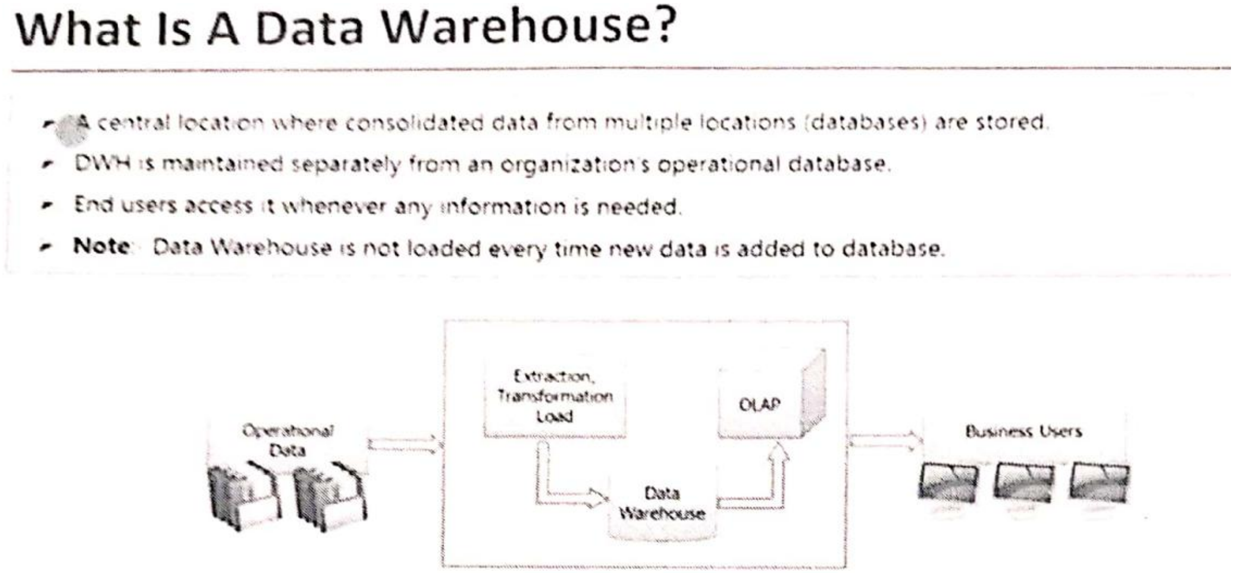
First sessional Paper Solution

Data Warehousing & Data Mining

SECTION – A

1.



What Is A Data Warehouse?

- A central location where consolidated data from multiple locations (databases) are stored.
- DWH is maintained separately from an organization's operational database.
- End users access it whenever any information is needed.
- **Note**: Data Warehouse is not loaded every time new data is added to database.

