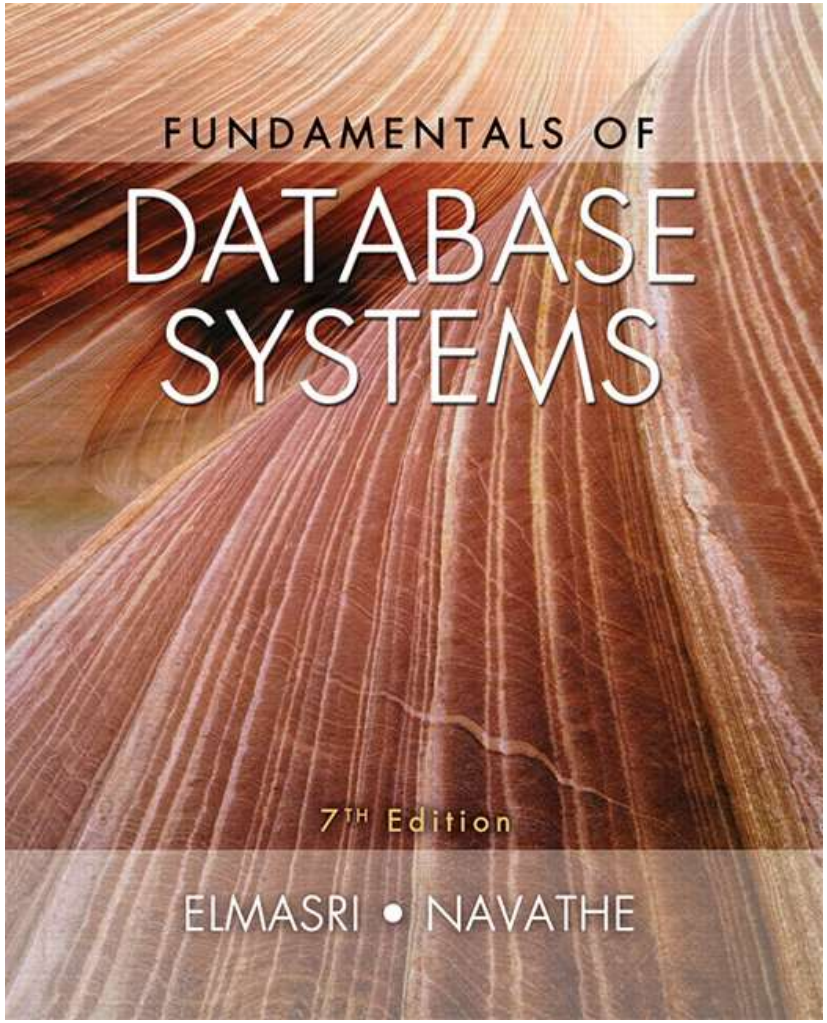


CHAPTER 29

Overview of Data Warehousing and OLAP

Chapter Outline

- 1 Purpose of Data Warehousing
- 2 Introduction, Definitions, and Terminology
- 3 Comparison with Traditional Databases
- 4 Characteristics of Data Warehouses
- 5 Classification of Data Warehouses
- 6 Data Modeling for Data Warehouses
- 7 Multi-dimensional Schemas
- 8 Building a Data Warehouse
- 9 Typical Functionality of a Data Warehouse
- 10 Data Warehouse vs. Data Views
- 11 Implementation difficulties and open issues

Introduction, Definitions, and Terminology (1)

- Data Warehouse (DW) was proposed as a new type of database management system which would keep no transactional data but only summarized historical information for decision making purposes.
- DWs are modeled and structured differently, use different techniques for storage and retrieval and cater to a different set of users.
- W. H Inmon characterized a data warehouse as:
  - **“A subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management’s decisions.”**

Purpose of Data Warehousing

- Traditional databases are not optimized for data access - they have to balance the requirement of data access with the need to ensure integrity of data.
- DWs provide access for complex analysis of data, knowledge discovery and decision support both through **ad-hoc** and **canned** queries.
- Most of the times the data warehouse users need only read access but, need the access to be fast over a large volume of data.
- Most of the data required for data warehouse analysis comes from multiple sources that may include databases from different data models and sometimes files acquired from independent systems and platforms.

Introduction, Definitions, and Terminology (2)

- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support
  - Traditional databases support online transaction processing - **OLTP .**
  - Data Warehouses are for analytical applications- largely **OLAP**.
- Applications that data warehouse supports are:
  - **OLAP** (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse.
  - **DSS** (Decision Support Systems) also known as EIS (Executive Information Systems) supports organization's leading decision makers for making complex and important decisions.
  - **Data Mining** is used for knowledge discovery, the process of searching data for unanticipated new knowledge (See Chapter 28).

