Data Mining and Warehousing (03105430)

Dheeraj Kumar Singh, Assistant Professor
Department of Information Technology

The Course Outline

Chapter 1 : Introduction to data mining

Chapter 2: Overview and concepts Data Warehousing and Business Intelligence

Chapter 3: Data Warehousing and Online Analytical Processing

Chapter 4: Data Pre-processing

Chapter 5: Mining Frequent Patterns, Associations, and Correlations:

Chapter 6: Classification

Chapter 7: Clustering

Chapter 8: Applications

CHAPTER-3

Data Warehousing and OLAP

Introduction of Data Warehousing

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

Figure3.1: Representation of Data Warehouse

Subject- Oriented

- A data warehouse can be used to analyze a particular subject area.
- For example, "sales" can be a particular subject.
- Organized around major subjects, such as customer, product, sales

Integrated

- **A data warehouse integrates data from multiple data sources**.
  - Relational databases, flat files, on-line transaction records
- **Data cleaning and data integration techniques are applied.**
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
  - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Time- Variant

- **The time horizon for the data warehouse is significantly longer than that of operational systems**
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- **Every key structure in the data warehouse**
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

Non Volatile

- **A physically separate store of data transformed from the operational environment**
- **Operational update of data does not occur in the data warehouse environment**

  - Does not require transaction processing, recovery, and concurrency control mechanisms

  - Requires only two operations in data accessing:

  - Initial loading of data and access of data

Why a Separate Data Warehouse?

- **High performance for both systems**
  - DBMS tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse tuned for OLAP: complex OLAP queries, multidimensional view, consolidation

- **Different functions and different data**
  - Missing data: Decision support requires historical data which operational DBs do not typically maintain
  - Data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources

Why a Separate Data Warehouse?

- Data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources

- Data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled

- **Note: There are more and more systems which perform OLAP analysis directly on relational databases**