Operational Data

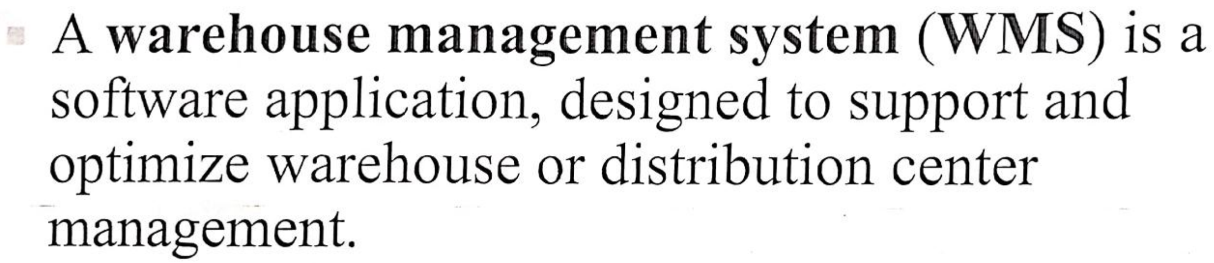
Extraction, Transformation Load

OLAP

Data Warehouse

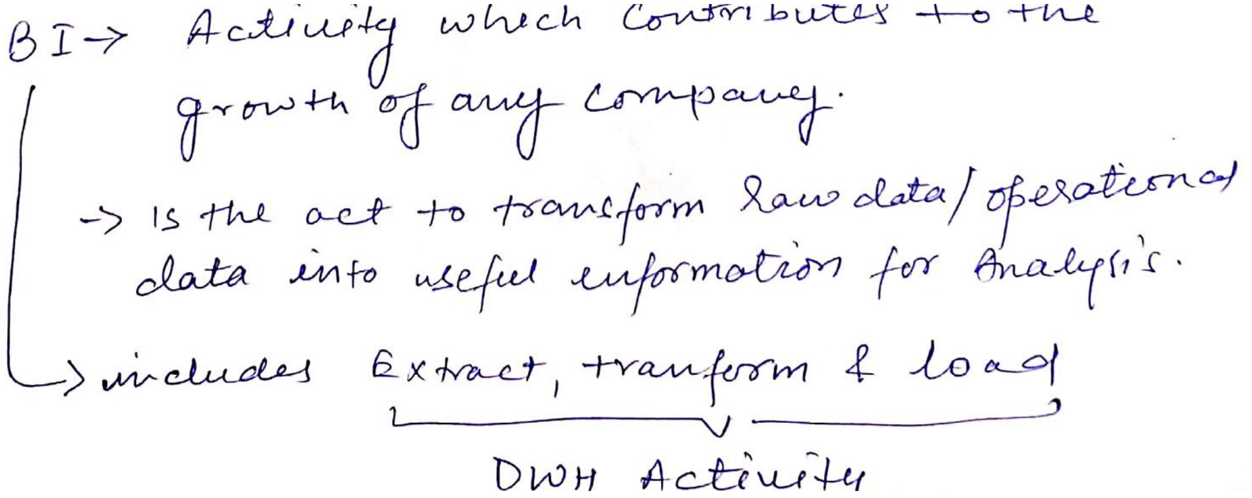
Business Users

2.



- A **warehouse management system (WMS)** is a software application, designed to support and optimize warehouse or distribution center management.

3.

