

Chapter 2

1. Data marts:

- 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.
- A data mart is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse.
- Data marts are small in size and are more flexible compared to a Datawarehouse.

2. Why do we need Data Mart?

- Data Mart helps to enhance user's response time due to reduction in volume of data
- It provides easy access to frequently requested data.
- Data mart are simpler to implement when compared to corporate Datawarehouse. At the same time, the cost of implementing Data Mart is certainly lower compared with implementing a full data warehouse.
- Compared to Data Warehouse, a datamart is agile. In case of change in model, datamart can be built quicker due to a smaller size.
- A Datamart is defined by a single Subject Matter Expert. Hence, Data mart is more open to change compared to Datawarehouse.
- Data is partitioned and allows very granular access control privileges.
- Data can be segmented and stored on different hardware/software platforms.

3. Type of Data Mart

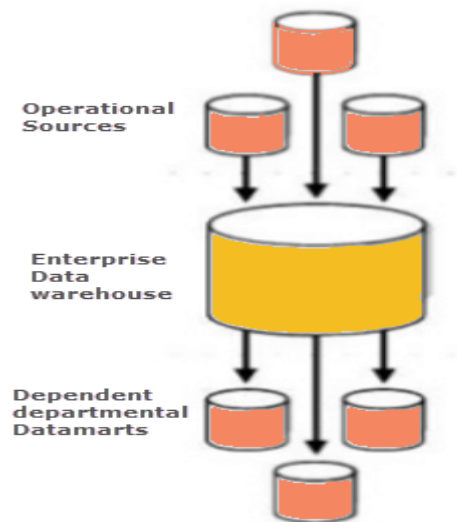
There are three main types of data marts are:

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

❖ Dependent Data Mart

A dependent data mart allows sourcing organization's data from a single Data Warehouse. It offers the benefit of centralization. If you need to develop one or more physical data marts, then you need to configure them as dependent data marts.

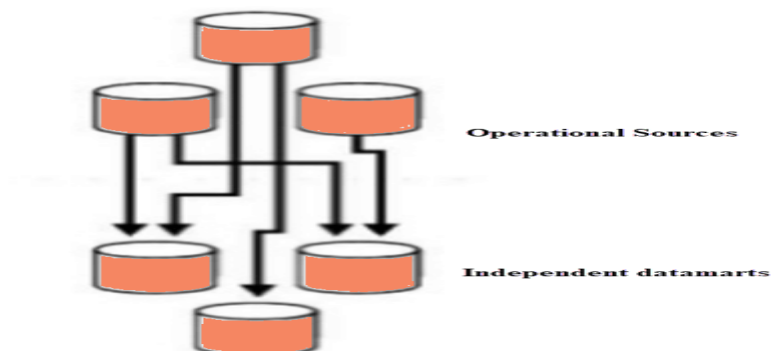
Dependent data marts can be built in two different ways. Either where a user can access both the data mart and data warehouse, depending on need, or where access is limited only to the data mart. The second approach is not optimal as it produces sometimes referred to as a data junkyard. In the data junkyard, all data begins with a common source, but they are scrapped, and mostly junked.



❖ Independent Data Mart

An independent data mart is created without the use of central Data warehouse. This kind of Data Mart is an ideal option for smaller groups within an organization.

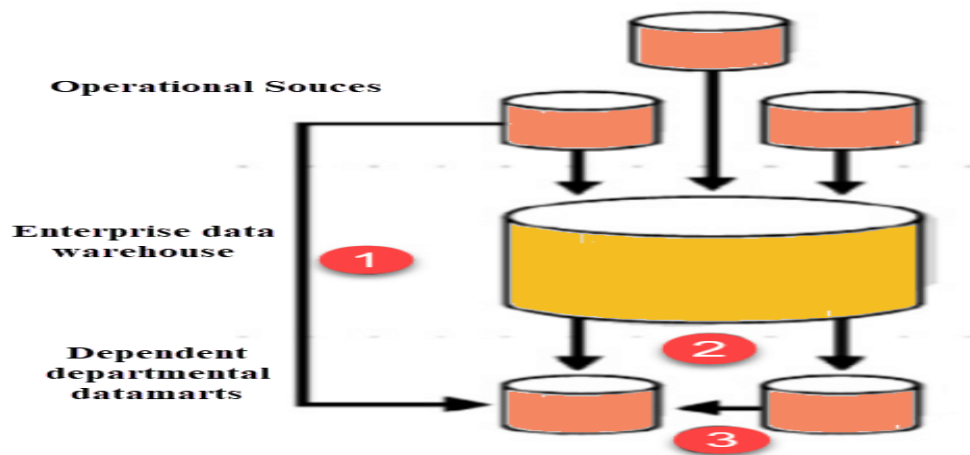
An independent data mart has neither a relationship with the enterprise data warehouse nor with any other data mart. In Independent data mart, the data is input separately, and its analyses are also performed autonomously.



