Data Mining and Data Warehousing 5

Data Warehousing

Chittaranjan Pradhan
School of Computer Engineering,
KIIT University

Data Warehouse

Data Warehouse

- Data warehouse is a single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context
- It is a collection of data designed to support management decision- making
- The traditional database is designed for transaction processing, whereas a data warehouse is a relational database that is designed for query and analysis
- It is a collection of methods, techniques, and tools used to support knowledge workers to conduct data analyses that help with performing decision-making processes and improving information resources

Data Warehouse...

Data Warehouse...

- A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of manager's decision making process
- **Subject-Oriented Data**
  - A data warehouse is organized around major subjects, such as customer, sales
  - It focuses on the modeling and analysis of data for decision makers, rather than doing the day-to-day transaction processing of the organization
  - It provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process
- **Integrated Data**
  - A data warehouse is constructed by integrating multiple heterogeneous data sources into one consistent database
  - Data cleaning and data integration techniques are applied when data is moved to the warehouse
  - It ensures consistency in naming conventions, encoding structures, attribute measures etc. among different data sources

Data Warehouse...

Data Warehouse...

- **Time-Variant Data**
  - A data warehouse focuses on change over time
  - It provides information from a historical perspective
  - Every key structure in the data warehouse Contains an element of time, explicitly or implicitly
- **Nonvolatile Data**
  - Data is never deleted from data warehouses and updates are normally carried out when data warehouses are offline. This means that data warehouses can be essentially viewed as read-only databases
  - Data warehouse does not require transaction processing, recovery, and concurrency control mechanisms
  - It requires only two operations in data accessing: initial loading of data and access of data

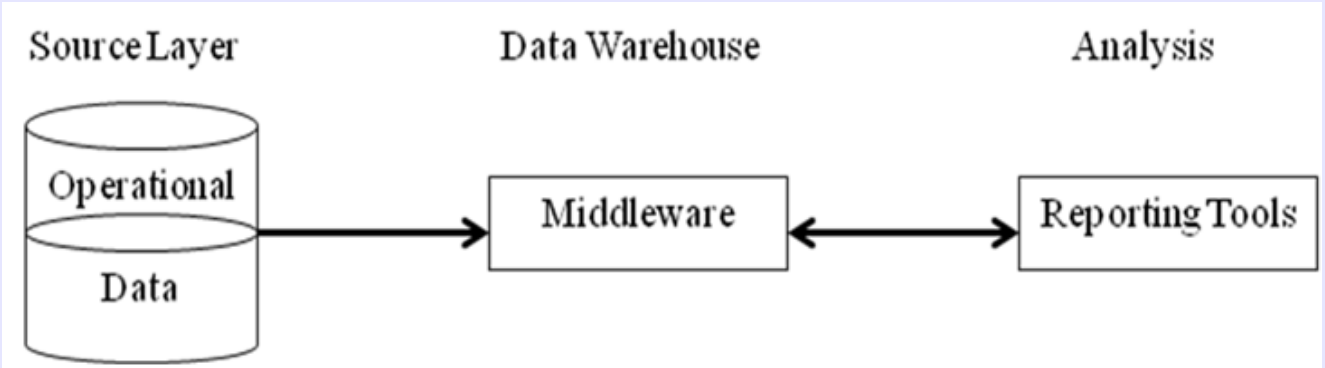
Operational Database (OLTP) vs. Data Warehouse (OLAP)

Operational Database (OLTP) vs. Data Warehouse (OLAP)

Feature	Operational Databases	Data Warehouses
Users	Thousands	Hundreds
Workload	Preset transactions	Specific analysis queries
Access	To hundreds of records, write and read mode	To millions of records, mainly read-only mode
Goal	Depends on applications	Decision-making support
Data	Detailed, both numeric and alphanumeric	Summed up, mainly numeric
Data integration	Application-based	Subject-based
Quality	In terms of integrity	In terms of consistency
Time coverage	Current data only	Current and historical data
Updates	Continuous	Periodical
Model	Normalized	Denormalized, multidimensional
Optimization	For OLTP access to a database Part	For OLAP access to most of the database