Conceptual Structure of Data Warehouse

- Data Warehouse processing involves
  - Cleaning and reformatting of data
  - ETL (Extract, Transform, Load)
  - OLAP – Data Analytics
  - Data Mining

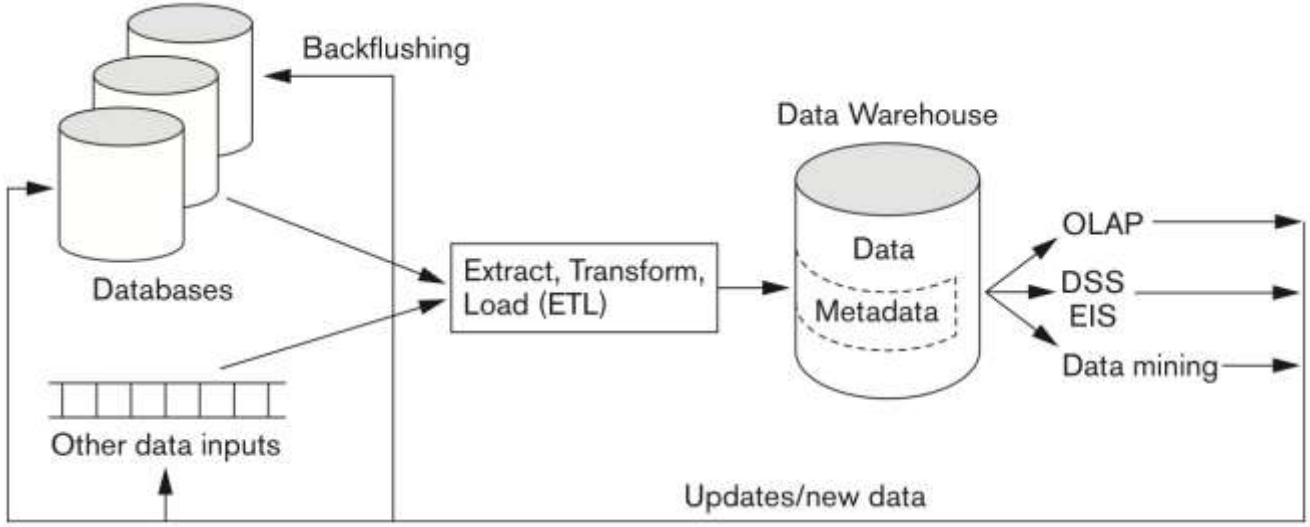


Figure 29.1
Overview of the general architecture of a data warehouse.

Comparison with Traditional Databases

- Data Warehouses are mainly optimized for appropriate data access.
  - Traditional databases are transactional and are optimized for both transaction processing and integrity assurance.
- Data warehouses emphasize more on historical data as their main purpose is to support time-series and trend analysis.
- In transactional databases transaction is the mechanism of change to the database. By contrast, information in data warehouse is relatively coarse grained and DWs are regarded as non-real time. The periodic refresh policy is carefully chosen, usually incremental.
- Compared with transactional databases, data warehouses are nonvolatile.

Characteristics of Data Warehouses

Based on Codd and Salley (1993) article on providing OLAP to users, the following characteristics of Data Warehouses were identified:

- Multidimensional conceptual view
- Unlimited dimensions and aggregation levels
- Unrestricted cross-dimensional operations
- Dynamic sparse matrix handling
- Client-server architecture
- Multiuser support
- Accessibility
- Transparency
- Intuitive data manipulation
- Inductive and deductive analysis
- Flexible distributed reporting

Classification of Data Warehouses

- Generally, Data Warehouses are an order of magnitude larger than the source databases.
- The sheer volume of data is an issue, based on which Data Warehouses could be classified as follows.
  - Enterprise-wide data warehouses
    - They are huge projects requiring massive investment of time and resources.
  - Virtual data warehouses
    - They provide views of operational databases that are materialized for efficient access.
  - Logical Data Warehouses
    - Use data federation, distribution and virtualization techniques
  - Data marts
    - These are generally targeted to a subset of organization, such as a department, and are more tightly focused.

Other Concepts common with Data Warehouses

- ODS: – Operational Data Store : This term is commonly used for intermediate form of databases before they are cleansed, aggregated and transformed into a warehouse.

