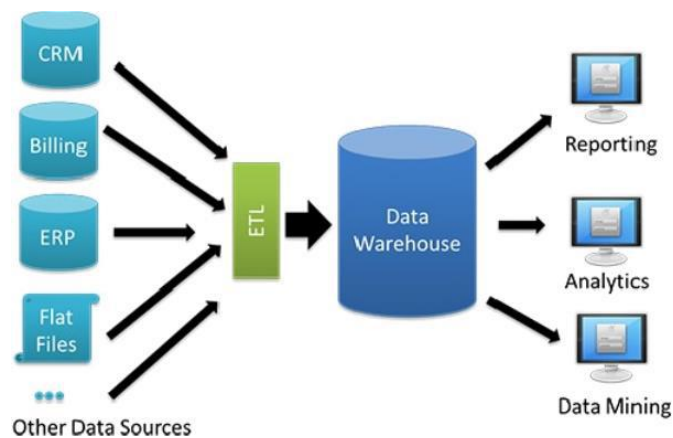
Information depends on data.

Data is measured in bits and bytes.

Information is measured in meaningful units like time, quantity, etc.

Data Warehouse:

Data warehouse is like a relational database designed for analytical needs. It functions on the basis of OLAP (Online Analytical Processing). It is a central location where consolidated data from multiple locations (databases) are stored.



What is Data warehousing?

Data warehousing is the act of organizing & storing data in a way so as to make its retrieval efficient and insightful. It is also called as the process of transforming data into information.

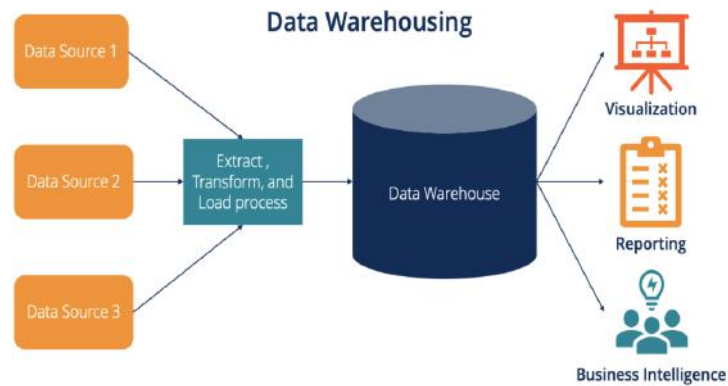


Fig: Data warehousing Process

Data Warehouse Characteristics:

A Data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process.

Subject-oriented:

A Data warehouse can be used to analyze a particular subject area

Ex: "Sales" can be particular subject

Integrated:

A Data warehouse integrates data from multiple data sources.

Time Variant:

Historical data is kept in a data warehouse.

Ex: one can retrieve data from 3 months, 6months, 12 months or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept.

Non-Volatile:

Once data is in the data warehouse, it will not change. So historical data in a data warehouse should never be altered.

Data warehouse Architecture:

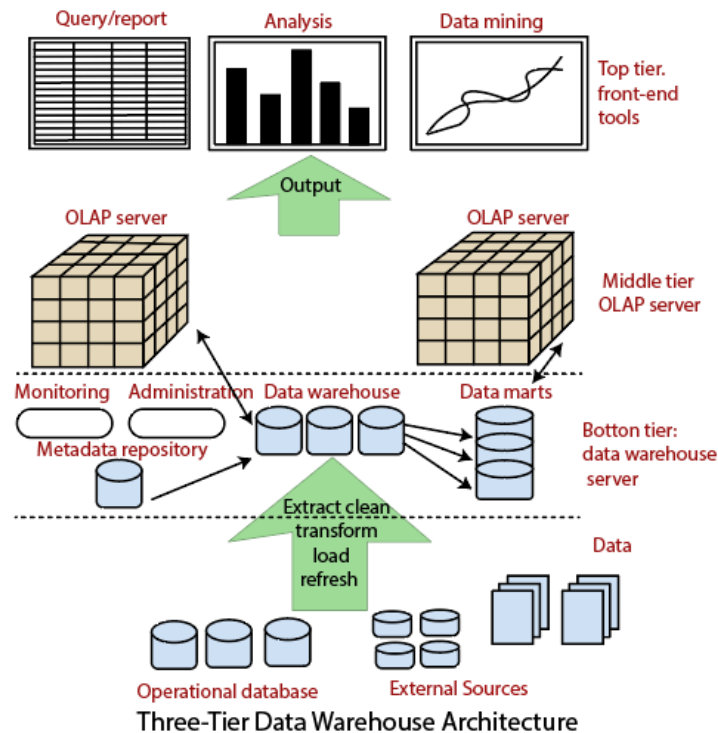


Fig: Data ware housing Architecture

Data warehouses often adopt a three-tier architecture

- 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 the data into the bottom tier from operational database or other external sources. These tools and utilities perform data extraction, cleaning and transformation(ex. To merge similar data from different sources into a unified format), 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.

