Data Warehouse-Basic Concepts

- **What is Data Warehouse?**

- Data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.

- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

- According to William H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process.

- The four keywords—subject-oriented, integrated, time-variant, and nonvolatile—distinguish data warehouses from other data repository systems

1. **Subject-oriented**:

- A data warehouse is organized around major subjects such as customer, supplier, product, and sales.

- **Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers.**

- Hence, data warehouses typically provide a simple and concise view of particular subject issues by excluding data that are not useful in the decision support process.

2. **Integrated**:

- A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.

- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

3. **Time-variant**:

- Data are stored to provide information from an historic perspective (e.g., the past 5–10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, a time element.

4. **Nonvolatile**:

- A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.

- Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

- Data warehousing provides alternative to traditional approach. Rather than using a **query-driven approach**, data warehousing employs an **update driven approach** in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis.

- **Differences between Operational Database Systems and Data Warehouses**

- The major task of online operational database systems is to perform online transaction and query processing. These systems are called **online transaction processing (OLTP) systems**. They cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

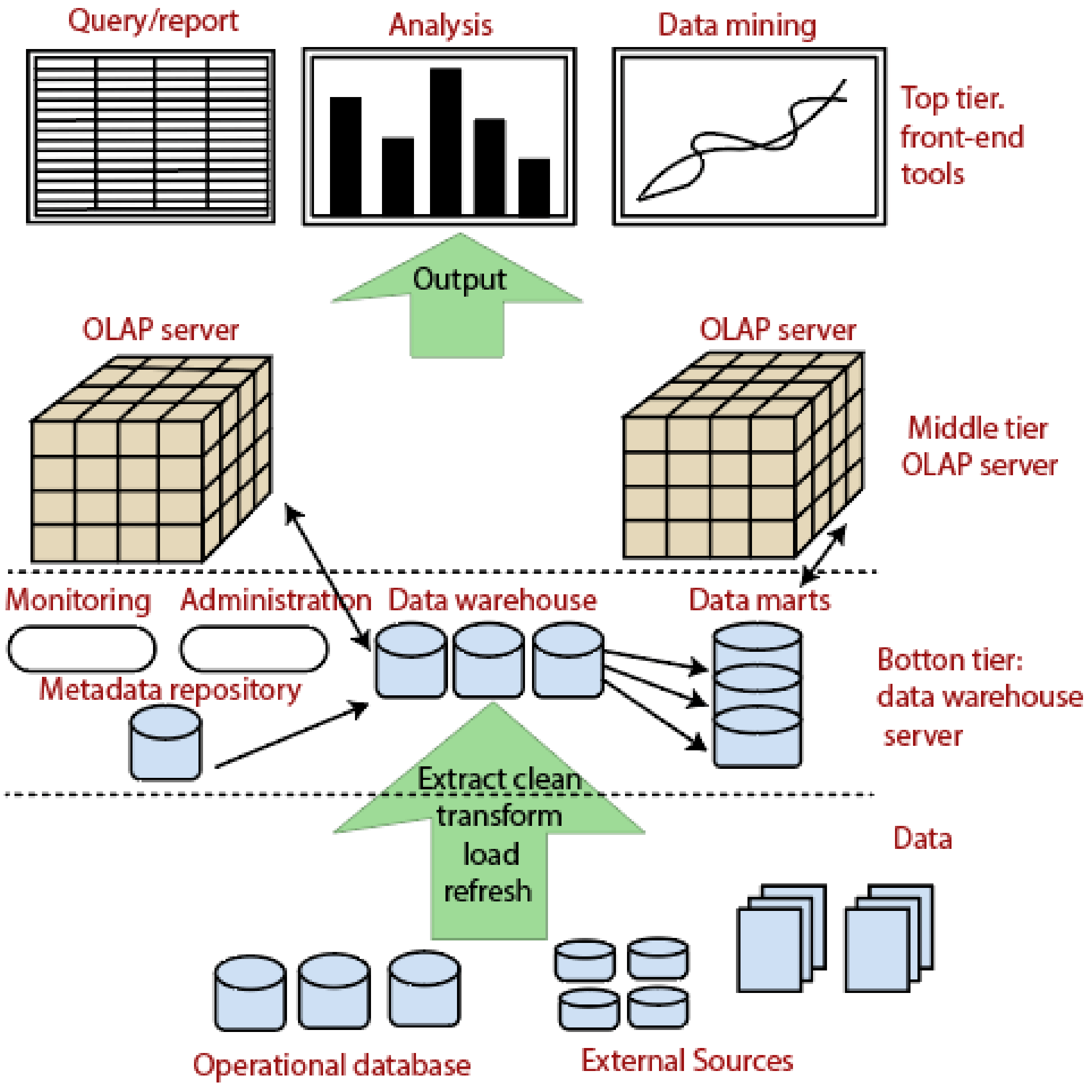
- Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making .These systems are known as **online analytical processing (OLAP) systems.**

