Data Ware Housing Concepts

Data Warehousing - Overview

- The term "Data Warehouse" was first coined by Bill Inmon in 1990.
- A data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data.
- A data warehouses provides us generalized and consolidated data in multidimensional view.
- Along with generalized and consolidated view of data, a data warehouses also provides us Online Analytical Processing (OLAP) tools.

Use case scenario..

- An operational database undergoes frequent changes on a daily basis on account of the transactions that take place. Suppose a business executive wants to analyze previous feedback on any data such as a product, a supplier, or any consumer data, then the executive will have no data available to analyze because the previous data has been updated due to transactions.

- A data warehouses provides us generalized and consolidated data in multidimensional view

Understanding a Data Warehouse

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

Why it is Separated from Operational Databases

- An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data.
- Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database.
- An operational database query allows to read and modify operations, while an OLAP query needs only read only access of stored data.
- An operational database maintains current data. On the other hand, a data warehouse maintains historical data.

Data Warehouse Features

- **Subject Oriented - A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.**
- **Integrated - A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.**

Data Warehouse Features (contd.)

- **Time Variant - The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.**
- **Non-volatile - Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database is not reflected in the data warehouse.**

Data Warehouse Applications

As discussed before, a data warehouse helps business executives to organize, analyze, and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess "closed-loop" feedback system for the enterprise management. Data warehouses are widely used in the following fields:

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

Types of Data Warehouse

- 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.
- 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.
- 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.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities:

- Data Extraction - Involves gathering data from multiple heterogeneous sources.
- Data Cleaning - Involves finding and correcting the errors in data.
- Data Transformation - Involves converting the data from legacy format to warehouse format.
- Data Loading - Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- Refreshing - Involves updating from data sources to warehouse.

Data Warehousing - Terminologies

- **Metadata**

Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

In terms of data warehouse, we can define metadata as following:

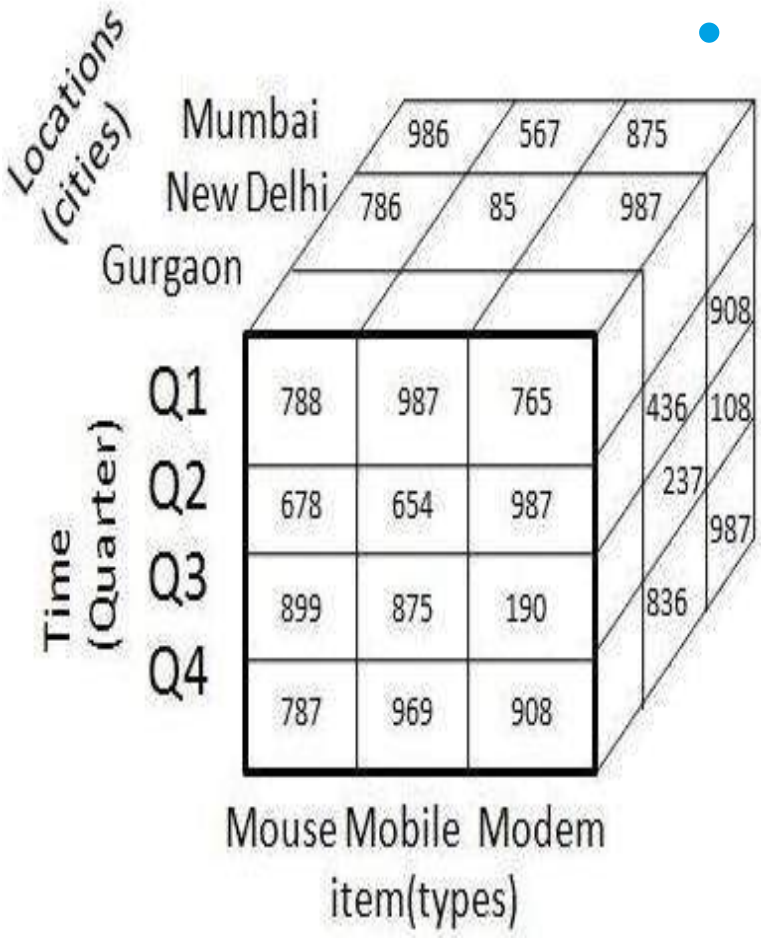
- **Metadata is a road-map to data warehouse.**
- **Metadata in data warehouse defines the warehouse objects.**
- **Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.**

Data Warehousing – Terminologies (contd.)

