Unit 1

DATA WAREHOUSING

UNIT 1

Introduction to Data Warehousing: ***Introduction, Necessity, Framework*** *of the datawarehouse, options, developing datawarehouses, end points.*

Data Warehousing Design Consideration and Dimensional Modeling: *Defining Dimensional Model, Granularity of Facts, Additivity of Facts, Functional dependency of the Data, Helper Tables, Implementation manyto-many relationships between fact and dimensional modelling*

What Is A Data Warehouse?

- A data warehouse is a powerful database model that significantly enhances the user's ability to quickly analyze large, multidimensional data sets.
- It cleanses and organizes data to allow users to make business decisions based on facts.
- Creating data to be analytical requires that it be **subject-oriented**, **integrated**, **time-referenced**, and **non-volatile**.

- Subject-Oriented Data
  - Data warehouses group data by subject rather than by activity.
  - subjects— employees, accounts, sales, products.
  - This subject specific design helps in reducing the query response time
- **Integrated Data**
  - Integrated data refers to de-duplicating informationand merging it from many sources into one consistent location.
  - Much of the transformation and loading work that goes into the data warehouse is centered on integrating data and standardizing it.

- Time-Referenced Data
  - time-referenced data essentiallyrefers to its time-valued characteristic.
  - EG:the user may ask “What were the total sales of product ‘A’ for the past three years on New Year’s Day across region ‘Y ’?”.
  - This exploration activity is termed “data mining”
- **Non-Volatile Data**
  - The non-volatility of data, characteristic of data warehouse, enables users to dig deep into history and arrive at specific business decisions based on facts.

Why A Data Warehouse?

- **The Data Access Crisis**
  - Every day, organizations large and small, create billions of bytes of data about all aspects of their business;
  - millions of individual facts about their customers, products, operations and people. But for the most part, this is locked up in a maze of computer systems and is exceedingly difficult to get at.
  - This phenomenon has been described as “data in jail”.

Data Warehousing

- Data warehousing is a field that has grown from the integration of a number of different technologies and experiences over the past two decades. These experiences have allowed the IT industry to identify the key problems that need to besolved.

Operational vs. Informational Systems

- OPERATIONAL
  - Operational systems, as their name implies, are the systems that help the every day operation of the enterprise.
  - These are the backbone systems of any enterprise, and include order entry, inventory, manufacturing, payroll and accounting.
- INFORMATIONAL
  - Informational systems deal with analyzing data and making decisions
  - informational data needs often span a number of different areas and need large amounts of related operational data.

