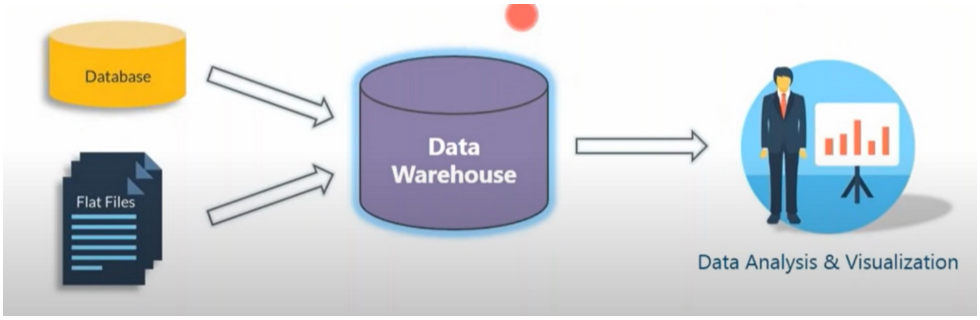
Data Warehouse & Modelling

Monday, January 3, 2022 4:27 PM

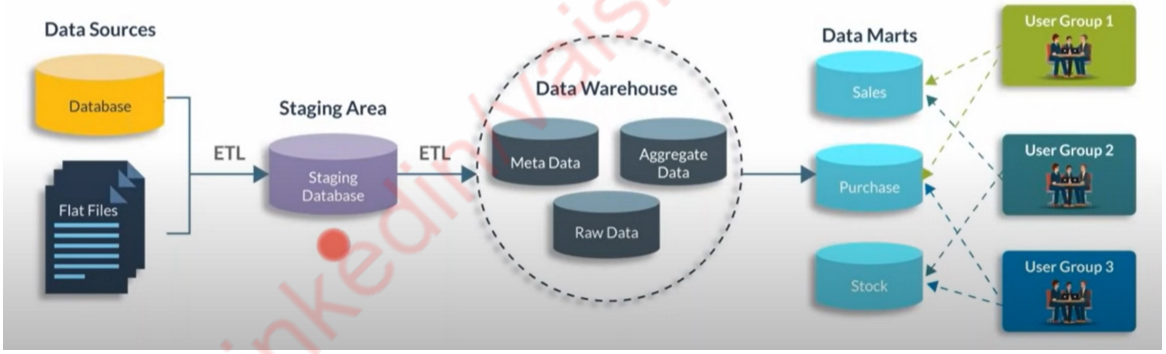
Data Warehouse:

What is a data warehouse ?
-->It is a analytical based collection of data from various sources
-->It is mainly used in analysis of historical data
-->It is a central location where data from various sources are stored and queried.



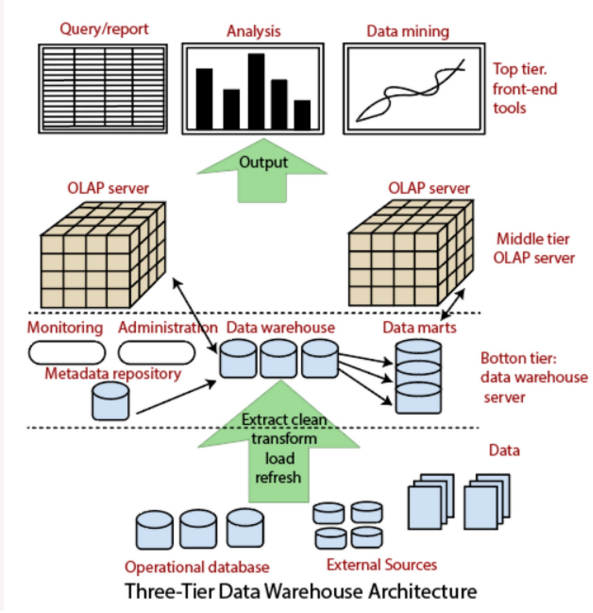
What is Data Warehousing ?
-->It is the process of organizing & storing data in such a way that it is efficient to query the data based on our needs.
-->Its also called as the process of transforming data into information