- ADS – Analytical Data Store: these are databases built for analysis. Typically, ODS's are reconfigured and repurposed into ADS's.

Data Modeling for Data Warehouses (1)

- Traditional Databases generally deal with two-dimensional data (similar to a spread sheet).
  - However, querying performance in a multi-dimensional data storage model (matrices) is much more efficient.
- Data warehouses can take advantage of this feature as generally these are
  - Non volatile
  - The degree of predictability of the analysis that will be performed on them is high.
- Typical Dimensions used in corporate DWs:
  - Fiscal Periods, Product Categories, Geographic Regions

Data Modeling for Data Warehouses (2)

- Example of Two- Dimensional vs. Multi-Dimensional (3D typically called “Data Cube”)

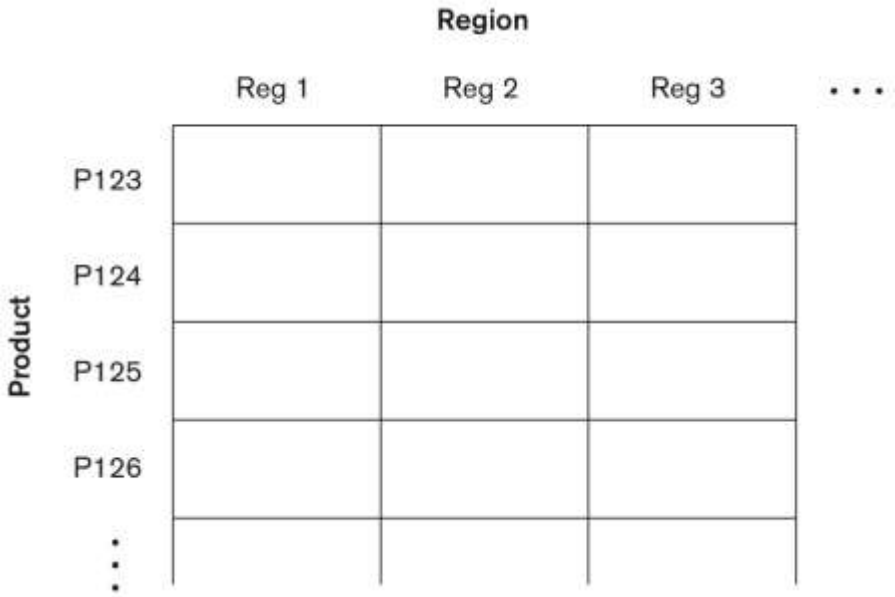


Figure 29.2 A two-dimensional matrix model.