Type of Date Warehousing Architecture

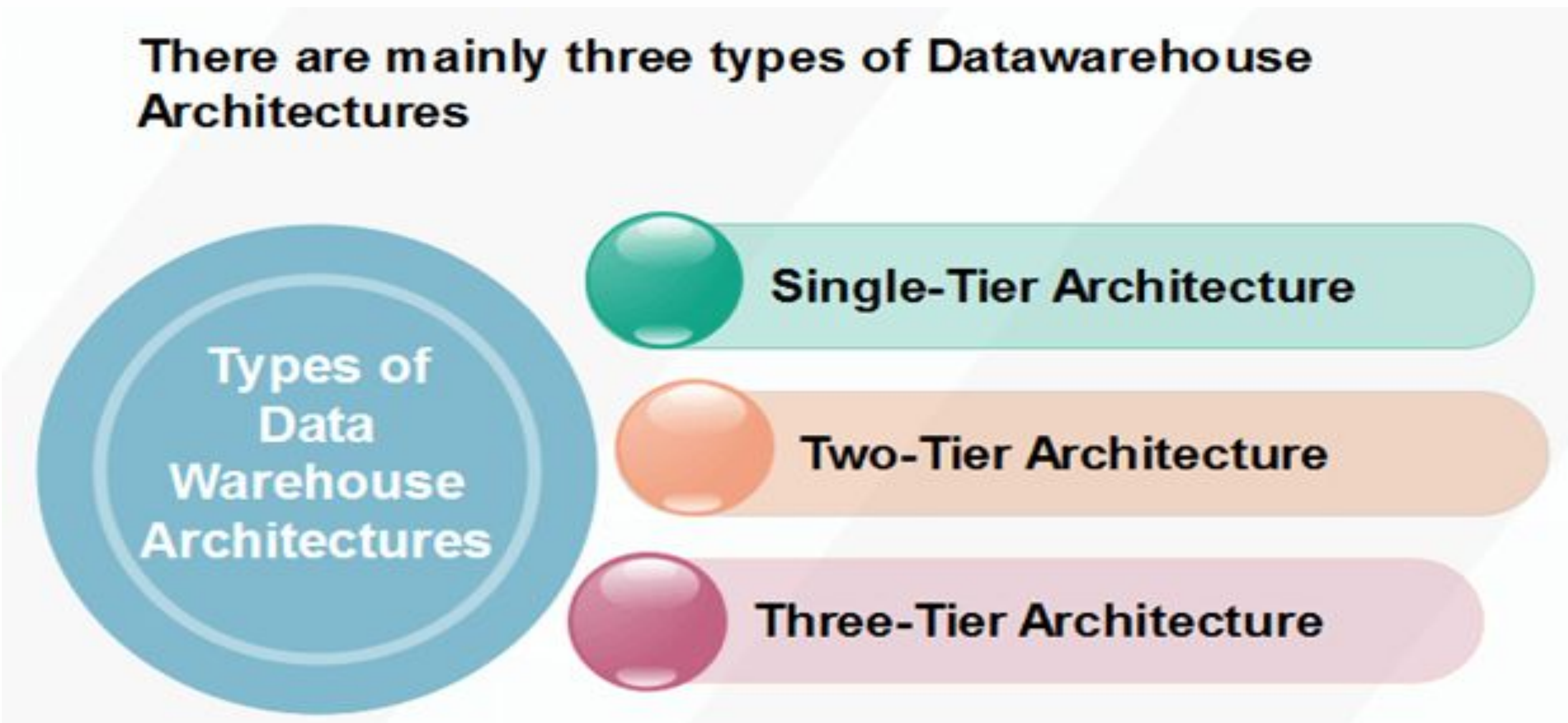

Figure 3.2 Type of Data Warehouse

Single- tier Data Warehouse Architecture

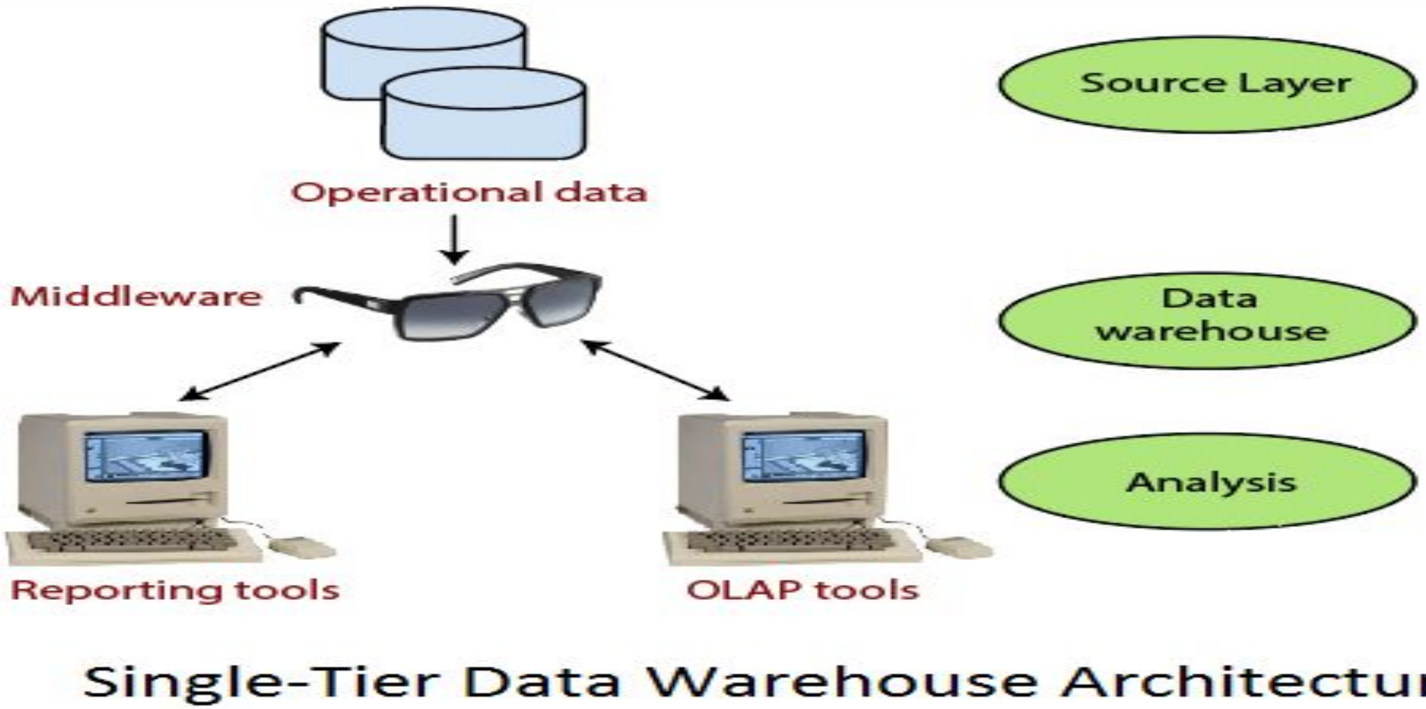

Figure 3.3
Single type of
Data
Warehouse

Two- tier Data Warehouse Architecture

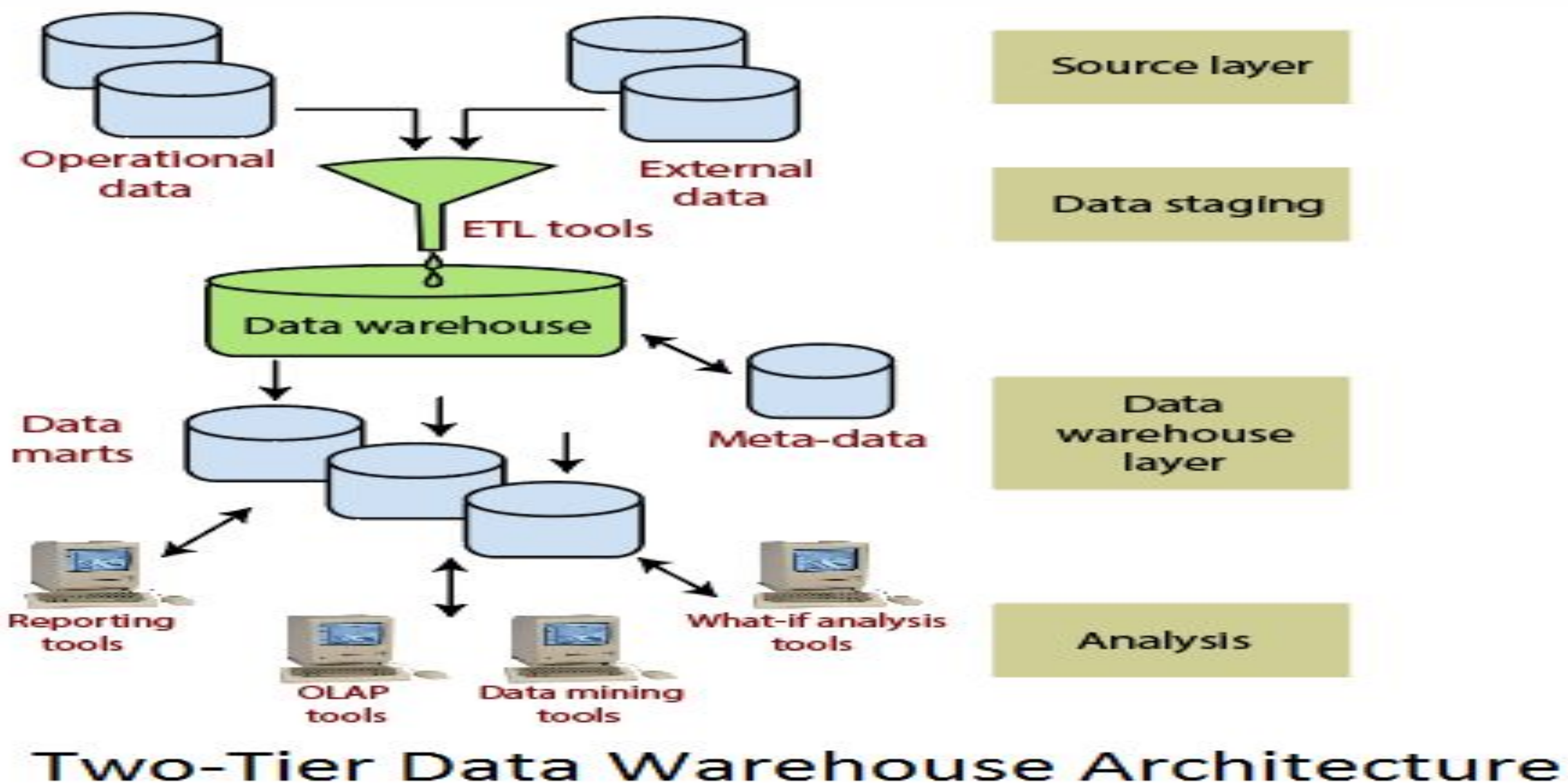

Figure 3.4 Two tier of Data Warehouse

Three- tier Data Warehouse Architecture

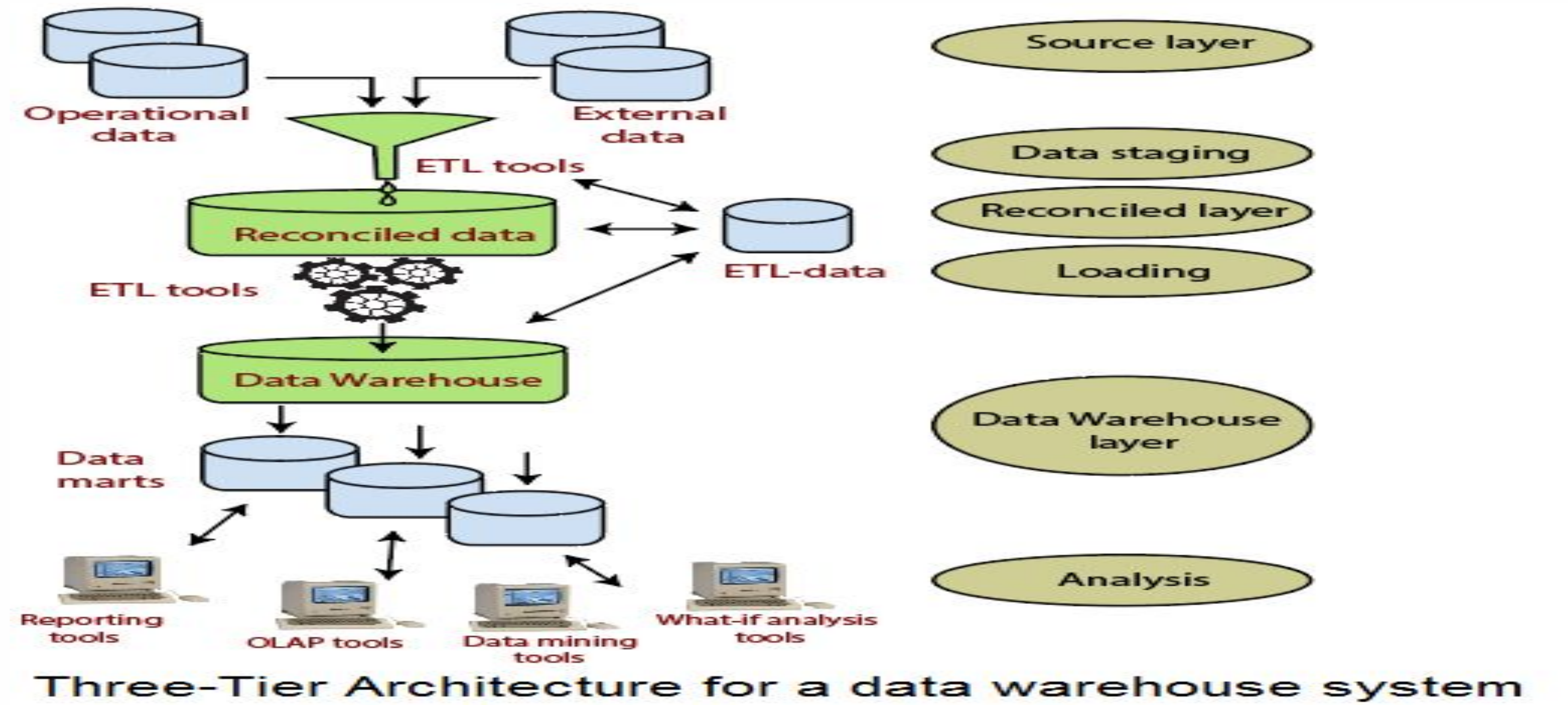