Data Warehouse Architecture

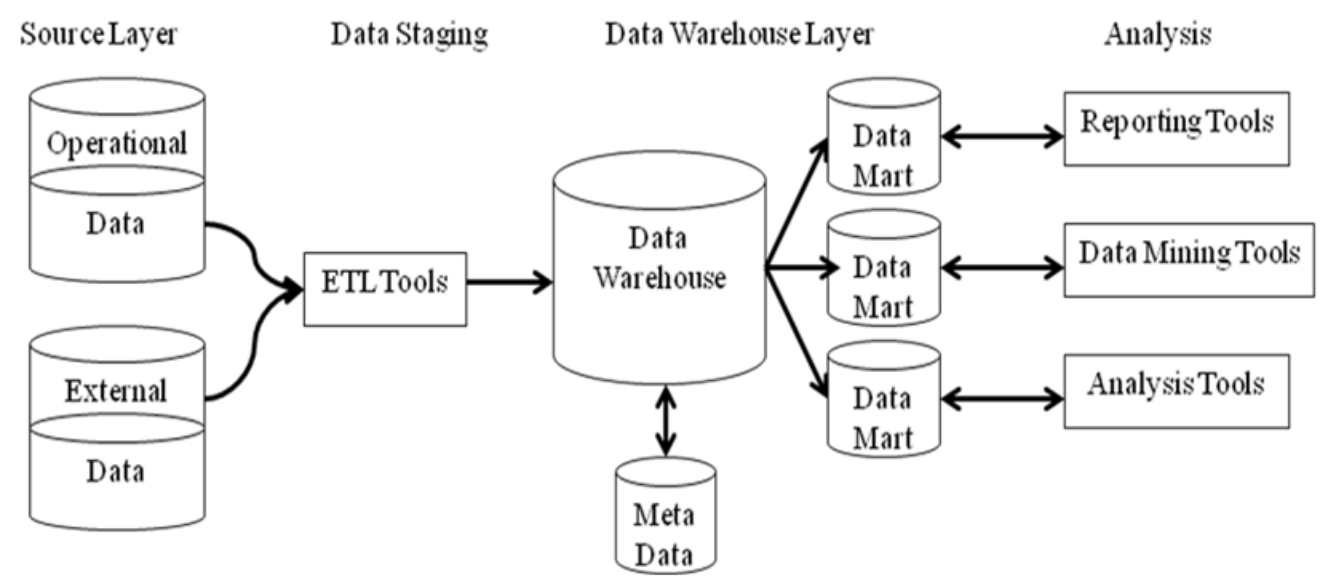
Single-Layer Architecture



- Here, the data warehouse is implemented as a multidimensional view of operational data created by specific middleware
- The weakness of this architecture lies in its failure to meet the requirements for separation between analytical and transactional processing

Data Warehouse Architecture...

Two-Layer Architecture



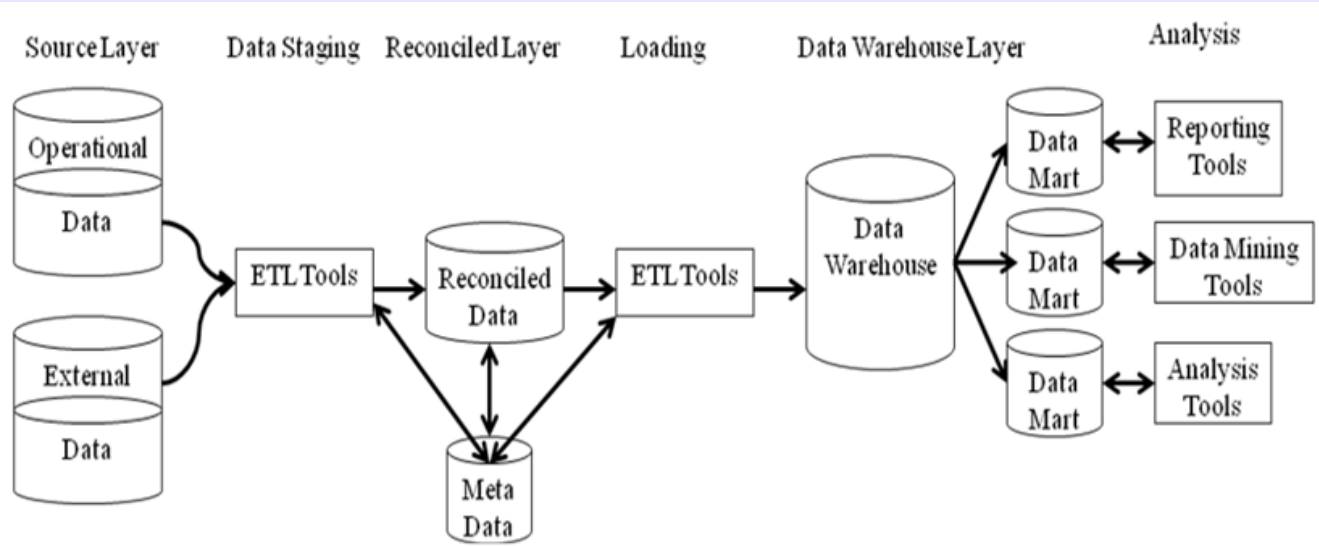
- *Meta-data repositories* store information on sources, access procedures, data staging, users, data mart schemata etc...
- The primary data warehouse acts as a centralized storage system for all the data being summed up, whereas data marts can be viewed as small, local data warehouses replicating the part of a primary data warehouse required for a specific application domain

Data Warehouse Architecture...

Two-Layer Architecture...

- Data warehouses are logically structured according to the multidimensional model, while operational sources are generally based on relational or semi-structured models

Three-Layer Architecture



Data Warehouse Architecture...

