Data Warehousing and Data Mining

Part I: Data Warehousing

Gao Cong
gaocong@cs.aau.dk

Slides adapted from Man Lung Yiu and Torben Bach Pedersen
Course Structure

- **Business intelligence**: Extract knowledge from large amounts of data collected in a modern enterprise
  - Data warehousing
  - Data mining

- **Purpose**
  - Acquire theoretical background in lectures and literature studies
  - Obtain practical experience on (industrial) tools in a mini-project

---

**Data warehousing**: construction of a database with only data analysis purpose

**Data Mining**: find patterns automatically in databases
Apple iPhone Smart Phone

$499 - $849  Compare prices (6)
★★★★★  User reviews (1233)
★★★★★  Expert reviews (1)
Quad Band - GSM 800, GSM 900, GSM 1800, GSM 1900 - Bluetooth, Wi-Fi - EDGE, GPRS - Polyphonic - 16.7 Million Colors - 4GB - Bar - Smartphone

User reviews  |  Product details  |  Expert reviews  |  Compare prices

All user reviews

General Comments

View by: Positive comments (85%)  |  Negative comments (15%)

The Iphone is the best cell phone I had so far and i think that if you buy it you will love it. More...

ruben_93 catalog.ebay.com 8/23/2008

I have decided to buy it because I like it very much and good product the apple iphone 16GB, has very good quality More...
davidxerach catalog.ebay.com 5/30/2008

We are very happy with the item, it is really a great product. More...

3703guns catalog.ebay.com 5/7/2008

Overall this is great item and I highly recommend this device to travelers More...

clink1381 catalog.ebay.com 4/26/2008
Contact Information

- **Data Warehousing**
  - Teacher: Gao Cong
  - Email: gaocong@cs.aau.dk

- **Data Mining**
  - Teacher: Thomas D. Nielsen and Manfred Jaeger
  - Email: tdn@cs.aau.dk, jaeger@cs.aau.dk

- **Course homepage:** [http://www.cs.aau.dk/~tdn/itev/](http://www.cs.aau.dk/~tdn/itev/)
  - Lecture slides, mini-project, ......

- **Seminar dates**
  - Nov 1st, 2008  Data Warehousing
  - Nov 22th, 2008  Data Warehousing, Data Mining
  - December 6th, 2008  Data Mining
Literature for Data Warehousing

- No textbook
- Books (selected pages available in the class)
  - *The Data Warehouse Lifecycle Toolkit*, Kimball et. al., Wiley 2008
- Additional references/articles:
  - To be given
Today’s Agenda

• 09.00 –10.00  Introduction (Mini-Project, and Exam)
• 10.00 –11.00  Multidimensional Model
• 11.00 –12.00  Multidimensional Model
• 12.00 –12.45  Lunch hour
• 12.45 –13.45  Multidimensional Model + Demo
• 13.45 –15.00  ETL
• 15.00 –15.45  ETL +Demo+Project
• 15.45 –16.00  Summary of the day
Overview

- Data Warehouse (DW) introduction
- DW Topics
  - Multidimensional modeling
  - Extract, Transformation and Load (ETL)
  - Data Warehousing architecture
  - Performance optimization
What is Business Intelligence (BI)?

- BI systems **help** the users make the **right** decisions, based on available data

- **Combination of technologies**
  - Data Warehousing (DW)
  - On-Line Analytical Processing (OLAP)
  - Data Mining (DM)
  - Customer Relationship Management (CRM)
  - …..

- **The Web makes BI more necessary**
  - Customers do not appear “physically” in the store
  - Customers can change to other stores more easily
Definition of a DW

• R. Kimball’s definition of a DW:
  - A data warehouse is a system that extracts, cleans, conforms and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making
  - Data Warehouse is the foundation for business intelligence, value add analytics

• Inmon’s Definition of a Data Warehouse
  - Subject oriented (versus function oriented)
  - Integrated (logically and physically)
  - Time variant (data can always be related to time)
  - Stable (data not deleted)
  - Supporting management decisions (different organization)

• A good DW is a prerequisite for successful BI
  - Ad-hoc analysis and reports
    - We will cover this soon ......
  - Data mining: discovery of hidden patterns and trends
Subject-Oriented Data Collections

Classical operation systems are organized around the applications of the company.