Framework Of The Data Warehouse

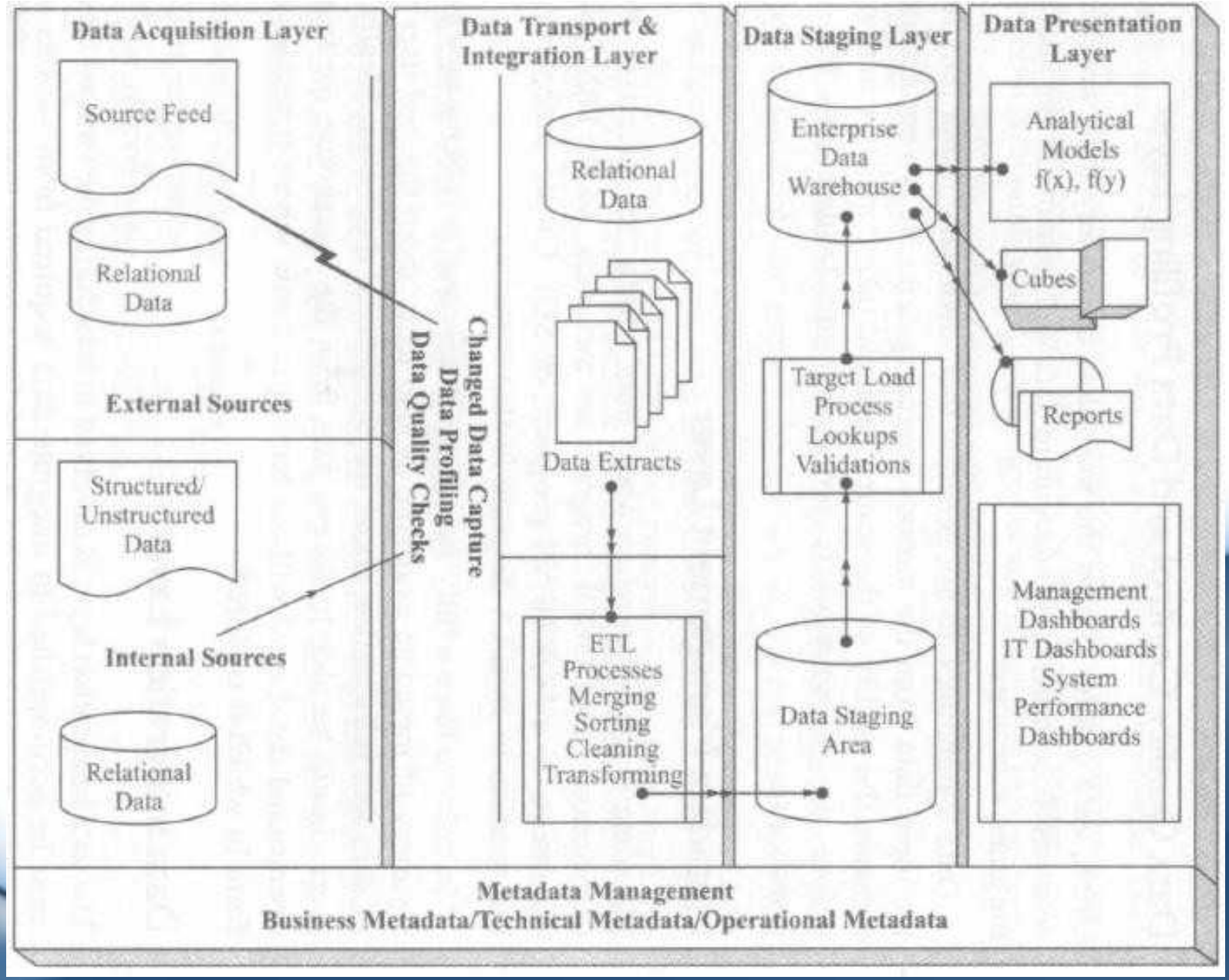
- One of the reasons that data warehousing has taken such a long time to develop is that it is actually a very comprehensive technology.
- In fact, it can be best represented as an enterprise-wide framework for managing informational data within the organization.
- In order to understand how all the components involved in a data warehousing strategy are related, it is essential to have a Data Warehouse Architecture. In order to understand how all the components involved in a data warehousing strategy are related, it is essential to have a Data Warehouse Architecture.

Data Warehouse Architecture

- A Data Warehouse Architecture (DWA) is a way of representing the overall structure of data, communication, processing and presentation that exists for end-user computing within the enterprise.

The architecture is made up of a number of interconnected parts

- Source system
- Source data transport layer
- Data quality control and data profiling layer
- Metadata management layer
- Data integration layer
- Data processing layer
- End user reporting layer


Data Acquisition Layer
Data Transport & Integration Layer
Data Staging Layer
Data Presentation Layer
Source Feed
Relational Data
External Sources
Structured/ Unstructured Data
Internal Sources
Relational Data
Changed Data Capture
Data Profiling
Data Quality Checks
Relational Data
Data Extracts
ETL Processes Merging Sorting Cleaning Transforming
Enterprise Data Warehouse
Target Load Process Lookups Validations
Data Staging Area
Analytical Models f(x), f(y)
Cubes
Reports
Management Dashboards IT Dashboards System Performance Dashboards
Metadata Management
Business Metadata/Technical Metadata/Operational Metadata

Data Warehouse Options

- number of key factors that need to be considered to develop data warehouses.

1. **Scope of the data warehouse**
   - The scope of a data warehouse may be as broad as all the informational data for the entire enterprise from the beginning of time, or it may be as narrow as a personal data warehouse for a single manager for a single year.
   - broader the scope, the more valuable the warehouse is to the enterprise and the more expensive and time consuming it is to create and maintain.

2. Data redundancy

- three levels of data redundancy
  - "Virtual" or "point-to-point" data warehouses
    - End users are allowed to get at operational databases directly
  - Central data warehouses
    - Central data warehouses are real. The data stored here is accessible from one place and must be loaded and maintained on a regular basis.
  - Distributed data warehouses
    - Distributed data warehouses are those in which certain components are distributed across a number of different physical databases.