Parameter	**Database**	**Data Warehouse**
Purpose	Is designed to record	Is designed to analyze
Processing Method	The database uses the Online Transactional Processing (OLTP)	Data warehouse uses Online Analytical Processing (OLAP).
Usage	The database helps to perform fundamental operations for your business	Data warehouse allows you to analyze your business.
Tables and Joins	Tables and joins of a database are complex as	Table and joins are simple in a data warehouse because

	they are normalized.	they are denormalized.
Orientation	Is an application-oriented collection of data	It is a subject-oriented collection of data
Storage limit	Generally limited to a single application	Stores data from any number of applications
Availability	Data is available real-time	Data is refreshed from source systems as and when needed
Usage	ER modeling techniques are used for designing.	Data modeling techniques are used for designing.
Technique	Capture data	Analyze data
Data Type	Data stored in the Database is up to date.	Current and Historical Data is stored in Data Warehouse. May not be up to date.
Storage of data	Flat Relational Approach method is used for data storage.	Data Ware House uses dimensional and normalized approach for the data structure. Example: Star and snowflake schema.
Query Type	Simple transaction queries are used.	Complex queries are used for analysis purpose.
Data Summary	Detailed Data is stored in a database.	It stores highly summarized data.

- **Data Warehousing: A Multitiered Architecture**

➢ Data warehouses often adopt a three-tier architecture as follows



Three-Tier 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.

▪ Data Warehouse Modeling :Data cube and OLAP

Data Cube: A Multidimensional Data Model

- **Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube.**

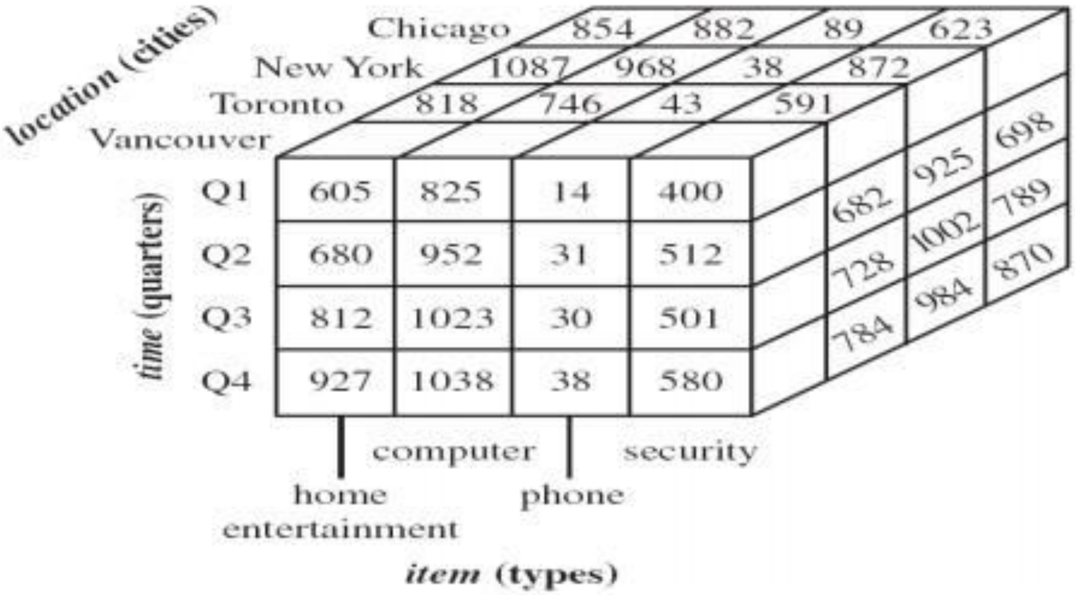
- Data cube is a multi-dimensional structure. Data cube is a data abstraction to view aggregated data from a number of perspectives.

- **Dimensions are** the perspectives or entities with respect to which an organization wants to keep records. **For example**, AllElectronics may create a sales data warehouse in order to keep records of the store’s sales with respect to the dimensions time, item, branch, and location.

- Each dimension may have a table associated with it, called a **dimension table**, which further describes the dimension. **For example**, a dimension table for item may contain the attributes item name, brand, and type.

- A multidimensional data model is typically organized around a central theme, such as sales. This theme is represented by a **fact table**. **Facts** are numeric measures. **The fact table** contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

- **3-D data cube representation**



- A 3-D data cube representation of the data in above diagram according to time, item, and location. The measure displayed is dollars sold (in thousands).