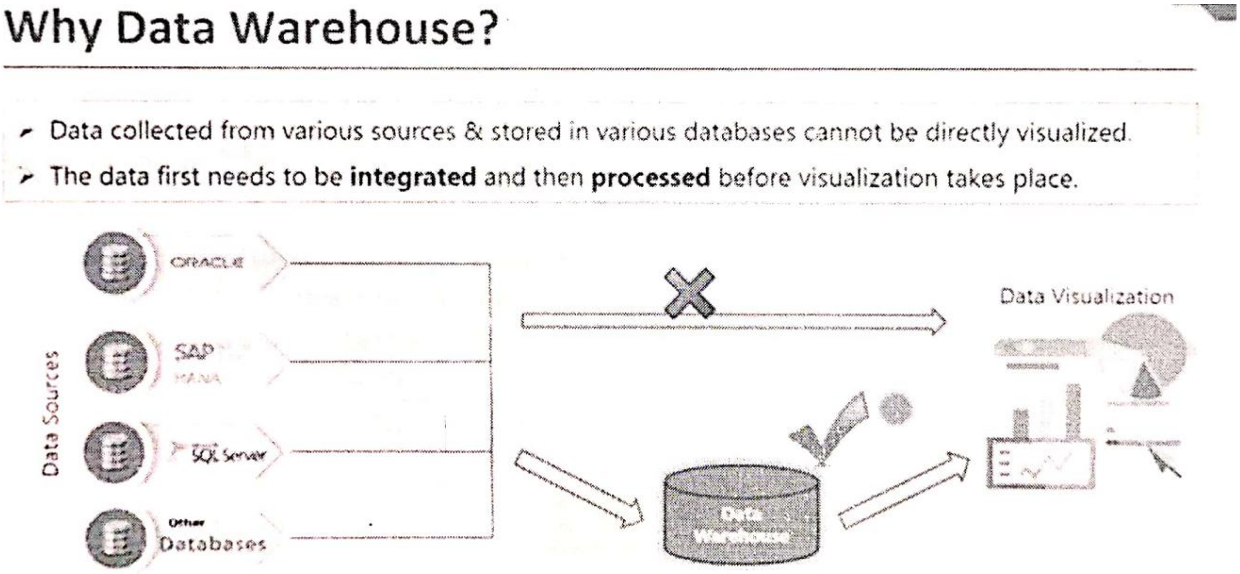


BI → Activity which contributes to the growth of any company.

→ Is the act to transform Raw data / operational data into useful information for Analysis.

→ includes $\underbrace{\text{Extract, tranform \& load}}_{\text{DWH Activity}}$

4.



Why Data Warehouse?

- Data collected from various sources & stored in various databases cannot be directly visualized.
- The data first needs to be **integrated** and then **processed** before visualization takes place.

5.

Metadata is simply defined as data about data. The data that is 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.

SECTION – B

1.

S.NO	STAR SCHEMA	FACT CONSTELLATION SCHEMA
1.	A star schema depicts each dimension with only one-dimension table.	While in this, dimension tables are shared by many fact tables.
2.	In star schema, tables can be maintained easily in comparison of fact constellation schema.	While in fact constellation schema, tables cannot be maintained easily comparatively.
3.	Star schema does not use normalization.	Whereas it is a normalized form of star and snowflake schema.
4.	In star schema, simple queries are used to access data from the database.	While in this, heavily complex queries are used to access data from the database.
5.	Star schema is easy to operate as compared to fact constellation schema as it has less number of joins between the tables.	While fact constellation schema is not easy to operate as compared to star schema as it has many joins between the tables.
6.	Star schema uses less space as compared to fact constellation schema.	While fact constellation schema uses more space comparatively.
7.	It is very simple to understand due to its simplicity.	While it is very difficult to understand due to its complexity.

2.

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

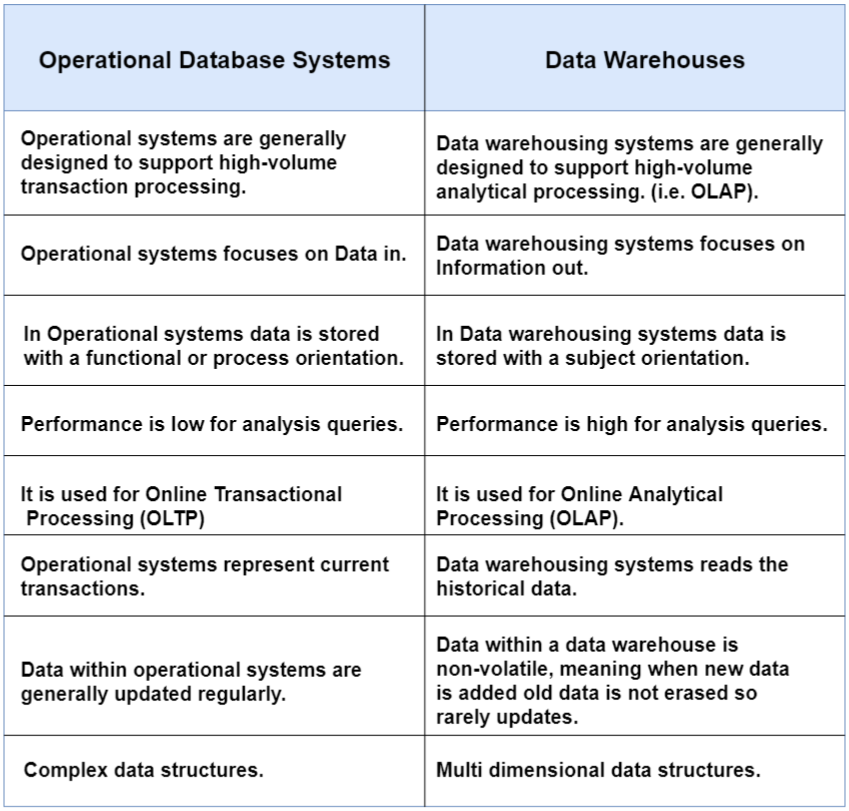
Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

An international manufacturing company can have many subject areas like raw material, location, transport, site of company, size, number of employees, branches, investment, product, consumer, marketing, advertisement, delivery, offers n discounts, profit etc. The subject areas can be all the dimensions related to the manufacturing business.

