# LECTURE 9

# DATA WAREHOUSE

Muhammad Hamiz Mohd Radzi
Faiqah Hafidzah Halim

# Content

- Introduction to Data Warehouse
- Data Warehouse Architecture
- Data Warehouse Tools and Techniques
- Data Marts
- Data Mining
- Designing a Data Warehousing Database
- Dimensionality Modeling
- Multidimensional Data Representation

# Introduction to Data Warehouse

- A data warehouse is a collection of integrated databases designed to support a Decision Support System.

- A central data repository where data from operational database and other sources are integrated, cleaned, and standardized to support decision making.

- Objective of DW is to evaluate future strategy of a company.

- For example, it increased Capital One’s customers from 1 million to approximately 9 millions in 8 years.

- Just like a muscle: DW increases in strength with active use.
  - *With each new test and product, valuable information is added to the DW, allowing the analyst to learn from the success and failure of the past.*

- The key to survival:
  - *Is the ability to analyze, plan, and react to changing business conditions in a much more rapid fashion.*


Which are our lowest/highest margin customers ?
What is the most effective distribution channel?
Who are my customers and what products are they buying?
What product prom- -otions have the biggest impact on revenue?
Which customers are most likely to go to the competition ?

- Characteristic of Data Warehouse

| | |
|---|---|
| Subject-oriented | Can be used to analyze a particular subject area. |
| Integrated | Integrates data from multiple data sources. |
| Time-variant | Historical data is kept in a data warehouse. |
| Non-volatile | Once data is in the data warehouse, it will not change. |

# DW VS OPERATIONAL DB (OLTP)

| Characteristic | Operational Data Store | Data Warehouse |
|---|---|---|
| How is it built? | One application or subject area at a time. | Typically multiple subject areas at a time |
| Area of support? | Day-to-day business operations. | Decision support for managerial activities. |
| Currency of data? | Up-to-the-minute, real time. | Typically represents a static point in time. |
| Typical unit for analysis? | Small, manageable, transaction level units. | Large, unpredictable, variable units. |
| Design focus? | High-performance, limited flexibility. | High flexibility, high performance. |

# To summarize ...

- OLTP Systems are used to *"run"* a business

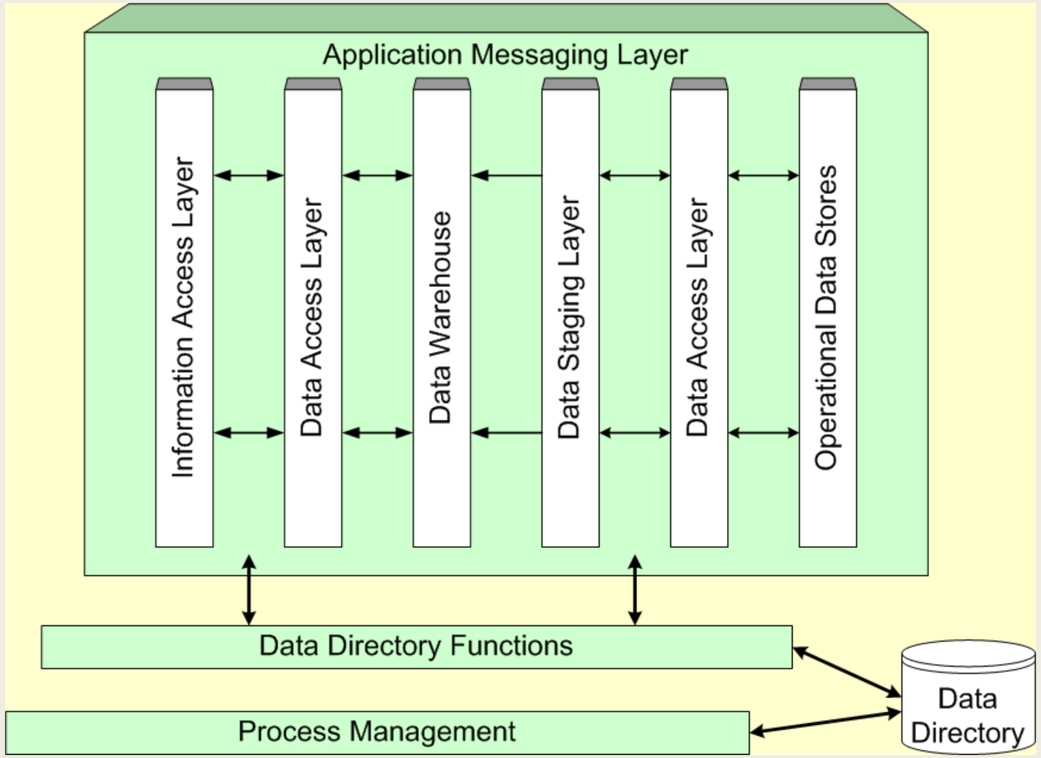
- The Data Warehouse helps to *"optimize"* the business

# Data Warehouse Architecture

The architecture consists of various interconnected elements:

- *Operational and external database layer – the source data for the DW*
- *Information access layer – the tools the end user access to extract and analyze the data*
- *Data access layer – the interface between the operational and information access layers*
- *Metadata layer – the data directory or repository of metadata information*