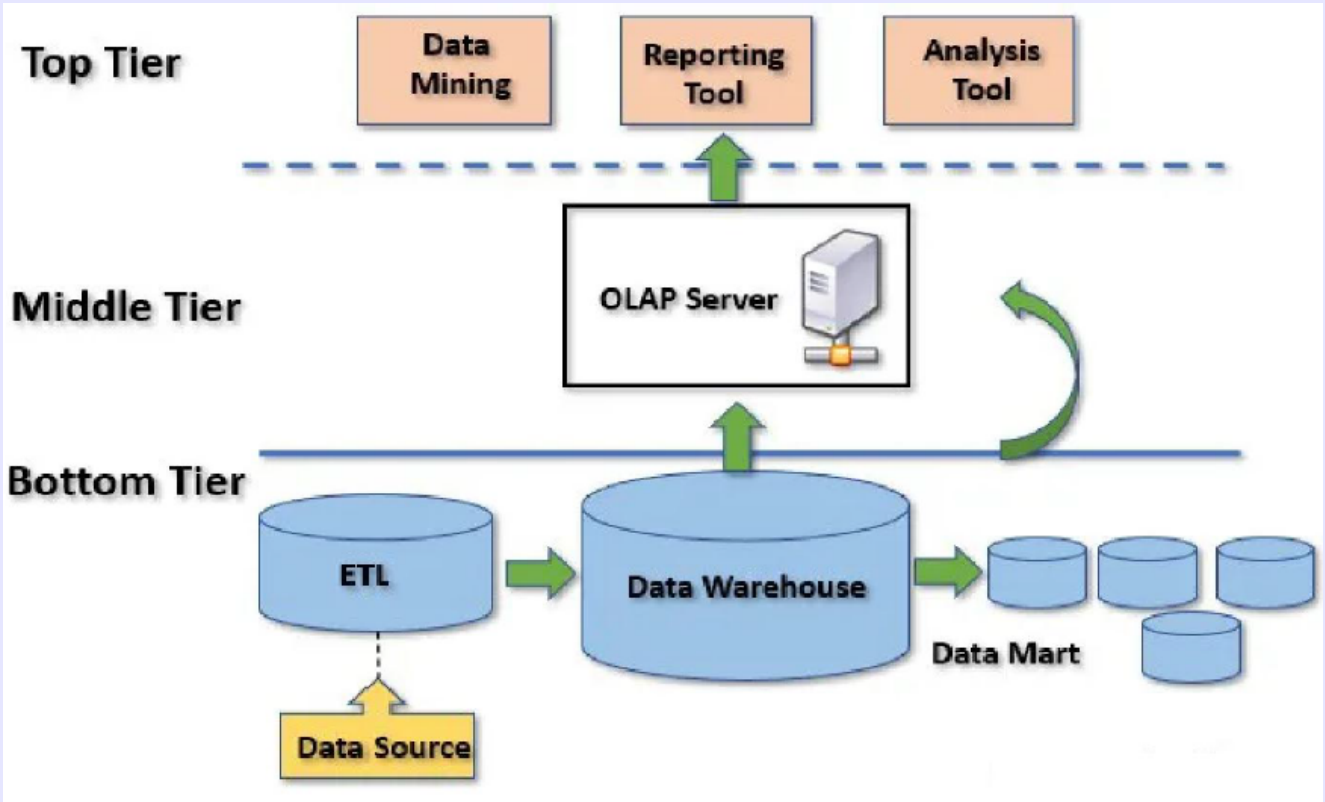
Three-Layer Architecture...

- The Reconciled layer materializes operational data obtained after integrating and cleansing source data. As a result, those data are integrated, consistent, correct, current, and detailed
- The main advantage of the reconciled data layer is that it creates a common reference data model for a whole enterprise. At the same time, it sharply separates the problems of source data extraction and integration from those of data warehouse population
- However, reconciled data leads to more redundancy of operational source data

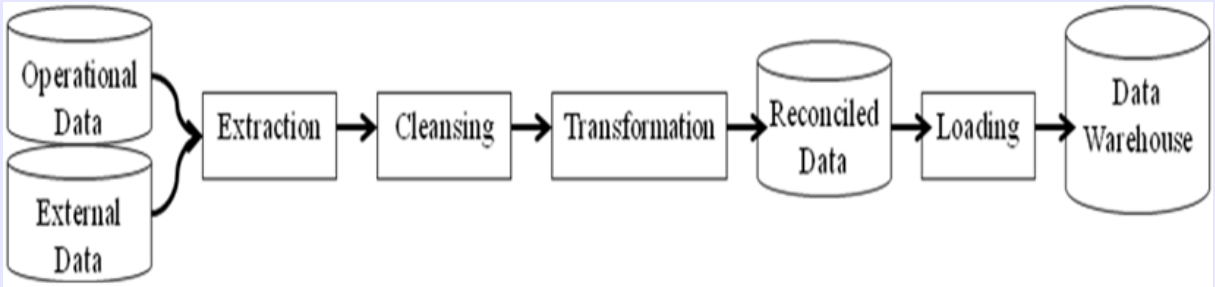
Data Warehouse Architecture...

Three-Tier Architecture...