3. Type of End-user
   - Executives and managers
   - Power users (business and financial analysts, engineers)
   - Support users (clerical, administrative)

Developing Data Warehouses

- Developing a good data warehouse is no different from any other IT project— it requires careful planning, requirements definition, design, prototyping and implementation.
- **Developing Strategy**
  - There are a number of strategies by which organizations can get into data warehousing.
    - Installing a set of data access, data directory and process management facilities
    - Training the end-users
    - Monitoring how the data warehouse facilities are actually used
    - Based on actual usage, creating a physical data warehouse to support the high-frequency requests

Evolving DWA

- **The DWA (Data Warehouse Architecture)** is simply a framework for understanding data warehousing and how the components of data warehouse fit together.
- One of the keys to data warehousing is flexibility

Designing Data Warehouses

- Designing data warehouses is very different from designing traditional operational systems.
  1. needs as operational users.
  2. thinking in terms of much broader, and more difficult to define
  3. quite close to Business Process Reengineering (BPR).
  4. design strategy for a data warehouse

Managing Data Warehouses

- how they want their warehouses to perform.
- also recognize that the maintenance of the data warehouse structure
- IT management must understand that if they embark on a data warehousing program

End Points

- Data warehousing is growing by leaps and bounds and it is becoming increasingly difficult to estimate what new developments are most likely to affect it.
- development of parallel DB servers with improved query engines is likely to be one of the most important.
- Parallel servers will make it possible to access huge data bases in much less time.
- data warehouse planners and developers have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility fortomorrow.

Goals

- **Provide Easy Access to Corporate Data**
  - must be easy to use
  - Access should be graphic
  - They must easily get answers to their questions and ask new questions, all without getting the IT team involved
  - The process of getting and analyzing data must be fast.
- **Provide Clean and Reliable Data for Analysis**
  - For consistent analysis, the data environment must be stable
  - One department doing an analysis must get the same result as any other
  - Source conflicts must be resolved.
  - Historical analysis must be possible, so that data can be analyzed across a span of time

2.Data Warehouse Design Consideration and Dimensional Modeling

- Warehouses support business decisions by collecting, consolidating, and organizing data for reporting and analysis with tools such as online analytical processing (OLAP) and data mining models.
- Although data warehouses are built on relational database technology, the design of a data warehouse data model and subsequent physical implementation differs substantially from the design of an online transaction processing (OLTP) system.

How do these two systems differ and what design considerations should be kept in mind while designing a data warehouse data model?