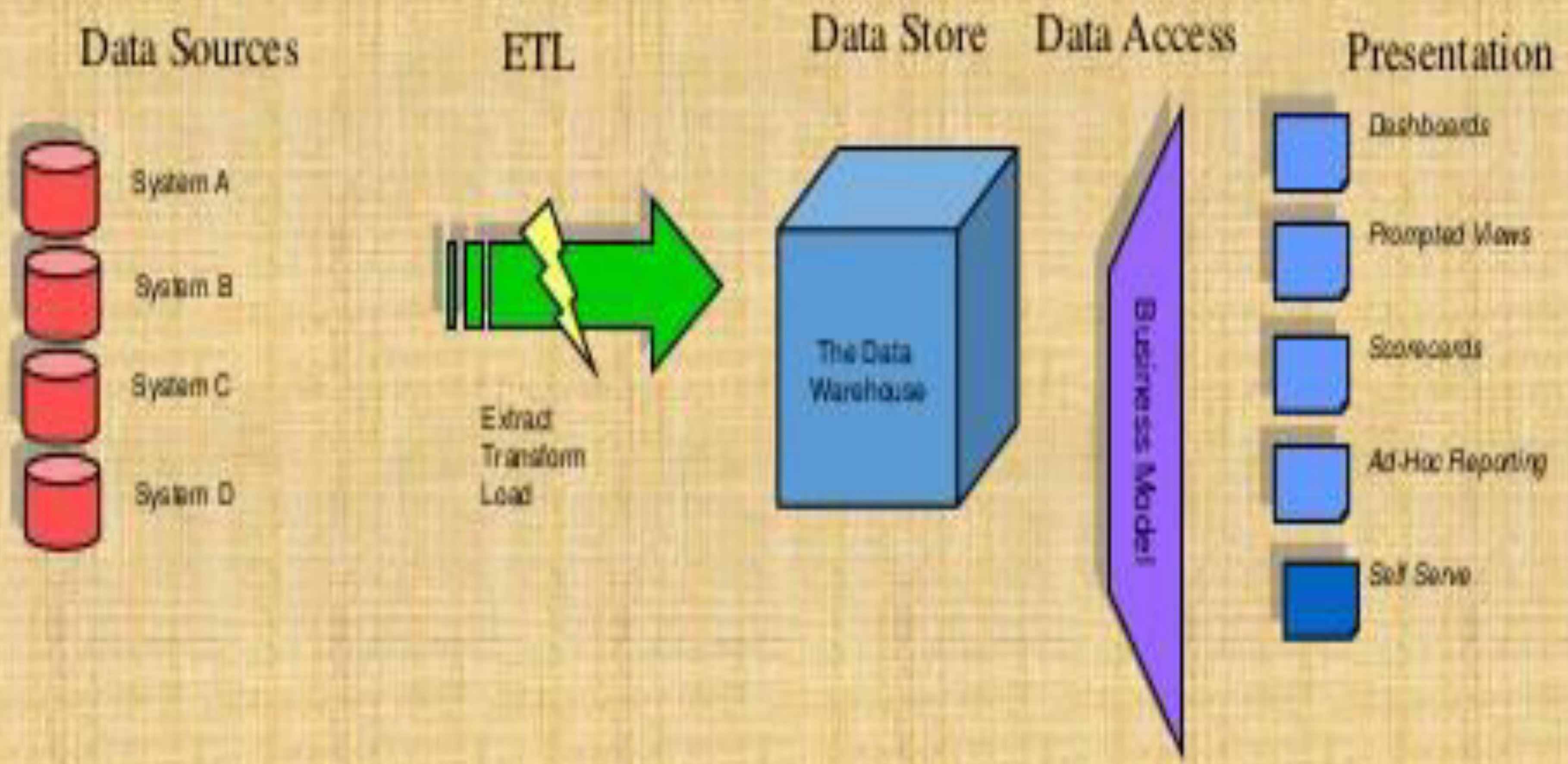
# Components of the Data Warehouse Architecture



Additional layers are:

- *Process management layer – the scheduler or job controller*
- *Application messaging layer – the "middleware" that transports information around the firm*
- *Physical data warehouse layer – where the actual data used in the DSS are located*
- *Data staging layer – all of the processes necessary to select, edit, summarize and load warehouse data from the operational and external data bases*

# Typical DW Architecture



# Data Warehouse Tools and Techniques

- ETL Process
  - *Extract – take the data from operational DB or any reliable sources to be integrated.*
  - *Transform – change the data format such as data merging, calculation and data summarization. But before transform, it must be cleaned to free from any data error.*
  - *Loading – once the data is summarized, it must be loaded into the DW.*

- DW DBMS
  - *Choose the right software to run a data warehouse.*
  - *Requirements needed:*
    - Load Performance – must be fast as the data is huge.
    - Load Processing – load data into a data warehouse must be done periodically and only involve changed data.
    - Data Quality Management – must give user precise result eventhough data is huge.
    - Query Performance – ad hoc analysis must not be slow.
    - Terabyte scalability – size can be scale up when needed.

- *Requirements needed:*
  - Mass user scalability – number of users can be increased without affecting the performance.
  - Networked DW – should be able to cooperating in larger network environment.
  - Warehouse administration – the DBA must be ready all the time to support any operation in the warehouse.
  - Integrated dimensional analysis – data representation can be done from the DW.
  - Advanced Query Functionality – any advance query by the user for calculation and comparative analysis can be done.

- Data Warehouse Metadata:

- For example, a line in a sales database may contain:

    4056  KJ596  223.45

- This is mostly meaningless until we consult the metadata that tells us it was store number 4056, product KJ596 and sales of $223.45

- The metadata are essential ingredients in the transformation of raw data into knowledge. They are the “keys” that allow us to handle the raw data.