- **A 4-D data cube representation**

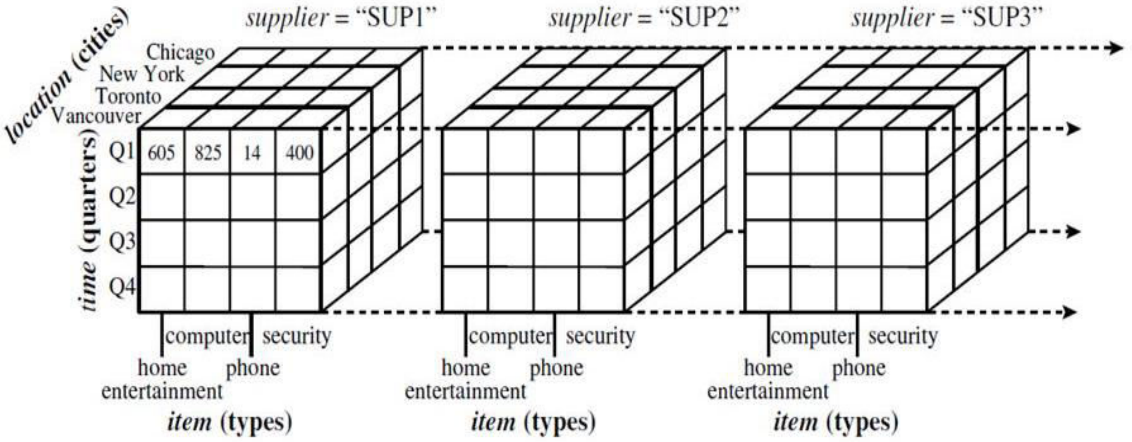
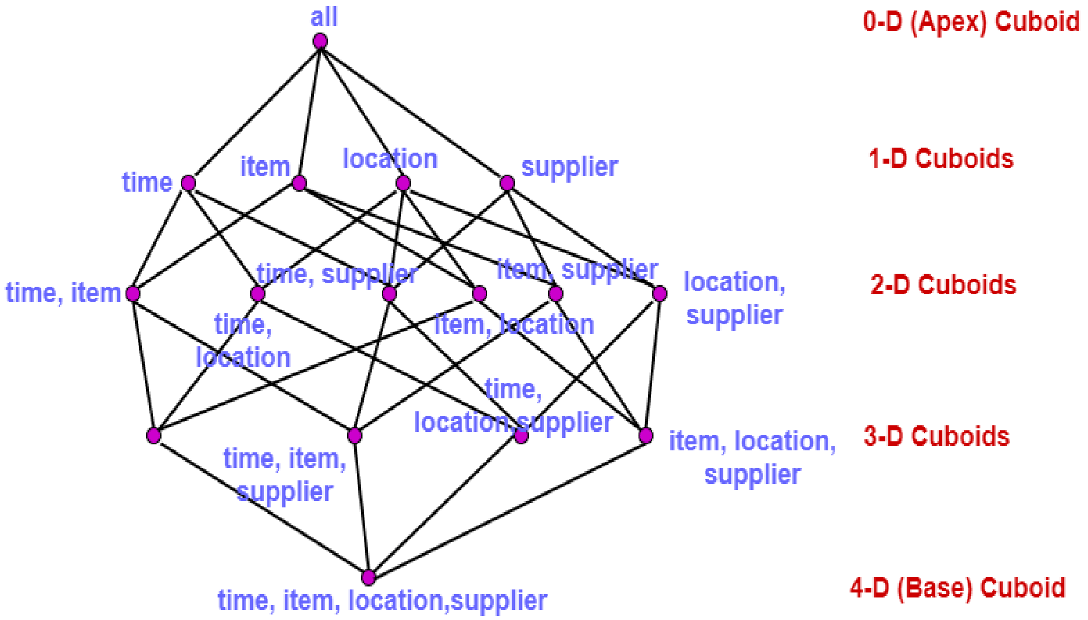


Figure: A 4-D data cube representation of sales data, according to the dimensions *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars sold* (in thousands). For improved readability, only some of the cube values are shown.

- A data cube like those shown in above diagram is often **referred to as a cuboid.** Given a set of dimensions, we can generate a cuboid for each of the possible subsets of the given dimensions. The result would form a lattice of cuboids, each showing the data at a different level of summarization, or group-by.

- **The cuboid that holds the lowest level of summarization is called the" base cuboid"**. **For example**, the 4-D cuboid in above diagram is the base cuboid for the given time, item, location, and supplier dimensions.

- **The 0-D cuboid, which holds the highest level of summarization, is called the "apex cuboid"**. In our example, this is the total sales, or dollars sold, summarized over all four dimensions. The apex cuboid is typically denoted by all.



- Lattice of cuboids, making up a 4-D data cube for time, item, location, and supplier. Each cuboid represents a different degree of summarization.