Architecture:
-->Data Sources: Here we have the data from various sources in various forms which has to be moved into data warehouses for analytical processing.
-->Staging Area: Here ETL is performed (extract data, transform it and load into data warehouse)
-->Data Warehouse: It consists of transformed data from ETL which comprises of meta data (data about data), aggregated data (from tables/DB's) and raw data
-->Data Marts: It contains only certain part of data from data warehouse which is a subset of DWH and provides secure access to users. It can be divided based on logical grouping of data. As entire organization data is present in data warehouse and there were few restrictions placed to share the data with everyone, we have data marts created which has restricted the access to data.



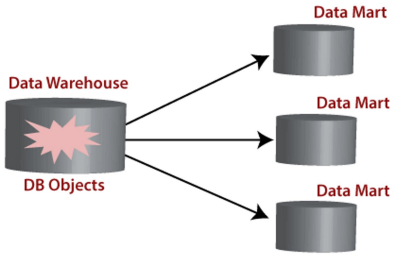
Three-Tier Architecture:
DWH have 3-level architecture that consists of

1. Bottom Tier (DWH server)
2. Middle Tier (OLAP server)
3. Top Tier (front end tools)



Three-Tier Data Warehouse Architecture

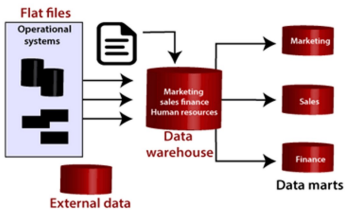
What is a Data Mart?
-->Data Marts are analytical record store derived from a data warehouse.
-->They are used to serve a particular business function(sales, marketing, products etc) to a specific community within the organization.
-->They are subsets of data warehouse
-->Its main usage is in BI(business intellegence) applications



Advantages:
-->Proper distinction of business use cases using data marts.
-->Frequently queried data can be easily accessed
-->Easy to implement rather than entire DWH
-->Only essential bussiness data is present
-->Limit the access of sensitive data by putting constraints
-->Secure access to the data

What are different types of data marts ?

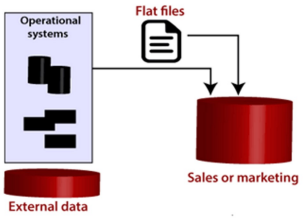
Dependent Data Mart:
-->Dependent Data marts are logical subset of the physical DWH
-->Here DWH is firstly created from which data marts are created
-->Data marts get the data required from DWH due to which they are dependent.
-->As DWH has all the data there is no need for integration



Dependent Data Mart

Independent Data Mart:
-->Here data marts are firstly created

-->Later DWH are designed based on the independent data marts
-->Integration of data marts are required as it has all the data.



Independent Data Mart

Hybrid Data Marts:
-->It allows us to combine input from data sources apart from DWH

Steps to implement a Data Mart:
-->It involves design of schema, constructing storage space, populating data from different sources, access/query data for business usage and manage it over time.

Designing:
-->It involves designing physical and logical architecture of the system. It invovles the below steps
1. To gather business and technical requirements
2. To dentify the data sources

Constructing:
-->It involves creating the physical database.
1. To create logical structures like tablespace(db storage & management objects)
2. Creating db schema objects like indexes, tables

Populating:
-->It involves getting the access from various data sources to be integrated into it.
1. Extract, clean and transform the data
2. Load data into data mart
3. Create and store metadata

Accessing:
-->It involves to finally start using the data mart created for business use cases.
-->we can access and query the data

Managing:
-->It involves managing data over time and making changes to it when necessary.
1. Providing secure access to data
2. Optimizing the system for better performance