General metadata issues associated with Data Warehouse use:

- *What tables, attributes and keys does the DW contain?*
- *Where did each set of data come from?*
- *What transformations were applied with cleansing?*
- *How have the metadata changed over time?*
- *How often do the data get reloaded?*
- *Are there so many data elements that you need to be careful what you ask for?*

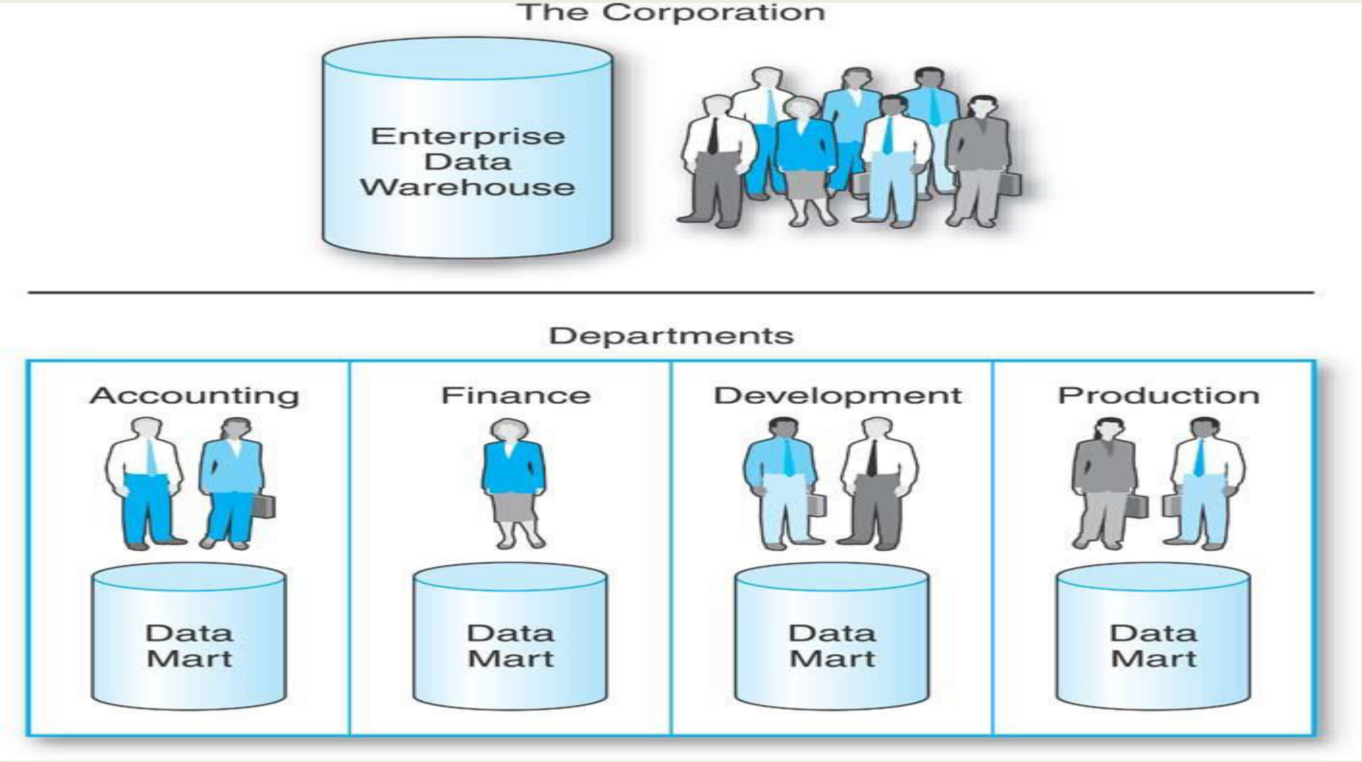
- Component of DW Metadata
  - *Transformation maps – records that show what transformations were applied*
  - *Extraction & relationship history – records that show what data was analyzed*
  - *Algorithms for summarization – methods available for aggregating and summarizing*
  - *Data ownership – records that show origin*
  - *Patterns of access – records that show what data are accessed and how often*

# Example of DW Applications

| Industry | Key Applications |
| --- | --- |
| Airline | Yield management, route assessment |
| Telecommunications | Customer retention, network design |
| Insurance | Risk assessment, product design, fraud detection |
| Retail | Target marketing, supply-chain management |

# Data Marts

- It is a subset or view of a data warehouse
- Typically at a department or functional level.



# Data Mining

- The process of discovering implicit patterns in data and using these patterns for business advantage.
- Facilitates the ability to detect, understand, and predict patterns.
- Example of data mining application is on mobile data usage for telecommunication industry.

# Designing a Data Warehousing Database

Designing a DW is highly complex. These questions must be answered before start designing:

1. Which user requirements are most important?
2. Which data should be considered first?
3. Should the project be scaled down?

Why? Because a failed DW project could lead to company bankruptcy.

# Methodology

- There are 2 different methodologies suggested by Inmon and Kimball.
- Both Kimball and Inmon's architectures share a same common feature that each has a single integrated repository of atomic data.
- In Inmon's architecture, it is called *enterprise data warehouse.* And in Kimball's architecture, it is known as the *dimensional data warehouse*.

- Both architectures have an enterprise focus that supports information analysis across the organization.
- This approach enables to address the business requirements not only within a subject area but also across subject areas.
- However there are some differences in the data warehouse architectures of both experts:

1. Kimball uses the dimensional model such as star schemas or snowflakes to organize the data in *dimensional data warehouse* while Inmon uses ER model in enterprise data warehouse.

Inmon only uses dimensional model for data marts only while Kimball uses it for all data

2. Inmon uses data marts as physical separation from enterprise data warehouse and they are built for departmental uses.

While in Kimball’s architecture, it is unnecessary to separate the data marts from the dimensional data warehouse.