Figure 3.5
Three Tier of
Data

ETL Process in Data Warehouse

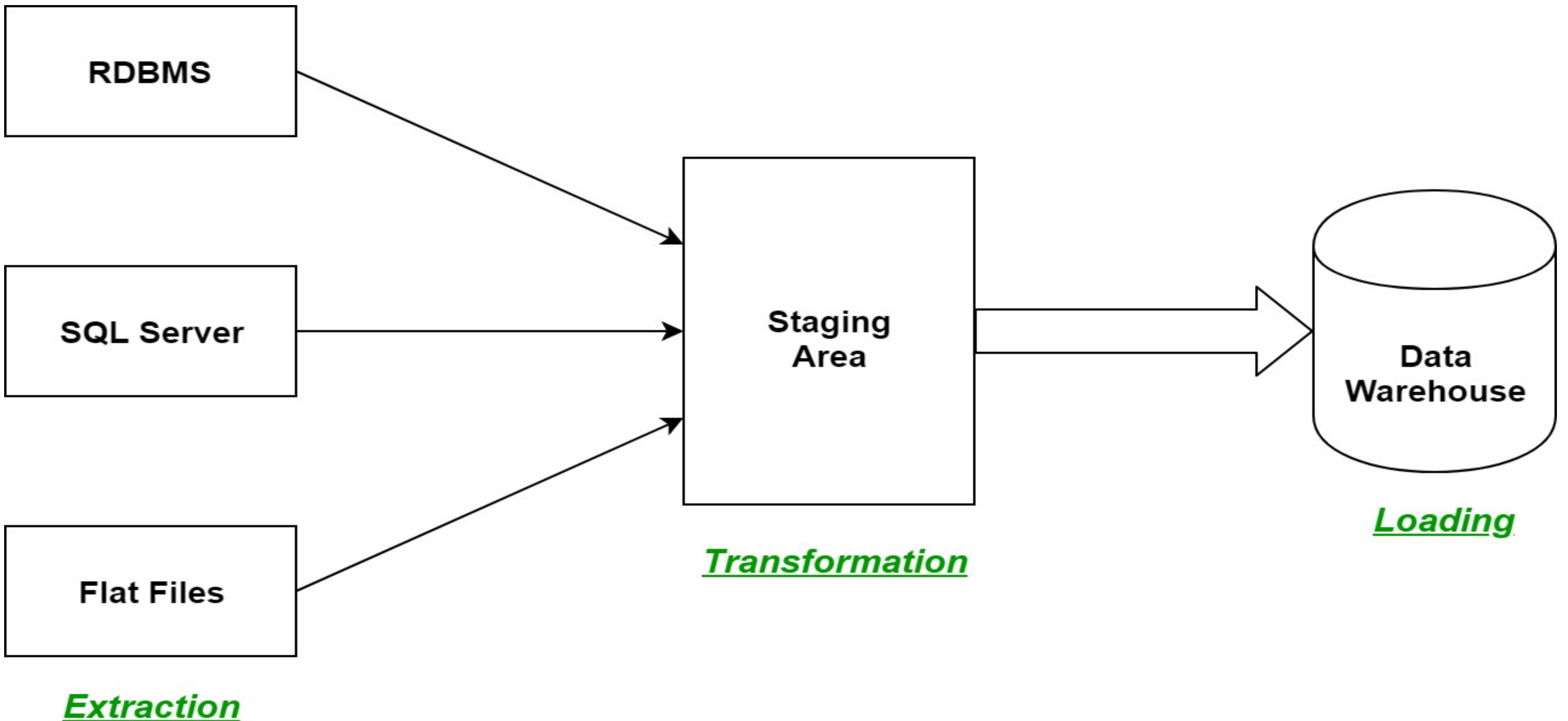

Figure 3.6 ETL Process

ETL Process (Contd...)

- **Data extraction**
  - Get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - Detect errors in the data and rectify them when possible
- **Data transformation**
  - Convert data from legacy or host format to warehouse format

ETL Process (Contd.....)