An International Manufacturing companies.

- Shipments (shipment is the major subject for an internationalManufacturing companies)
- Manufacturing(it contains product,order etc dimension)
- Financial Management(it contains sales, price,account,contract)
- Customers
- Time
- Supplier

3.



Operational Database Systems	Data Warehouses
Operational systems are generally designed to support high-volume transaction processing.	Data warehousing systems are generally designed to support high-volume analytical processing. (i.e. OLAP).
Operational systems focuses on Data in.	Data warehousing systems focuses on Information out.
In Operational systems data is stored with a functional or process orientation.	In Data warehousing systems data is stored with a subject orientation.
Performance is low for analysis queries.	Performance is high for analysis queries.
It is used for Online Transactional Processing (OLTP)	It is used for Online Analytical Processing (OLAP).
Operational systems represent current transactions.	Data warehousing systems reads the historical data.
Data within operational systems are generally updated regularly.	Data within a data warehouse is non-volatile, meaning when new data is added old data is not erased so rarely updates.
Complex data structures.	Multi dimensional data structures.

4.

Granularity

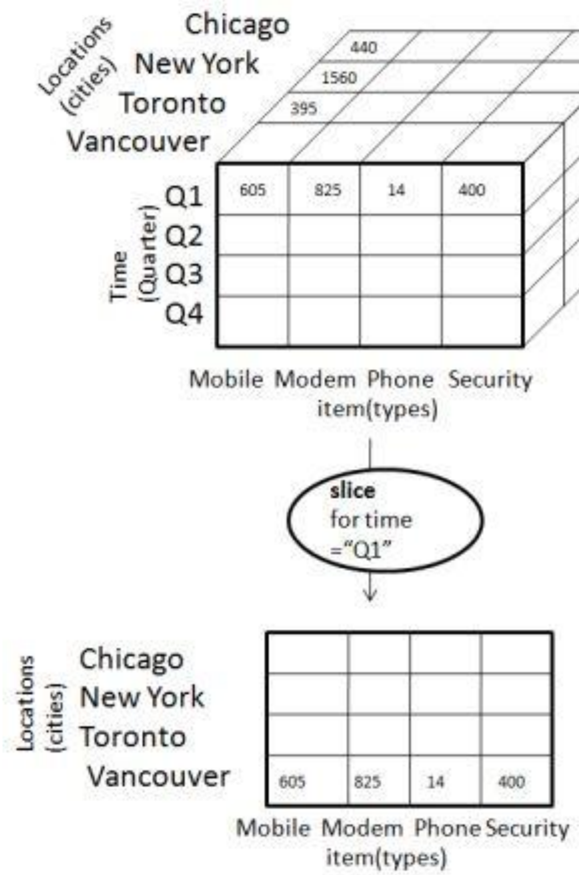
Granularity means the level of detail of your data within the data structure. In a typical Data Warehouse one might find very detailed data (such as seconds, single product, and one specific attribute) and aggregated data (such as total number of, monthly orders, all products).

The higher the granularity of a fact table the more data (or in an excel sheet: rows) you will have. But the granularity of your data also determines what kind of information you can get out of the stored data. So to aggregate data you need of course the same granularity. (A weekly report can only be generated when you have time related data stored. At least it should be a “week”, better it is to have “day”.)