3. In dimensional data warehouse of Kimball, analytic systems can access data directly.

While in Inmon’s architecture, analytic systems can only access data in enterprise data warehouse via data marts.

| Characteristics | Favours Kimball | Favours Inmon |
| --- | --- | --- |
| **Business decision support requirements** | Tactical | Strategic |
| **Data integration requirements** | Individual business requirements | Enterprise-wide integration |
| **The structure of data** | KPI, business performance measures, scorecards… | Data that meet multiple and varied information needs and non-metric data |
| **Persistence of data in source systems** | Source systems are quite stable | Source systems have high rate of change |
| **Skill sets** | Small team of generalists | Bigger team of specialists |
| **Time constraint** | Urgent needs for the first data warehouse | Longer time is allowed to meet business' needs. |
| **Cost to build** | Low start-up cost | High start-up costs |

# Dimensionality Modeling

- It is a logical design technique that aims to represent the data in standard, intuitive form that allows for high-performance access.

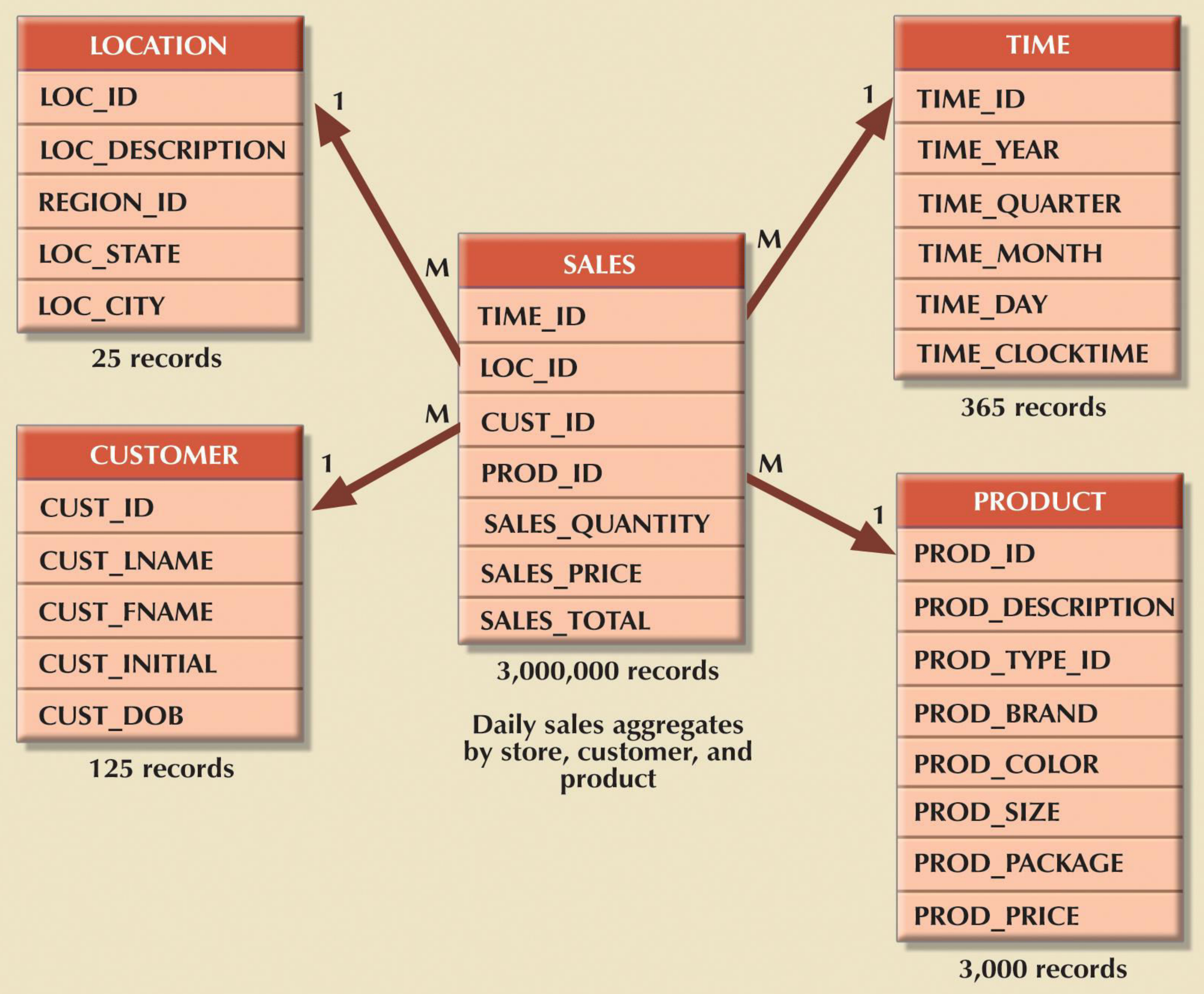
- It will look like ERD, but it can be called as either:
  1. Star Schema
  2. Snowflake Schema
  3. Constellation Schema

- Each schema must contain:
  1. Fact table – numeric table
  2. Dimension table – descriptive table

# Star Schema

- Data modeling representation of multidimensional database
- Star schema diagram looks like star with one large central table is fact table while other is dimension tables
- There is 1-M relationship from each dimension table to facts table

FIGURE 13.17 Star schema for SALES



Product

| Product_Code | Description | Color | Size |
|---|---|---|---|
| 100 | Sweater | Blue | 40 |
| 110 | Shoes | Brown | 10 1/2 |
| 125 | Gloves | Tan | M |
| • • • | | | |

Period

| Period_Code | Year | Quarter | Month |
|---|---|---|---|
| 001 | 2004 | 1 | 4 |
| 002 | 2004 | 1 | 5 |
| 003 | 2004 | 1 | 6 |
| • • • | | | |