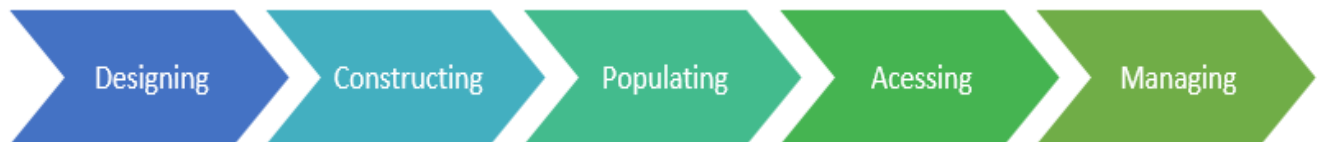
❖ Hybrid data Mart:

A hybrid data mart combines input from sources apart from Data warehouse. This could be helpful when you want ad-hoc integration, like after a new group or product is added to the organization.

It is best suited for multiple database environments and fast implementation turnaround for any organization. It also requires least data cleansing effort. Hybrid Data mart also supports large storage structures, and it is best suited for flexible for smaller data-centric applications.



4. Steps in Implementing a Datamart



Implementing a Data Mart is a rewarding but complex procedure. Here are the detailed steps to implement a Data Mart:

Designing

Designing is the first phase of Data Mart implementation.

The design step involves the following tasks:

- Gathering the business & technical requirements and Identifying data sources.
- Selecting the appropriate subset of data.
- Designing the logical and physical structure of the data mart.

Data could be partitioned based on following criteria:

- Date
- Business or Functional Unit
- Geography
- Any combination of above

Data could be partitioned at the application or DBMS level. Though it is recommended to partition at the Application level as it allows different data models each year with the change in business environment.

What Products and Technologies Do You Need?

A simple pen and paper would suffice. Though tools that help you create UML or ER diagrams would also append meta data into your logical and physical designs.

Constructing

This is the second phase of implementation. It involves creating the physical database and the logical structures.

This step involves the following tasks:

- Implementing the physical database designed in the earlier phase. For instance, database schema objects like table, indexes, views, etc. are created.

What Products and Technologies Do You Need?

You need a relational database management system to construct a data mart. RDBMS have several features that are required for the success of a Data Mart.

- **Storage management:** An RDBMS stores and manages the data to create, add, and delete data.
- **Fast data access:** With a SQL query you can easily access data based on certain conditions/filters.
- **Data protection:** The RDBMS system also offers a way to recover from system failures such as power failures. It also allows restoring data from these backups in case of the disk fails.
- **Multiuser support:** The data management system offers concurrent access.
- **Security:** The RDMS system also provides a way to regulate access by users to objects and certain types of operations.

Populating:

In the third phase, data is populated in the data mart.

The populating step involves the following tasks:

- Source data to target data Mapping
- Extraction of source data
- Cleaning and transformation operations on the data
- Loading data into the data mart
- Creating and storing metadata

What Products and Technologies Do You Need?

You accomplish these population tasks using an ETL(Extract Transform Load)Tool. This tool allows you to look at the data sources, perform source-to-target mapping, extract the data, transform, cleanse it, and load it back into the data mart.

Accessing

Accessing is a fourth step which involves putting the data to use: querying the data, creating reports, charts, and publishing them. End-user submit queries to the database and display the results of the queries

What Products and Technologies Do You Need?

You can access the data mart using the command line or GUI. GUI is preferred as it can easily generate graphs and is user-friendly compared to the command line.

Managing

This is the last step of Data Mart Implementation process. This step covers management tasks such as-

- Ongoing user access management.
- System optimizations and fine-tuning to achieve the enhanced performance.
- Adding and managing fresh data into the data mart.
- Planning recovery scenarios and ensure system availability in the case when the system fails.

What Products and Technologies Do You Need?

You could use the GUI or command line for data mart management.