- **Load**
  - Sort, summarize, consolidate, compute views, check integrity, and build indices and partitions
- **Refresh**
  - Propagate the updates from the data sources to the warehouse

Multi Dimensional Model

- Sales volume as a function of product, month, and region

- Dimensions: Production, Location, time

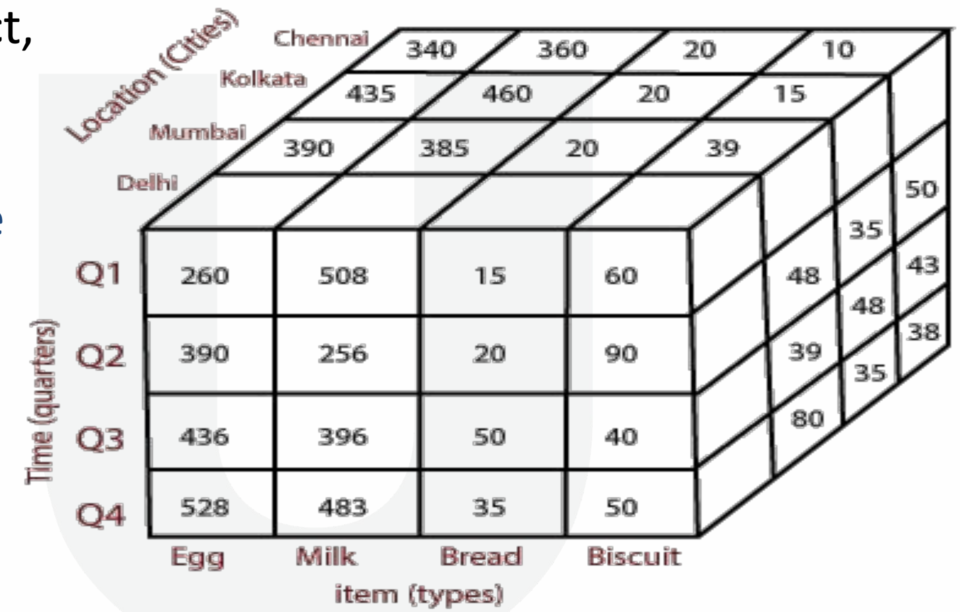

Figure 3.7 Multi Dimensional Model

Conceptual Modelling in Data Warehouses

- **Three are three type of Schema**

  - Star schema
  - Snowflake Schema
  - Fact constellations Schema

Star Schema

- **Two different type of table in Star schema**
  - **- Fact Table**
  - **- Dimension table**
  - - A fact table in the middle connected to a set of dimension tables