Sales

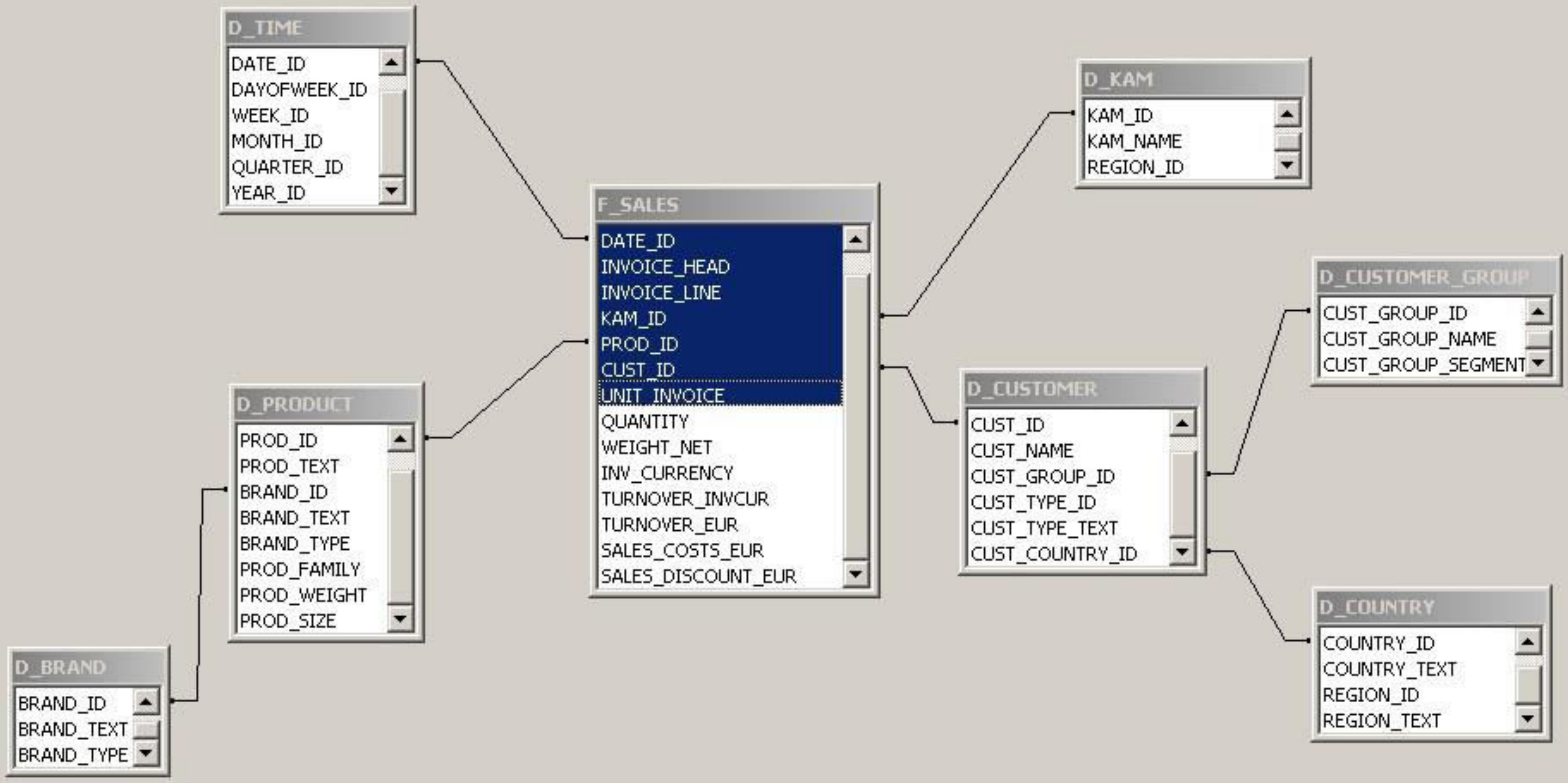
| Product_Code | Period_Code | Store_Code | Units_Sold | Dollars_Sold | Dollars_Cost |
|---|---|---|---|---|---|
| 110 | 002 | S1 | 30 | 1500 | 1200 |
| 125 | 003 | S2 | 50 | 1000 | 600 |
| 100 | 001 | S1 | 40 | 1600 | 1000 |
| 110 | 002 | S3 | 40 | 2000 | 1200 |
| 100 | 003 | S2 | 30 | 1200 | 750 |
| • • • | | | | | |

Store

| Store_Code | Store_Name | City | Telephone | Manager |
|---|---|---|---|---|
| S1 | Jan's | San Antonio | 683-192-1400 | Burgess |
| S2 | Bill's | Portland | 943-681-2135 | Thomas |
| S3 | Ed's | Boulder | 417-196-8037 | Perry |
| • • • | | | | |