Typical OLAP Operations

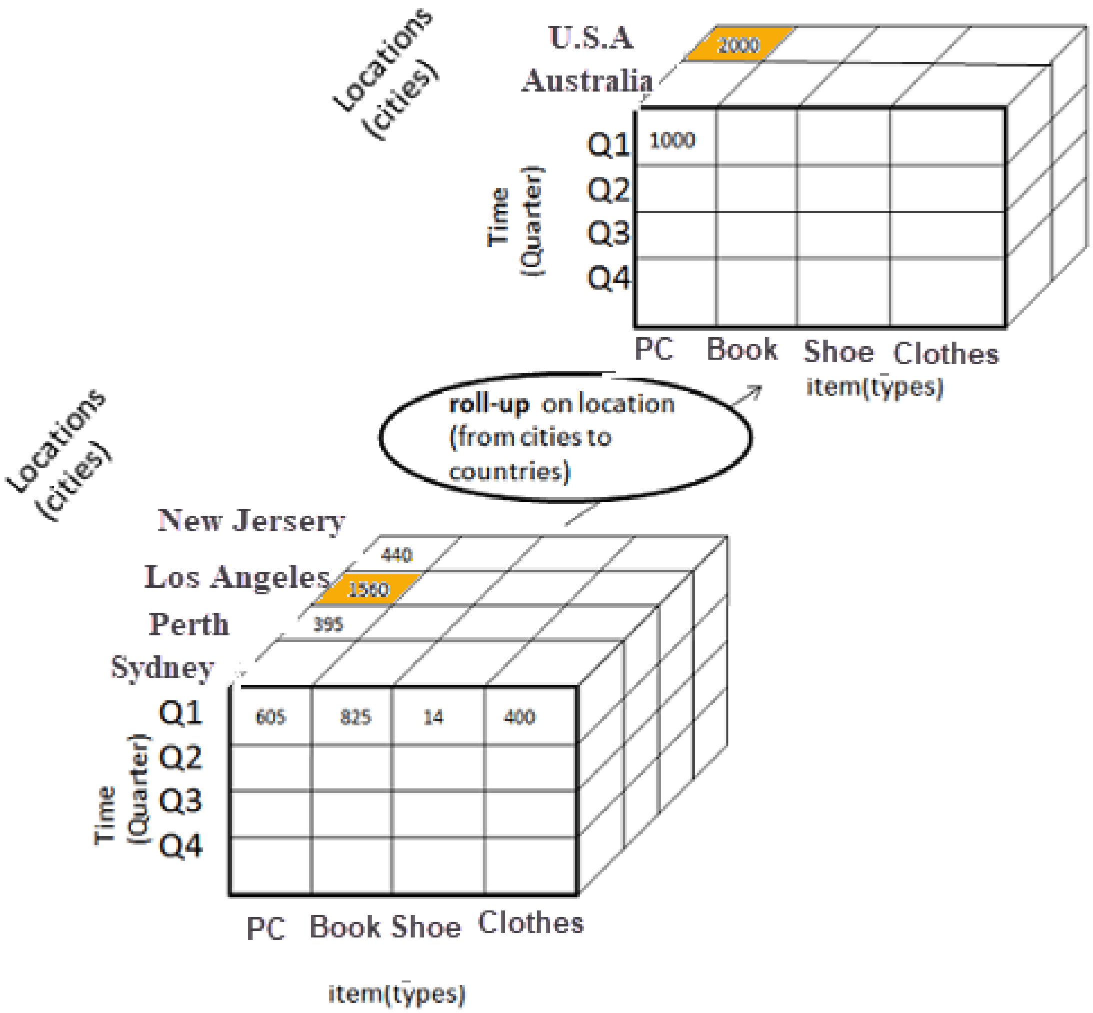
1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)

1) Roll-up:

- Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways

1. Reducing dimensions
2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

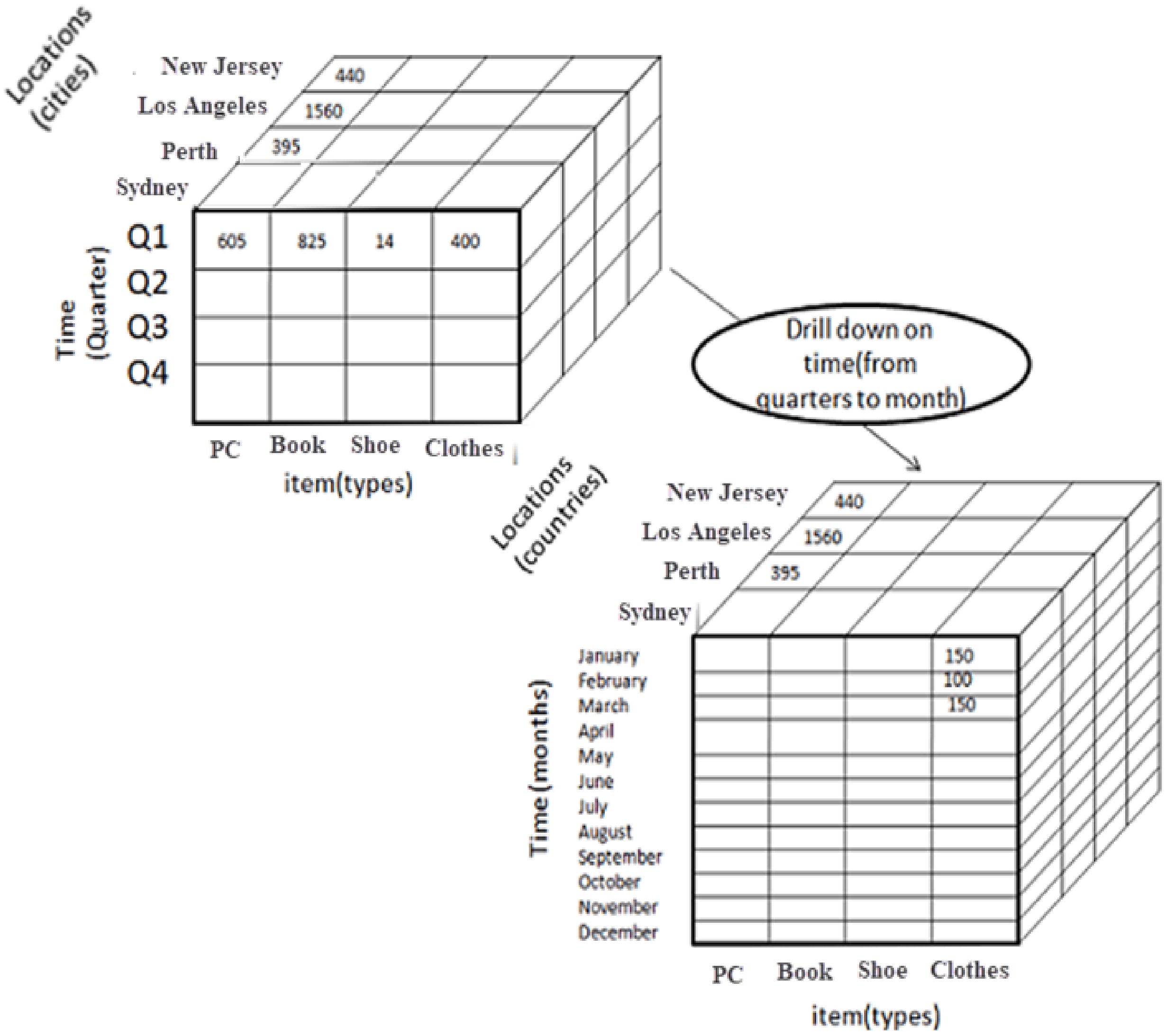
Consider the following diagram



- In this example, cities New jersey and Lost Angles and rolled up into country USA
- The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up
- In this aggregation process, data is location hierarchy moves up from city to the country.
- In the roll-up process at least one or more dimensions need to be removed. In this example, Quater dimension is removed.

2) Drill-down

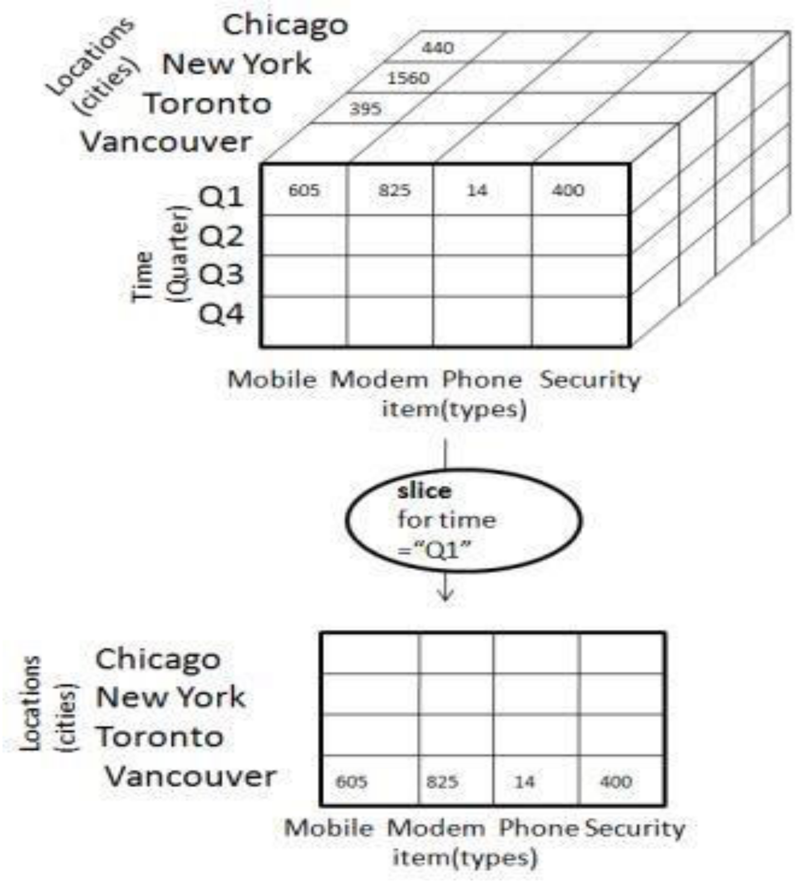
- In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via
    - Moving down the concept hierarchy
    - Increasing a dimension



- In this example Quarter Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.
- In this example, dimension months are added.