SECTION - C

3.1.Slice

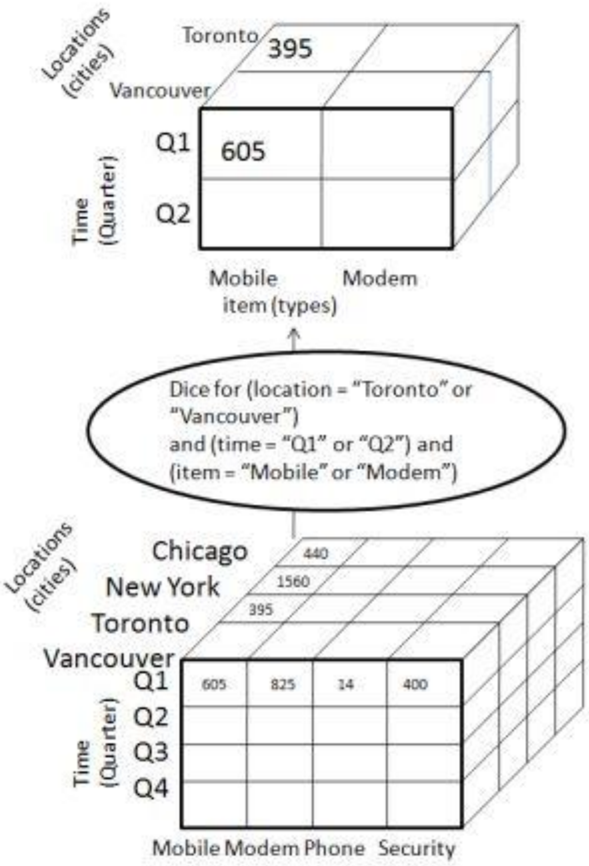
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



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

Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.



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

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

3.2.

A **dimension table** contains a higher granular information so have less no of records and it needs to have all the necessary details (More columns) related to the grain of the table. On the other side, A **fact table** has the lowest level grain of a subject area. Lower grain cause more number of rows in the Fact table.

Visualize it…

You need more lookup data to fit in a single row in dimensions.. Its lookup data, it wont have as many records as the fact, but it would have far more columns. A Visually, if you think of dimension table as a excel sheet, you would be scrolling horizontally to see all the data, thus the phrase, “wide”

On the other hand, The fact stores actual measures, transactional data at the lowest possible granularity.. resulting in far more no of rows. Visually speaking, if you represent a fact table as an excel table it would have too many rows … making you scroll down…. thus the “deep” phrase.

4.1

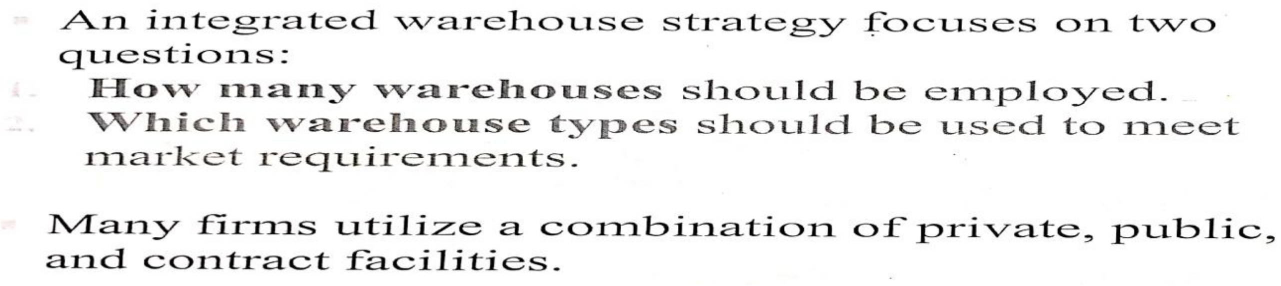
A **data mart** is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse.
A data mart is a condensed version of Data Warehouse and is designed for use by a specific department, unit or set of users in an organization. E.g., Marketing, Sales, HR or finance. It is often controlled by a single department in an organization.
Data Mart usually draws data from only a few sources compared to a Data warehouse. Data marts are small in size and are more flexible

There are three main types of data marts are:
Dependent: Dependent data marts are created by drawing data directly from operational, external or both sources.
Independent: Independent data mart is created without the use of a central data warehouse.
Hybrid: This type of data marts can take data from data warehouses or operational systems.