The middle tier is an **OLAP** server that is typically implemented using either

(a) a relational OLAP(**ROLAP**) model, that is an extended relational DBMS that maps operations on multidimensional data to standard relational operations, or

(b) a multidimensional OLAP(**MOLAP**) model that is a special-purpose server that directly implements multidimensional data and operations.

The top tier is a front end client layer, which contains query and reporting tools, analysis tools and data mining tools(ex: trend analysis, prediction....)

Multi-dimensional Data Model:

- A multidimensional model views data in the form of a data-cube.
- When data is grouped or combined in multidimensional matrices called Data Cubes.
- A data cube enables data to be modeled and viewed in multiple dimensions.

It is defined by dimensions and facts.

- A multidimensional data model is organized around a central theme, for example, sales. This theme is represented by a fact table. Facts are numerical measures.
- The fact table contains the names of the facts or measures of the related dimensional tables.

FACT VS DIMENSION

Fact/Measure(What you want to analyse is your fact)

Ex: What is My sales, What is my profit, What is my customers preferences.

Dimensions(By Which you want to Analyze is your Dimensions)

Sales By Location/Product/Period

Total Profit By Location/Product/Period

- These Dimensions allow the store to keep track of things like monthly sales of items and branches and locations at which the items were sold.
- Each dimension may have a table associated with it called a **dimension table**, which further describes the dimension.

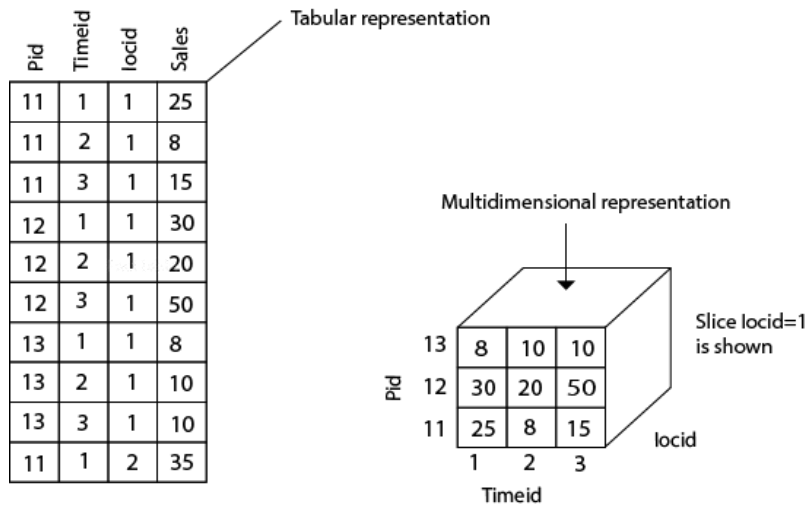


Fig: Multidimensional Representation

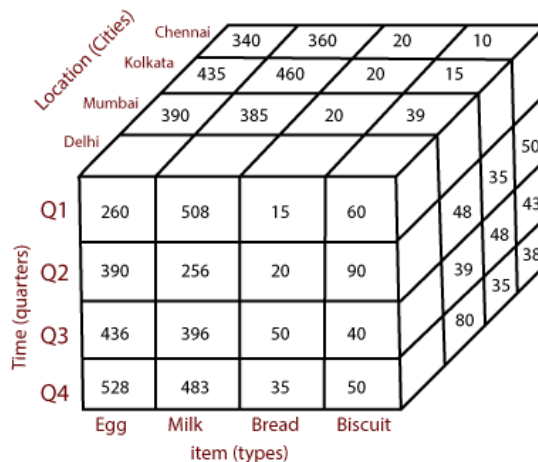
- Consider the data of a shop for items sold per quarter in the city of Delhi. The data is shown in the table.
- In this 2D representation, the sales for Delhi are shown for the time dimension (organized in quarters) and the item dimension (classified according to the types of an item sold).
- The fact or measure displayed in rupee_sold (in thousands).

Location="Delhi"				
Time (quarter)	item (type)			
	Egg	Milk	Bread	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	528	483	35	50

Now, if we want to view the sales data with a third dimension, For example, suppose the data according to time and item, as well as the location is considered for the cities Chennai, Kolkata, Mumbai, and Delhi. These 3D data are shown in the table. The 3D data of the table are represented as a series of 2D tables.