Best practices for Implementing Data Marts

Following are the best practices that you need to follow while in the Data Mart Implementation process:

- The source of a Data Mart should be departmentally structured
- The implementation cycle of a Data Mart should be measured in short periods of time, i.e., in weeks instead of months or years.
- It is important to involve all stakeholders in planning and designing phase as the data mart implementation could be complex.
- Data Mart Hardware/Software, Networking and Implementation costs should be accurately budgeted in your plan
- Even though if the Data mart is created on the same hardware they may need some different software to handle user queries.
- A data mart may be on a different location from the data warehouse. That's why it is important to ensure that they have enough networking capacity to handle the Data volumes needed to transfer data to the data mart.

Advantages and Disadvantages of a Data Mart

Advantages

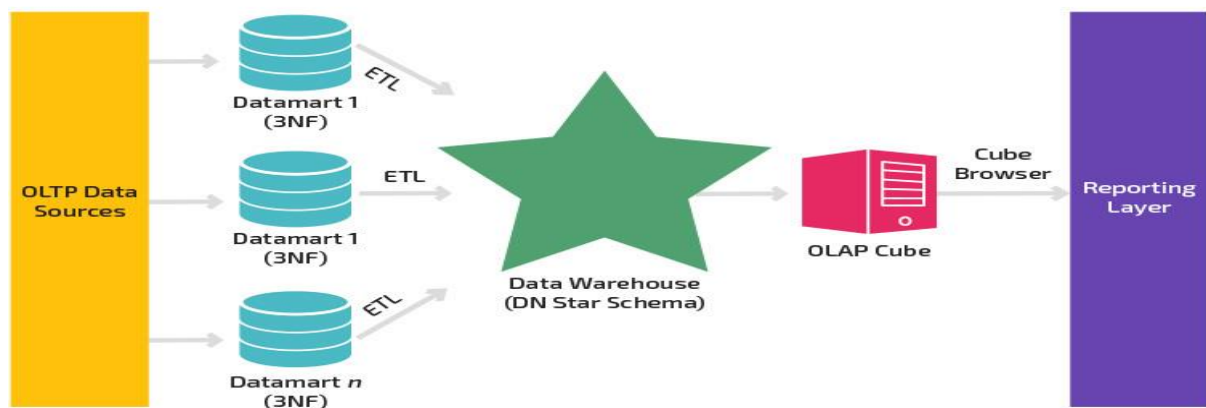
- Data marts contain a subset of organization-wide data. This Data is valuable to a specific group of people in an organization.
- It is cost-effective alternatives to a data warehouse, which can take high costs to build.
- Data Mart allows faster access of Data.
- Data Mart is easy to use as it is specifically designed for the needs of its users. Thus a data mart can accelerate business processes.
- Data Marts needs less implementation time compare to Data Warehouse systems.
- It contains historical data which enables the analyst to determine data trends.

Disadvantages

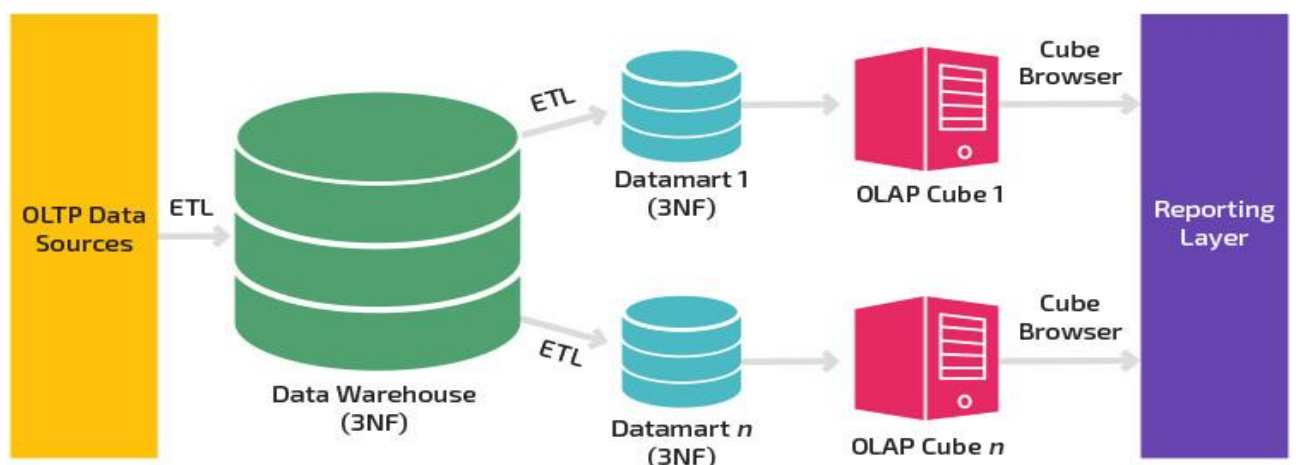
- Many a times enterprises create too many disparate and unrelated data marts without much benefit. It can become a big hurdle to maintain.
- Data Mart cannot provide company-wide data analysis as their data set is limited.