- **Metadata Repository**

Metadata repository is an integral part of a data warehouse system. It contains the following metadata:

- **Business metadata - It contains the data ownership information, business definition, and changing policies.**
- **Operational metadata - It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.**
- **Data for mapping from operational environment to data warehouse - It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.**
- **The algorithms for summarization - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.**



- A data cube helps us represent data in multiple dimensions. It is defined by dimensions and facts. The dimensions are the entities with respect to which an enterprise preserves the records.

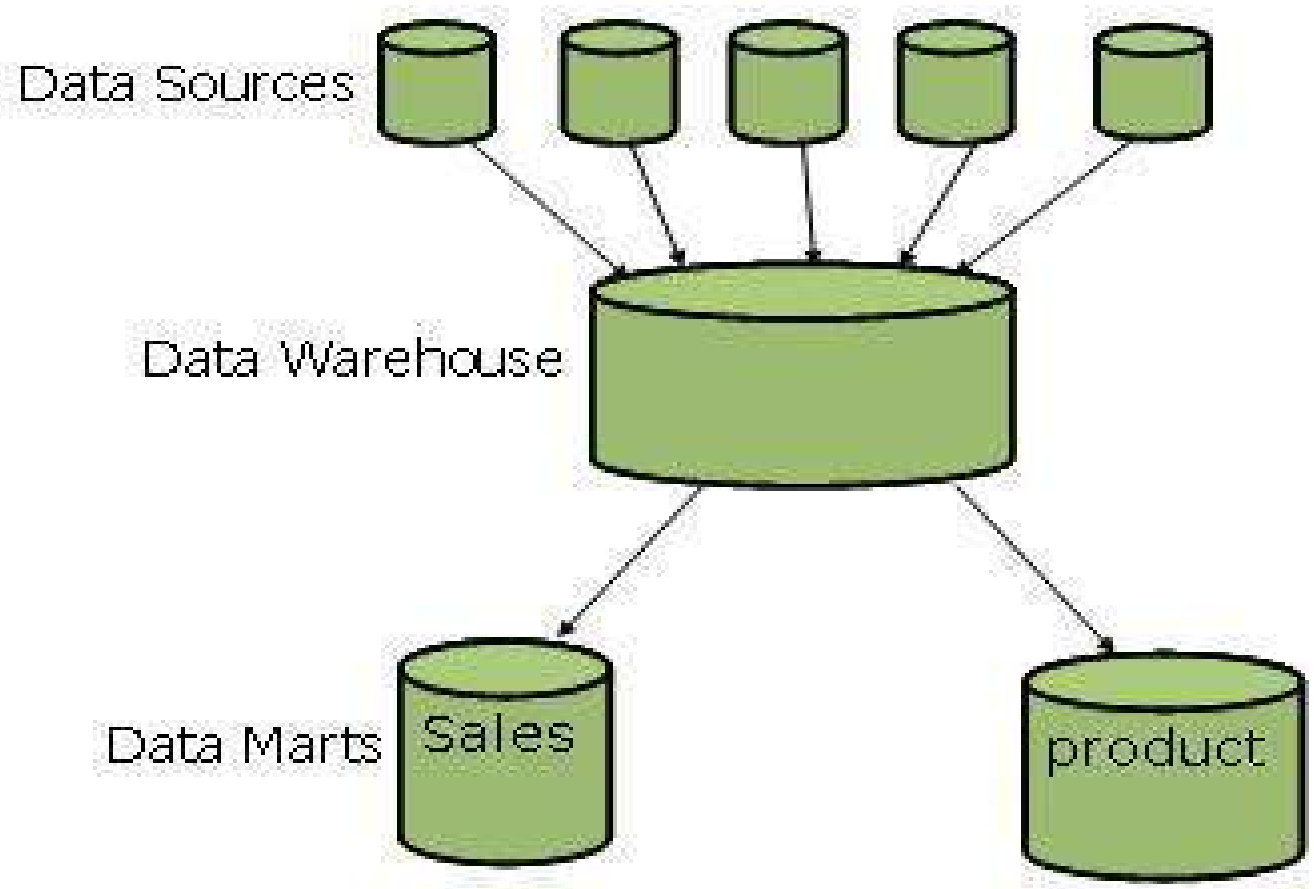
Data Mart

Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization. In other words, a data mart contains only those data that is specific to a particular group. For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.