3) Slice

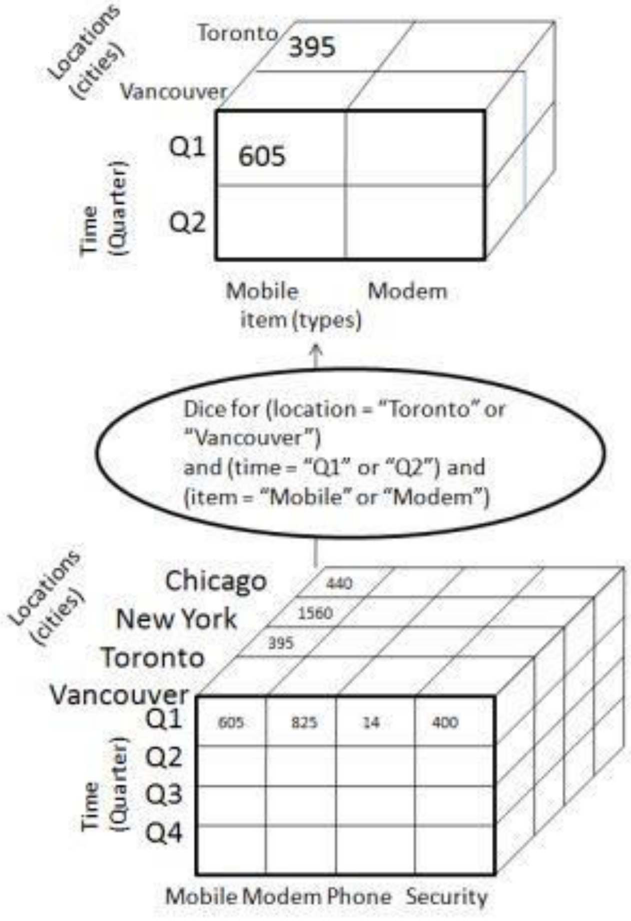
- The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



- 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.

4) Dice

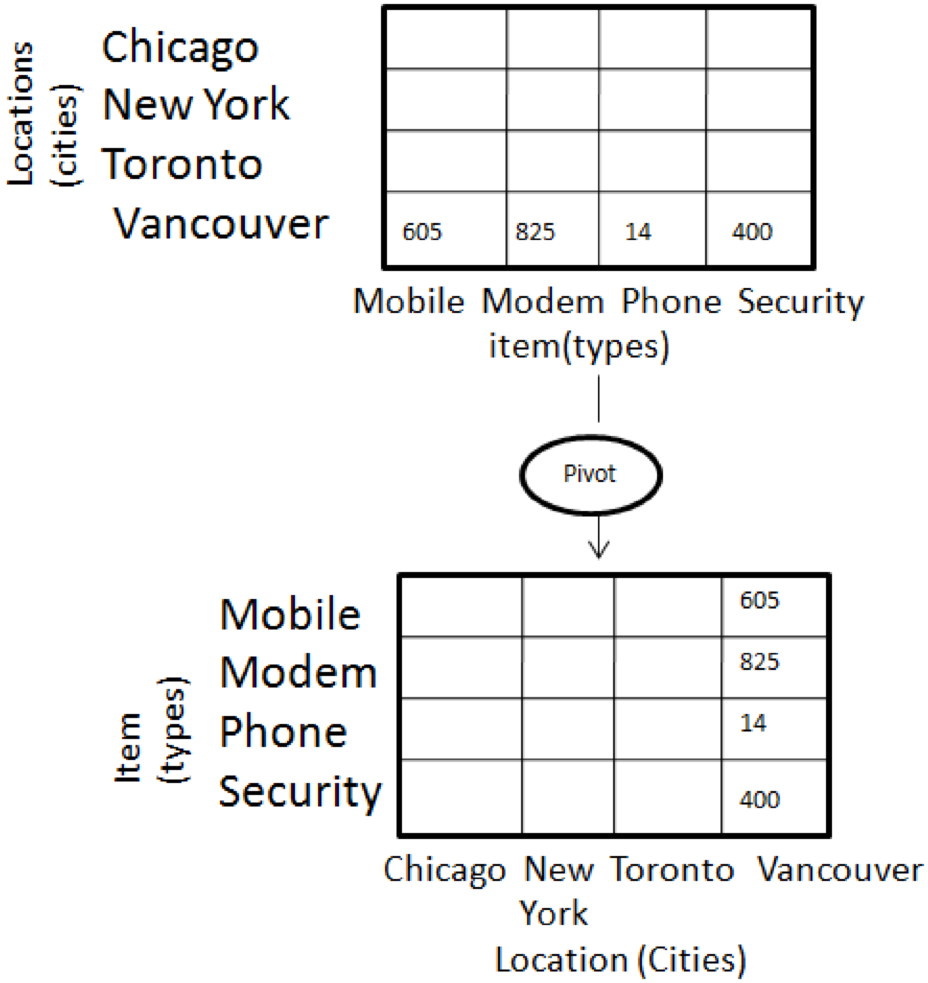
- Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.



- 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")

5) 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. Consider the following diagram that shows the pivot operation.