Example of Star Schema

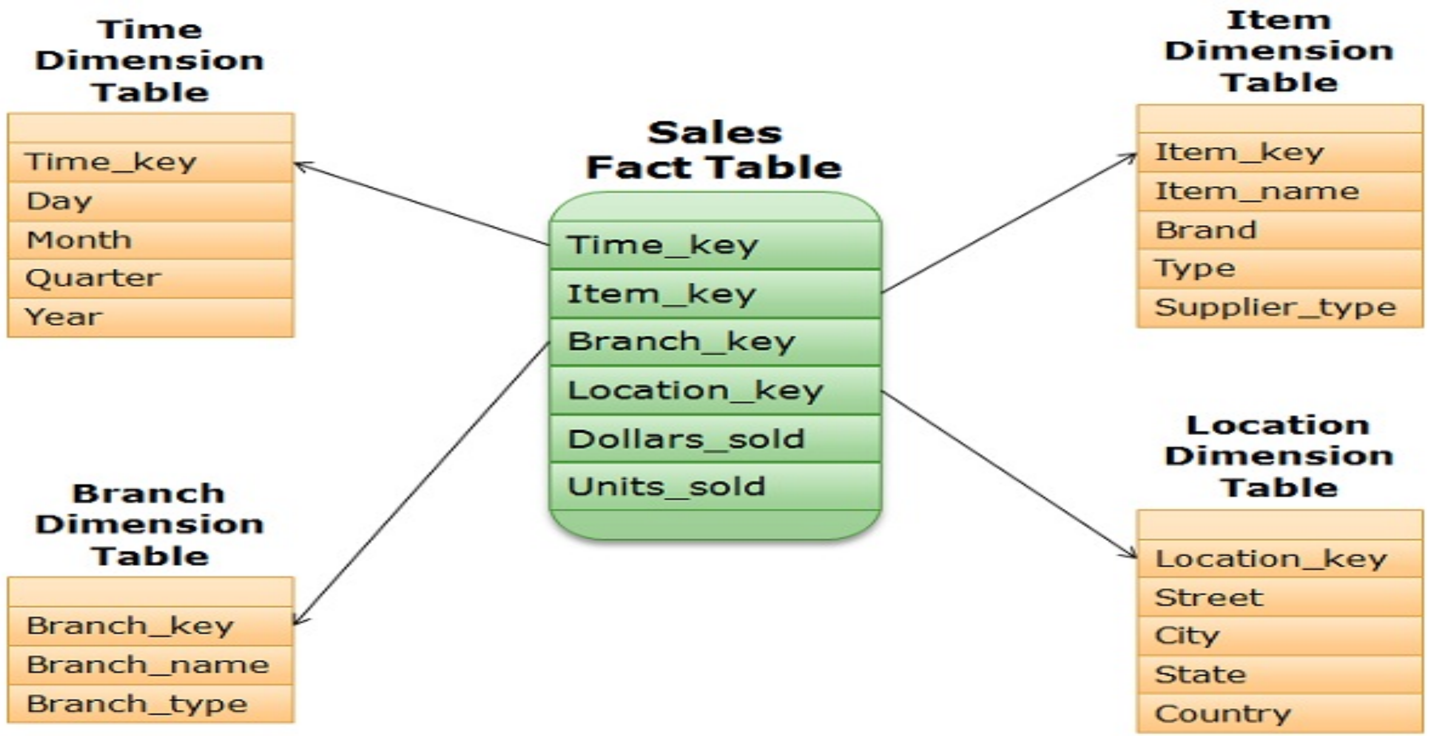

Figure 3.8 Star Schema

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

Example of Snowflake Schema

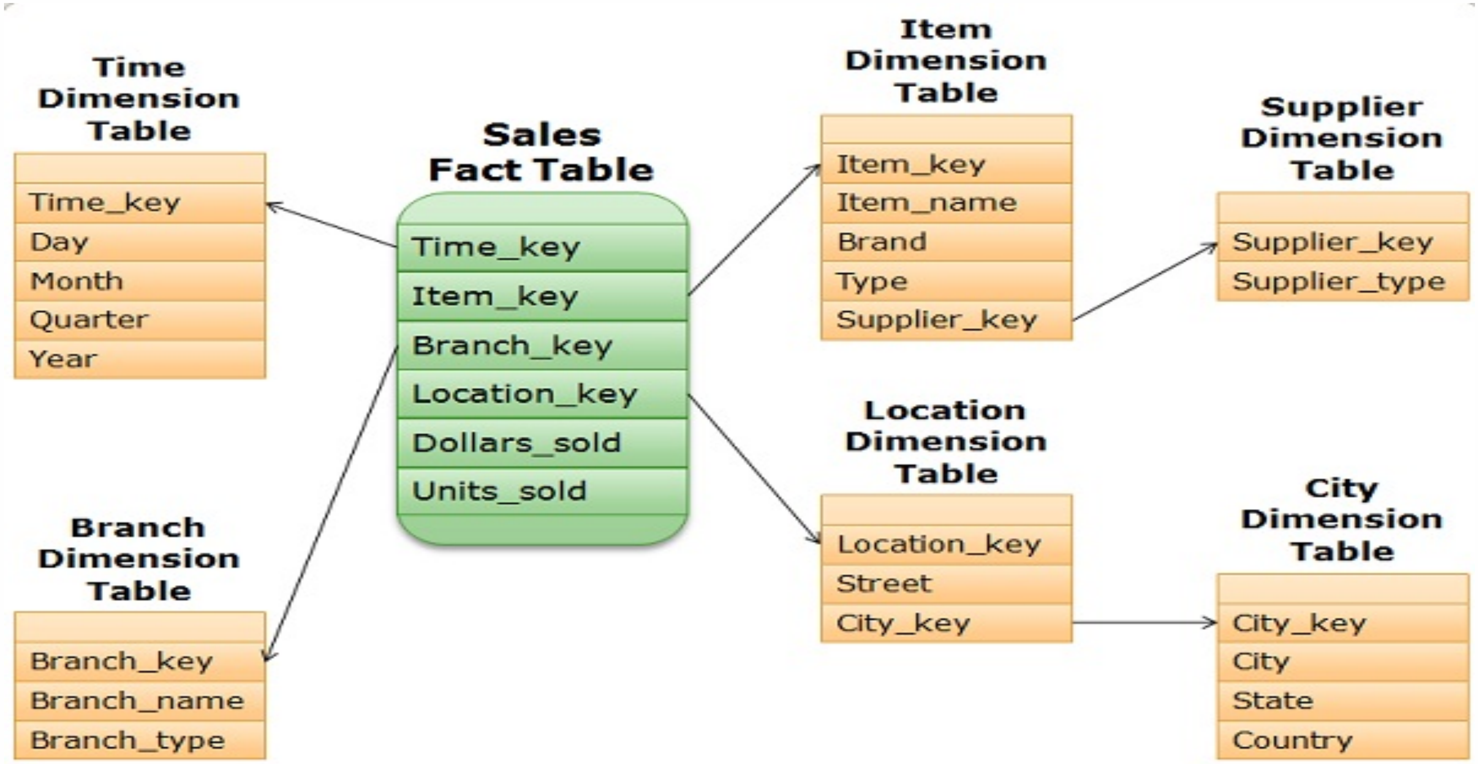

Figure 3.9 Snowflake Schema

Fact Constellations Schema

- Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Fact Constellations Schema (Contd.....)