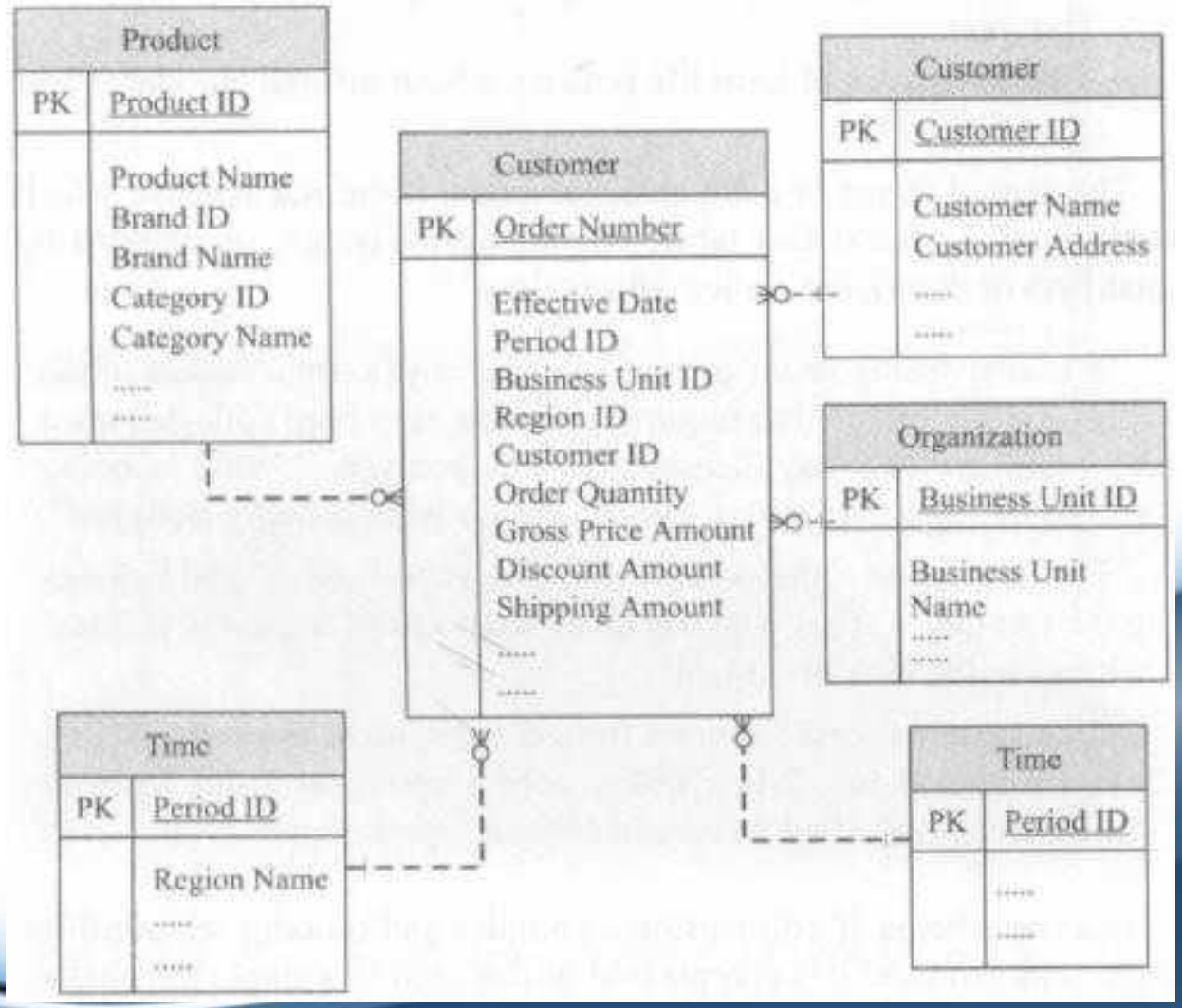
OLTP Database	*Data Warehouse Database*
Designed for real-time business transactions and processes.	Designed for analysis o f business measures by subject area, category and attributes.
Optimized for a common and known set o f transactions.	Optimized for bulk loads and large complex, unpredictable queries.
Designed for validation o f data during transactions,	Designed to be loaded with consistent, valid data; uses very minimal validation
Supports few concurrent users relative to the OLTP environment.	Supports large user bases often distributed across geographies.
Houses very minimal historical data	Houses a mix o f most current information as well as historical data