Points to Remember About Data Marts

- Windows-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation cycle of a data mart is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of data marts may be complex in the long run, if their planning and design are not organization-wide.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data marts are flexible.

Points to Remember About Data Marts (cont.)



Data Warehousing - Architecture

- Business Analysis Framework
- Three-Tier Data Warehouse Architecture

Business Analysis Framework

The business analyst get the information from the data warehouses to measure the performance and make critical adjustments in order to win over other business holders in the market. Having a data warehouse offers the following advantages:

- Since a data warehouse can gather information quickly and efficiently, it can enhance business productivity.
- A data warehouse provides us a consistent view of customers and items, hence, it helps us manage customer relationship.
- A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period in a consistent and reliable manner.

Three-Tier Data Warehouse Architecture

- **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.**
- **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.
- **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 Schema Architecture

Data Warehouse environment usually transforms the relational data model into some special architectures. There are many schema models designed for data warehousing but the most commonly used are:

- Star schema

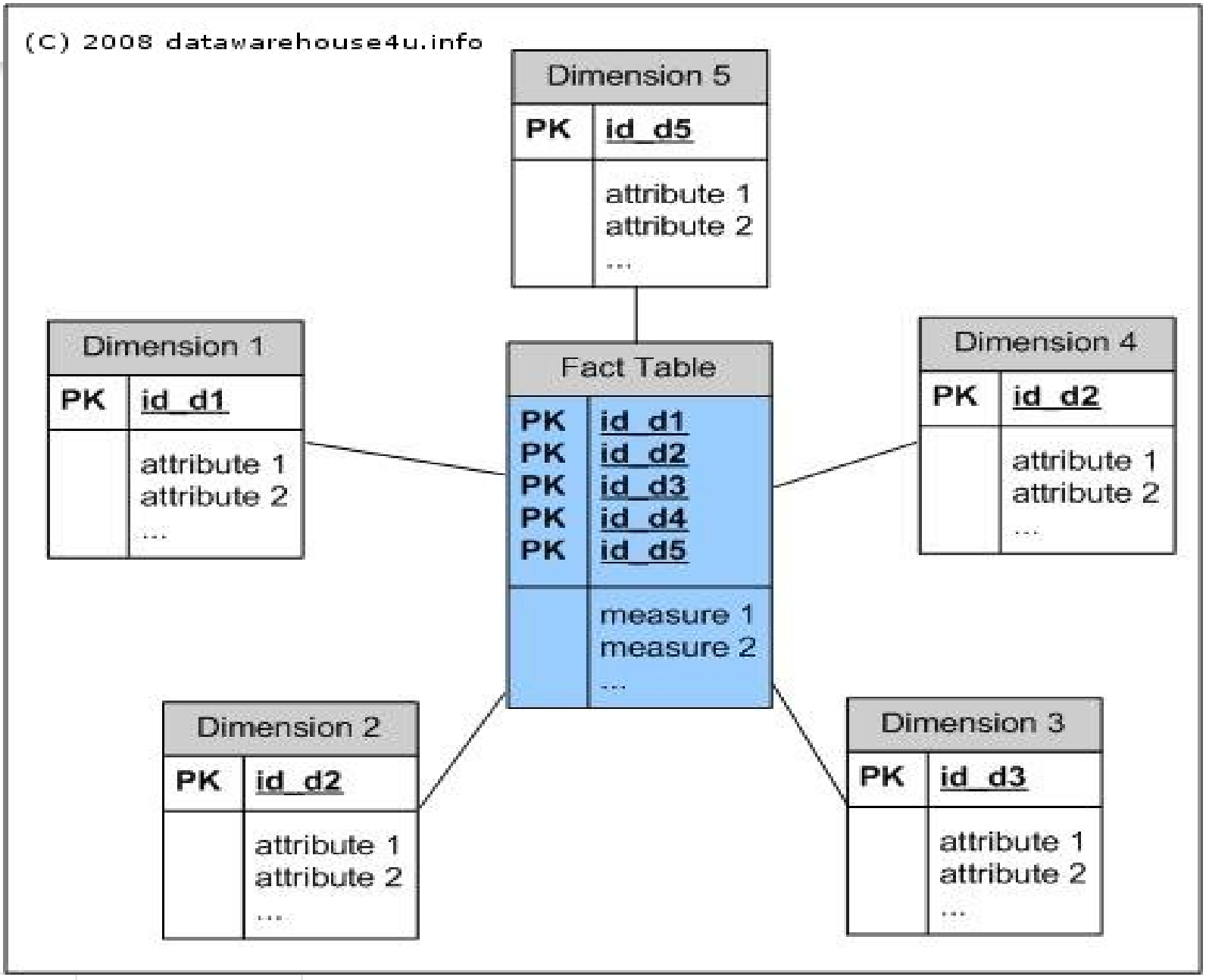