	Location="Chennai"				Location="Kolkata"				Location="Mumbai"				Location="Delhi"			
	item				item				item				item			
Time	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit
Q1	340	360	20	10	435	460	20	15	390	385	20	39	260	508	15	60
Q2	490	490	16	50	389	385	45	35	463	366	25	48	390	256	20	90
Q3	680	583	46	43	684	490	39	48	568	594	36	39	436	396	50	40
Q4	535	694	39	38	335	365	83	35	338	484	48	80	528	483	35	50

Conceptually, it may also be represented by the same data in the form of a 3D data cube, as shown in fig:



What is Schema?

- 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.
- Modeling data warehouses: dimensions & measures

- Star schema: A fact table in the middle connected to a set of dimension tables
- Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
- Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

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.

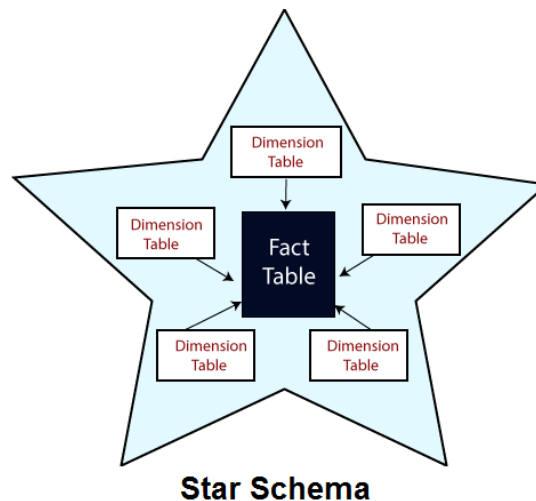
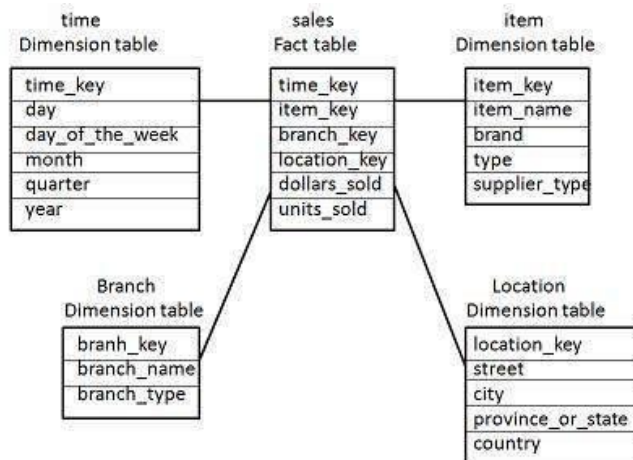


Fig: Star Schema Representation

Star Schema:

- 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.
- 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.

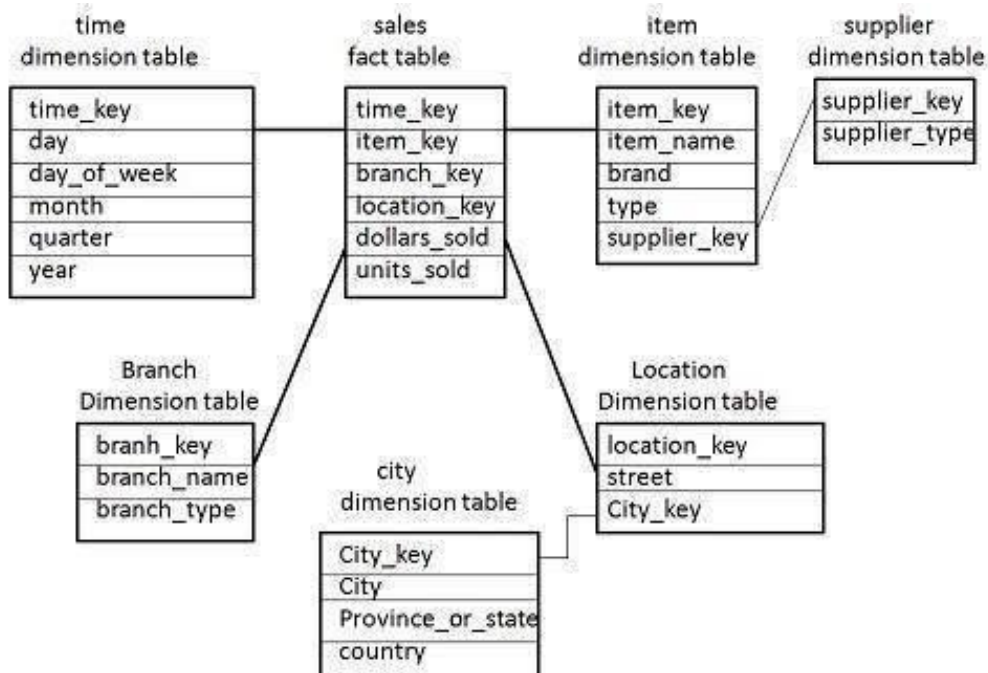
Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.