	Mobile	Modem	Phone	Security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

	Chicago	New York	Toronto	Vancouver
Mobile				605
Modem				825
Phone				14
Security				400

- **Data warehouse Design and usage**

- **A Business Analysis Framework for Data Warehouse Design**

➢ To design an effective and efficient data warehouse, we need to understand and analyze the business needs and construct a business analysis framework. Each person has different views regarding the design of a data warehouse. These views are as follows −

- **The top-down view** −This view allows the selection of relevant information needed for a data warehouse.
- **The data source view** −This view presents the information being captured, stored, and managed by the operational system.
- **The data warehouse view** −This view includes the fact tables and dimension tables. It represents the information stored inside the data warehouse.
- **The business query view** −It is the view of the data from the viewpoint of the end-user.

- **Data Warehouse Design Process**

➢ A data warehouse can be built using following approaches.

1. **"Top-down" approach**
2. **"Bottom-up" approach**

1. Top-down Design Approach

➢ In the "Top-Down" design approach, a data warehouse is described as a subject-oriented, time-variant, non-volatile and integrated data repository for the entire enterprise data from different sources are validated, reformatted and saved in a normalized (up to 3NF) database as the data warehouse.

- The data warehouse stores "atomic" information, the data at the lowest level of granularity, from where dimensional data marts can be built by selecting the data required for specific business subjects or particular departments.

- An approach is a **data-driven approach** as the information is gathered and integrated first and then business requirements by subjects for building data marts are formulated. The advantage of this method is which it supports a single integrated data source. Thus data marts built from it will have consistency when they overlap.

Advantages of top-down design

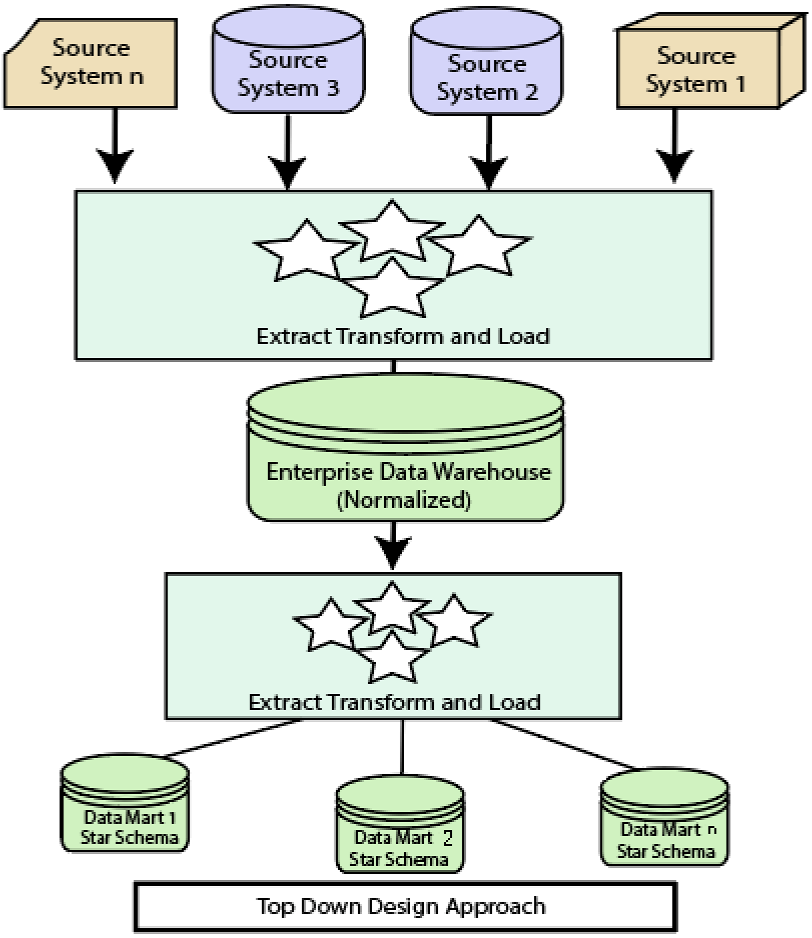
- Data Marts are loaded from the data warehouses.

- Developing new data mart from the data warehouse is very easy.

Disadvantages of top-down design

- This technique is inflexible to changing departmental needs.

- The cost of implementing the project is high.