- Snowflake schema

- Fact constellation schema

The determination of which schema model should be used for a data warehouse should be based upon the analysis of project requirements, accessible tools and project team preferences.

Star Schema

- The star schema architecture is the simplest data warehouse schema.
- It is called a star schema because the diagram resembles a star, with points radiating from a center.
- The center of the star consists of fact table and the points of the star are the dimension tables.
- Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized. Despite the fact that the star schema is the simplest architecture, it is most commonly used nowadays and is recommended by Oracle.


(C) 2008 datawarehouse4u.info
Dimension 5
PK
id_d5
attribute 1
attribute 2
...
Dimension 1
PK
id_d1
attribute 1
attribute 2
...
Fact Table
PK
id_d1
PK
id_d2
PK
id_d3
PK
id_d4
PK
id_d5
measure 1
measure 2
...
Dimension 4
PK
id_d2
attribute 1
attribute 2
...
Dimension 2
PK
id_d2
attribute 1
attribute 2
...
Dimension 3
PK
id_d3
attribute 1
attribute 2
...

Fact Tables

Fact tables contain the data corresponding to a particular business process. Each row represents a single event associated with that process and contains the measurement data associated with that event. For example, a retail organization might have fact table related to customer purchases, customer service telephone calls and product returns. The customer purchases table would likely contain information about the amount of the purchase, any discounts applied and the sales tax paid.

Star Schema (cont.)