Each type of company has its own unique set of subjects
- For an insurance company,
- For a manufacturer, the major subject areas might be product, order, vendor, bill of material, and raw goods.
- For a retailer, the major subject areas may be product, SKU, sale, vendor, and so forth.
Integrated Data Collections

- Integration is the important.

- Data is fed from multiple disparate sources into the data warehouse. As the data is fed it is converted, reformatted,…

- The result is that data—once it resides in the data warehouse—has a single physical corporate image.
Stable: Non-volatile Data Collections

- Data is updated in the operational environment as a regular matter of course.

- Data warehouse data is loaded (usually en masse) and accessed, but it is often updated.
  - When data in the data warehouse is loaded, it is loaded in a snapshot, static format.
  - When subsequent changes occur, a new snapshot record is written.
Time-variant Data Collections

- Time variancy implies that every unit of data in the data warehouse is accurate as of some one moment in time.
  - a record is time stamped.
  - a record has a date of transaction.
- there is some form of time marking to show the moment in time during which the record is accurate.
- the data warehouse contains much more history than any other environment.
  - A 60-to-90-day time
  - a 5-to-10-year time
DW Architecture – Data as Materialized Views

Existing databases and systems (OLTP)

Appl.

DB

Appl.

DB

Appl.

DB

Appl.

DB

New databases and systems (OLAP)

Trans.

DW

DB

DM

DM

OLAP

Data mining

Visualization

Data mining (Local)

Data Marts (Global) Data Warehouse

(Local) Data Marts

Analogy: (data) suppliers ↔ warehouse ↔ (data) consumers
Business Dimensional Lifecycle

Queries Hard/Infeasible for OLTP

• Business analysis
  ■ Which **public holiday** we have the largest sales?
  ■ Which **week** we have the largest sales?
  ■ Does the sales of **dairy products** increase over time?

• Difficult to represent these queries by using SQL
  ■ second query: extract the “week” value using a function
    ◆ But the user has to learn many transformation functions …
  ■ third query: use a “special” table to store IDs of all dairy products, in advance
    ◆ We have many other product types as well …

• The need of multidimensional modeling
Multidimensional Modeling

- Example: sales of supermarkets
- Facts and measures
  - Each sales record is a fact, and its sales value is a measure
- Dimensions
  - Each sales record is associated with its values of Product, Store, Time
  - Correlated attributes grouped into the same dimension → easier for analysis tasks

<table>
<thead>
<tr>
<th>Product</th>
<th>Type</th>
<th>Category</th>
<th>Store</th>
<th>City</th>
<th>County</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>Beer</td>
<td>Beverage</td>
<td>Trøjborg</td>
<td>Århus</td>
<td>DK</td>
<td>25</td>
<td>Maj</td>
<td>1997</td>
<td>5.75</td>
</tr>
</tbody>
</table>
Multidimensional Modeling

- How do we model the *Time* dimension?
  - A tree structure, with multiple levels
  - Attributes, e.g., holiday, event

```
T
/   \
/     \
Week   Year
   /   \
  /     \
Day  Month
```

- Advantage of this model?
  - Easy for query (more about this later)

- Disadvantage?
  - Data redundancy (controlled redundancy is acceptable)

<table>
<thead>
<tr>
<th>tid</th>
<th>day</th>
<th>week</th>
<th>month</th>
<th>year</th>
<th>work</th>
<th>day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2008</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2008</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Aalborg University 2008 - DWDM course
## OLTP vs. OLAP