Top Down Design Approach

2. Bottom-Up Design Approach

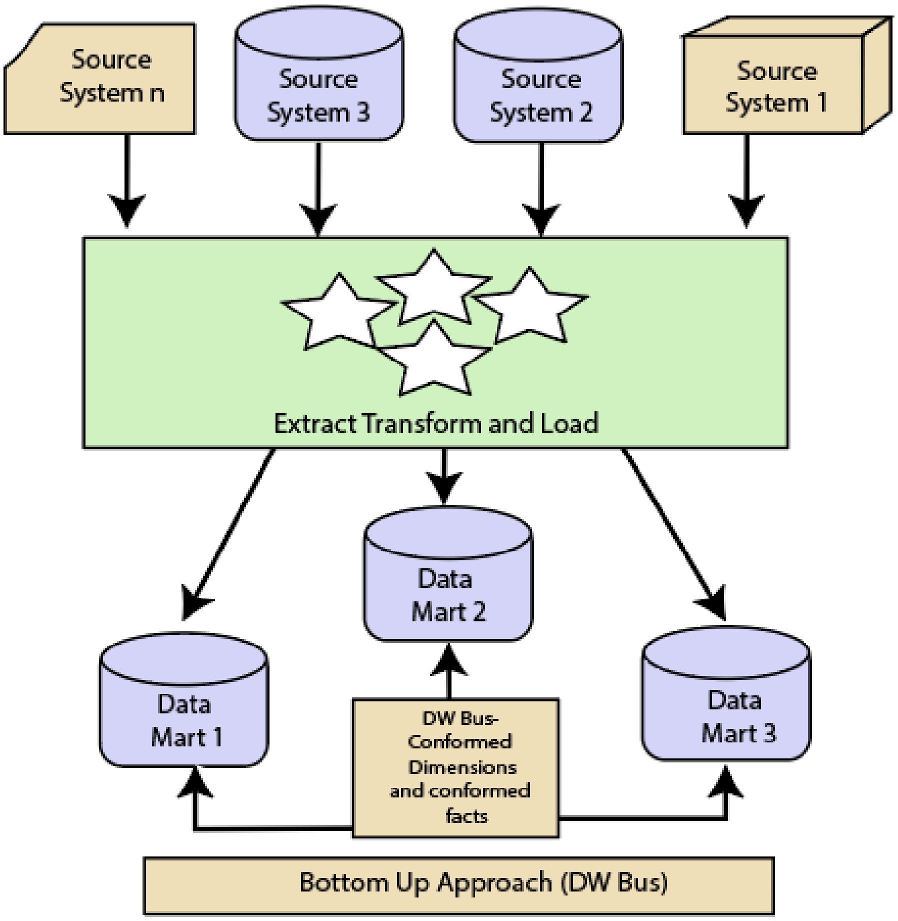
➢ In the "Bottom-Up" approach, a data warehouse is described as "a copy of transaction data specifical architecture for query and analysis," term the star schema.

➢ In this approach, a data mart is created first to necessary reporting and analytical capabilities for particular business processes (or subjects). Thus it is needed to be a business-driven approach in contrast to Inmon's data-driven approach.

➢ Data marts include the lowest grain data and, if needed, aggregated data too. Instead of a normalized database for the data warehouse, a denormalized dimensional database is adapted to meet the data delivery requirements of data warehouses.

- Using this method, to use the set of data marts as the enterprise data warehouse, data marts should be built with conformed dimensions in mind, defining that ordinary objects are represented the same in different data marts.

- The conformed dimensions connected the data marts to form a data warehouse, which is generally called a virtual data warehouse.


Source System n
Source System 3
Source System 2
Source System 1
Extract Transform and Load
Data Mart 2
Data Mart 1
Data Mart 3
DW Bus-Conformed Dimensions and conformed facts
Bottom Up Approach (DW Bus)

Bottom Up Design Approach

Advantages of bottom-up design

- Documents can be generated quickly.

- The data warehouse can be extended to accommodate new business units.

- It is just developing new data marts and then integrating with other data marts.

Disadvantages of bottom-up design

- The locations of the data warehouse and the data marts are reversed in the bottom-up approach design.

- In general, the warehouse design process consists of the following steps:

1) **Choose a business process to model** (e.g., orders, invoices, shipments, inventory, account administration, sales, or the general ledger).

2) **Choose the business process grain**, which is the fundamental, atomic level of data to be represented in the fact table for this process (e.g., individual transactions, individual daily snapshots, and so on).

3) **Choose the dimensions** that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.

4) **Choose the measures** that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

- **Data Warehouse Usage for Information Processing**

- Data warehouses and data marts are used in a wide range of applications. Business executives use the data in data warehouses and data marts to perform data analysis and make strategic decisions.

- There are three kinds of data warehouse applications

    1) Information processing
    2) analytical processing
    3) data mining

1) **Information Processing** - A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.
2) **Analytical Processing** - A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic OLAP operations, including slice-and-dice, drill down, drill up, and pivoting.
3) **Data Mining** - Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using the visualization tools.

- **From Online Analytical Processing to Multidimensional Data Mining**

- Multidimensional data mining is particularly important for the following reasons:

High quality of data in data warehouses:

- Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data integration and data transformation as preprocessing steps.

- A data warehouse constructed by such preprocessing serves as a valuable source of high-quality data for OLAP as well as for data mining.

Available information processing infrastructure surrounding data warehouses:

- Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple heterogeneous databases, ODBC/OLEDB connections, Web accessing and service facilities, and reporting and OLAP analysis tools.

OLAP-based exploration of multidimensional data:

- Multidimensional data mining provides facilities for mining on different subsets of data and at varying levels of abstraction—by drilling, pivoting, filtering, dicing, and slicing on a data cube and/or intermediate data mining results. This, together with data/knowledge visualization tools, greatly enhances the power and flexibility of data mining.

Online selection of data mining functions:

- By integrating OLAP with various data mining functions, multidimensional data mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.