Defining Dimensional Model

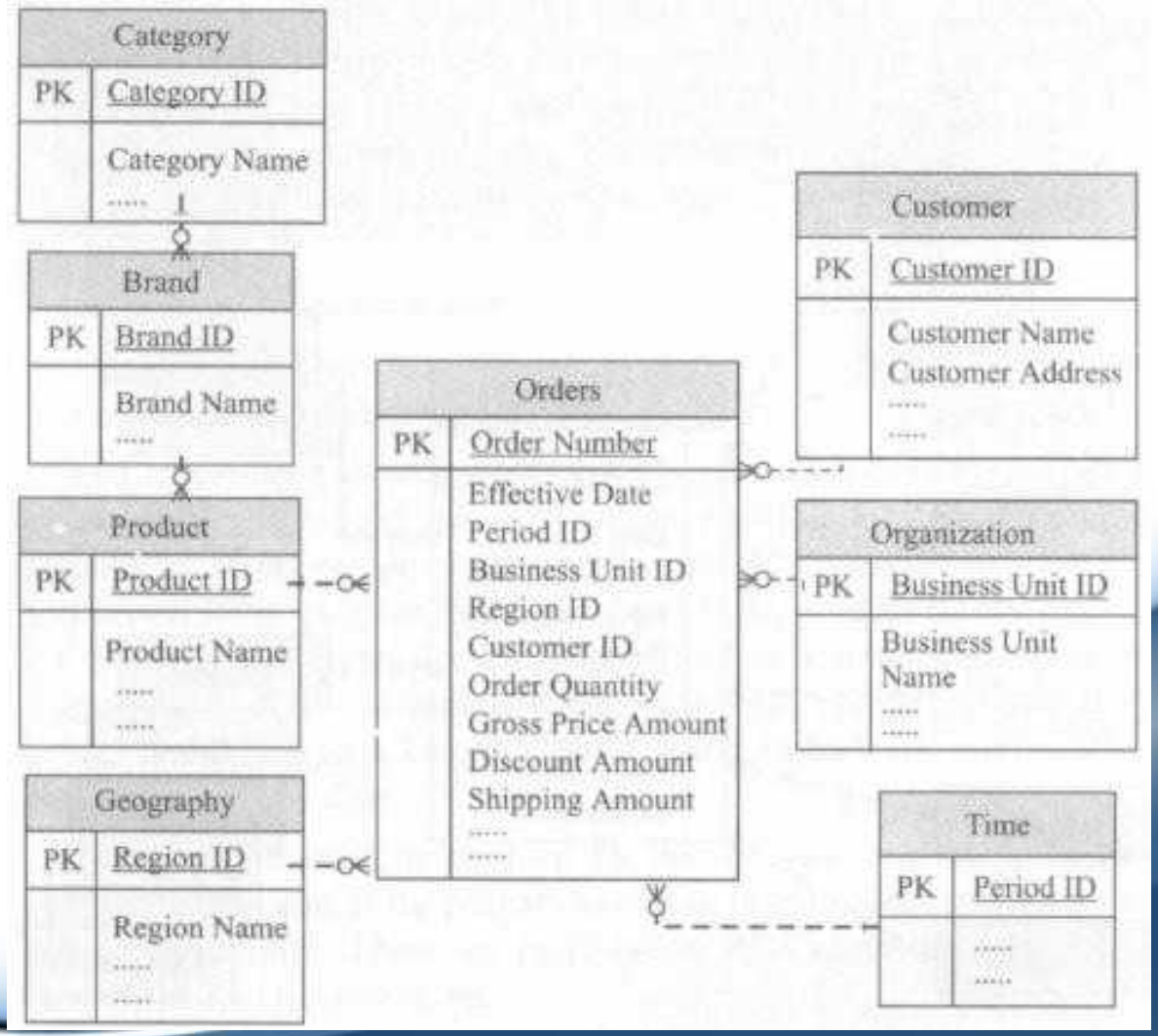
- The purpose of dimensional model is to improve performance by matching data structures to queries.
- Users query the data warehouse looking for data like
  - Total sales in volume and revenue for the NE region for product 'XYZ' for a certain period this year compared to the same period last year
- *The central theme of a dimensional model is the star schema, which consists of a central fact table, containing measures, surrounded by qualifiers or descriptors called 'dimensions'.*

- In a star schema, if a dimension is complex and contains relationships such as hierarchies, it is compressed or flattened to a single dimension.
- Another version of star schema is a snowflake schema. In a snowflake schema complex dimensions are normalized. Here, dimensions maintain relationships with other levels of the same dimension.