4.2

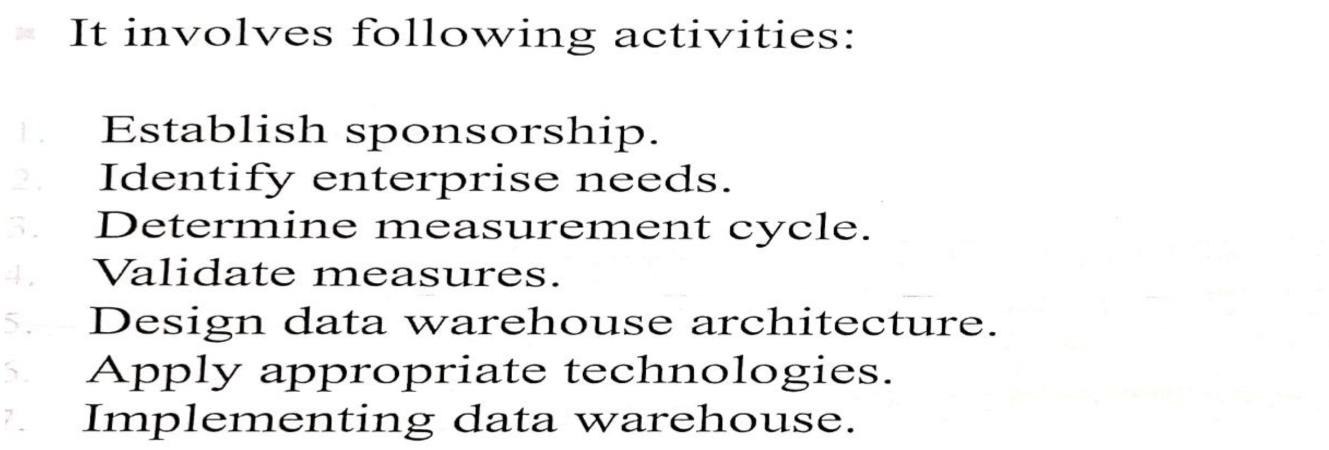
ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. **Full form of ETL is Extract, Transform and Load**. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.
In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

5.1



- An integrated warehouse strategy focuses on two questions:
  1. **How many warehouses** should be employed.
  2. **Which warehouse types** should be used to meet market requirements.

- Many firms utilize a combination of private, public, and contract facilities.



- It involves following activities:

1. Establish sponsorship.
2. Identify enterprise needs.
3. Determine measurement cycle.
4. Validate measures.
5. Design data warehouse architecture.
6. Apply appropriate technologies.
7. Implementing data warehouse.

5.2

OLAP is a category of software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view. OLAP stands for Online Analytical Processing.

OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy.

OLAP cube:

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

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

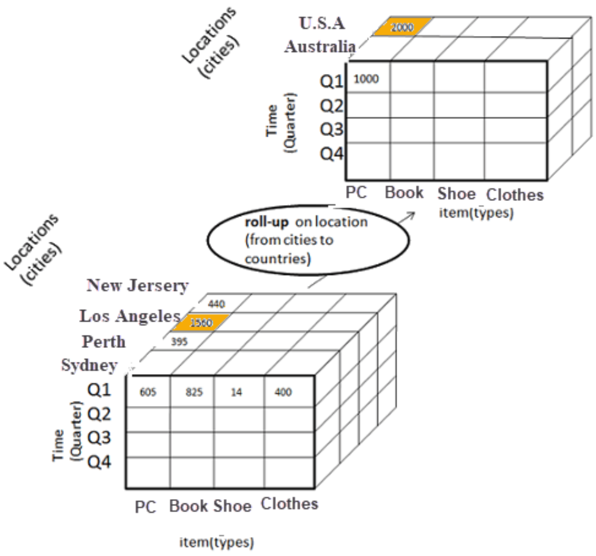
Four types of analytical operations in OLAP are:

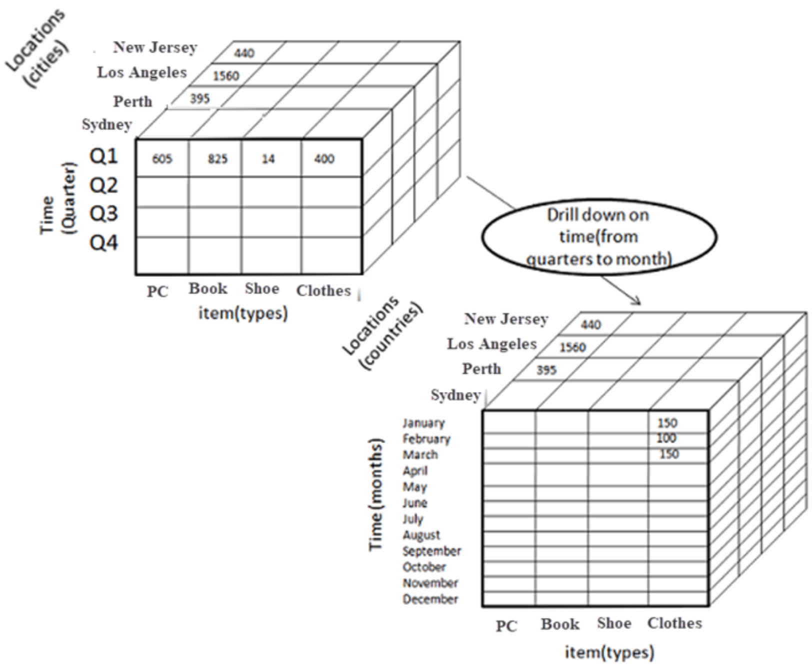
Roll-up

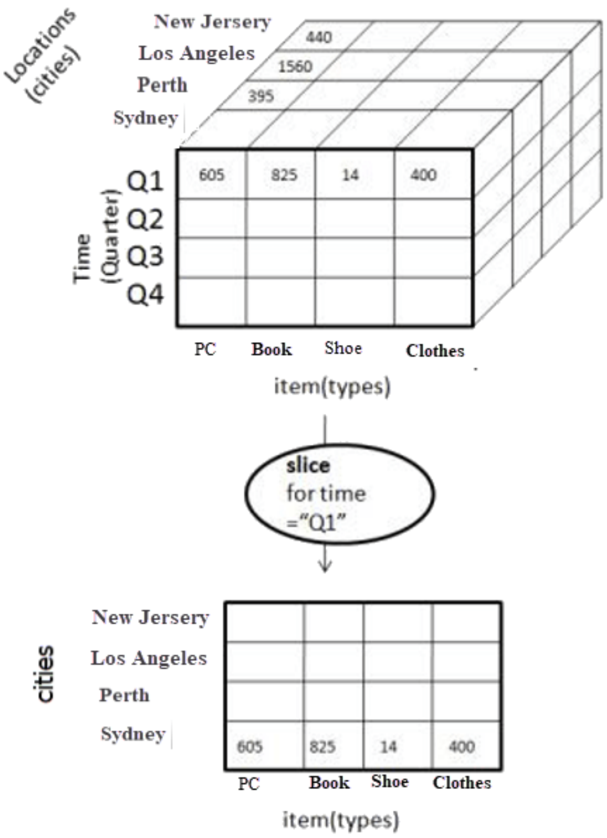
Drill-down

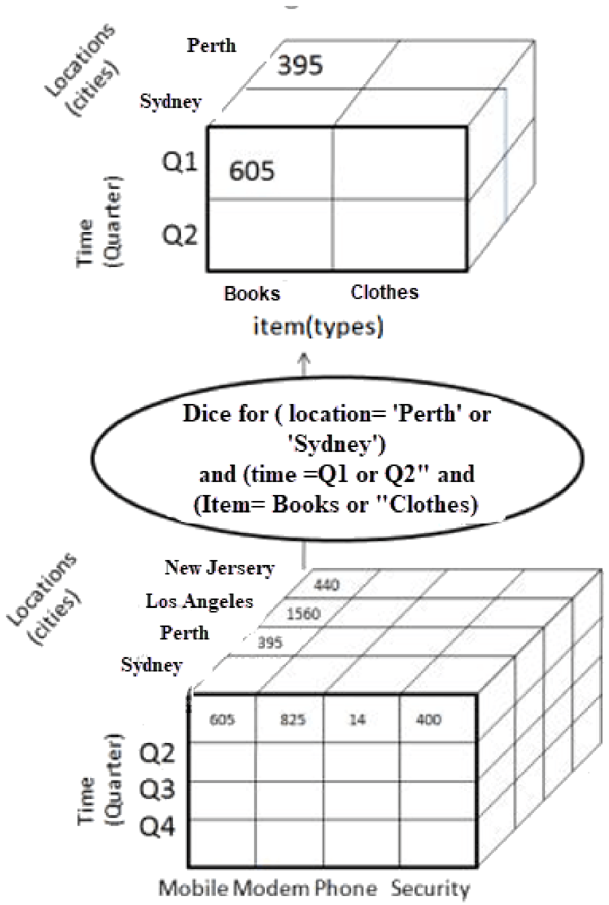
Slice and dice

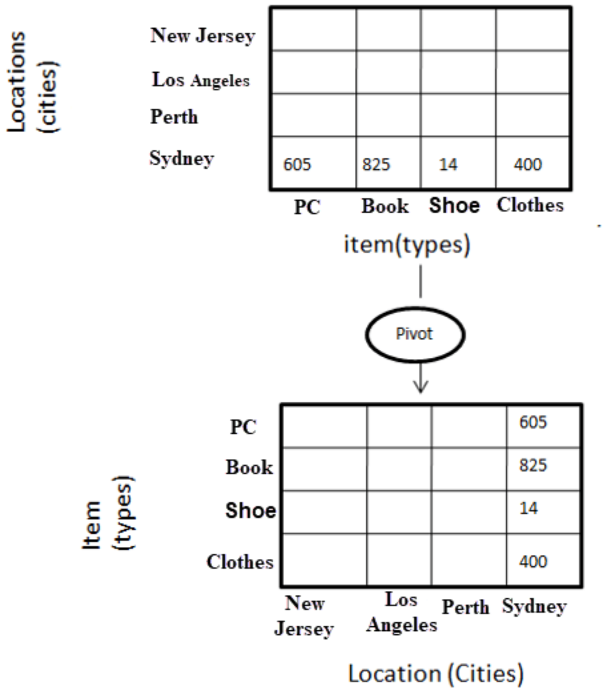
Pivot (rotate)




Locations (cities)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
Drill down on time(from quarters to month)
Locations (countries)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
PC
Book
Shoe
Clothes
item(types)


Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
slice for time ="Q1"
cities
New Jersery
Los Angeles
Perth
Sydney
605
825
14
400
PC
Book
Shoe
Clothes
item(types)


Locations (cities)
Perth
Sydney
395
Q1
605
Q2
Time (Quarter)
Books
Clothes
item(types)
Dice for (location= 'Perth' or 'Sydney')
and (time =Q1 or Q2" and
(Item= Books or "Clothes)
Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
605
825
14
400
Time (Quarter)
Q2
Q3
Q4
Mobile
Modem
Phone
Security


Locations (cities)
New Jersey
Los Angeles
Perth
Sydney
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
Pivot
Item (types)
PC
Book
Shoe
Clothes
605
825
14
400
New Jersey
Los Angeles
Perth
Sydney
Location (Cities)