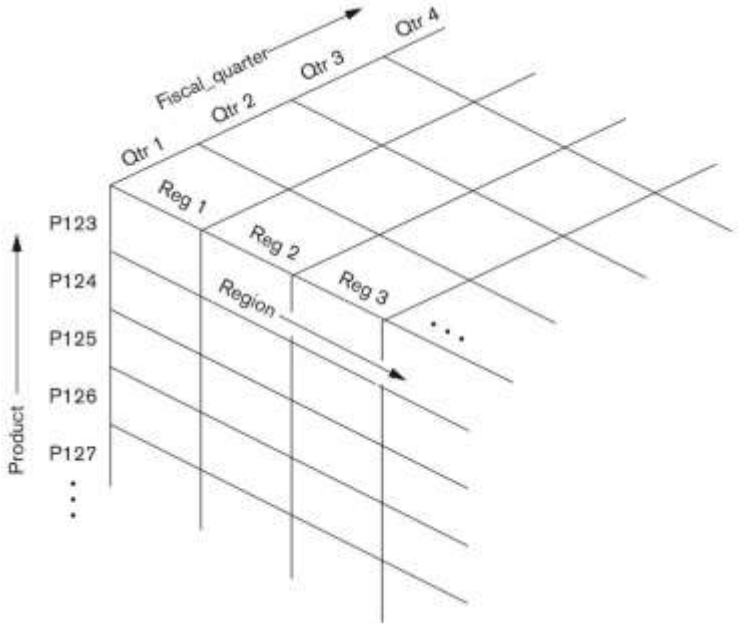


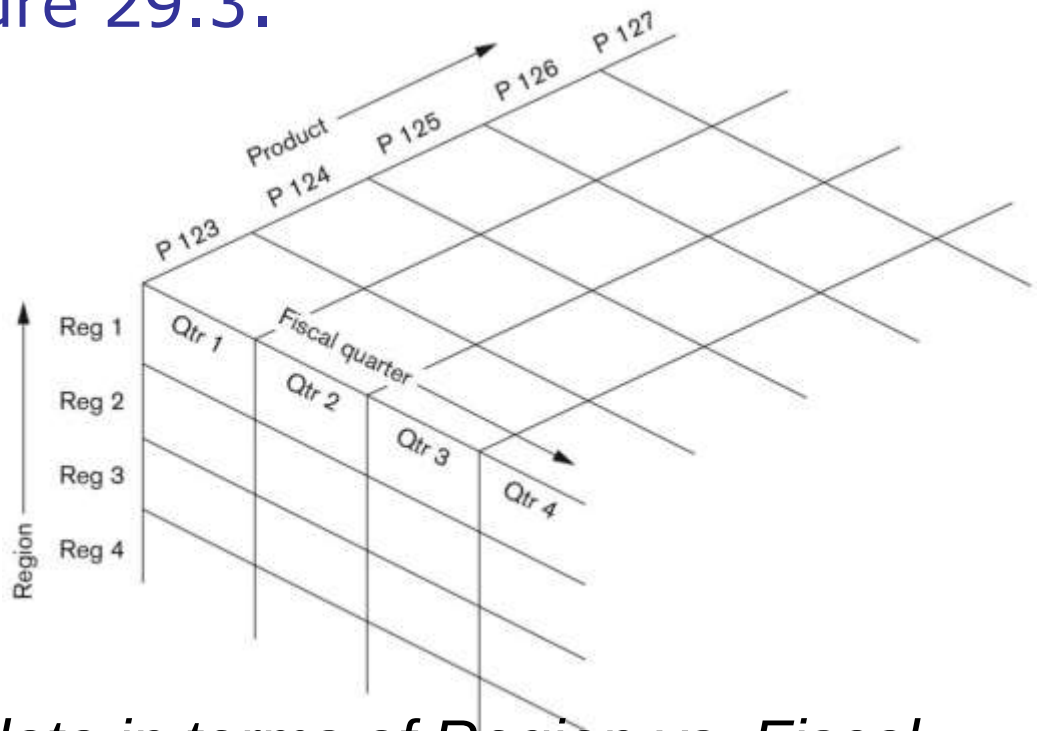
Figure 29.3 A three-dimensional data cube model.

Functionality of a Data Warehouse

- Functionality that can be expected:
  - **Pivot:** Cross tabulation (also referred to as rotation) is performed.
  - **Roll-up (also Drill-up):** Data is summarized with increasing generalization (for example, weekly to quarterly to annually).
  - **Drill-down:** Increasing levels of detail are revealed (the complement of roll-up).
  - **Slice and dice:** Projection operations are performed on the dimensions.
  - **Sorting:** Data is sorted by ordinal value.
  - **Selection:** Data is filtered by value or range.
  - **Derived (computed) attributes:** Attributes are computed by operations on stored and derived values.

The Pivot operation in a Data Warehouse

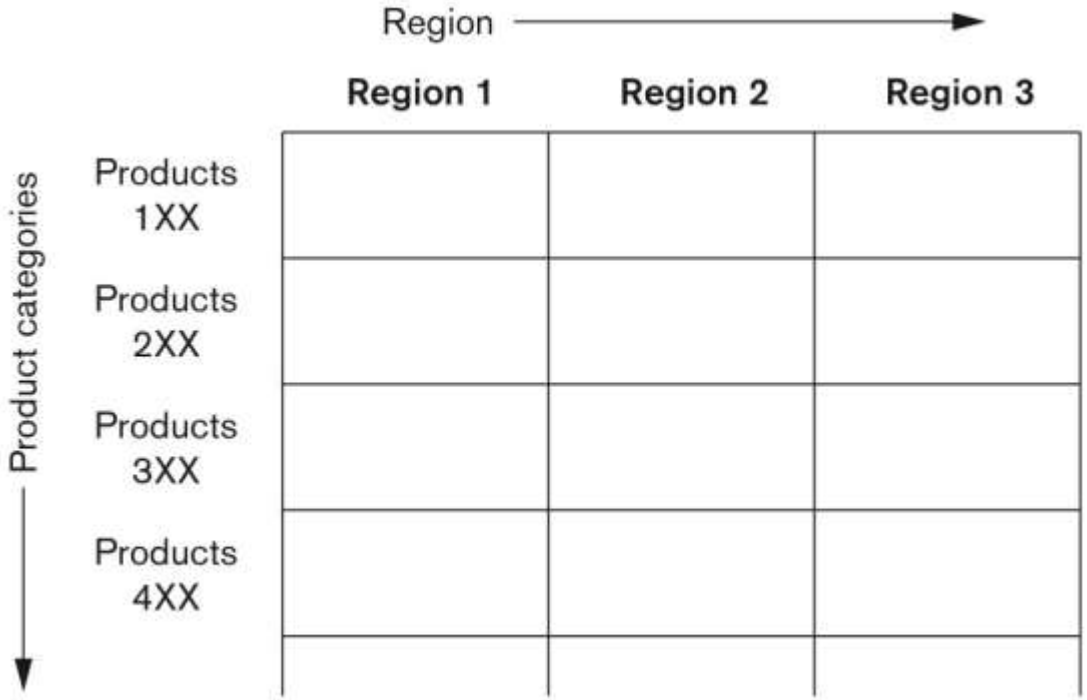
- **Figure 29.4** Pivoted version of the data cube from Figure 29.3.