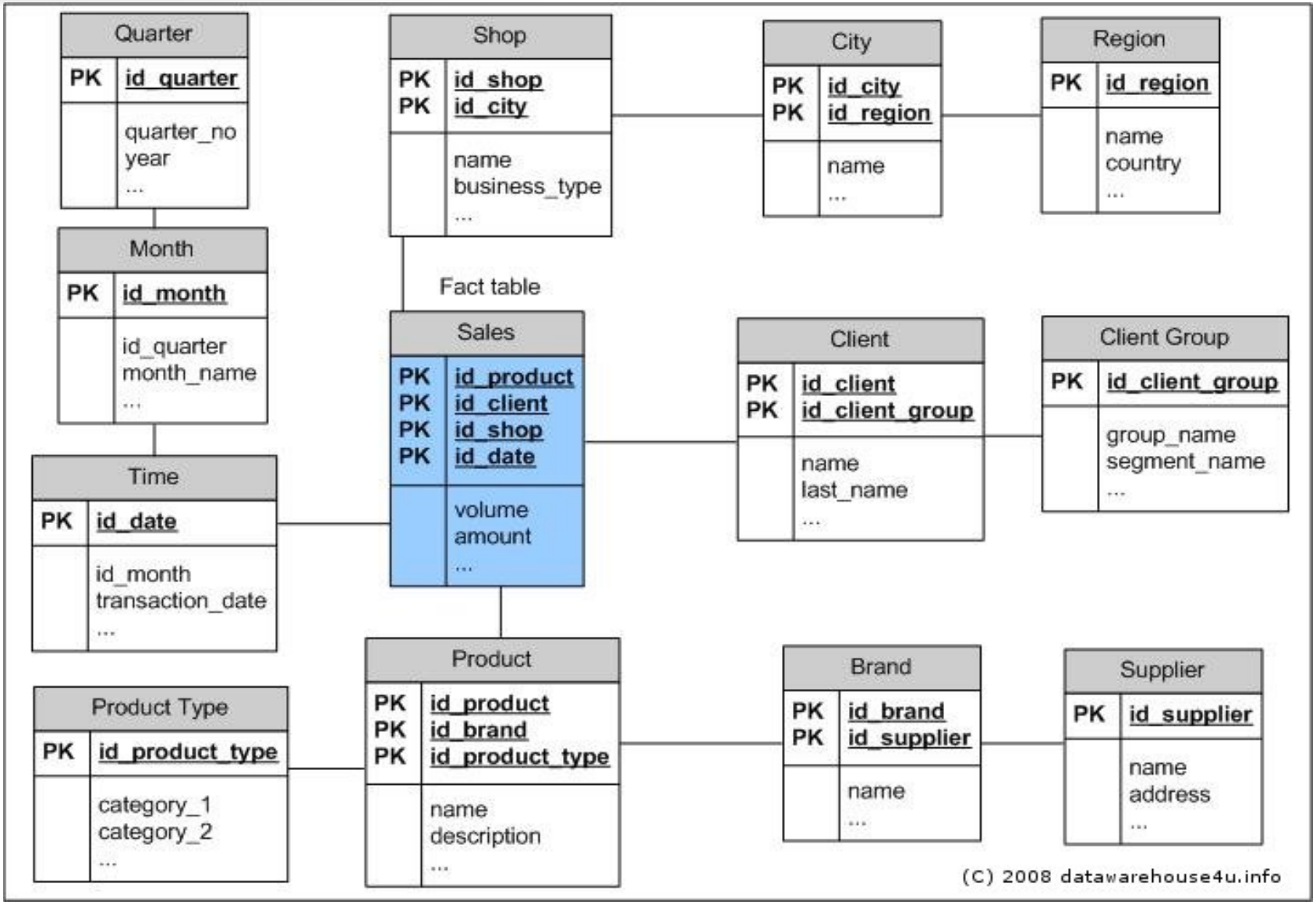
The information contained within a fact table is typically numeric data and it is often data that can be easily manipulated, particularly by summing together many thousands of rows. For example, the retailer described above may wish to pull a profit report for a particular store, product line or customer segment. The retailer can do this by retrieving information from the fact table that relates to those transactions meeting the specific criteria and then adding those rows together.

Dimension Table

Dimensions describe the objects involved in a business intelligence effort. While facts correspond to events, dimensions correspond to people, items, or other objects. For example, in the retail scenario, we discussed that purchases, returns and calls are facts. On the other hand, customers, employees, items and stores are dimensions and should be contained in dimension tables.