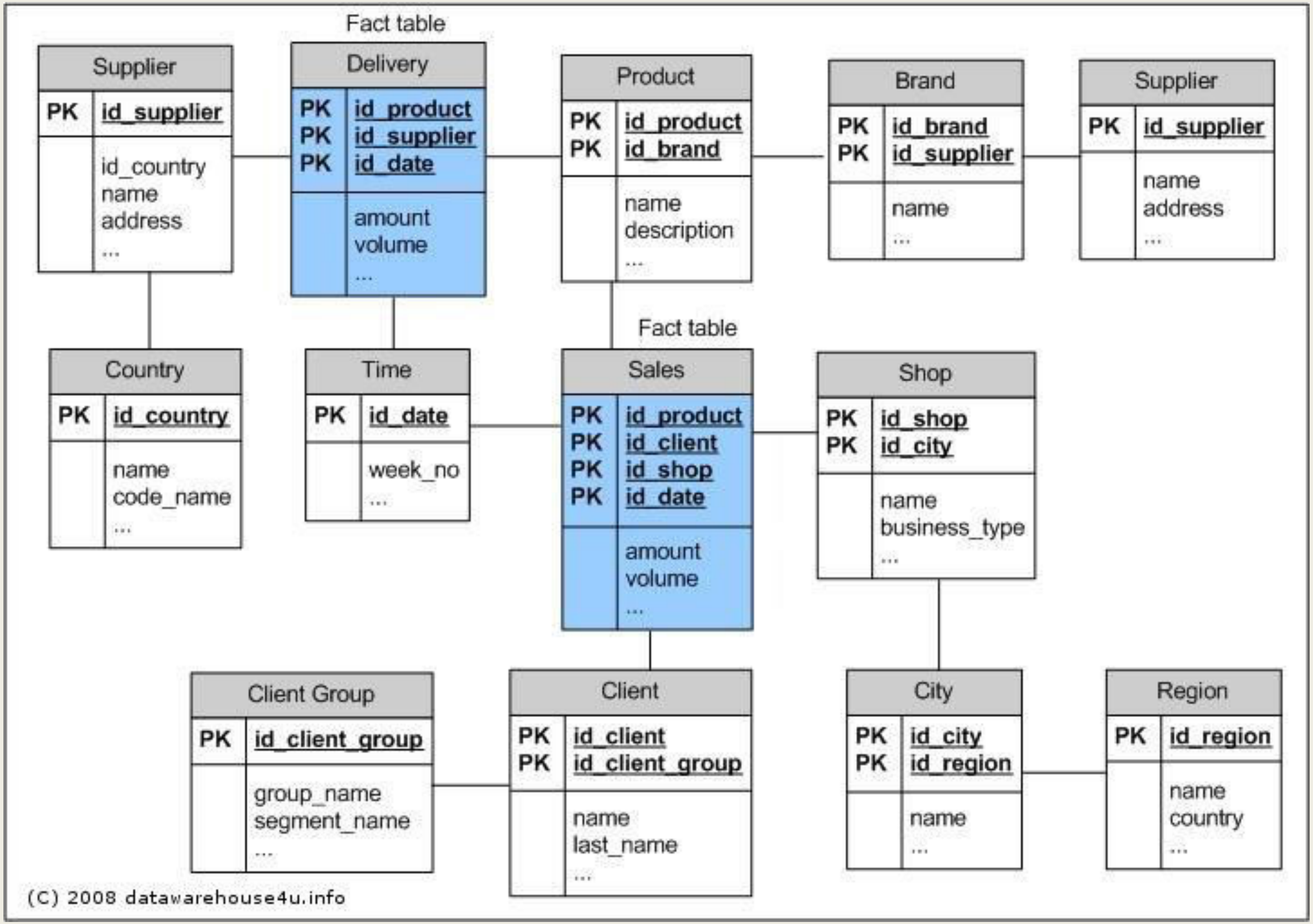
# Snowflake Schema

- Data modeling representation of multidimensional database
- Snowflake schema has multiple levels of dimension tables related to one or more facts tables
- The snowflake schema instead of the star schema for small dimension tables that are not in 3NF
- However, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed


D_TIME
DATE_ID
DAYOFWEEK_ID
WEEK_ID
MONTH_ID
QUARTER_ID
YEAR_ID
F_SALES
DATE_ID
INVOICE_HEAD
INVOICE_LINE
KAM_ID
PROD_ID
CUST_ID
UNIT_INVOICE
QUANTITY
WEIGHT_NET
INV_CURRENCY
TURNOVER_INVCUR
TURNOVER_EUR
SALES_COSTS_EUR
SALES_DISCOUNT_EUR
D_KAM
KAM_ID
KAM_NAME
REGION_ID
D_CUSTOMER_GROUP
CUST_GROUP_ID
CUST_GROUP_NAME
CUST_GROUP_SEGMENT
D_CUSTOMER
CUST_ID
CUST_NAME
CUST_GROUP_ID
CUST_TYPE_ID
CUST_TYPE_TEXT
CUST_COUNTRY_ID
D_COUNTRY
COUNTRY_ID
COUNTRY_TEXT
REGION_ID
REGION_TEXT
D_PRODUCT
PROD_ID
PROD_TEXT
BRAND_ID
BRAND_TEXT
BRAND_TYPE
PROD_FAMILY
PROD_WEIGHT
PROD_SIZE
D_BRAND
BRAND_ID
BRAND_TEXT
BRAND_TYPE

# Constellation Schema

- Data modeling representation of multidimensional database
- A constellation schema contains multiple facts table in the center related to the dimension table
- Typically, the facts table share some dimension tables
- Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation



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

| Year | Region | Agent | Product | Quantity |
|---|---|---|---|---|
| 2009 | East | Carlos | Erasers | 50 |
| 2009 | East | Tere | Erasers | 12 |
| 2009 | North | Carlos | Widgets | 120 |
| 2009 | North | Tere | Widgets | 100 |
| 2009 | North | Carlos | Widgets | 30 |
| 2009 | South | Victor | Balls | 145 |
| 2009 | South | Victor | Balls | 34 |
| 2009 | South | Victor | Balls | 80 |
| 2009 | West | Mary | Pencils | 89 |
| 2009 | West | Mary | Pencils | 56 |
| 2010 | East | Carlos | Pencils | 45 |
| 2010 | East | Victor | Balls | 55 |
| 2010 | North | Mary | Pencils | 60 |
| 2010 | North | Victor | Erasers | 20 |
| 2010 | South | Carlos | Widgets | 30 |
| 2010 | South | Mary | Widgets | 75 |
| 2010 | South | Mary | Widgets | 50 |
| 2010 | South | Tere | Balls | 70 |
| 2010 | South | Tere | Erasers | 90 |
| 2010 | West | Carlos | Widgets | 25 |
| 2010 | West | Tere | Balls | 100 |