- The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools.
- **Advantages:**
 - (i) Simplest and Easiest
 - (ii) It optimizes navigation through database
 - (iii) Most suitable for Query Processing

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.

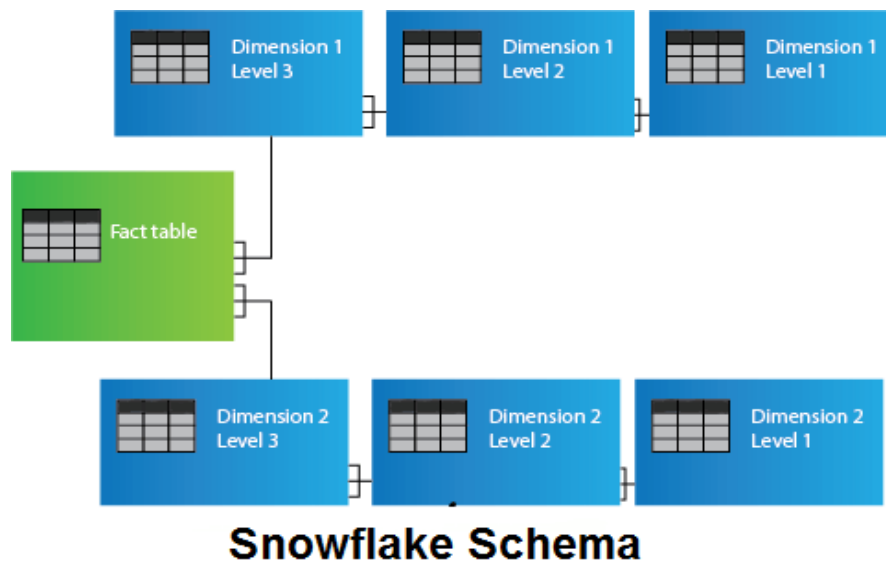
Note : Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.



Fig: Snowflake image

A snowflake schemas can have any number of dimension, and each dimension can have any number of levels.

The following diagram shows a snowflake schema with two dimensions, each having three levels.



Advantages:

(i) Less redundancies due to normalization Dimension Tables.

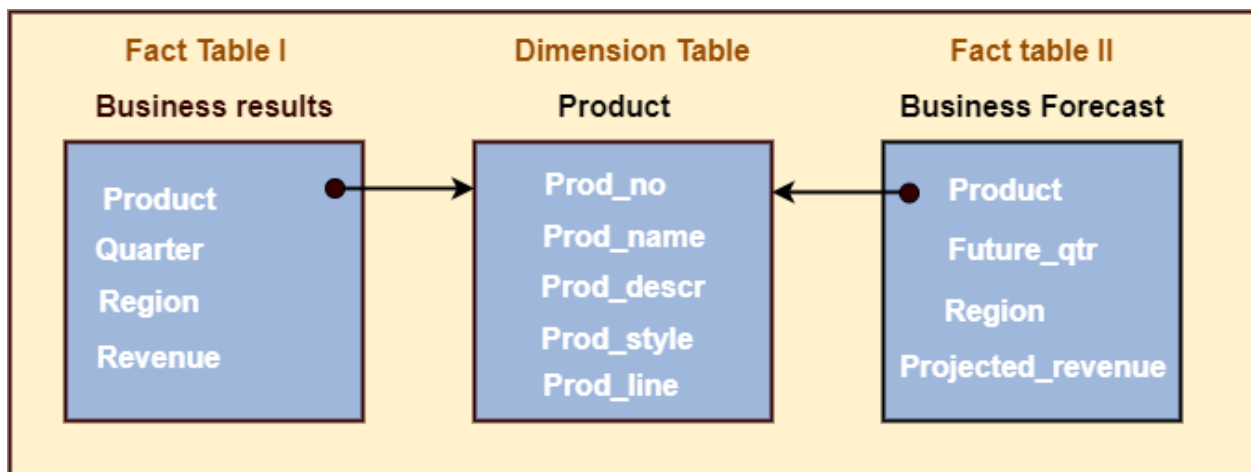
(ii) Dimension Tables are easier to update.

Disadvantages:

It is complex schema when compared to star schema.

