Data Warehousing

Grigorios Loukides
Email: grigorios.loukides@kcl.ac.uk

Session Objectives

In this session, you will learn:

- Purpose of Data Warehousing
- Introduction, Definitions, and Terminology
- Comparison with Traditional Databases
- Characteristics and Classification of Data Warehouses
- Multi-dimensional Schemas
- Building a Data Warehouse

Purpose of Data Warehousing

- Traditional databases are not optimized for data access
  - They ensure integrity of data
- Data warehouse users need only read access but, need the access to be fast over a large volume of data
- Data required for data warehouse analysis comes from multiple databases
- There is a great need for tools that provide decision makers with information to make decisions quickly and reliably based on historical data

Data Warehousing & OLAP

The above functionality is achieved by Data Warehousing and Online Analytical Processing (OLAP)

Data Warehouse

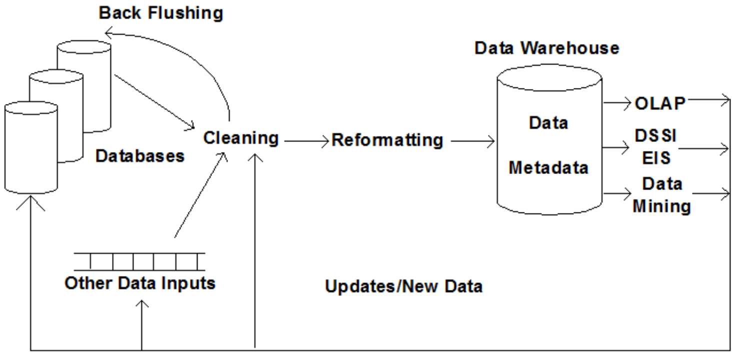
Data Warehouse (DW)

A subject-oriented, integrated, nonvolatile, time-variant collection of data in **support of management's decisions**

OLAP

OLAP (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse

Structure of a Data Warehouse



Data warehouse

Data warehouse is a collection of data:

- Subject-Oriented: organized around the major subjects of the enterprise (e.g., customers, products, sales) rather than the major application areas
- Integrated: integrates application-oriented data (OLTP) from different source systems, which often includes data that is inconsistent (e.g. in different formats).
- Time-variant: historical data is kept in a data warehouse
- Non-volatile: data is not real-time updated but is refreshed on a regular basis. New data is always added as a supplement to the database (contains historical data)

Exercise: Comparison among DW and OLTP Systems[1]

	OLTP Systems	DW Systems
Holds historical data		
Data is dynamic		
Transaction-driven		
Subject-oriented		
High-level of transactions		
Supports Strategic Decisions		
Ad hoc processing		

[1] On-line transaction processing systems

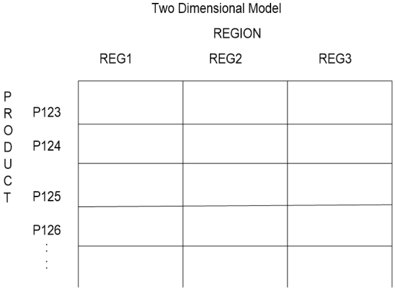
Data Mart

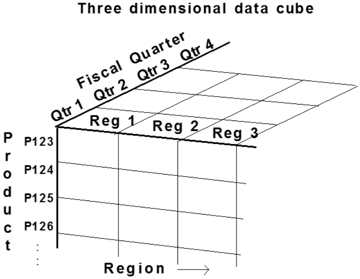
- A Data Mart (DM) is a subset of a DW that supports the requirements of a particular department or business function (e.g., Sales dept. instead of across an enterprise)
- DMs contain less data compared to DW
- Approaches for building data marts:
  - Build several DMs with a view to the eventual integration into a DW
  - Build the infrastructure for the DW while at the same time building one or more DMs to satisfy immediate business needs