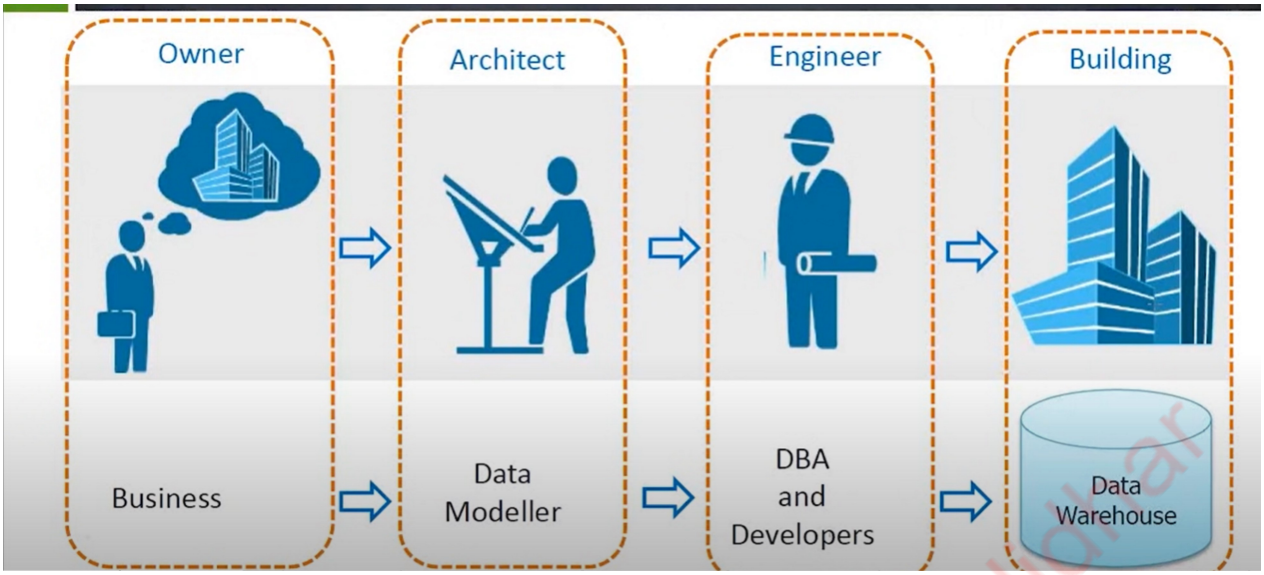
Data WareHouse VS Data Mart

DWH	DM
1. DWH is a complete repository of organization data 2. It holds data of multiple subject lines 3. It holds very detailed information of data 4. It uses fact constellation/galaxy schema 5. It is data oriented	1. DM is a subset of DWH 2. It holds data from one particular subject line like products, sales 3. It holds mainly summarized data 4. It uses star or snowflake schema 5. It is project/subject line oriented

What is a data model ?
--> Data models are used to organize the data well so that it is easy to understand and query the data for analysis and report generation.
--> Data model acts as a blueprint to developers in building the data warehouse.
-->In the data model we have the data tables, primary keys, foreign keys and relationship between them.

Who are involved in building data models ?
-->Business : They define the requirement to solve a particular bussiness problems
-->Architect: They produce the logical data models considering various constraints

-->Engineer: They build/develop the physical data model based on inputs from data architect
-->Data warehouse: Final data system is used by the business for their usage.



ER Diagrams (Entity Relationship)
-->Here we see how the objects are related to each other.
Ex. One-to-one , one-to-many, many-to-one and many-to-many

Types of Data Models:

Conceptual Data Model	Logical Data Model	Physical Data Model
1. High level overview of data modelling 2. Mainly comprises of requirement gathering 3. Mainly designed for business consumers to get an idea about the data warehouse	1. Entity and their relationship is described in detail in this model 2. Mainly comprises of in-depth description of the data tables 3. It is has facts and dimensions 4. Mainly designed to describe the logical view of data warehouse	1. It represents how the model will be built in a database. 2. It shows up entities, column names, PK, FK, constraints 3. Physical model is then converted into DDL statements 4. Mainly deigned to show the final view of a data warehouse as developed by the developers

OLAP (online analytical processing)
--> It helps in analysis of historical data and multidimensional data (data which can be viewed in various forms/angles)
-->Data warehouses are built based on OLAP
-->OLTP uses data in 2-dimensional format comprising of rows and columns

Types of OLAP Cubes:

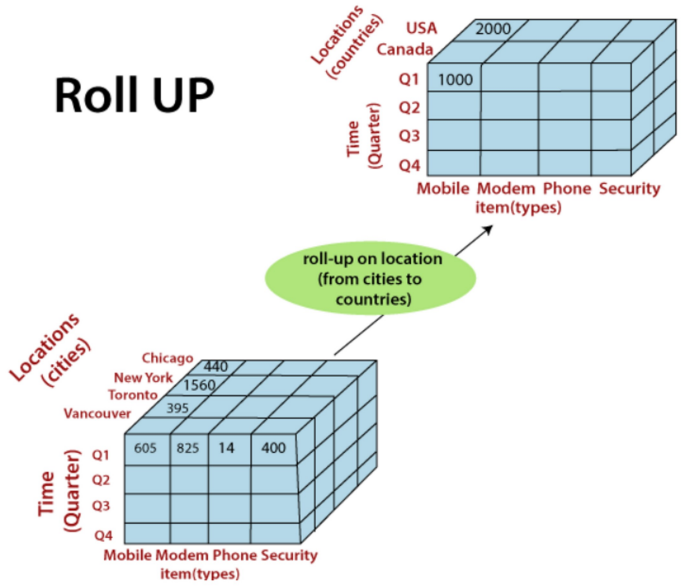
MOLAP(multi-dimensional)	ROLAP (relational)	HOLAP (hybrid model)
1. Used to store multi-dimensional data directly into a multi-dimensional database. 2. Pro: Excellent performance 3. Con: Limited amount of data can be handled	1. Used to store multi-dimensional data directly into a relational database. 2. Pro: Huge amounts of data can be processed 3. Con: Requires high amount of processing and resources	1. It is a hybrid model of both MOLAP and ROLAP 2. Pro: It can drill through the cube into underlying relational data

OLAP Operations: Roll-Up
-->Roll-Up is also known as aggregation operation.
-->Data is aggregated on a data cube by dimension reduction.

Here, the temperature of cities is aggregated into countries
Sum(city temp)=country temp
Ie (chicago+newyork)=USA and (toronto+vancouver)=canada

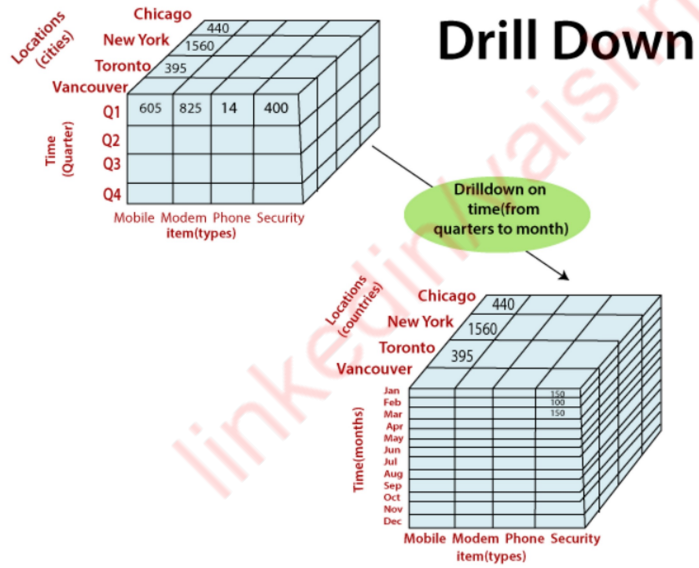
The roll-up operation groups the information by levels of temperature.

The following diagram illustrates how roll-up works.