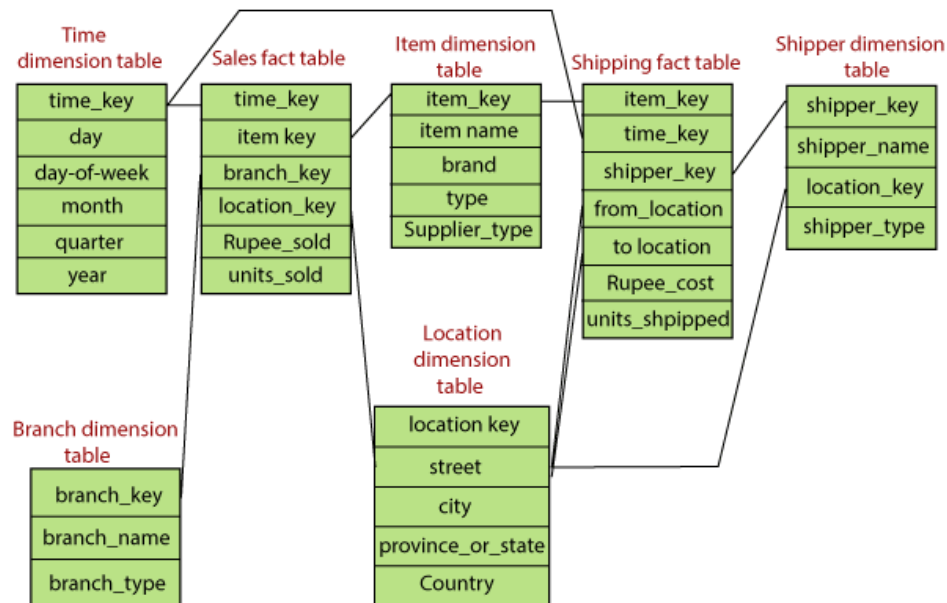
Fact Constellation Schema:

- 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

A fact constellation schema is shown in the figure below.



- This schema defines two fact tables, sales, and shipping. Sales are treated along four dimensions, namely, time, item, branch, and location.
- The schema contains a fact table for sales that includes keys to each of the four dimensions, along with two measures: Rupee_sold and units_sold.
- The shipping table has five dimensions, or keys: item_key, time_key, shipper_key, from_location, and to_location, and two measures: Rupee_cost and units_shpipped.
- 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.

Disadvantages:

- (i) Complex due to multiple fact tables
- (ii) It is difficult to manage
- (iii) Dimension Tables are very large.

OLAP OPERATIONS:

- 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 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.

OLAP having 5 different operations

- (i) Roll-up
- (ii) Drill-down
- (iii) Slice
- (iv) Dice
- (v) Pivot

Roll-up:

- The roll-up 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.
- **It is also known as drill-up or aggregation operation**
- 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.

Roll UP

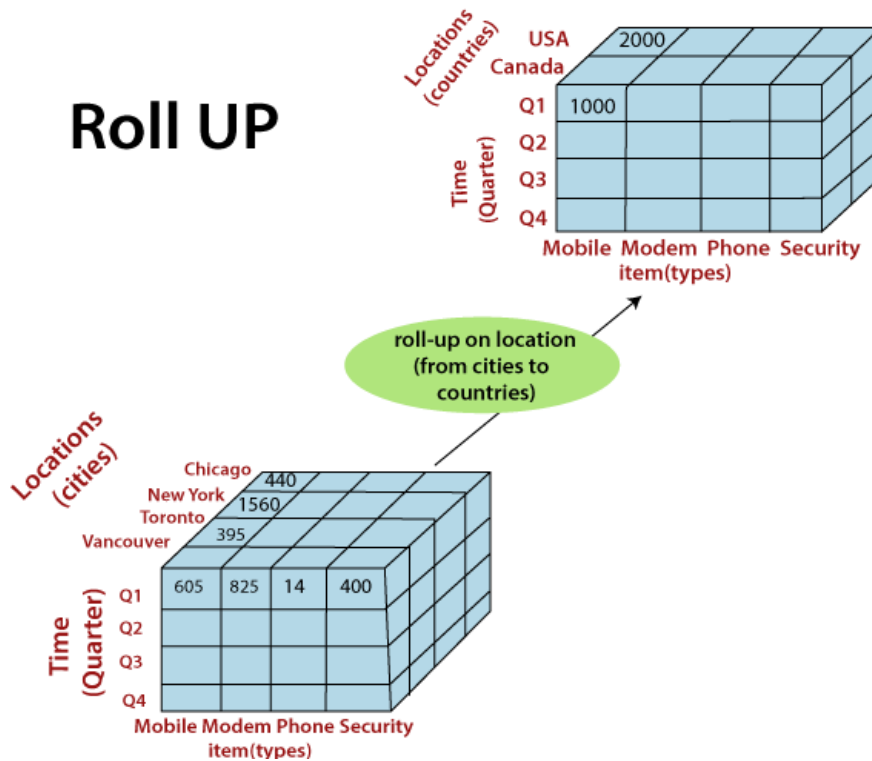


Fig: Roll-up operation on Data Cube

Drill-Down

- The drill-down operation is the reverse operation of **roll-up**.
- It is also called **roll-down operation**.
- 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.