Data Modelling for Data Warehouses

- Traditional Databases generally deal with two-dimensional data (tables with rows and columns):
  - This is good for avoiding inconsistencies with frequent updates
  - However, querying performance in a multi-dimensional data storage model is much more efficient
- Data warehouses can take advantage of this feature as generally these are non-volatile

Example of Two- Dimensional vs. Multi-Dimensional





Advantages of a multi-dimensional mode

- Multi-dimensional models lend themselves readily to hierarchical views in what is known as roll-up display and drill-down display.
- The data can be directly queried in any combination of dimensions, bypassing complex database queries

Dimensionality modelling

- A logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access
- Uses the concepts of Entity-Relationship modelling with some important restrictions

Dimensionality modelling: Tables

Every dimensional model is composed of:

- One table with a composite primary key, called the fact table
- A set of smaller tables called dimension tables

Introduction

DW

Data Modelling for DW

Concepts

Methodology

Example

Building a DW

Acquisition of data for the warehouse

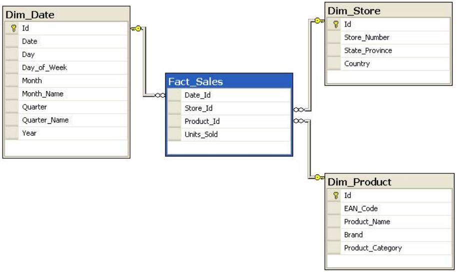
Homework: Data warehouses and SQL

Tutorial Exercises

Conclusion

Suggested

Dimensionality modelling: Keys



- Each dimension table has a simple (non-composite) primary key
- The fact table has a composite key that made up by the primary keys of dimension tables
- Forms 'star-like' structure, which is called a star schema

Star schema

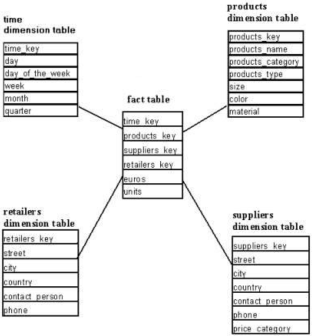
Star schema:

- Logical structure that has a fact table containing factual data in the center, surrounded by dimension tables containing reference data
- Facts are generated by events that occurred in the past, and are unlikely to change

Content of tables

- A fact table usually contains one or more numerical measures, or 'fact' that occur for each record and are numeric and additive
- Dimension tables usually contain descriptive textual information
  - Dimension attributes are used as the constraints in data warehouse queries

Star schema: Denormalization

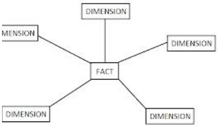


Star schemas can be used to speed up query performance by denormalizing reference information into a single dimension table.

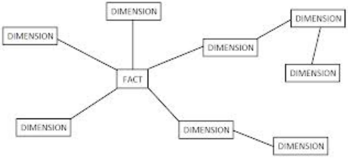
Also, multiple dimension tables may contain the same information (e.g., country)

Snowflake schema

Snowflake schema is a variant of the star schema where dimension tables do not contain denormalized data



Star



Snow

Comparison Star & Snowflake Schemas

Star schemas

- They can be used to speed up query performance by denormalizing reference information into a single dimension table
- Denormalization is appropriate when there are a number of entities related to the dimension table that are accessed often, avoiding the overhead of having to join additional tables

Snowflake schemas

- When the denormalization is not appropriate we create a snowflake schema
- Denormalization is not appropriate when the additional data is not accessed very often, because the overhead of scanning the expanded dimension table may not be compensated by any gain on the query performance

DW Design Methodology

Four-Ordered-Step Methodology includes:

1. Select the business process to model
2. Choosing the grain of the business process
3. Identifying and conforming the dimensions
4. Choosing the facts (Storing pre-calculations in the fact table)

Select the business process to model

- The process (function) refers to the subject matter of a particular data mart
- First DM built should be the one that is most likely to be delivered on time, within budget, and to answer the most commercially important business questions

(e.g., do we select property sales, or property maintenance?)