- The three-tier data warehouse architecture is the commonly used design that includes
  - Bottom Tier (Data sources and data storage)
  - Middle Tier (OLAP Engine)
  - Top Tier (Front-End Tool)



Data Staging & ETL



- The data staging layer hosts the ETL processes that extract, integrate, and clean data from operational sources to feed the data warehouse layer
- ETL takes place once when a data warehouse is populated for the first time, then it occurs every time the data warehouse is regularly updated
- **Extraction**
  - This phase gathers data from multiple, heterogeneous and external sources
  - *Static extraction* is used when a data warehouse needs to be populating for the first time
  - *Incremental extraction* is used to update data warehouses regularly

Data Staging & ETL...

- **Cleansing**
  - This phase detects errors in the data and rectifies them when possible
  - The most frequent errors and inconsistencies that make data unclean:
    - Duplicate data
    - Missing data
    - Impossible or Wrong values
    - Inconsistent values
- **Transformation**
  - This phase converts data from legacy or host format to warehouse format
  - It is the core of the reconciliation phase
  - It converts data from its operational source format into a specific data warehouse format
  - The main transformation processes are:
    - *Conversion* and *normalization* that operate on both storage formats and units of measure to make data uniform
    - *Matching* that associates equivalent fields in different sources
    - *Selection* that reduces the number of source fields and records

Data Staging & ETL...

- **Loading**
  - This phase sorts, consolidates, checks integrity, and builds indices and partitions
  - It can be carried out in two ways:
    - *Refresh*: Data warehouse data is completely rewritten. This means that older data is replaced. Refresh is normally used in combination with static extraction to initially populate a data warehouse
    - *Update*: Only those changes applied to source data are added to the data warehouse. Update is typically carried out without deleting or modifying preexisting data. This technique is used in combination with incremental extraction to update data warehouses regularly

Data Warehouse Model

From architecture point of view, there are three warehouse models: Enterprise warehouse, Data mart and Virtual warehouse

Enterprise Warehouse

- An enterprise warehouse collects all information topics spread throughout the organization
- It provides corporate-wide data integration, typically from one or several operational systems or external information providers, and is cross-functional in scope
- It usually contains detailed data as well as summarized data and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond
- The traditional mainframe, computer super server, or parallel architecture has been implemented on platforms
- This requires extensive commercial modeling and may take years to design and manufacture

Data Warehouse Model...