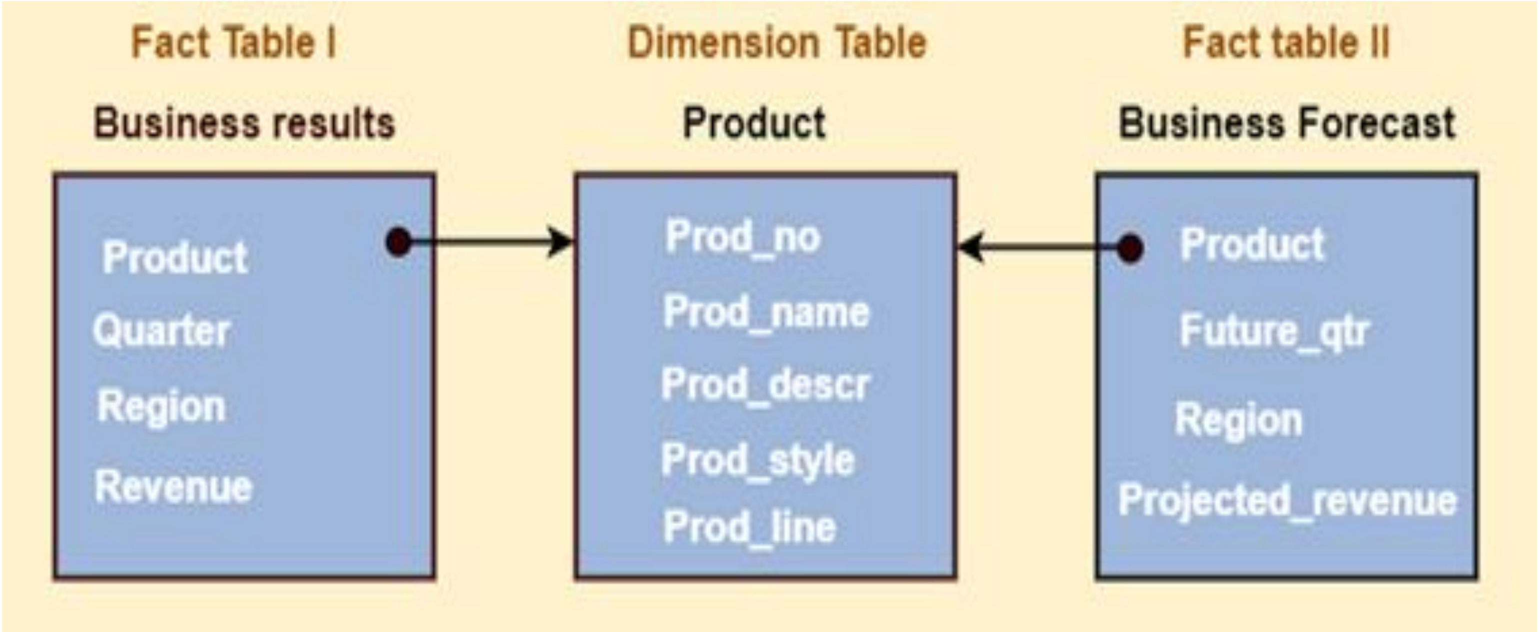

Figure 3.10
Snowflake
Schema

Example of Fact Constellations Schema

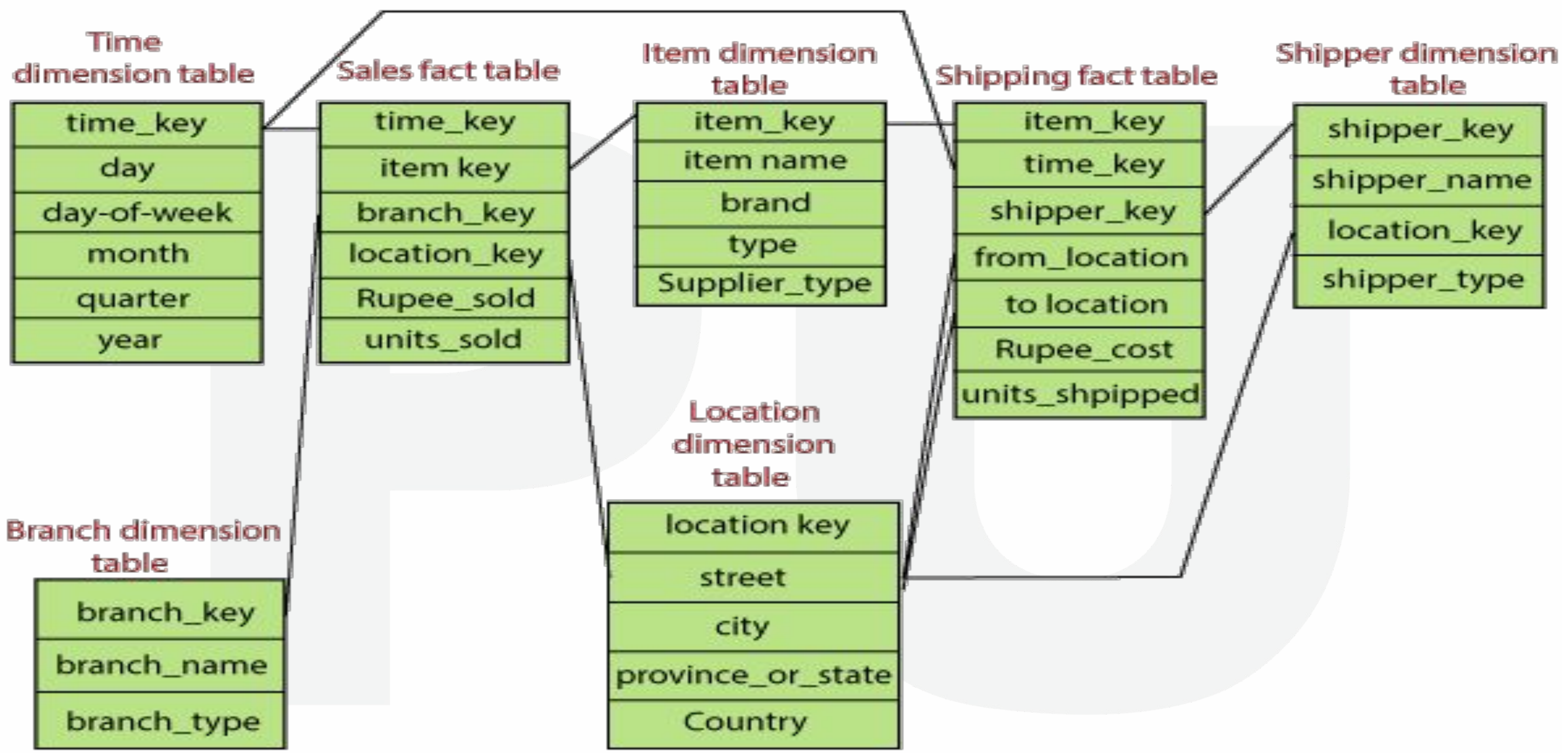

Figure 3.11 Fact Constellations Schema

Data Warehouse Model

- Data Warehouse model has categorized into three parts:

  - Enterprise warehouse

  - Data Mart

  - Virtual warehouse

Data Warehouse Model (Contd.....)

- **Enterprise warehouse**
  - Collects all of the information about subjects spanning the entire organization

- **Data Mart**
  - A subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
  - Independent vs. dependent (directly from warehouse) data mart

Data Warehouse Model (Contd.....)

- **Virtual warehouse**
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Concept Hierarchies

- Reduces the data size by collecting and then replacing the low-level concepts (such as 43 for age) to high-level concepts concepts (categorical variables such as middle age or Senior).

- For numeric data following techniques can be followed:
  - Binning
  - Histogram analysis

Binning (Contd.....)

- Binning is the process of changing numerical variables into categorical counterparts.

- The number of categorical counterparts depends on the number of bins specified by the user.

Histogram (Contd.....)

- The histogram is used to partition the value for the attribute X, into disjoint ranges called brackets.

- There are several partitioning rules:
  - Equal Frequency partitioning
  - Equal Width Partitioning
  - Clustering

Partitioning Rule of Histogram Analysis (Contd...)

- **Equal Frequency partitioning:**
  - Partitioning the values based on their number of occurrences in the data set.

- **Equal Width Partitioning:**
  - Partitioning the values in a fixed gap based on the number of bins i.e. a set of values ranging from 0-20.

- **Clustering:**
  - Grouping the similar data together.

OLAP Server

- **Relational OLAP (ROLAP)**

  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware

  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services

  - Greater scalability

OLAP Server

- **Multidimensional OLAP (MOLAP)**
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data

- **Hybrid OLAP (HOLAP)**
  - (e.g., Microsoft SQL Server)
  - Flexibility, e.g., low level: relational, high-level: array

- **Specialized SQL servers**
  - (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Roll Up