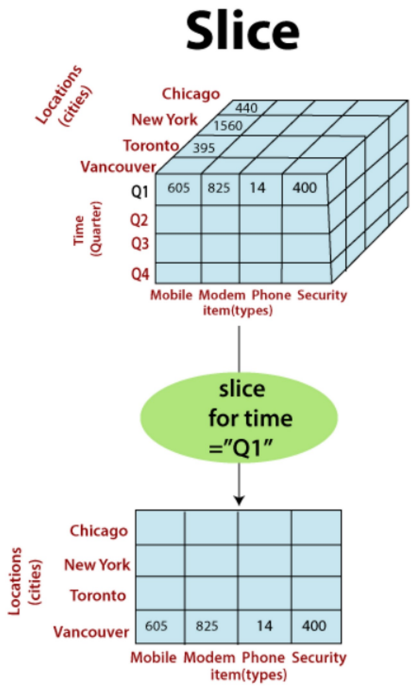
Roll Down:
-->Data is rolled down based on adding more details to given data which can also result in adding new dimensions.
Ex. Here Q1(quarter) is broken down into months
First quarter(Q1) ==> jan+feb+mar+apr

The following diagram illustrates how Drill-down works.



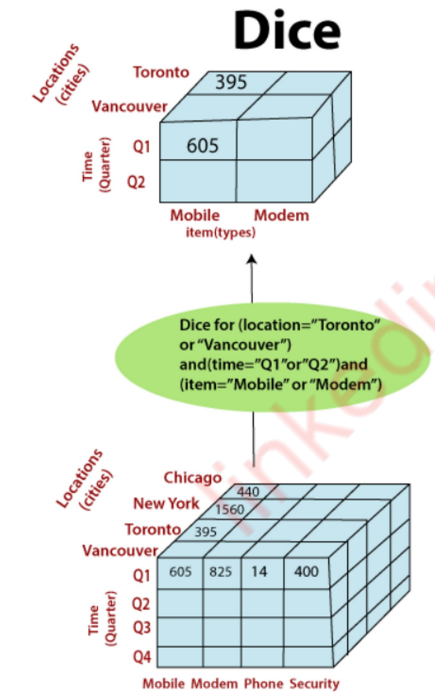
Slice :
-->Slice operation results in slicing of the data cube which results in sub-cube.
-->Ex. When we perform a selection of one dimension on a 3-dimensional it becomes 2-dimensional

The following diagram illustrates how Slice works.

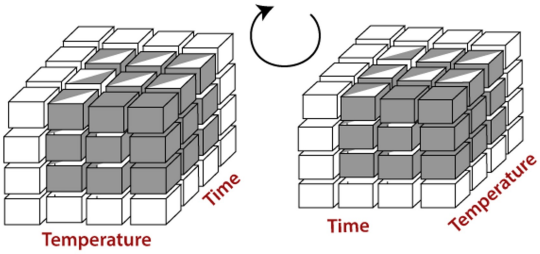


Dice:
-->It is used to describe a subcube by performing selection on two or more dimensions
-->Ex. Selection is done based on 3 dimensions namely (location, time, item)

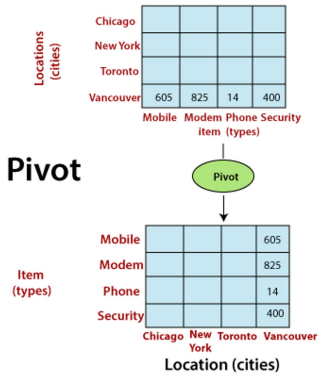
Consider the following diagram, which shows the dice operations.



Pivot:
-->It is also known as rotation operation where data axes are rotated to provide different visual representation of data.
-->It also includes swapping of rows into columns and vice versa.



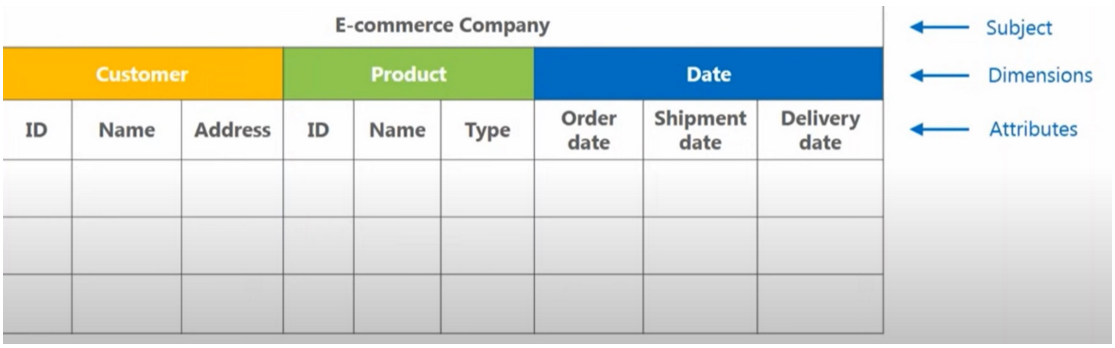
Consider the following diagram, which shows the pivot operation.