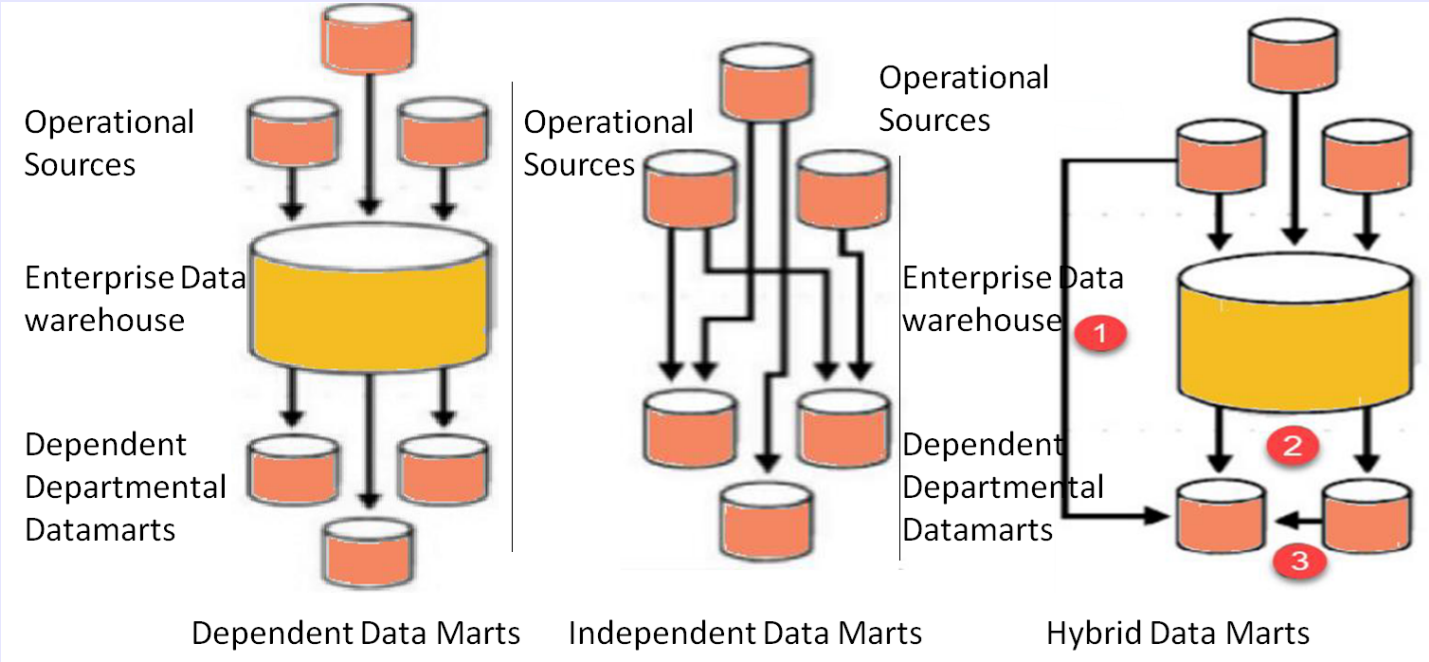
Data Mart

- A data mart contains a subset of organization-wide data that is important to a specific group of an organization
- The scope is limited to specific selected subjects. e.g. a marketing data mart may limit its topics to customers, goods, and sales
- The data contained in the data marts are summarized, small in size and flexible
- Data marts are customized by department to the departmentally structured data warehouse
- Data marts are typically applied to low-cost departmental servers that are Unix/Linux or Windows-based
- The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years

Data Warehouse Model...

Data Mart Types

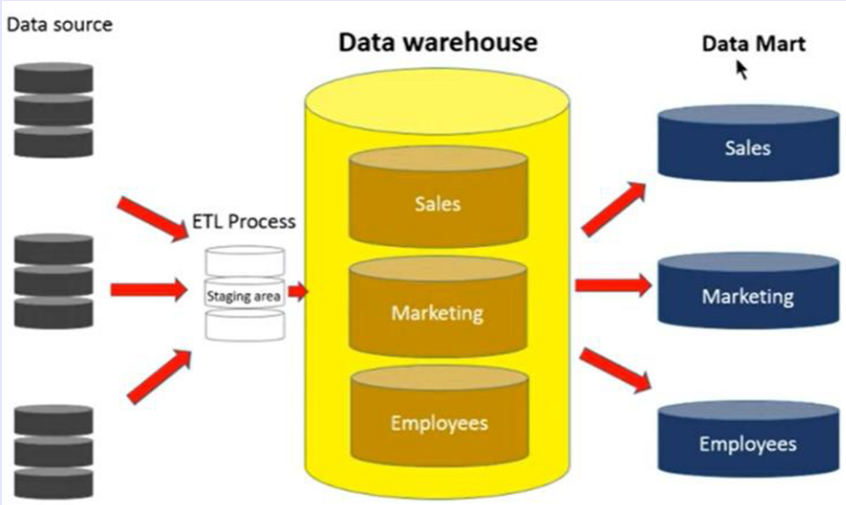
- **Dependent data mart**: are created by drawing data from operational, external or from both sources through data warehouse
- **Independent data mart**: is created without the use of a central data warehouse
- **Hybrid data mart**: can take data from data warehouse or operational systems



Data Warehouse Model...

Advantages of Data Mart

- Improve end-user response time
- Lower implementation cost
- Fast easy access data
- Frequently requested data is fastly provided to the end-user
- Data mart store only single subject area data



Data Mart	Data Warehouse
Smaller size area (GB)	Larger size area (TB)
Focus on single object	Focus on multiple objects
Simple to build	Complex to build
Data contained in summarized form	Data contained in detailed form
Decentralized	Centralized

Data Warehouse Model...

Virtual Warehouse

- A virtual warehouse is a group of views on an operational database
- For efficient query processing, only a few possible summary views can be physical
- Creating a virtual warehouse is easy, but requires additional capacity on operational database servers