Choosing the grain of the business process

- Declaring the grain means specifying what an individual fact table row represents
- *How you describe a single row in the fact table?*
- The recommendation is to build the dimensional model using the lower level of detail available

(e.g., For a fact table *PropertySale* the grain is attributes representing an individual property sales)

Identifying and conforming the dimensions

- *How do business people describe the data that results from the business process?*
- Dimensions set the context for asking questions about the facts in the fact table

(e.g. to build an understandable dimensional model, we can select attributes that describe clients performing sales in the fact table)

Choosing the facts (Storing pre-calculations in the fact table)

- The grain of the fact table determines which facts can be used in the DM
- Facts should be numeric and additive
  (e.g., totalRent for an individual property salve)
- Unusable facts include:
  - non-numeric facts
  - non-additive facts (cannot be summed up for any of the dimensions present in the fact table)
  - fact at different granularity from other facts in table

(e.g., staffName, monthlyRent, lastYearsRevenue)?

Example

An online order wine company requires the designing of a data warehouse to record the quantity and sales of its wines to its customers by day of the week and year. Part of the original database is composed by the following tables:

- CUSTOMER (Code, Name, Address, Phone, BirthDay, Gender)
- WINE (Code, Name, Type, Vintage, BottlePrice, ClassCode)
- CLASS (Code, Region)
- ORDER (CustomerCode, WineCode, TimeStamp, nrBottles)

Draw a star schema and a snowflake schema:

Example: Star Schema

1. Select the business process to model
   - `In this example we need to model the sales`
2. Choosing the grain of the business process
   - `Individual product sales`
3. Identifying and conforming the dimensions:
   - `CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)`
   - `WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)`
   - `TIME (TimeCode, Day, Year)`
4. Choosing the facts:
   - `Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)`

Example: Star Schema

1. Select the business process to model
   - In this example we need to model the sales
2. Choosing the grain of the business process
   - Individual product sales
3. Identifying and conforming the dimensions:
   - CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)
   - WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)
   - TIME (TimeCode, Day, Year)
4. Choosing the facts:
   - Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)

Example: Star Schema

1. Select the business process to model
   - In this example we need to model the sales
2. Choosing the grain of the business process
   - Individual product sales
3. Identifying and conforming the dimensions:
   - CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)
   - WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)
   - TIME (TimeCode, Day, Year)
4. Choosing the facts:
   - Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)

Example: Star Schema

1. Select the business process to model
   - `In this example we need to model the sales`
2. Choosing the grain of the business process
   - `Individual product sales`
3. Identifying and conforming the dimensions:
   - `CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)`
   - `WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)`
   - `TIME (TimeCode, Day, Year)`
4. Choosing the facts:
   - `Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)`

Example: Star Schema

1. Select the business process to model
   - `In this example we need to model the sales`
2. Choosing the grain of the business process
   - `Individual product sales`
3. Identifying and conforming the dimensions:
   - `CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)`
   - `WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)`
   - `TIME (TimeCode, Day, Year)`
4. Choosing the facts:
   - `Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)`

Example: Star Schema

1. Select the business process to model
   - `In this example we need to model the sales`
2. Choosing the grain of the business process
   - `Individual product sales`
3. Identifying and conforming the dimensions:
   - `CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)`
   - `WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)`
   - `TIME (TimeCode, Day, Year)`
4. Choosing the facts:
   - `Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)`

Example: Star Schema

1. Select the business process to model
   - In this example we need to model the sales
2. Choosing the grain of the business process
   - Individual product sales
3. Identifying and conforming the dimensions:
   - CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)
   - WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)
   - TIME (TimeCode, Day, Year)
4. Choosing the facts:
   - Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)

Example: Star Schema

1. Select the business process to model
   - `In this example we need to model the sales`
2. Choosing the grain of the business process
   - `Individual product sales`
3. Identifying and conforming the dimensions:
   - `CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)`
   - `WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)`
   - `TIME (TimeCode, Day, Year)`
4. Choosing the facts:
   - `Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)`