Star Schema Model



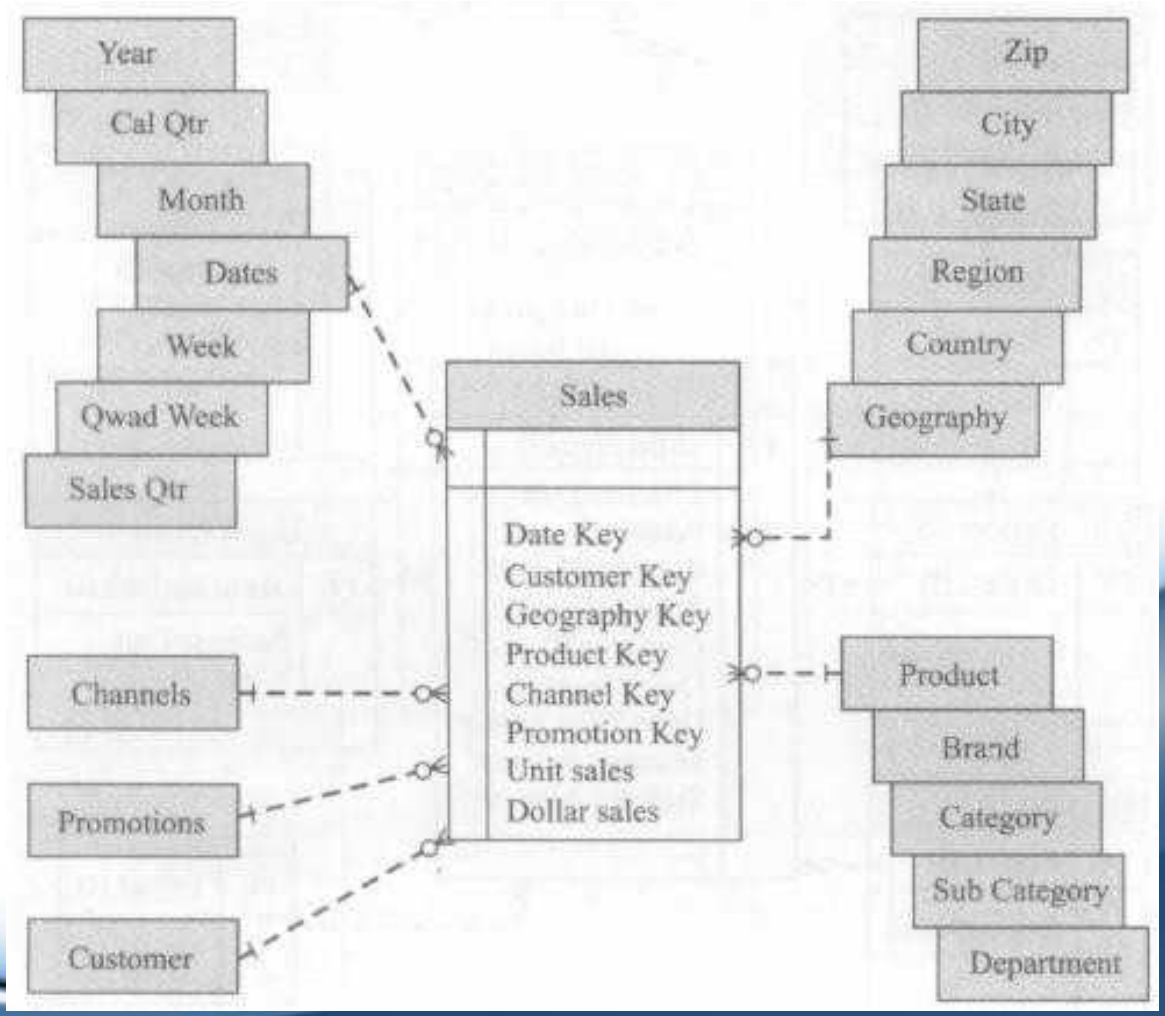
Snowflake Model



Granularity Of Facts

- The granularity of a fact is the level of detail at which it is recorded. If data is to be analyzed effectively, it must be all at the same level of granularity.
- As a general rule, data should be kept at the highest (most detailed) level of granularity.

Heavily Snow Flaked Model



- Granularity is determined by:
  - Numberof parts to a key
  - Granularity of those parts
- Adding elements to an existing key alwaysincreases the granularity of thedata.
- removing any part of an existing keydecreases its granularity
- Using customer sales to illustrate this, a key of customer ID and period ID is less granularthan customer ID, product ID and period ID.

Additivity Of Facts

- A fact is additive over a particular dimension if adding it through or over the dimension results in a fact with the same essential meaning as the original, but is now relative to the **new granularity**.
  - A fact is said to be fully additive if it is additive over every dimension of its dimensionality
  - partially additive if additive over at least one but not all of the dimensions
  - non-additive if not additive over any dimension