- **Roll up (drill-up)**
  - Summarize data
  - By climbing up hierarchy

  or by dimension reduction

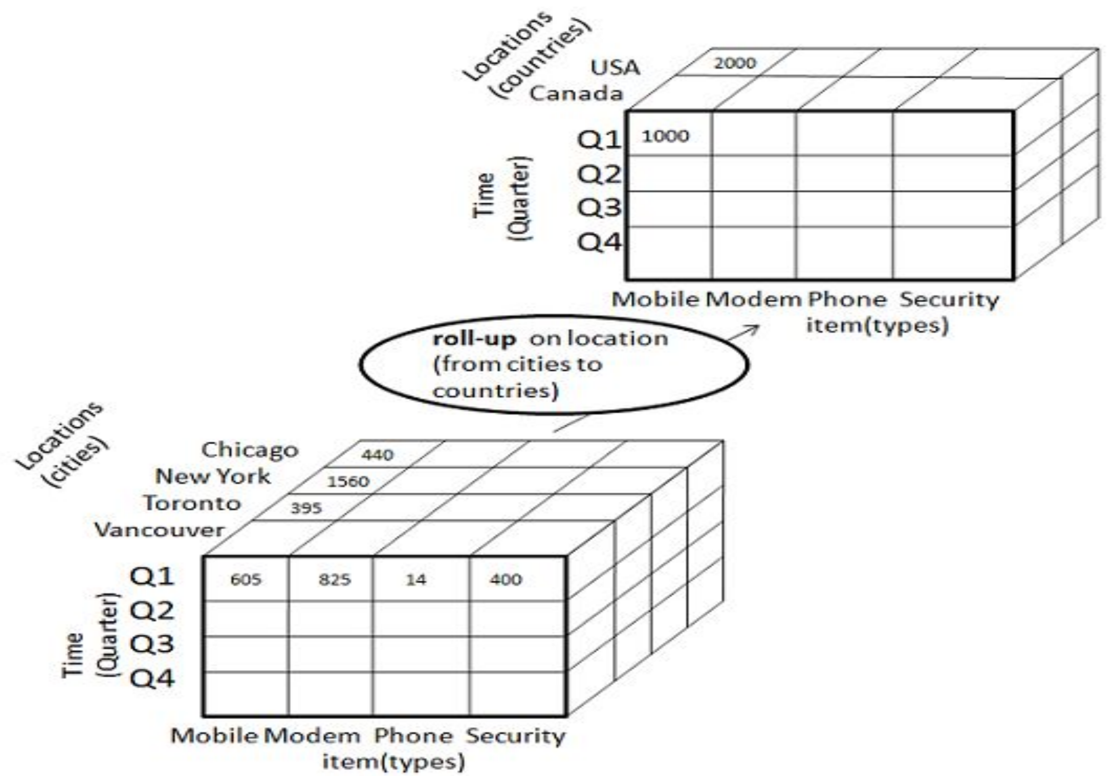

Figure 3.12 Roll Up

Drill Down

- **Drill down (roll down):** reverse of roll-up
- From higher level summary to lower level summary or detailed data, or introducing new dimensions

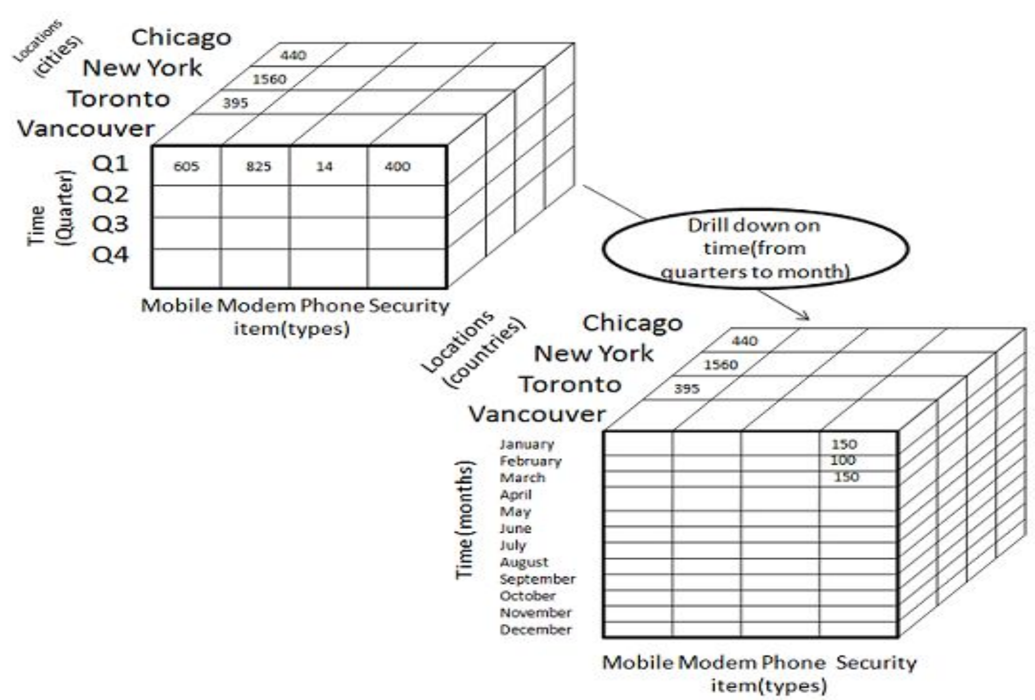

Figure 3.13 Drill Down

Slice and Dice

- **Slice and dice:** project and select

  - Here Slice is performed for the dimension "time" using the criterion time = "Q1".