What is Dimensional Modelling:
-->To model the data in various dimensions making it suitable for OLAP data management. It mainly comprises of 2 key components namely facts and dimensions.

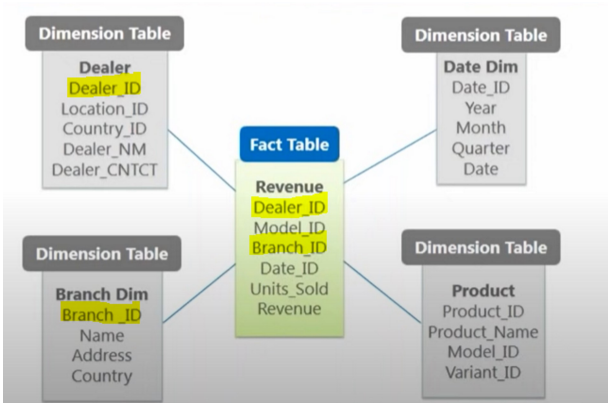
Fact : They are numerical transactional data. (Ex. productsSold, totalSalesAmount, totalCustomers)
Fact Table: It contains numerical value data which we measure.
-->It contains the foreign keys of dimensional tables
-->It contains less columns and more rows as attributes are not present in them
-->Fact Table= Dimension key(foreign key for dimension table) + measure
-->Ex facts: totalproductsSold, customersVisited,totalAmount

Dimension : They provide reference information for the facts (Ex. productName, customerCity, customerProduct)
Dimensional Table: It contains the details about the facts
-->It contains descriptive data about the the facts and attributes about the facts
-->It has more columns as tables are denormalized
-->Ex. itemSold,customerCity,timeofSell

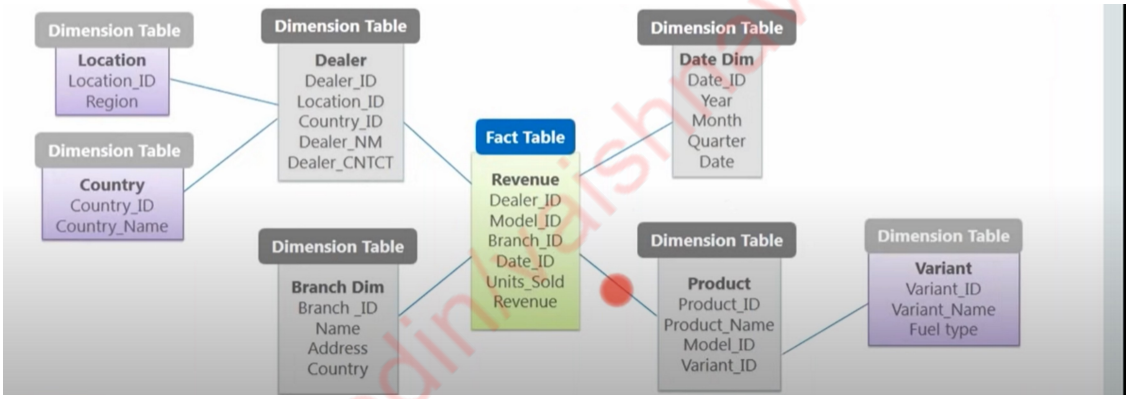


Schema in a DWH:
-->Schema describes the logical representation of data in a table belonging to a data base.
-->It consists of primary key, foreign key, constraints, data types, reltationship between tables etc
-->Schema in a DWH is classified into 2 major types namely star schema and snowflake schema and fact constellation schema