Multidimensional Model

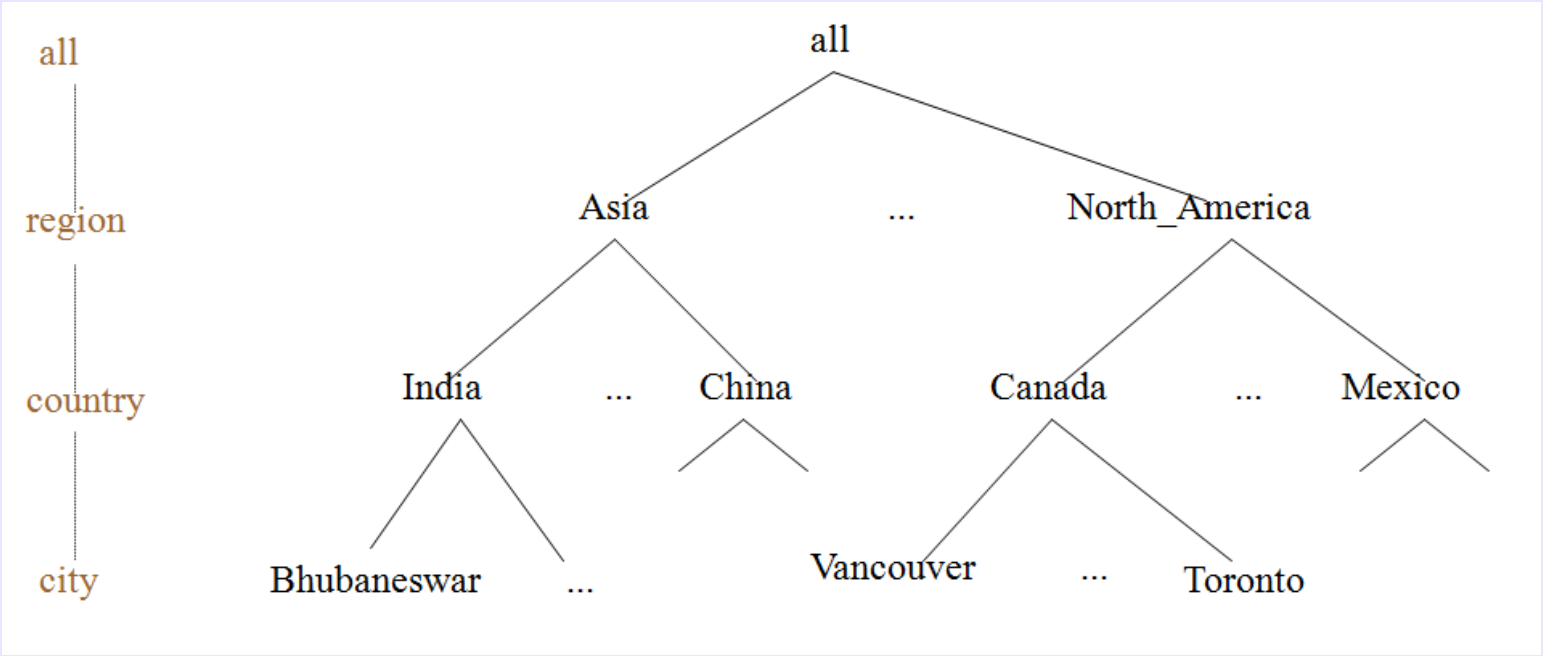
- Data is divided into Dimensions and Facts
- **Dimensions**
  - Dimensions are the perspectives or entities with respect to which an organization wants to keep records
  - Ex: Sales data warehouse may keep records of the store's sales w.r.t. the dimensions *time, item, branch and location*
  - Each dimension may have a table associated with it, called a *dimension table*
- **Facts**
  - Facts are numerical measures which are used to analyse the relationship between dimensions
  - Ex: Facts for a Sales data warehouse include *dollars_sold, units_sold*
  - The *fact table* contains the names of the facts or measures as well as keys to each of the related dimension tables
- Dimensions describe facts
- Facts have measures that can be aggregated: sales price

Multidimensional Model...

- Goal for dimensional modeling:
  - Surround facts with as much context (dimensions) as possible
- A data warehouse is based on a *multidimensional data model* which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - *Dimension tables*, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - *Fact table* contains *measures* (such as dollars_sold) and keys to each of the related dimension tables
- *ER models describe entities and relationships where as Dimensional models describe measures and dimensions*
- Each dimension is associated with a hierarchy of aggregation levels, called as *roll - up hierarchy*

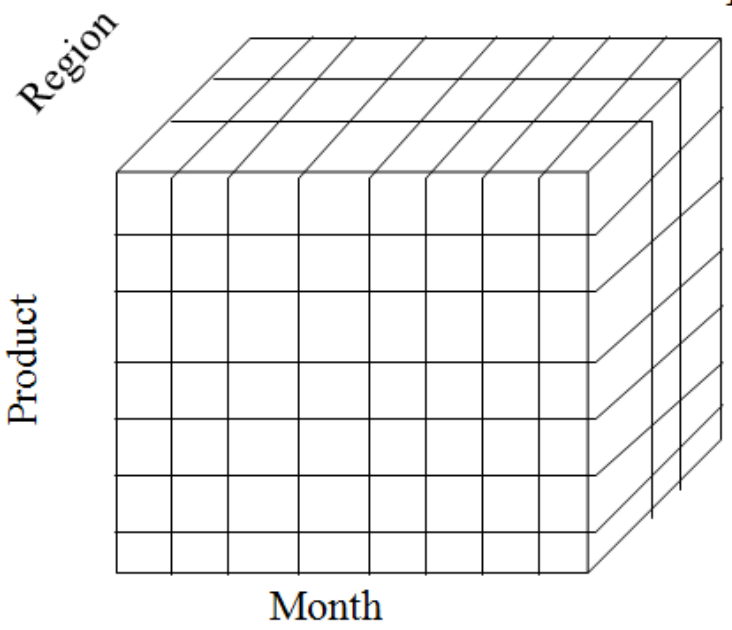
Multidimensional Model...

- Hierarchies consist of levels called *dimensional attributes*

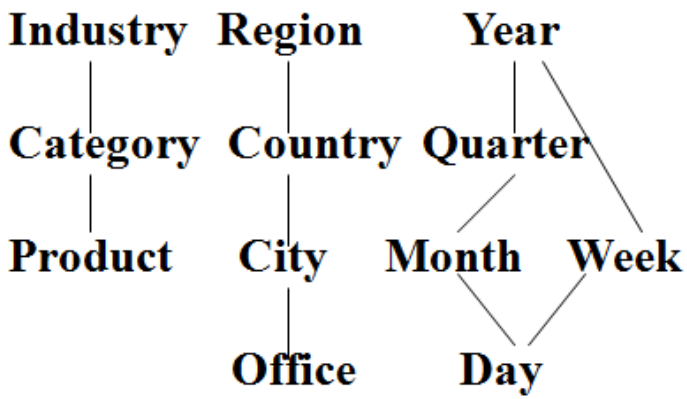


Multidimensional Model...