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target</strong></td>
<td>operational needs</td>
<td>business analysis</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>small, operational data</td>
<td>large, historical data</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>normalized</td>
<td>denormalized/multidimensional</td>
</tr>
<tr>
<td><strong>Query language</strong></td>
<td>SQL</td>
<td>not unified</td>
</tr>
<tr>
<td><strong>Queries</strong></td>
<td>small</td>
<td>large</td>
</tr>
<tr>
<td><strong>Updates</strong></td>
<td>frequent and small</td>
<td>infrequent and batch</td>
</tr>
<tr>
<td><strong>Transactional recovery</strong></td>
<td>necessary</td>
<td>not necessary</td>
</tr>
<tr>
<td><strong>Optimized for</strong></td>
<td>update operations</td>
<td>query operations</td>
</tr>
</tbody>
</table>
Quick Review: Normalized Database

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Product</th>
<th>Category</th>
<th>Price</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>3301</td>
<td>Beer</td>
<td>Beverage</td>
<td>6.00</td>
<td>02-02-2008</td>
</tr>
<tr>
<td>3301</td>
<td>Rice</td>
<td>Cereal</td>
<td>4.00</td>
<td>02-02-2008</td>
</tr>
<tr>
<td>3302</td>
<td>Beer</td>
<td>Beverage</td>
<td>6.00</td>
<td>05-02-2008</td>
</tr>
<tr>
<td>3303</td>
<td>Wheat</td>
<td>Cereal</td>
<td>5.00</td>
<td>07-02-2008</td>
</tr>
</tbody>
</table>

- Normalized database avoids
  - Redundant data
  - Modification anomalies
- How to get the original table? (join them)
- No redundancy in OLTP, controlled redundancy in OLAP
OLAP Data Cube

• Data cube
  ■ Useful data analysis tool in DW
  ■ Generalized GROUP BY queries
  ■ Aggregate facts based on chosen dimensions
    ◆ Product, store, time dimensions
    ◆ Sales measure of sale facts

• Why data cube?
  ■ Good for visualization (i.e., text results hard to understand)
  ■ Multidimensional, intuitive
  ■ Support interactive OLAP operations

• How is it different from a spreadsheet?
On-Line Analytical Processing (OLAP)

- Interactive analysis
- Explorative discovery
- Fast response times required

OLAP operations/queries
- Aggregation, e.g., SUM
- Starting level, (Year, City)
- Roll Up: Less detail
- Drill Down: More detail
- Slice/Dice: Selection, Year=2000
Extract, Transform, Load (ETL)

- “Getting multidimensional data into the DW”
- Problems
  - Data from different sources
  - Data with different formats
  - Handling of missing data and erroneous data
- ETL
  - Extract
  - Transformations / cleaning
  - Load
- The most time-consuming process in DW development
  - 80% of development time spent on ETL
Data’s Way To DW

- **Extraction**
  - Extract from many heterogeneous systems

- **Staging area**
  - Large, sequential bulk operations ➔ flat files best?

- **Cleaning**
  - Data checked for missing parts and erroneous values
  - Default values provided and out-of-range values marked

- **Transformation**
  - Data transformed to decision-oriented format
  - Data from several sources merged, optimize for querying

- **Aggregation?**
  - Are individual business transactions needed in the DW?

- **Loading into DW**
  - Large bulk loads rather than SQL INSERTs (Why?)
  - Fast indexing (and pre-aggregation) required
Performance Optimization

• Performance optimization
  ■ Fine tune performance for important queries
  ■ Aggregates, indexing, other optimizations (environment, partitioning)

• Using aggregates
  ■ How can aggregates improve performance?

• Choosing aggregates
  ■ Which aggregates should we materialize?
Materialization Example

- Imagine 1 billion sales rows, 1000 products, 100 locations
- CREATE VIEW TotalSales (pid, locid, total) AS
  SELECT s.pid, s.locid, SUM(s.sales)
  FROM Sales s
  GROUP BY s.pid, s.locid
- The materialized view has 100,000 rows
- Rewrite the query to use the view
  - SELECT p.category, SUM(s.sales) FROM Products p, Sales s
    WHERE p.pid = s.pid GROUP BY p.category
    - can be rewritten to
  - SELECT p.category, SUM(t.total) FROM Products p,
    TotalSales t WHERE p.pid = t.pid GROUP BY p.category
    - Query becomes 10,000 times faster!
Common DW Issues