Functional Dependency Of The Data

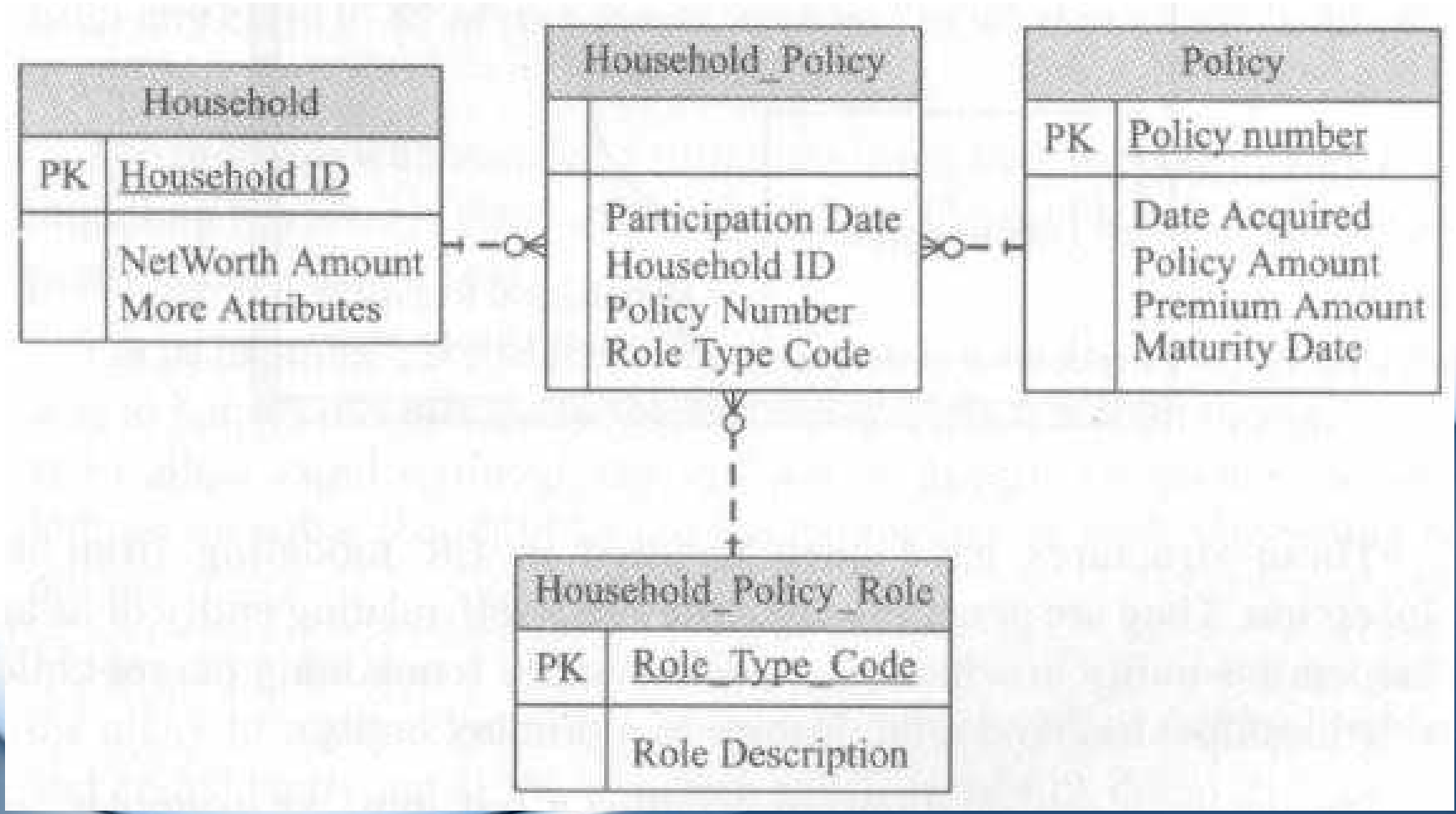
- Functional dependency of data means that the attributes within a given entity are fully dependent on the entire primary key of the entity— no more, no less.
- (**cust_id**,cust_name,cust_add,...,...)

Helper Tables

- Helper tables usually take one of two forms:
  - Help for multi-valued dimensions
  - Helper tables for complex hierarchies