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

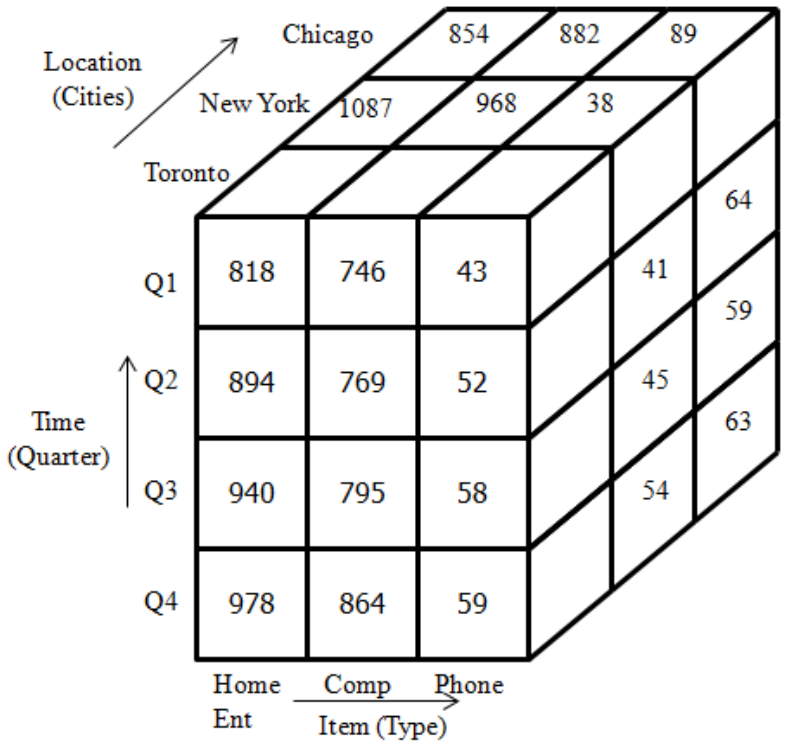


Dimensions: *Product, Location, Time*
Hierarchical summarization paths

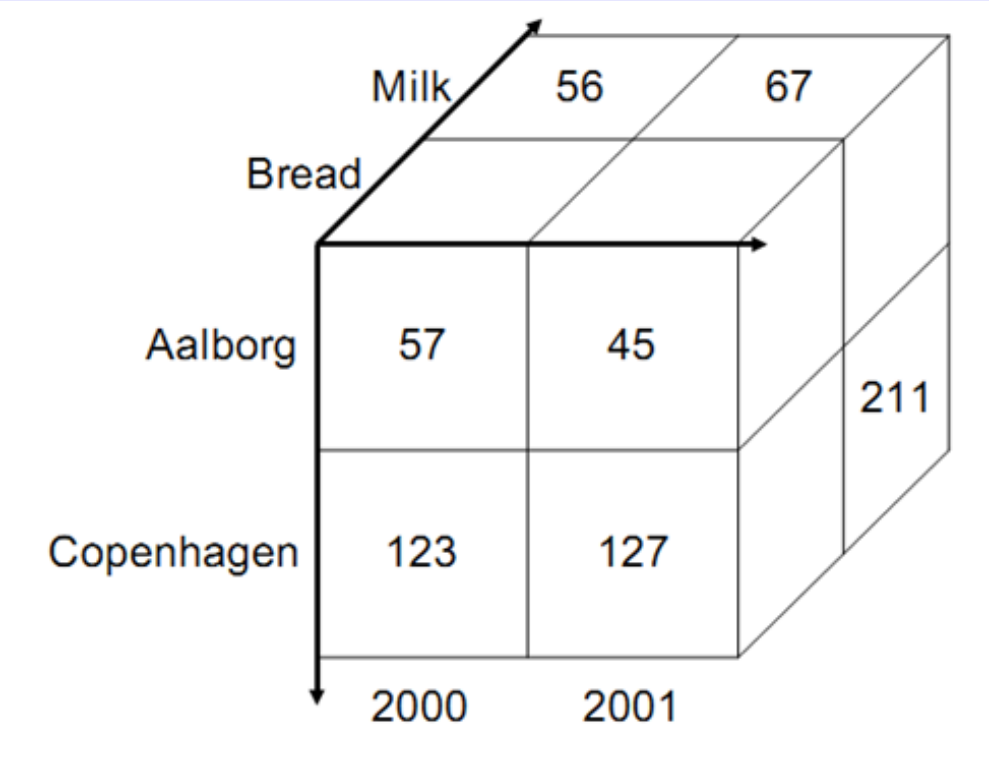


Multidimensional Model...