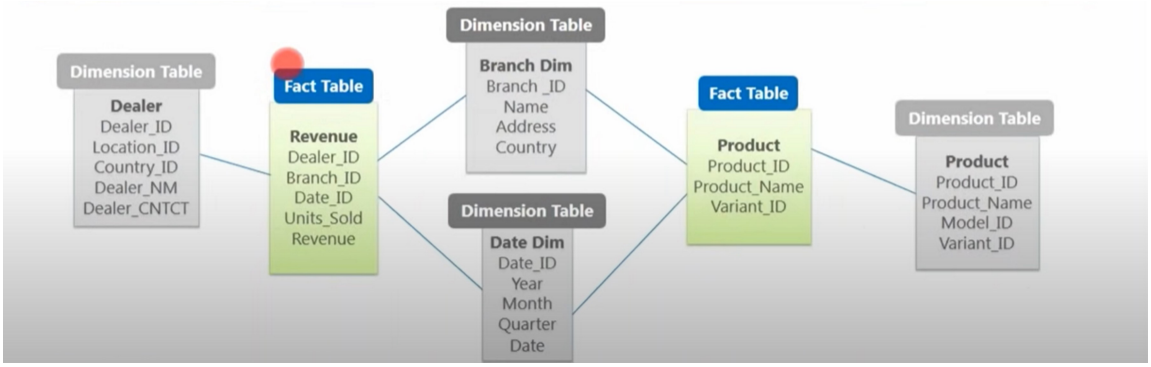
Star Schema:
-->In this schema, each dimension is represented in a single table with a set of attribues.
-->Fact table is at the centre which contains foreign keys of dimension table + numerical measures (facts)
-->Every dimension table is connected to a single fact table



Snowflake Schema:
-->In this schema dimension tables are normalized (divided into smaller tables)
-->In 'dealer' table we have foreign keys (location, country) from the normalized dimension tables 'location' and 'country_id'
-->Dealer ==> location + country



Fact constellation /Galaxy Schema:
-->In this schema we can have more than 1 fact table
-->A single dimension table can be shared by more than 1 fact table
-->Dimensions which are shared are called conformed dimensions

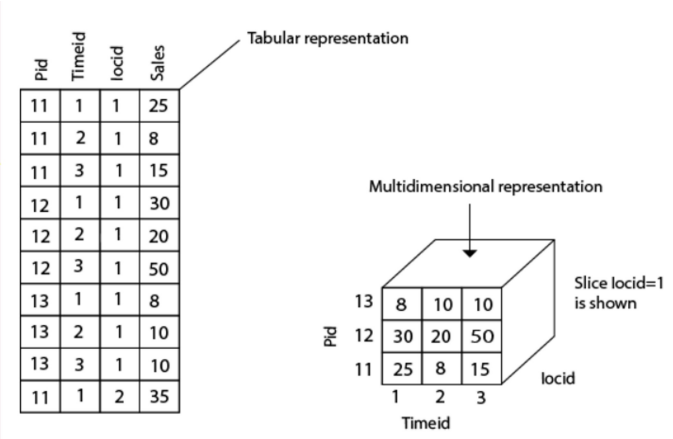


What is Multi-Dimensional data model ?

-->Multi dimensional data model is the one which can be used to model and view the data in different dimensions. It is defined by facts and dimensions.

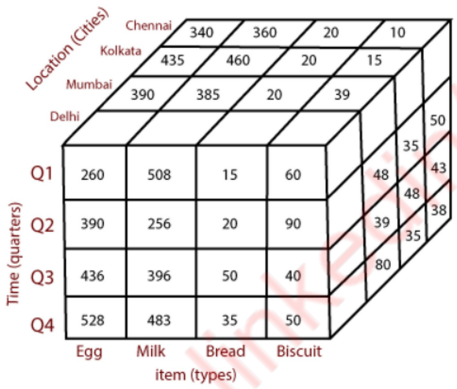
-->It is built around a central theme ie sales, marketing, cardTransactions.

-->Ex. Conversion of 2-dimensional model into 3-dimensional model



	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 a data cube ?

-->when data is grouped/combined in multi-dimensional models it is termed as data cube.