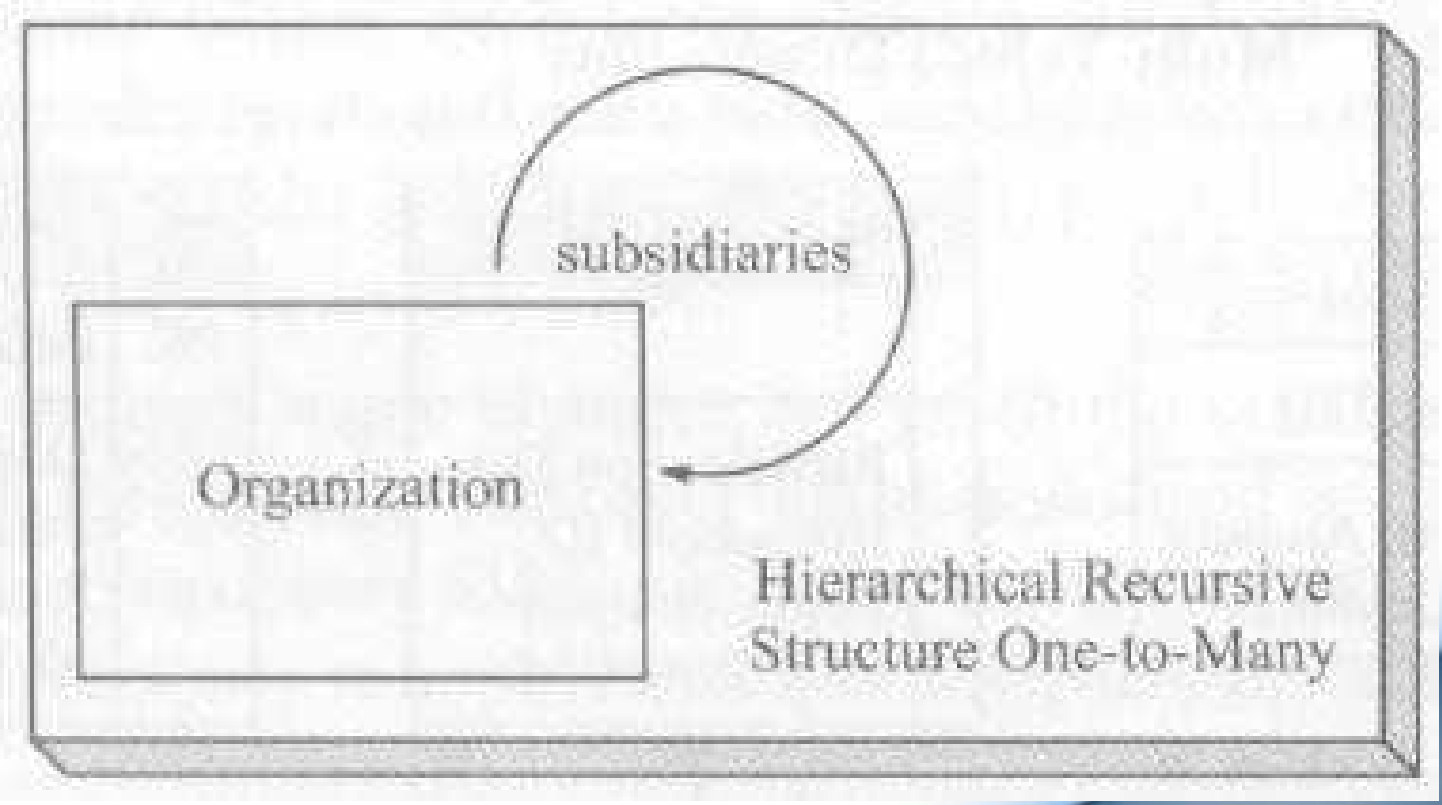
Multi-Valued Dimensions

- Take a situation where a household can own many insurance policies, yet any policy could be owned by multiple households.
- The simple approach to this is the traditional resolution of the many-to-many relationship, called an associative entity.
- The traditional way to resolve a many-to-many relationship is to create an associative entity whose key is formed from the keys of each participating parent.


Household
PK
Household ID
NetWorth Amount
More Attributes
Household_Policy
Participation Date
Household ID
Policy Number
Role Type Code
Policy
PK
Policy number
Date Acquired
Policy Amount
Premium Amount
Maturity Date
Household_Policy_Role
PK
Role_Type_Code
Role Description

Complex Hierarchies

- A hierarchy is a tree structure, such as an organization chart. Hierarchies can involve some form of recursive relationship.
- Recursive relationships come in two forms— “self” relationships (1:M) and “bill of materials” relationships (M: M). A self relationship involves one table whereas a bill of materials involves two.


subsidiaries
Organization
Hierarchical Recursive
Structure One-to-Many

DATA WAREHOUSING

Thank You