	Location = "Chicago"			Location = "New York"			Location = "Toronto		
	Item			Item			Item		
Time	Home Ent	Comp	Phone	Home Ent	Comp	Phone	Home Ent	Comp	Phone
Q1	854	882	89	1087	968	38	818	746	43
Q2	943	890	64	1130	1024	41	894	769	52
Q3	1032	924	59	1034	1048	45	940	795	58
Q4	1129	992	63	1142	1091	54	978	864	59



Multidimensional Model...



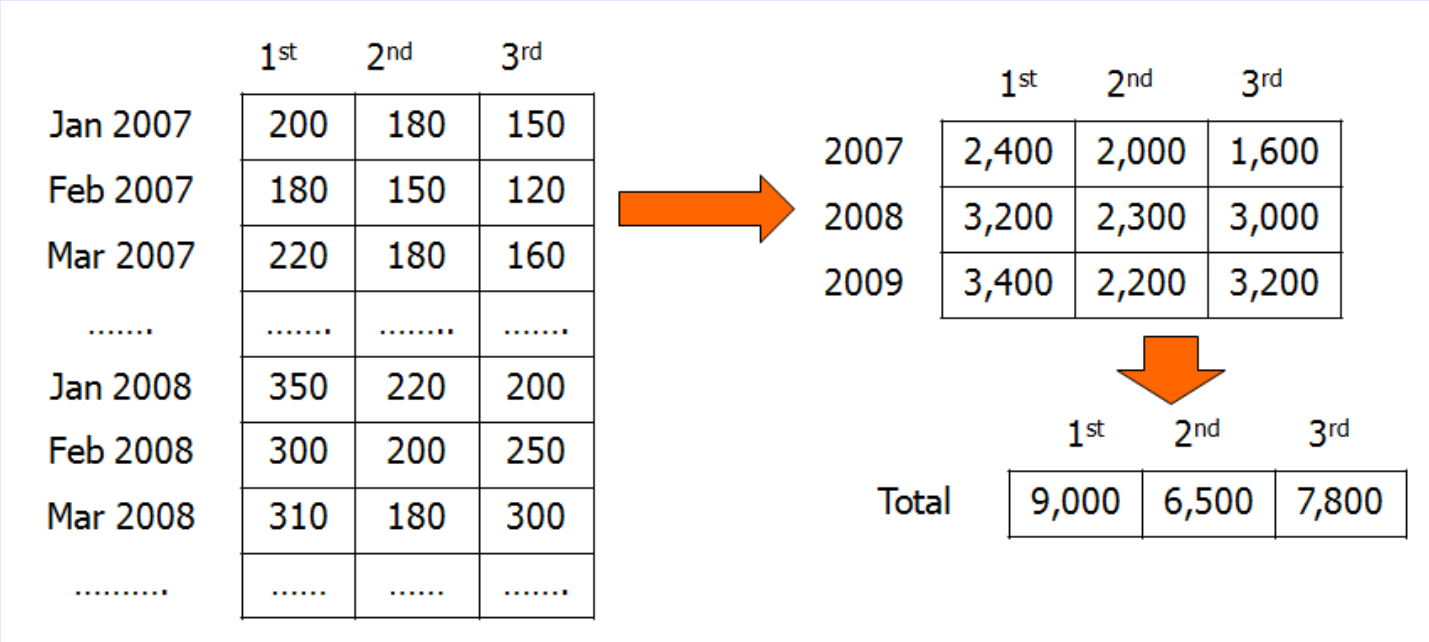
- Each cube axis shows a possible analysis dimension
- Each dimension can be analyzed at different detail levels specified by hierarchically structured attributes

Multidimensional Model...

- **Restriction**:
  - Restricting data means separating part of the data from a cube to mark out an analysis field
  - **Slicing**: Slicing means decreasing cube dimensionality by setting one or more dimensions to a specific value
  - **Dicing**: Dicing is the generalization of slicing. It puts the conditions on dimensional attributes to scale down the size of a cube
  - **Projection**: Projection can be referred to as a choice to keep just one subgroup of measures for every event and reject other measures

Multidimensional Model...

- **Aggregation**
  - Every aggregate event will sum up the data available in the events it aggregates
  - You can aggregate along various dimensions at the same time



Meta - Data

- It specifies source, values, usage, and features of data ware house data and defines how data can be changed and processed at every architecture layer
- Applications use it intensively to carry out data - staging and analysis tasks
- In data warehouse, metadata is used for building, maintaining, managing, and using the data warehouses. Metadata helps users to easily access, understand the content and find data in data warehouse
- Metadata includes the following:
  - The location and descriptions of warehouse systems and components
  - Names, definitions, structures, and content of data-warehouse and endusers views
  - Integration and transformation rules used to populate data, to deliver information to end-user analytical tools
  - Metrics used to analyze warehouses usage and performance
  - Security authorizations, access control list, etc

Meta - Data...

Meta - Data...

- **Internal Meta-Data**
  - It defines sources, transformation processes, population policies, logical and physical schema, constraints and user profiles
  - System administrator is interested
- **External Meta-Data**
  - It is about definitions, quality standards, units of measure, relevant aggregations
  - It is relevant to end users

Ref: J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 3rd edition