# Multidimensional Data Representation

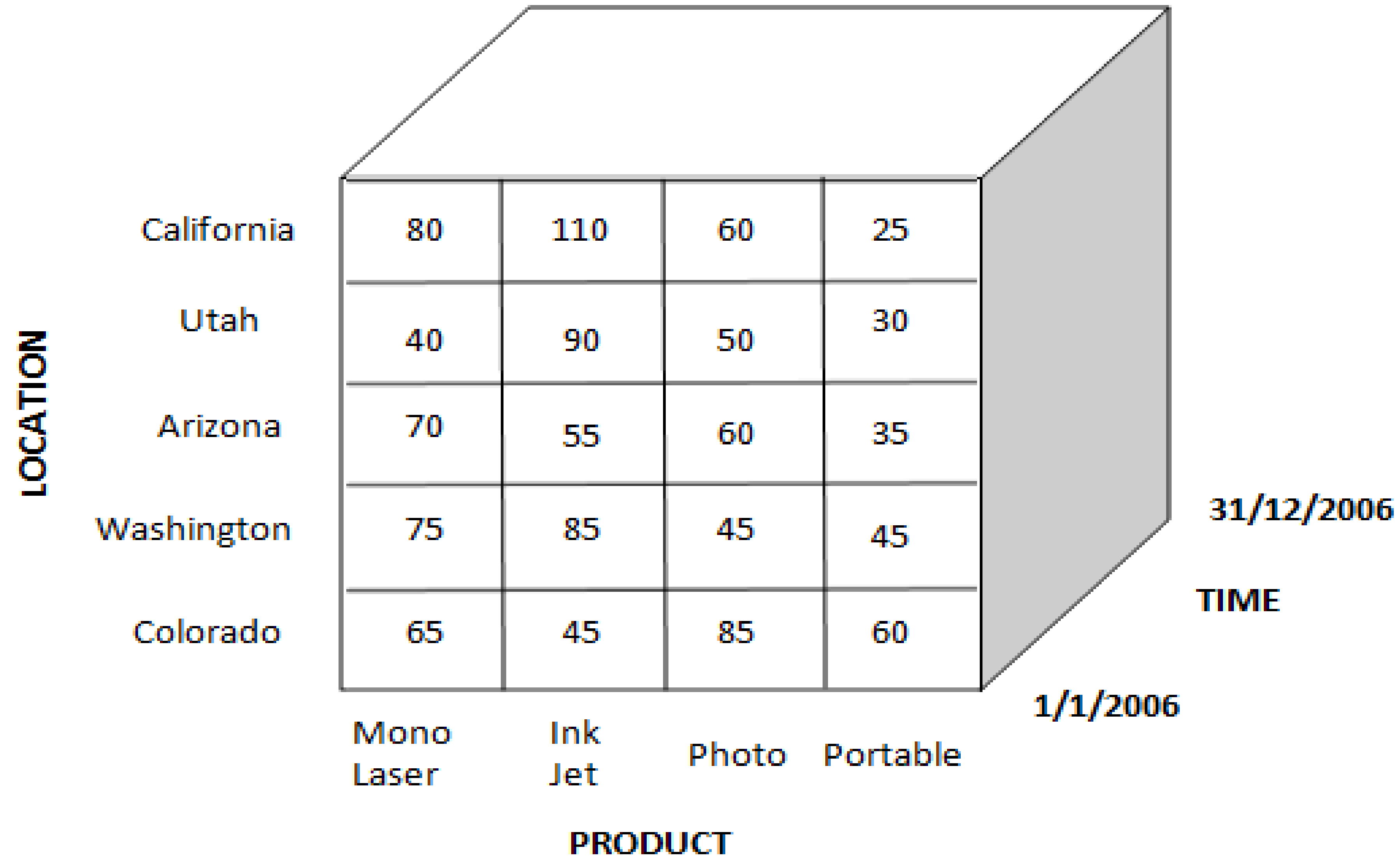
- Multidimensional data model supports data representation and operations for decision support processing in data warehouse

- Users think about decision support data as data cubes

- Data cube or hypercube is multidimensional format consists of cells containing measures (eg sales amount) and dimensions to label numeric data (eg product, location, time). Each dimension contains values known as members (eg location members are California, Utah etc)

| PRODUCT | LOCATION | SALES |
|---|---|---|
| Mono Laser | California | 80 |
| Mono Laser | Utah | 40 |
| Mono Laser | Arizona | 70 |
| Mono Laser | Washington | 75 |
| Mono Laser | Colorado | 65 |
| Ink Jet | California | 110 |
| Ink Jet | Utah | 90 |
| Ink Jet | Arizona | 55 |
| Ink Jet | Washington | 85 |
| Ink Jet | Colorado | 45 |
| Photo | California | 60 |
| Photo | Utah | 50 |
| Photo | Arizona | 60 |
| Photo | Washington | 45 |
| Photo | Colorado | 85 |
| Portable | California | 25 |
| Portable | Utah | 30 |
| Portable | Arizona | 35 |
| Portable | Washington | 45 |
| Portable | Colorado | 60 |

| | Product | | | |
|---|---|---|---|---|
| Location | Mono Laser | Ink Jet | Photo | Portable |
| California | 80 | 110 | 60 | 25 |
| Utah | 40 | 90 | 50 | 30 |
| Arizon | 70 | 55 | 60 | 35 |
| Washington | 75 | 85 | 45 | 45 |
| Colorado | 65 | 45 | 85 | 60 |