	<u>Data Mart</u>	<u>Data Warehouse</u>
Size	< 100 GB	100 GB +
Subject	Single Subject	Multiple Subjects
Scope	Line-of-Business	Enterprise-wide
Data Sources	Few Sources	Many Source Systems
Data Integration	One Subject Area	All Business Data
Time to Build	Minutes, Weeks, Months	Many Months to Years

Kimball Model



Inmon Model



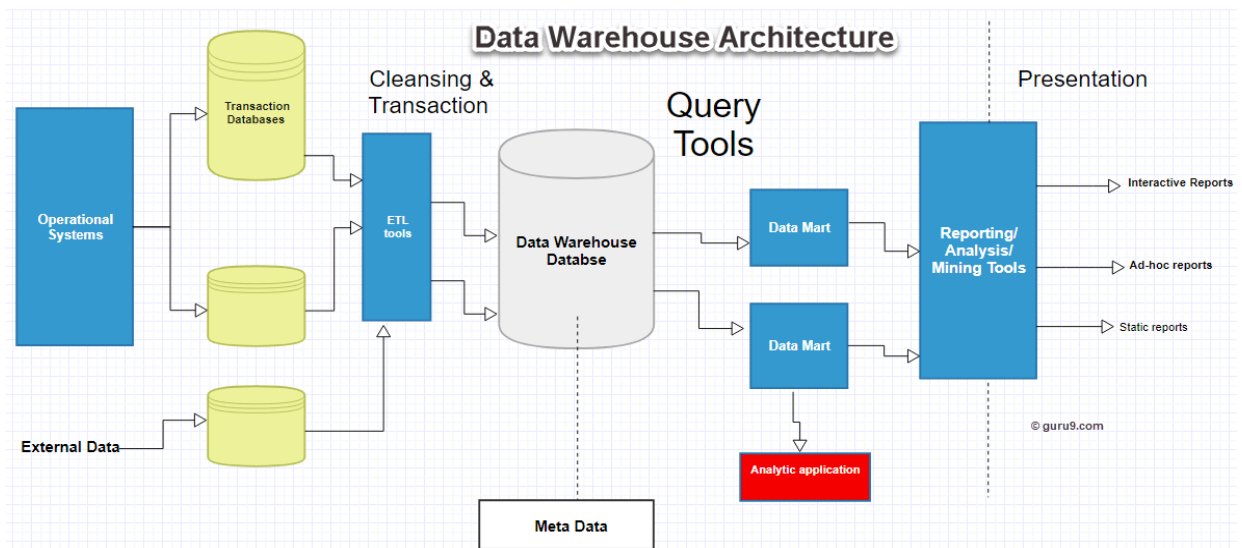
- **Data Cube: The data cube is used to represent data along some measure of interest.**

1. Data cube is a structure that enables OLAP to achieve the multidimensional functionality.
2. Data cube are an easy way to look at the data.
3. Cube is comparable to table in RDBMS
4. It is 2 and 3 dimensional or higher dimensional
5. Data Cube have two Categories of data
 - i. Dimensions :Some descriptions ex time, locations
 - ii. Measure: some facts Ex: Cost, Unit of service

- **Data Warehouse Architecture: Three-Tier Data Warehouse Architecture**

Generally a data warehouse adopts three-tier architecture. Following are the three tiers of the data warehouse architecture.

1. **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.
2. **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.
3. **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.



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:

1. **Data Warehouse Database:** The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology
2. **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.
3. **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.
4. **Query Tools:** One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions.
 - Query and reporting tools
 - Application Development tools
 - Data mining tools
 - OLAP tools
5. **Data warehouse Bus Architecture:** Data warehouse Bus determines the flow of data in your warehouse. The data flow in a data warehouse can be categorized as Inflow, Upflow, Downflow, Outflow and Meta flow.

- **Virtual Datawarehouse:**

1. A virtual warehouse is essentially a business database. The data found in a virtual warehouse is usually copied from multiple sources throughout a production system.
2. A virtual warehouse is another term for a data warehouse.
3. Virtual data warehouse refers to a layer that sits on top of existing data bases and enables the user to query all of them as if they were one entity.
4. Data Virtualization makes all data, regardless of where it's located and regardless of what format it's in, look as if it is one place and in a consistent format.
5. It provides access to data directly from one or more disparate data sources, without physically moving the data and provides it in such a manner that the technical aspects of location, structure, and access language are transparent to the analyst.