This presents data in terms of Region vs. Fiscal Quarter : product by product. "Pivoting" is also called as "Rotation"

The “roll-up” operation in a Data Warehouse

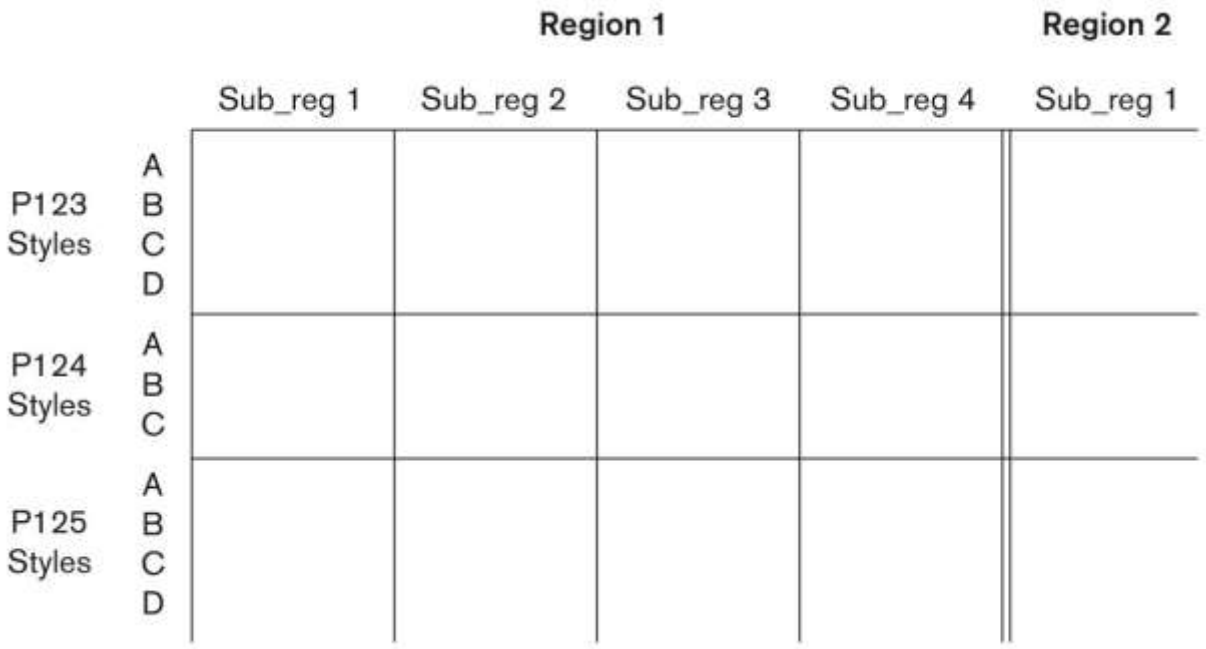
- **Figure 29.5** The roll-up operation.



The roll-up in above figure aggregates data for all products numbered 123, 124, …… into 1XX, etc.

The “drill-down” operation in a Data Warehouse

- **Figure 29.6** The drill-down operation.



A drill-down expands aggregate data into a finer grained view. In this example , products are divided into styles and regions into sub-regions.

Multi-dimensional Schemas (1)

- Multi-dimensional model (also called ”dimensional model”) includes two types of tables:
  - **Dimension table**
    - It consists of tuples of attributes of the dimension.
  - **Fact table**
    - Each tuple is a recorded fact. This fact contains some measured or observed variable (s) and identifies it with pointers to dimension tables. The fact table contains the data, and the dimensions to identify each tuple in the data.
    - A fact table is as an agglomerated view of transaction data whereas each dimension table represents “master data” that those transactions belonged to.