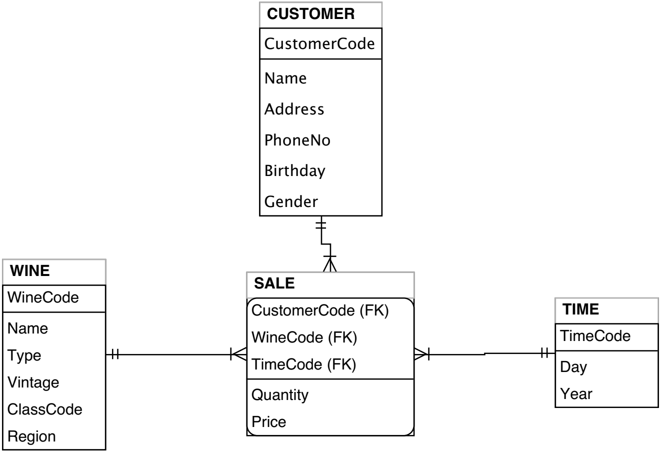
Example: Star Schema

1. Select the business process to model
   - `In this example we need to model the sales`
2. Choosing the grain of the business process
   - `Individual product sales`
3. Identifying and conforming the dimensions:
   - `CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)`
   - `WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)`
   - `TIME (TimeCode, Day, Year)`
4. Choosing the facts:
   - `Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)`

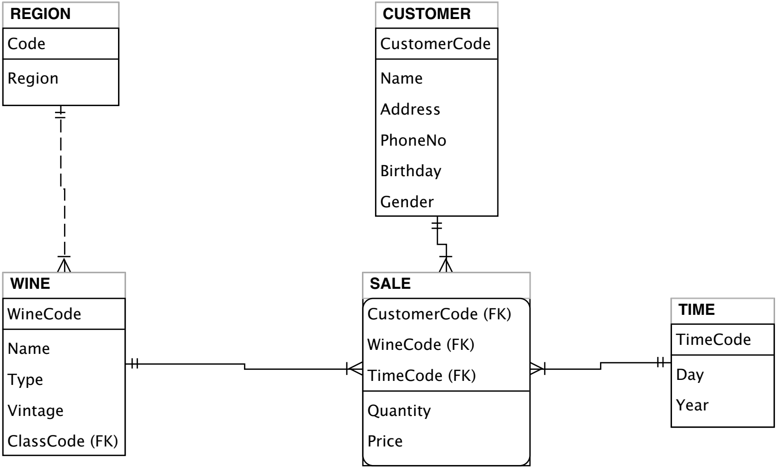
Example: Star Schema

1. Select the business process to model
   - In this example we need to model the sales
2. Choosing the grain of the business process
   - Individual product sales
3. Identifying and conforming the dimensions:
   - CUSTOMER (CustomerCode, Name, Address, PhoneNo, Birthday, Gender)
   - WINE (WineCode, Name, Type, Vintage, ClassCode, Name, Region)
   - TIME (TimeCode, Day, Year)
4. Choosing the facts:
   - Sale(CustomerCode, WineCode, TimeCode, Quantity, Price)

Example: Star Schema



Example: Snowflawe Schema



Comparison among DM[2] and ER models[3]

	ERs	DM
Describe	Database of OLTP	Database of DW
Goal	Remove redundancy	Data retrieval
	Simple transactions	*Ad hoc* end-user queries
	Fast transactions	Complex transactions
Decomposition	One ER model	Several DM

[2]Dimensionality Modeling
[3]Entity Relationship

Building A Data Warehouse

The builders of Data warehouse should take a broad view of the anticipated use of the warehouse

- The design should support ad-hoc querying
- An appropriate schema should be chosen that reflects the anticipated usage