6. The data stored in a virtual warehouse is static. This means new data is stored alongside existing data rather than over it, allowing you to access historical information as well as current information.
- **OLTP (Online Transactional Processing):** is a category of data processing that is focused on transaction-oriented tasks. OLTP typically involves inserting, updating, and/or deleting small amounts of data in a database. OLTP mainly deals with large numbers of transactions by a large number of users.

Examples of OLTP Transactions

1. Online banking
 2. Purchasing a book online
 3. Booking an airline ticket
 4. Sending a text message
 5. Order entry
- **Online Analytical Processing Server (OLAP)** is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

We have four types of OLAP servers –

1. Relational OLAP (ROLAP): ROLAP uses relational or extended-relational DBMS.
2. Multidimensional OLAP (MOLAP): MOLAP uses array-based multidimensional storage engines for multidimensional views of data.

3. Hybrid OLAP (HOLAP): Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP.
4. Specialized SQL Servers: Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

- **OLAP Operations**

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

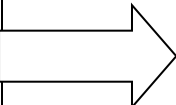
- 1. Roll-up**

Roll-up performs aggregation on a data cube in any of the following ways –

- a) By climbing up a concept hierarchy for a dimension
- b) By dimension reduction

Ex1:

Country	Medal
Delhi	5
New York	2
Patiala	3
Los Angeles	5

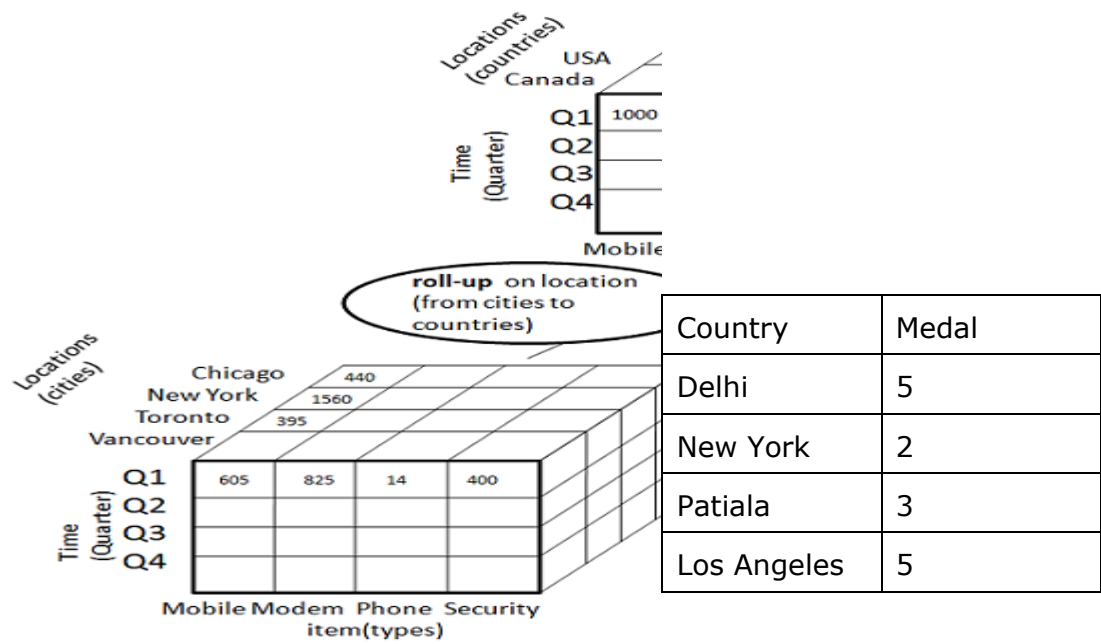


Country	Medal
India	8
America	7

Delhi, New York, Patiala and Los Angeles wins 5, 2, 3 and 5 medals o in this example, we roll up on Location from cities to countries.

More detailed data to less detailed data.

concept hierarchy was "street < city < province < country".