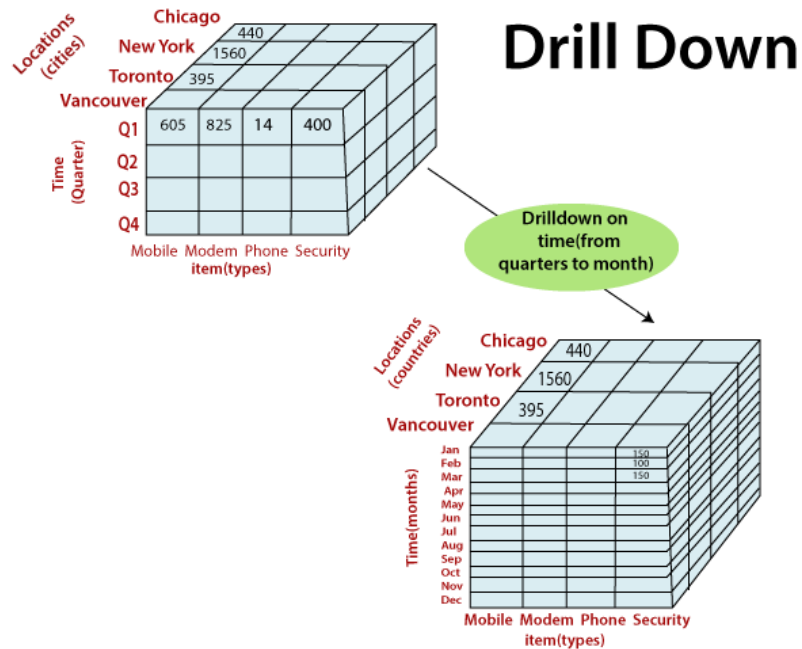


Fig: Drill-down operation

Slice:

- A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension.
- The slice operation provides a new sub cube from one particular dimension in a given cube.
- 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 sub cube.
- 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.

Slice

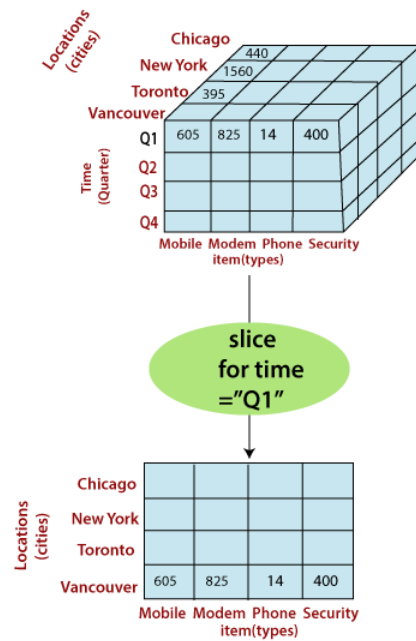


Fig: Slice operation

Dice:

- The dice operation describes a sub cube by operating a selection on two or more dimension.

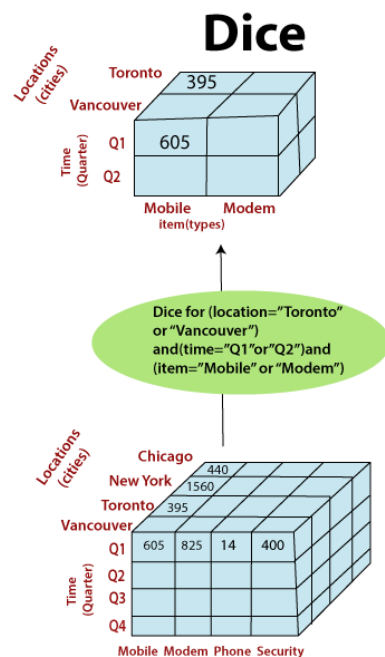


Fig: Dice operation

- 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")

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.