Multi-dimensional Schemas (2)

- Multidimensional DW systems implemented the multidimensional model as is.
- The more popular approach is to implement the multidimensional model on top of the relational model.
- Two common multi-dimensional schemas are
  - **Star schema:**
    - Consists of a fact table with a single table for each dimension
  - **Snowflake Schema:**
    - It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.

Multi-dimensional Schemas (3)

- **Star schema**:
  - Consists of a fact table with a single table for each dimension.

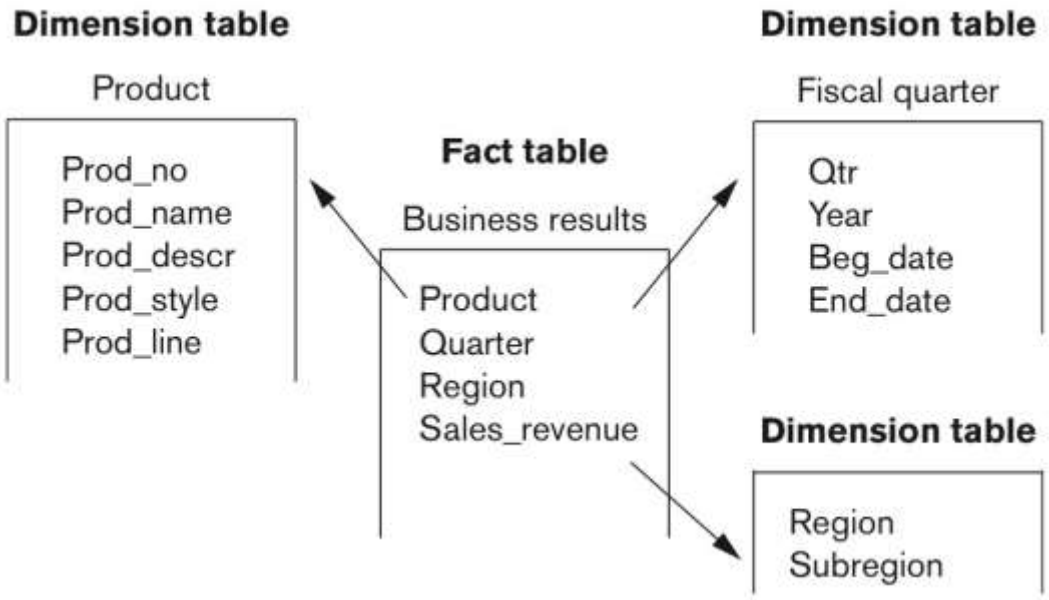


Figure 29.7 A star schema with fact and dimensional tables.

Multi-dimensional Schemas (4)

- **Snowflake Schema:**
  - It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.