LOCATION
California
Utah
Arizona
Washington
Colorado
80
110
60
25
40
90
50
30
70
55
60
35
75
85
45
45
65
45
85
60
Mono Laser
Ink Jet
Photo
Portable
PRODUCT
31/12/2006
TIME
1/1/2006

*Multidimensional Terminology:*

- *Dimension: subject label for a row or column*
- *Member: value of dimension*
- *Measure: quantitative data stored in cells*
- *Sparsity: - large empty cells in a data cube.*
  *- Waste space and be slow to process.*

We have a multidimensional data model with the fact table Sales and the dimension tables Customers, Products, and Salespeople. The sales table below represents sales data for 1st January 2010:

| CUSTID | PRODID | STAFFID | QUANTITY |
|---|---|---|---|
| 101 | 1 | 10 | 20 |
| 101 | 2 | 11 | 10 |
| 101 | 3 | 10 | 5 |
| 102 | 2 | 11 | 8 |
| 102 | 3 | 10 | 24 |
| 103 | 1 | 11 | 50 |
| 103 | 2 | 11 | 45 |
| 103 | 3 | 10 | 30 |

- Draw a 3D picture of a data cube. Assume that all values that are missing from the Sales table are 0.

# OLAP Servers

- Online Analytical Processing Server (OLAP) is based on the multidimensional data model.
- It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.
- This part will cover the types of OLAP, operations on OLAP, difference between OLAP, and statistical databases and OLTP.

- There are four types of OLAP servers:
  - *Relational OLAP (ROLAP)*
  - *Multidimensional OLAP (MOLAP)*
  - *Hybrid OLAP (HOLAP)*
  - *Specialized SQL Servers (not covered)*

# ROLAP (STAR SCHEMA)

- Relational OLAP (ROLAP) servers are placed between relational back-end server and client front-end tools.

- To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

- ROLAP includes the following:
  - *Implementation of aggregation navigation logic.*
  - *Optimization for each DBMS back end.*
  - *Additional tools and services.*

# MOLAP (DATA CUBE)

- MOLAP uses array-based multidimensional storage engines for multidimensional views of data.

- With multidimensional data stores, the storage utilization may be low if the data set is sparse.

- Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

# HOLAP

- Hybrid OLAP is a combination of both ROLAP and MOLAP.
- It offers higher scalability of ROLAP and faster computation of MOLAP.
- HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

# OLAP OPERATION

| Operator | Purpose | Description |
|---|---|---|
| Slice | Focus attention on a subset of dimensions | Replace a dimension with a single member value or with a summary of its measure values |
| Dice | Focus attention on a subset of member values | Replace a dimension with a subset of members |
| Drill-down | Obtain more detail about a dimension | Navigate from a more general level to a more specific level |
| Roll-up | Summarize details about a dimension | Navigate from a more specific level to a more general level |
| Pivot | Present data in a different order | Rearrange the dimensions in a data cube |

# SLICE OPERATOR

- Focus on a subset of dimensions
- Similar to restriction operator
- Dimension are set to specific value
- Set dimension to specific value:  1/1/2006

# Example Slice Operation

| | Product | | | |
|---|---|---|---|---|
| **Location** | **Mono Laser** | **Ink Jet** | **Photo** | **Portable** |
| California | 80 | 110 | 60 | 25 |
| Utah | 40 | 90 | 50 | 30 |
| Arizon | 70 | 55 | 60 | 35 |
| Washington | 75 | 85 | 45 | 45 |
| Colorado | 65 | 45 | 85 | 60 |

# Example Slice-Summarize Operation

| | Time | | | |
|---|---|---|---|---|
| **Location** | **1/1/2006** | **1/2/2006** | | **Total Sales** |
| California | 80 | 110 | ... | 16,250 |
| Utah | 40 | 90 | ... | 11,107 |
| Arizon | 70 | 55 | ... | 21,500 |
| Washington | 75 | 85 | ... | 20,900 |
| Colorado | 65 | 45 | ... | 21,336 |

# DICE OPERATOR

- Focus on a subset of member values
- Replace dimension with a subset of values
- Dice operation often follows a slice operation

| Location | Mono Laser | Ink Jet | Photo | Portable |
|---|---|---|---|---|
| Utah | 40 | 90 | 50 | 30 |

Product

# DRILL-DOWN

- navigate from a more general level to more specific.
- Obtain more detail about dimension.

| | Product | | | |
|---|---|---|---|---|
| **Location** | **Mono Laser** | **Ink Jet** | **Photo** | **Portable** |
| California | 80 | 110 | 60 | 25 |
| Utah | | | | |
| Salt Lake | 20 | 20 | 10 | 15 |
| Park City | 5 | 30 | 10 | 5 |
| Ogden | 15 | 40 | 30 | 10 |
| Arizon | 70 | 55 | 60 | 35 |
| Washington | 75 | 85 | 45 | 45 |
| Colorado | 65 | 45 | 85 | 60 |

# ROLL-UP/DRILL-UP

- Remove detail from a dimension.

- Moving from a specific level to a more general level of a hierarchical dimension.

- eg:  roll up sales data from daily to quarterly level.

# PIVOT

- Rearrange dimensions so that data cube can be presented in a visually appealing order.
- Most typically used on data cube of more than two dimensions.

- Introduction to Data Warehouse
- Data Warehouse Architecture
- Data Warehouse Tools and Techniques
- Data Marts
- Data Mining
- Designing a Data Warehousing Database
- Dimensionality Modeling
- Multidimensional Data Representation

# References

*Database Systems: A Practical Approach to Design, Implementation, and Management,* Thomas Connolly and Carolyn Begg, 5th Edition, 2010, Pearson.