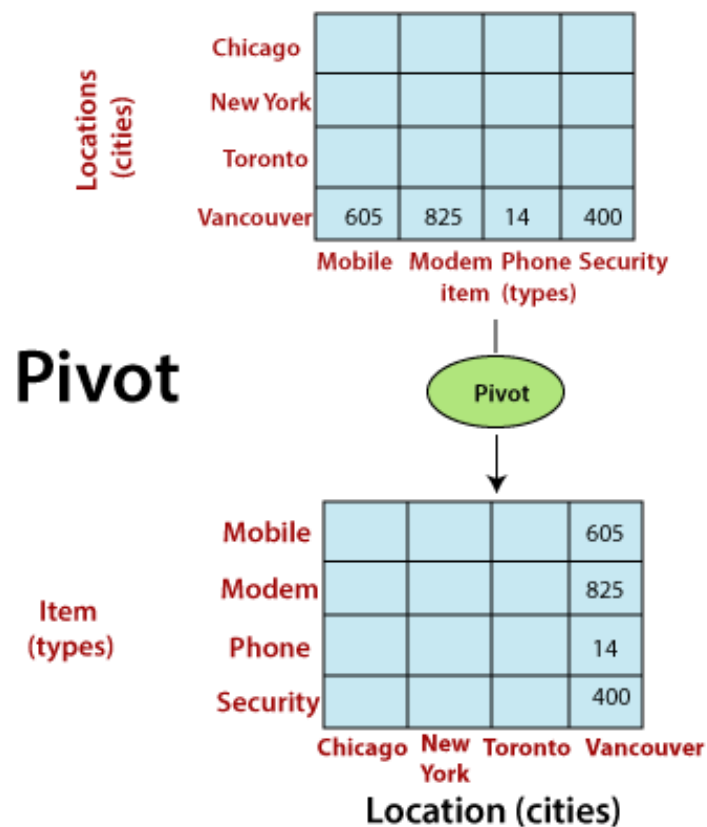


Fig: Pivot Operation

Differences between OLTP and OLAP:

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	Transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/warehouse, subject-oriented
Data	current, guaranteed up-to-date	historical, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of joins
Number of records accessed	Tens	millions
Number of users	Thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Parallel DBMS Vendors:

What is a DBMS vendor?

(Data base Management System) Software that controls the organization, storage, retrieval, security and integrity of data in a database.

The major DBMS vendors are Oracle, IBM, Microsoft and Sybase (see Oracle Database, DB2, SQL Server and ASE).

D B M S	V e n d o r	Type	Primary Market
Access (Jet, MSDE)	Microsoft	R	Desktop
Adabas D	Software AG	R	Enterprise
Adaptive Server Anywhere	Sybase	R	Mobile/Embedded
Adaptive Server Enterprise	Sybase	R	Enterprise
Advantage Database Server	Extended Systems	R	Mobile/Enterprise
Datcom	Computer Associates	R	Enterprise
DB2 Everyplace	IBM	R	Mobile
Filemaker	FileMaker Inc.	R	Desktop
IDMS	Computer Associates	R	Enterprise
Ingres ii	Computer Associates	R	Enterprise

Interbase	Inprise (Borland)	R	Open Source
MySQL	Freeware	R	Open Source
NonStop SQL	Tandem	R	Enterprise
Pervasive.SQL 2000 (Btrieve)	Pervasive Software	R	Embedded
Pervasive.SQL Workgroup	Pervasive Software	R	Enterprise (Windows 32)
Progress	Progress Software	R	Mobile/Embedded
Quadbase SQL Server	Quadbase Systems, Inc.	Relational	Enterprise
R:Base	R:Base Technologies	Relational	Enterprise
Rdb	Oracle	R	Enterprise
Red Brick	Informix (Red Brick)	R	Enterprise (Data Warehousing)
SQL Server	Microsoft	R	Enterprise
SQLBase	Centura Software	R	Mobile/Embedded
SUPRA	Cincom	R	Enterprise
Teradata	NCR	R	VLDB (Data Warehousing)
YARD-SQL	YARD Software Ltd.	R	Enterprise
TimesTen	TimesTen Performance Software	R	In-Memory
Adabas	Software AG	XR	Enterprise
Model 204	Computer Corporation of America	XR	VLDB
UniData	Informix (Ardent)	XR	Enterprise
UniVerse	Informix (Ardent)	XR	Enterprise
Cache'	InterSystems	OR	Enterprise
Cloudscape	Informix	OR	Mobile/Embedded

DB2	IBM	OR	Enterprise/ VLDB
Informix Dynamic Server 2000	Informix	OR	Enterprise
Informix Extended Parallel Server	Informix	OR	VLDB (Data Warehousing)
Oracle Lite	Oracle	OR	Mobile
Oracle 8I	Oracle	OR	Enterprise
PointBase Embedded	PointBase	OR	Embedded
PointBase Mobile	PointBase	OR	Mobile
PointBase Network Server	PointBase	OR	Enterprise
PostgreSQL	Freeware	OR	Open Source
UniSQL	Cincom	OR	Enterprise
Jasmine ii	Computer Associates	OO	Enterprise
Object Store	Exceleron	OO	Enterprise
Objectivity DB	Objectivity	OO	VLDB (Scientific)
POET Object Server Suite	Poet Software	OO	Enterprise
Versant	Versant Corporation	OO	Enterprise
Raima Database Manager	Centura Software	RN	Mobile/Embedded
Velocis	Centura Software	RN	Enterprise/ Embedded
Db.linux	Centura Software	RNH	Open Source/ Mobile/Embedded
Db.star	Centura Software	RNH	Open Source/ Mobile/Embedded

Types of Data Warehouse:

There are three main types of DWH. Each has its specific role in data management operations.