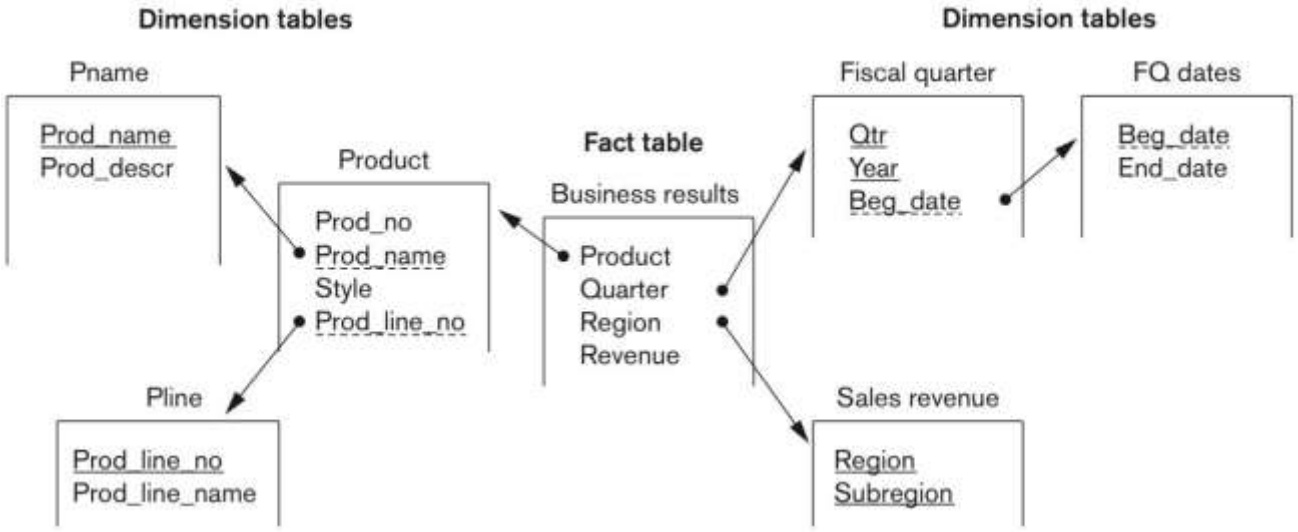


Figure 29.8 A snowflake schema.

Multi-dimensional Schemas (5)

- **Fact Constellation**
  - Fact constellation is a set of tables that share some dimension tables. However, fact constellations limit the possible queries for the warehouse.
  - Example shows the Product dimension table being shared by two Fact tables.

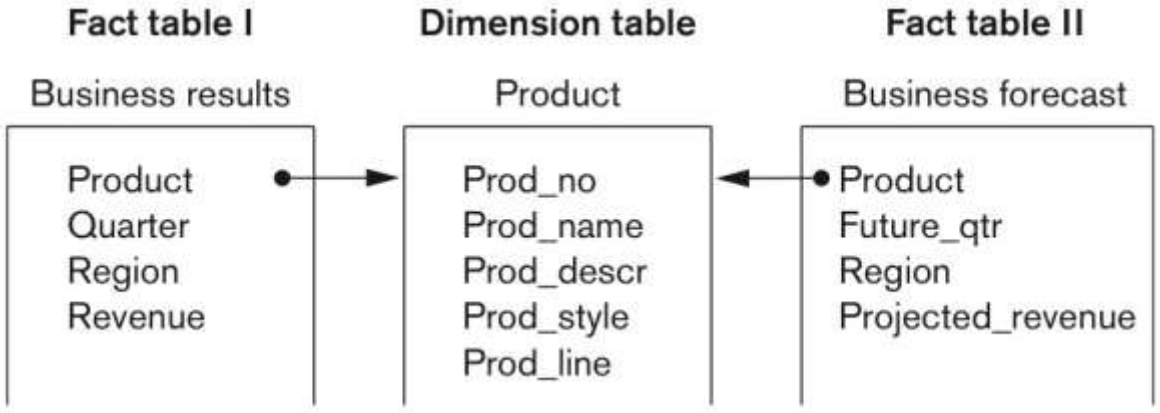


Figure 29.9 A fact constellation.

Multi-dimensional Schemas (6)

- **Indexing**
  - Data warehouse also utilizes indexing to support high performance access.
  - A technique called bitmap indexing constructs a bit vector for each value in the domain being indexed.
  - Indexing works very well for domains of low cardinality. (See example of using a bitmap index in Section 19.8)
- **Master Data Management (MDM)**
  - Purpose of MDM is to define standards, processes, policies and governance issues related to critical data elements entities of the organization

Building A Data Warehouse (1)

- The builders of Data warehouse should take a broad view of the anticipated use of the warehouse.
  - The design should harmonize dimensions across the whole enterprise and multiple data sources
  - The design should support ad-hoc querying
  - An appropriate schema should be chosen that reflects the anticipated usage and the business model of the organization.

Building A Data Warehouse (2)

- The Design of a Data Warehouse involves following steps.
  - Acquisition of data for the warehouse.
  - Ensuring that Data Storage meets the query requirements efficiently.
  - Giving full consideration to the environment in which the data warehouse resides.

Data Acquisition (1)

- Acquisition of data for the warehouse
  - The data must be extracted from multiple, heterogeneous sources.
  - Data must be formatted for consistency within the warehouse.
  - The data must be cleaned to ensure validity.
    - Difficult to automate cleaning process.
    - Back flushing: refers to upgrading the source data by returning cleaned data.

Data Acquisition (2)

- Acquisition of data for the warehouse (contd.)
  - The data must be fitted into the data model of the warehouse. Data may have to be converted from its source model into a multi-dimensional format.
  - The data must be loaded into the warehouse.
    - Proper design for refresh policy should be considered.
    - Data may come from different systems, language areas and time-zones.
    - Order of loading is critical and semantic constraints must be obeyed.