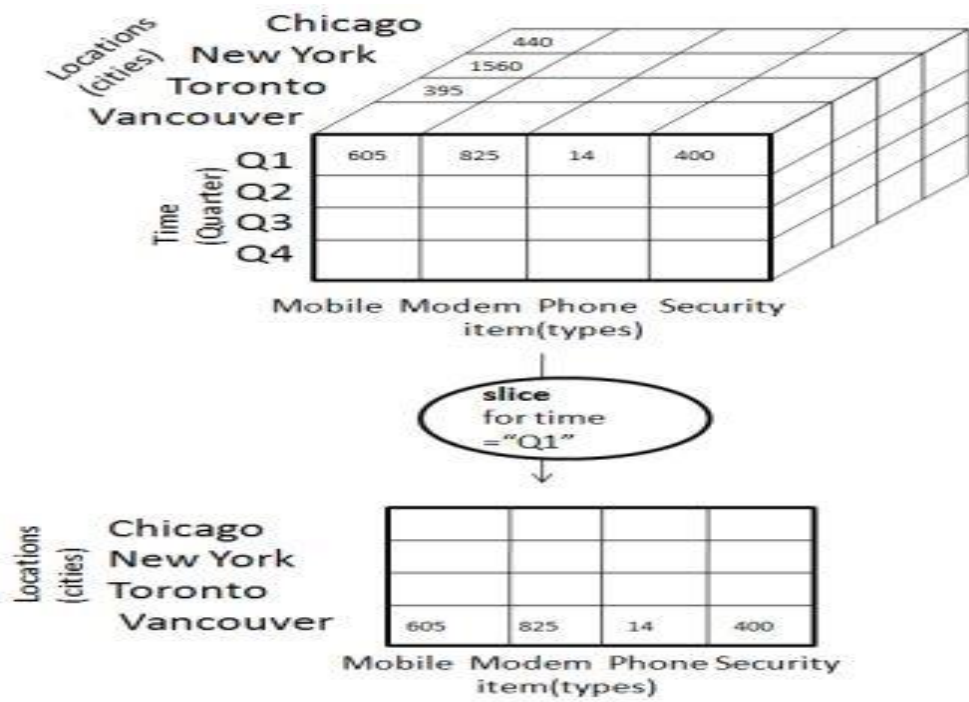


2. Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways –

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

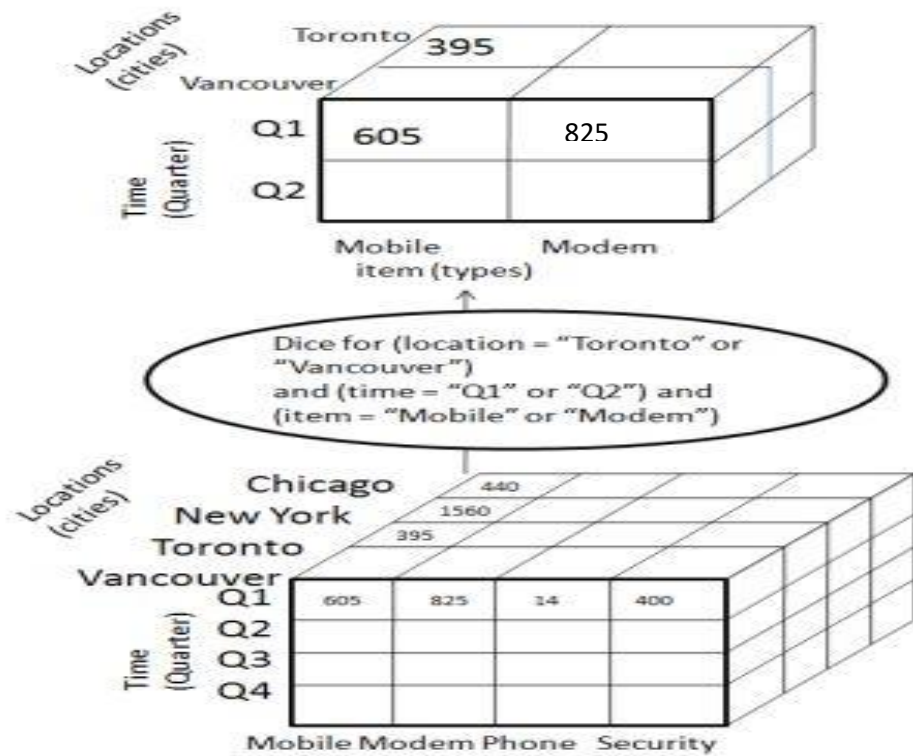
Country	Medal	➔
India	8	
America	7	



Dice: Dice selects two or more dimensions from a given cube and provides a new sub-cube.

For example, if we want to make a select where Medal = 3 or Location = New York

Country	Medal
Patiala	3
NewYork	2



4. Pivot(rotate)

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data.