Building A Data Warehouse: Steps

- Design of the DW
- Acquisition of data for the warehouse
- Ensuring that Data Storage meets the query requirements efficiently

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
- The data must be fitted into the data model of the warehouse
- The data must be loaded into the warehouse
  - Proper design for refresh policy should be considered
  - Creating and maintaining required data structures

Data warehouse benefits

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers

Data warehouse problems

- Problems related to data:
  - Underestimation of resources for data loading
  - Hidden problems with source systems
  - Required data not captured
  - Data homogenization
  - Data ownership
- Increased end-user demands
- High demand for resources (large amounts of disk space)
- High maintenance
- Long duration projects
- Complexity of integration

Data warehouses and SQL

SQL has operators for calculating aggregations (subtotals and totals) based on the values in the dimension tables. Read the info here:

`http://dev.mysql.com/doc/refman/5.7/en/group-by-modifiers.html`

Exercise I

Let us consider the case of a real estate agency whose database is composed by the following tables:

```
OWNER (IDOwner, Name, Surname, Address, City, Phone)
ESTATE (IDEstate, IDOwner, Category, Area, City, Province,
        Rooms, Bedrooms, Garage, Meters)
CUSTOMER (IDCust, Name, Surname, Budget, Address, City, Phone)
AGENT (IDAgent, Name, Surname, Office, Address, City, Phone)
AGENDA (IDAgent, Date, Hour, IDEstate, ClientName)
VISIT (IDEstate,IDAgent, IDCust, Date, Duration)
SALE (IDEstate,IDAgent, IDCust, Date, AgreedPrice, Status)
RENT (IDEstate,IDAgent, IDCust, Date, Price, Status, Time)
```

Apply the steps for designing a star schema for the data warehouse that provides strategic and tactical support to the sales department

Exercise II

Let us consider the case of a real estate agency whose database is composed by the following tables:

```
OWNER (IDOwner, Name, Surname, Address, City, Phone)
ESTATE (IDEstate, IDOwner, Category, Area, City, Province,
        Rooms, Bedrooms, Garage, Meters)
CUSTOMER (IDCust, Name, Surname, Budget, Address, City, Phone)
AGENT (IDAgent, Name, Surname, Office, Address, City, Phone)
AGENDA (IDAgent, Date, Hour, IDEstate, ClientName)
VISIT (IDEstate,IDAgent, IDCust, Date, Duration)
SALE (IDEstate,IDAgent, IDCust, Date, AgreedPrice, Status)
RENT (IDEstate,IDAgent, IDCust, Date, Price, Status, Time)
```

Apply the steps for designing a star schema for the data warehouse that provides strategic and tactical support to the hr department (to manage the agent's work)

Exercise III

Create a data warehouse for a supermarket company to provide strategic and tactical support to the sales department. The company database stores information about customers such as first name, surname, address, city and phone. For each product it stores its name, id, brand, category and price. For each supermarket it stores the address, id and city. For each sale, the company registers the date, client, products and supermarket. Apply the steps for designing a star schema for the data ware house

Exercise IV

Design a data warehouse for a hotel company to provide strategic and tactical support to the booking department. The company database stores information about customers such as first name, surname, address, city and phone number. For each hotel it stores its name, id, address, category and number of rooms. For each booking, the company registers the date, client, number of days, number of rooms, and price. Apply the four main steps on dimensional modeling to design a star schema for the data warehouse

Conclusion

In this session we have covered:

- The main concepts, benefits and problems problems associated with data warehousing
- How online transaction processing (OLTP) systems differ from data warehousing
- The architecture and main components of a data warehouse
- The issues associated with designing a data warehouse:
  - Dimensionality modelling
  - How ER models differ from DM models

Suggested Readings

- Chapter 29 of Fundamentals of Database Systems. Elmasri & Navathe.
- Chapters 32 and 33 of Database systems: a practical approach to design, implementation, and management. Connolly, Thomas M; Begg, Carolyn