Snowflake Schema

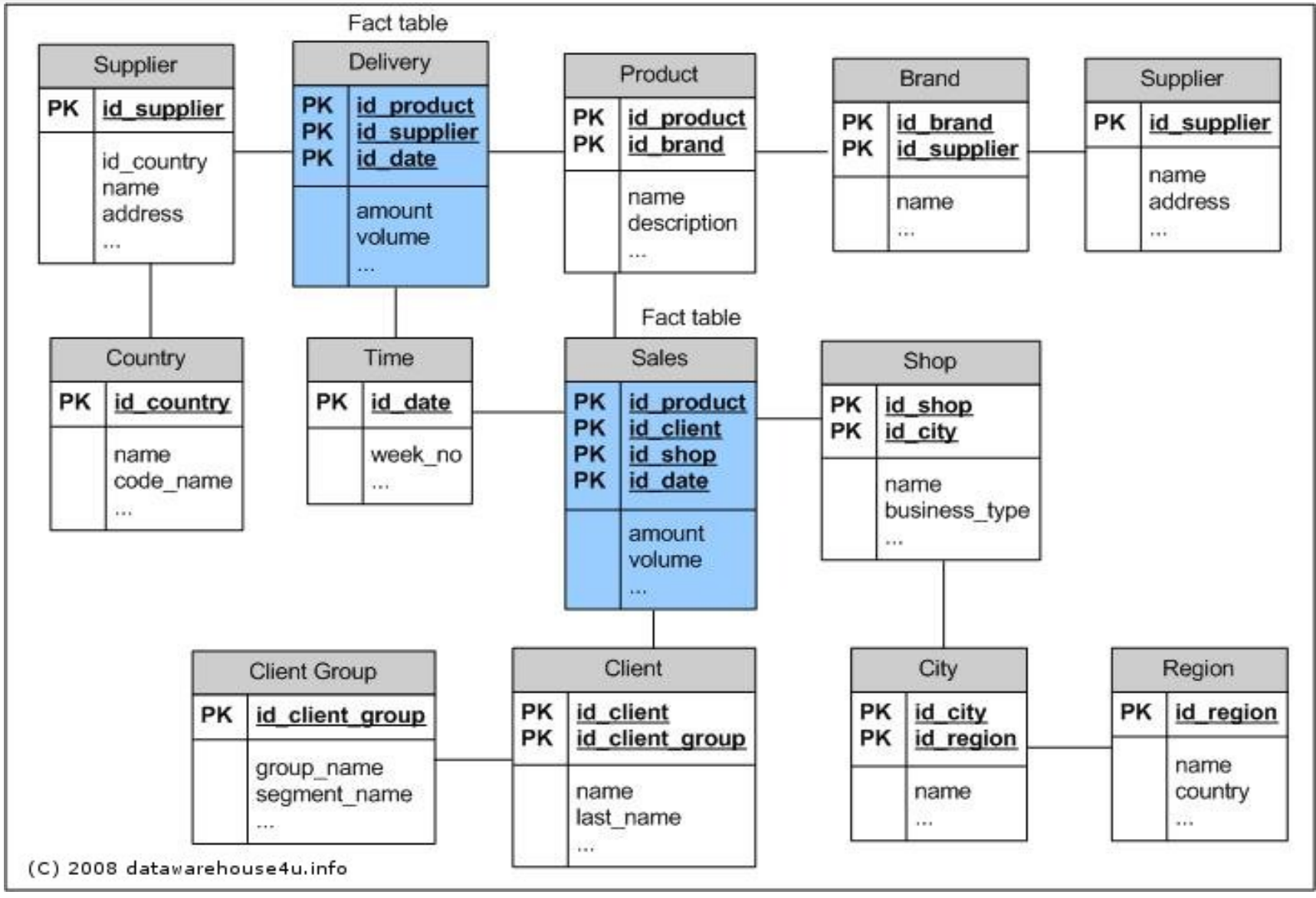
The snowflake schema architecture is a more complex variation of the star schema used in a data warehouse, because the tables which describe the dimensions are normalized.


Quarter
PK id_quarter
quarter_no
year
...
Shop
PK id_shop
PK id_city
name
business_type
...
City
PK id_city
PK id_region
name
...
Region
PK id_region
name
country
...
Month
PK id_month
id_quarter
month_name
...
Fact table
Sales
PK id_product
PK id_client
PK id_shop
PK id_date
volume
amount
...
Client
PK id_client
PK id_client_group
name
last_name
...
Client Group
PK id_client_group
group_name
segment_name
...
Time
PK id_date
id_month
transaction_date
...
Product
PK id_product
PK id_brand
PK id_product_type
name
description
...
Brand
PK id_brand
PK id_supplier
name
...
Supplier
PK id_supplier
name
address
...
Product Type
PK id_product_type
category_1
category_2
...
(C) 2008 datawarehouse4u.info

Fact Constellation Schema

For each Star schema it is possible to construct fact constellation schema(for example by splitting the original Star schema into more star schemes each of them describes facts on another level of dimension hierarchies). The Fact constellation architecture contains multiple Fact tables that share many dimension tables.

The main shortcoming of the fact constellation schema is a more complicated design because many variants for particular kinds of aggregation must be considered and selected. Moreover, dimension tables are still large.


Fact table
Supplier
PK id_supplier
id_country
name
address
...
Delivery
PK id_product
PK id_supplier
PK id_date
amount
volume
...
Product
PK id_product
PK id_brand
name
description
...
Brand
PK id_brand
PK id_supplier
name
...
Supplier
PK id_supplier
name
address
...
Fact table
Country
PK id_country
name
code_name
...
Time
PK id_date
week_no
...
Sales
PK id_product
PK id_client
PK id_shop
PK id_date
amount
volume
...
Shop
PK id_shop
PK id_city
name
business_type
...
Client Group
PK id_client_group
group_name
segment_name
...
Client
PK id_client
PK id_client_group
name
last_name
...
City
PK id_city
PK id_region
name
...
Region
PK id_region
name
country
...
(C) 2008 datawarehouse4u.info