1. Enterprise Data Warehouse

Enterprise data warehouse (EDW) serves as a central or main database to facilitate decision-making throughout the enterprise. Key benefits of having an EDW include access to cross-organizational information, the ability to run complex queries, and the enablement of enriched, far-sighted insights for data-driven decisions and early risk assessment.

2. ODS (Operational Data Store)

In ODS, the DWH refreshes in real-time. Therefore, organizations often used it for routine enterprise activities, such as storing records of the employees. Business processes also use ODS as a source for providing data to the EDW.

3. Data Mart

It is a subset of a DWH that supports a particular department, region, or business unit. Consider this: You have multiple departments, including sales, marketing, product development, etc. Each department will have a central repository where it stores data. This repository is called a data mart. The EDW stores the data from the data mart in the ODS on a daily/weekly (or as configured) basis. The ODS acts as a staging area for data integration. It then sends the data to the EDW to store it and use it for BI purposes.

DATAWAREHOUSE COMPONENTS:

The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key components to make the entire environment functional, manageable and accessible

There are mainly five components of Data Warehouse:

DATA WAREHOUSE DATABASE:

The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained

by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins, aggregates are resource intensive and slow down performance.

Hence, alternative approaches to Database are used as listed below-

In a data warehouse, relational databases are deployed in parallel to allow for scalability. Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.

New index structures are used to bypass relational table scan and improve speed.

Use of multidimensional database (MDDBs) to overcome any limitations which are placed because of the relational data model. Example: Essbase from Oracle.

SOURCING, ACQUISITION, CLEAN-UP AND TRANSFORMATION TOOLS (ETL):

The data sourcing, transformation, and migration tools are used for performing all the conversions, summarizations, and all the changes needed to transform data into a unified format in the datawarehouse. They are also called Extract, Transform and Load (ETL) Tools.

These Extract, Transform, and Load tools may generate cron jobs, background jobs, Cobol programs, shell scripts, etc. that regularly update data in datawarehouse. These tools are also helpful to maintain the Metadata.

These ETL Tools have to deal with challenges of Database & Data heterogeneity.

METADATA:

The name Meta Data suggests some high- level technological concept. However, it is quite simple. Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse.

In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.

QUERY TOOLS:

One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system.

These tools fall into four different categories:

Query and reporting tools

Application Development tools

Data mining tools

OLAP tools

Characteristics of OLAP:

The main characteristics of OLAP are as follows:

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.

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.

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.

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

Uniform documenting performance: Increasing the number of dimensions or database size should not significantly degrade the reporting performance of the OLAP system.

OLAP provides for distinguishing between zero values and missing values so that aggregates are computed correctly.

OLAP system should ignore all missing values and compute correct aggregate values.

OLAP facilitate interactive query and complex analysis for the users.

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.

OLAP provides the ability to perform intricate calculations and comparisons.

OLAP presents results in a number of meaningful ways, including charts and graphs.

OLAP Types:

Three types of OLAP servers are:-

1. Relational OLAP (ROLAP)
2. Multidimensional OLAP (MOLAP)
3. Hybrid OLAP (HOLAP)

1. Relational OLAP (ROLAP):

Relational On-Line Analytical Processing (ROLAP) work mainly for the data that resides in a relational database, where the base data and dimension tables are stored as relational tables. ROLAP servers are placed between the relational back-end server and client front-end tools. ROLAP servers use RDBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.

Advantages of ROLAP:

1. ROLAP can handle large amounts of data.
2. Can be used with data warehouse and OLTP systems.

Disadvantages of ROLAP:

1. Limited by SQL functionalities.
2. Hard to maintain aggregate tables.

2. Multidimensional OLAP (MOLAP):

Multidimensional On-Line Analytical Processing (MOLAP) support multidimensional views of data through array-based multidimensional storage engines. With multidimensional data stores, the storage utilization may be low if the data set is sparse.

Advantages of MOLAP

1. Optimal for slice and dice operations.
2. Performs better than ROLAP when data is dense.
3. Can perform complex calculations.

Disadvantages of MOLAP

1. Difficult to change dimension without re-aggregation.
2. MOLAP can handle limited amount of data.

3. Hybrid OLAP (HOLAP):

Hybrid On-Line Analytical Processing (HOLAP) is a combination of ROLAP and MOLAP. HOLAP provide greater scalability of ROLAP and the faster computation of MOLAP.

Advantages of HOLAP

1. HOLAP provide advantages of both MOLAP and ROLAP.
2. Provide fast access at all levels of aggregation.

Disadvantages of HOLAP

1. HOLAP architecture is very complex because it support both MOLAP and ROLAP servers.