• Metadata management
  - Need to understand data = metadata needed
  - Greater need in OLAP than in OLTP as “raw” data is used
  - Need to know about:
    - Data definitions, dataflow, transformations, versions, usage, security

• DW project management
  - DW projects are large and different from ordinary SW projects
    - 12-36 months and US$ 1+ million per project
    - Data marts are smaller and “safer” (bottom up approach)
  - Reasons for failure
    - Lack of proper design methodologies
    - High HW+SW cost
    - Deployment problems (lack of training)
    - Organizational change is hard... (new processes, data ownership,..)
    - Ethical issues (security, privacy,...)
Summary

• Data Warehouse (DW) introduction
• DW Topics
  ■ Multidimensional modeling
  ■ ETL
  ■ Performance optimization
• BI provide many advantages to your organization
  ■ A good DW is a prerequisite for BI

• Reading materials:
  ■ Jarke chapter 1. Page 1-10
  ■ Optional: Kimball chapter 1.
DW Software

• DW part of the mini-project

• DW software
  ■ Obtain from MSDNAA, and install
    ◆ MS SQL Server 2005 Standard Edition/ 2008 Enterprise
    ◆ MS Analysis Services, Integration Services, Reporting Services
  ■ Checking after installation
    ◆ Open “Component Services” and check whether all four services above have been started
    ◆ Open “SQL Server Management Studio” and see whether you can connect to “Database Engine”
  ■ Read the mini-project webpage for installation details
Study plan

- Reading materials and project each week
  - Week Nov 1-7 (Data Warehousing)
  - Week Nov 8-14 (Data Warehousing)
  - Week Nov 15-21 (Data Warehousing)
  - Week Nov 22-28 (Data Warehousing + Data Mining)
  - Week Nov 29-Dec 5 (Data Mining)
  - Week Dec 6-12 (Data Mining)
  - Week Dec 13-19 (Data Mining)


We also remind you what to do and read every week by email.
Exam

- Exam
  - Individual oral exam, for 30 minutes
    - 15 minutes of DW questions
    - 15 minutes of ML questions
  - Danish 7-point grading scale
  - Submitting project report is the prerequisite to take the exam

- Prepare Exam
  - Mini-project report as the basis for discussion
  - Exam also covers theoretical background in lectures and literature
    - Question will be related to mini-projects
Outline of DW Mini-Project

- Selecting business process and data sources
  - You are strongly encouraged to find problem and data from your work

- Dimensional data modeling
  - Both ”real” multidimensional and star/snowflake
  - Consider advanced aspects such as SCDs

- Implementation of multidimensional DB
  - Star/snowflake schema in SQL Server + cubes

- ETL: building an ETL flow to populate the DW
  - The hardest part 😊

- Building reports in Reporting Services

- Doing performance optimization

- Details given each week in study plan
Mini-Project (Data Warehousing Part)

• Topic and Data
  ■ Business process and data from your work
  ■ Fklub

• Mini-project
  ■ Performed in groups of ~3 persons
  ■ Documented in report of 10-20 pages
  ■ Firm Deadline: January 9, 2009, at 12.00 (Soft Deadline: Dec 14)
  ■ The mini project should be sent by email to Lene Even, even@cs.aau.dk
Project Supervision

- Lunch time and break in seminars
- Emails
  - Each week, reminding what to read/do
- Phones
- One hour group meeting using skype each week can be arranged if requested
  - For all students who want to be in (not just for a single group)
Demonstration on SQL server

- To be put on the Web
Today’s Agenda

- 09.00 –10.00 Introduction (Mini-Project, and Exam)
- 10.00 –11.00 Multidimensional Model
- 11.00 –12.00 Multidimensional Model
- 12.00 –12.45 Lunch hour
- 12.45 –13.45 Multidimensional Model + Demo
- 13.45 –15.00 ETL
- 15.00 –15.45 ETL +Demo+Project
- 15.45 –16.00 Summary of the day
Sharing your experience

• Why do you want to take the Data Warehousing course?

• Have you ever worked on something related to a Data Warehouse project?