Storing Data in a Data Warehouse

- Storing the data according to the data model of the warehouse
- Creating and maintaining required data structures
- Creating and maintaining appropriate access paths
- Providing for time-variant data as new data are added
- Supporting the updating of warehouse data.
- Refreshing the data
- Purging data

DW Design Considerations

- Usage projections
- The fit of the data model
- Characteristics of available resources
- Design of the metadata component
- Modular component design
- Design for manageability and change
- Considerations of distributed and parallel architecture
  - Distributed DWs: Replication, Partitioning, Communication, Consistency issues
  - Federated DWs : Decentralized federation of autonomous DWs.

Metadata Repositories

- The **metadata repository** is a key data warehouse component. It includes both technical and business metadata.
  - **technical metadata**, covers details of acquisition, processing, storage structures, data descriptions, warehouse operations and maintenance, and access support .
  - **business metadata**, includes the relevant business rules and organizational details supporting the warehouse.

Functionality of Data Warehouses

- We already presented operations of Pivot, Roll-up, Drill-down, etc.
- Other functions include-
  - intersection and union of indexes
  - SQL extensions for aggregation
  - Advanced join techniques for star schemas
- Among OLAP functions, we have:
  - ROLAP : Relational OLAP – keeping DW as a relational DB
  - MOLAP: OLAP on multidimensional model
  - HOLAP: Hybrid OLAP –combines the above two. HOLAP can “drill-through” the data to underlying relations.

Data Warehouse vs. Data Views

- Views and data warehouses are alike in that they both have read-only extracts from the databases.
- However, data warehouses are different from views in the following ways:
  - Data Warehouses exist as persistent storage instead of being materialized on demand.
  - Data Warehouses are not just relational, but rather multi-dimensional with multiple levels of aggregation.
  - Data Warehouses can be indexed for optimal performance. Views cannot be indexed directly.
  - Data Warehouses provide specific support of functionality.
  - Data Warehouses deals with large volumes of integrated data that is contained generally in more than one database.
  - Data warehouses bring in data periodically from multiple sources via a complex ETL process. Views do not go through cleaning and pruning.

Difficulties of implementing Data Warehouses

- Lead time is huge in building a data warehouse
  - Potentially it takes years to build and efficiently maintain a data warehouse.
- Both quality and consistency of data as well as master data management are major concerns.
- Revising the usage projections regularly to meet the current requirements.
  - The data warehouse should be designed to accommodate addition and attrition of data sources without major redesign
- DW schema and acquisition component must handle the changes in data sources in terms of structure and content.
- Administration of data warehouse would require far broader skills than are needed for a traditional database.

Open Issues in Data Warehousing

- Data cleaning, indexing, partitioning, and views could be given new attention with perspective to data warehousing.
- Automation of
  - data acquisition
  - data quality management
  - selection and construction of access paths and structures
  - self-maintainability
  - functionality and performance optimization
- Incorporating of domain and business rules appropriately into the warehouse creation and maintenance process more intelligently.
- Creating appropriate teams of technical experts that also understand the business.

Future of Data Warehousing

- Businesses are getting dissatisfied with the traditional data warehousing techniques and technologies.
- New analytic requirements are driving new analytic appliances such at IBM Netezza, EMC Greenplum, SAP Hana and ParAccel.
- Big data analytics have driven Hadoop and other specialized data bases such as graph and key-value stores into the next generation of data warehousing (see Chapter 25 for Big Data technology based on Hadoop).
- Data virtualization platforms such as the one from Cisco will enable such Logical Data Warehouses to be built in future

See Description of Cisco's Data Virtualization Platform at
http://www.compositesw.com/products-services/data-virtualization-platform/

Future of Data Warehousing

- Gartner reports, a prestigious source of guidance for industry says:

  "the Logical Data Warehouse (LDW) is a new data management architecture for analytics which combines the strengths of traditional repository warehouses with alternative data management and access strategy. The LDW will form a new best practices by the end of 2015."

- The term *"Logical Data Warehouse"* is a clear demarcation between centralized repository approaches and managed data services for analytics.

See: https://www.gartner.com/doc/2057915/understanding-logical-data-warehouse-emerging

Recap

- Purpose of Data Warehousing
- Introduction, Definitions, and Terminology
- Comparison with Traditional Databases
- Characteristics of data Warehouses
- Classification of Data Warehouses
- Data Modeling for Data Warehouses
- Multi-dimensional Schemas
- Building A Data Warehouse
- Functionality of a Data Warehouse
- Warehouse vs. Data Views
- Implementation difficulties, open issues, and future.