  - 032

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

Slice

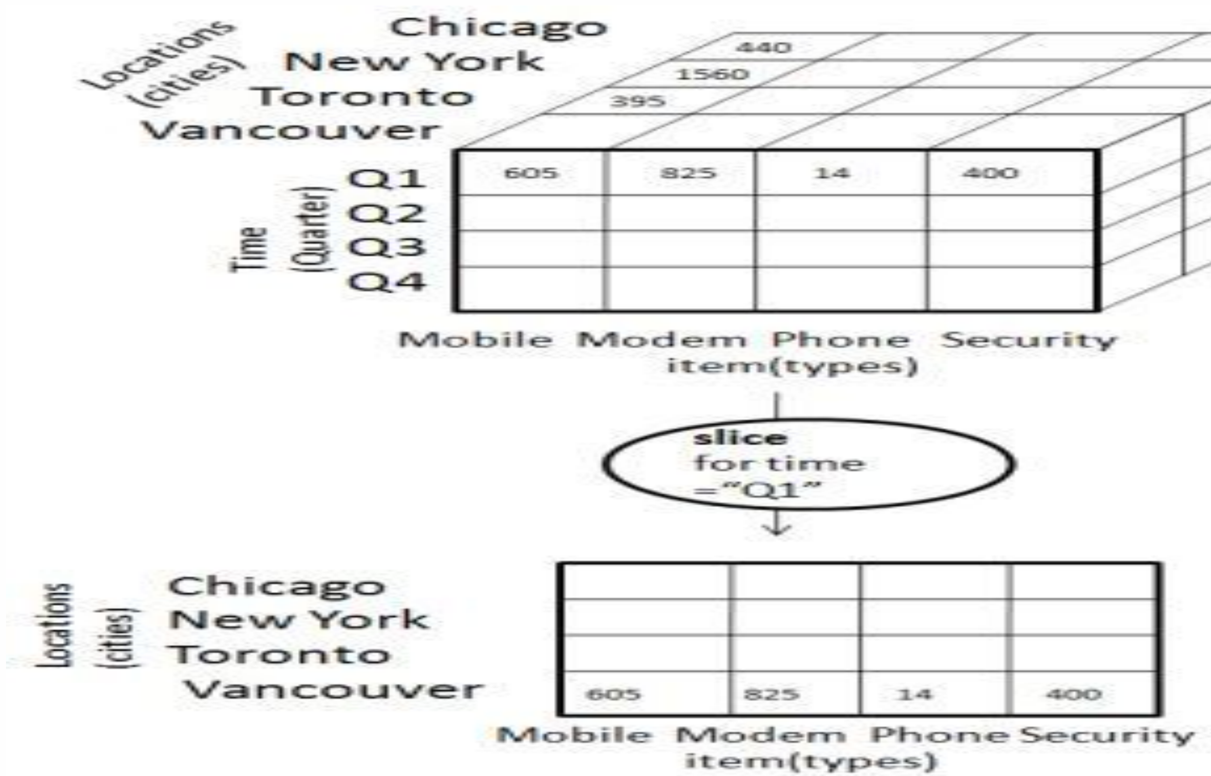

Figure 3.14
Slice

Dice

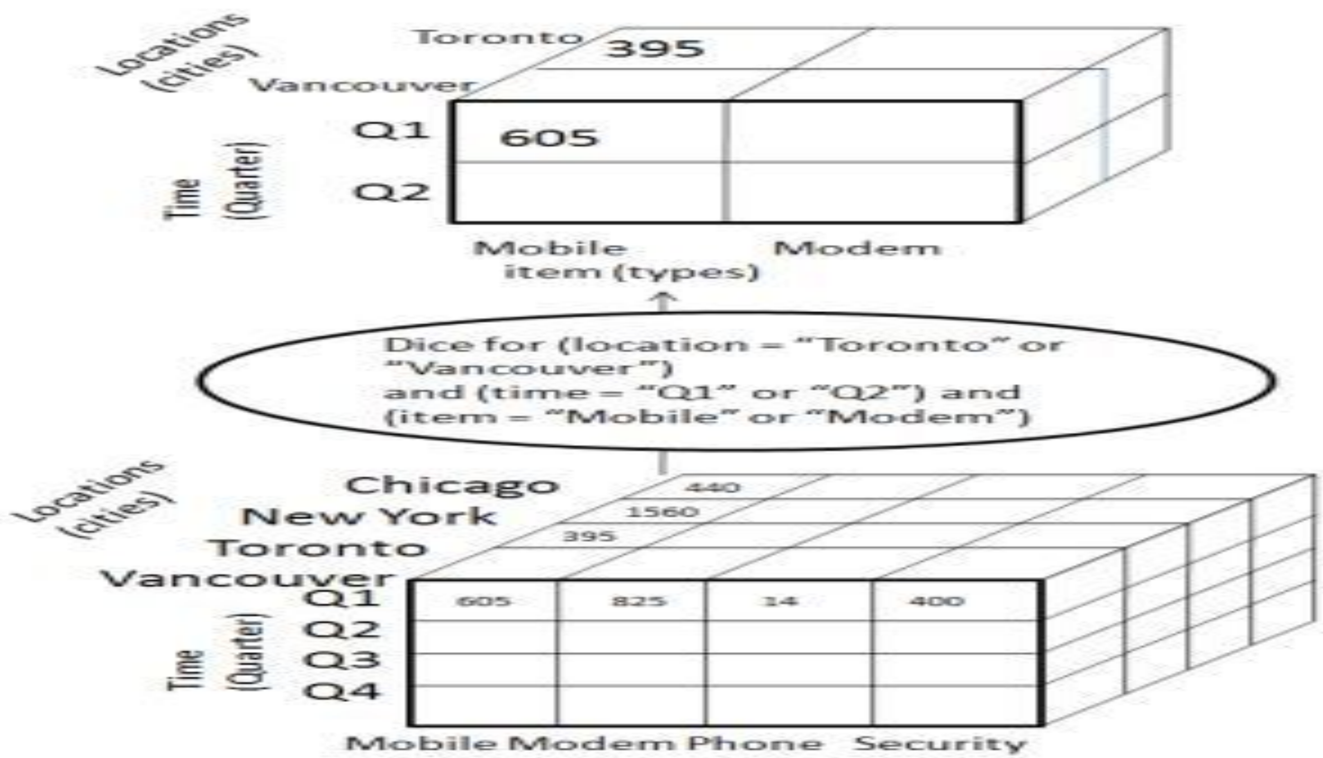

Figure 3.15 Dice

Pivot (Rotate)

- Reorient the cube, visualization, 3D to series of 2D planes

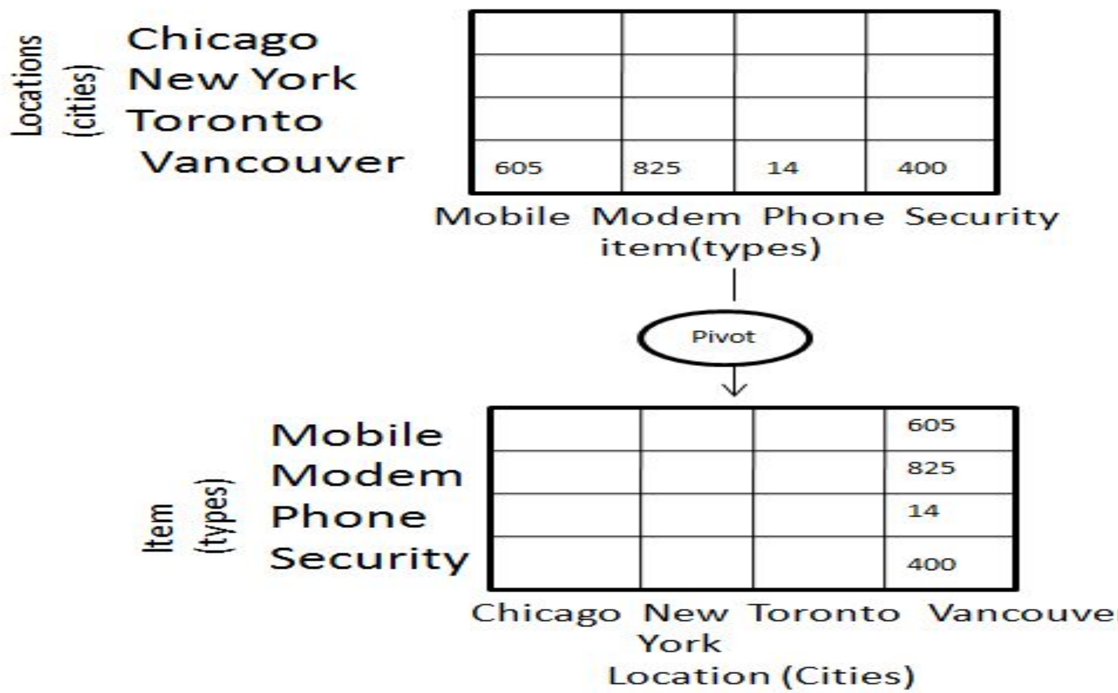

Figure 3.16
Pivot

OLAP AND OLTP

Table 3.1

	OLAP	OLTP
Definition	Multi-dimensional analysis of data arranged in cube format and modeled in a dimensional structure.	Analysis of data arranged in a relational database. Dimensional structuring of data is not necessary.
Schema	Star Schema, Snow Flake Schema.	Flat Schema.
Normalization	Renormalized.	Normalized.
Limitations	Inserts and updates are slow.	Read operations are slow.
Advantages	Fast read operations.	Fast updates and inserts.
Portability	Cubes can be created and accessed in an offline mode.	OLTP processing typically requires users to be online if they need to do any data analysis.
Use	Typically used where quick and efficient analysis is the primary use case.	Typically used where insertions and updates are to be performed frequently as part of day-to-day operations.
Data volumes	Typically involves large data volumes and historical data.	Typically data volumes are less and historical data is generally not maintained.
Pre-aggregated values	Computed at cube creation time itself for fast analysis, for example, total sales revenue for each quarter.	Computed at runtime, saving a lot of disk space but slower at the time of retrieval.

DIGITAL LEARNING CONTENT