Data Integration and ETL Process

Krzysztof Dembczyński

Intelligent Decision Support Systems Laboratory (IDSS)
Poznań University of Technology, Poland

Software Development Technologies
Master studies, first semester
Academic year 2014/15 (winter course)
Review of the previous lecture

• Mining of massive datasets
• Evolution of database systems: operational vs. analytical systems.
• Dimensional modeling:
  ▶ Three goals of the logical design of data warehouse: simplicity, expressiveness and performance.
  ▶ The most popular conceptual schema: star schema.
  ▶ Designing data warehouses is not an easy task.
Outline

1 Motivation
2 Data Extraction
3 Transformation and Integration of Data
4 Load of Data
5 Data Warehouse Refreshment
6 Summary
Outline

1 Motivation

2 Data Extraction

3 Transformation and Integration of Data

4 Load of Data

5 Data Warehouse Refreshment

6 Summary
Motivation

• OLAP queries are usually performed in a separate system, i.e., in a data warehouse.
• Data warehouses combine data from multiple sources.
• Data must be translated into a consistent format.
• Data integration represents 80% of effort for a typical data warehouse project!
ETL process

- ETL = Extraction, Transformation, and Load
  - Extraction of data from source systems,
  - Transformation and integration of data into a useful format for analysis,
  - Load of data into the warehouse and build of additional structures.

- Refreshment of data warehouse is closely related to ETL process.
- The ETL process is described by metadata stored in data warehouse.
- Architecture of data warehousing:
  
  Data sources → Data staging area → Data warehouse
ETL Tools

- Data extraction from heterogeneous data sources.
- Data transformation, integration, and cleansing.
- Data quality analysis and control.
- Data loading.
- High-speed data transfer.
- Data refreshment.
- Managing and analyzing metadata.

**Examples of ETL tools:**
  - MS SQL Server Integration Services (SSIS), IBM Infosphere DataStage, SAS ETL Studio, Oracle Warehouse Builder, Oracle Data Integrator, Business Objects Data Integrator, Pentaho Data Integration.
Outline

1 Motivation

2 Data Extraction

3 Transformation and Integration of Data

4 Load of Data

5 Data Warehouse Refreshment

6 Summary
Data Extraction

- Data warehouse needs extraction of data from different external data sources:
  - operational databases (relational, hierarchical, network, itp.),
  - files of standard applications (Excel, COBOL applications),
  - additional databases (direct marketing databases) and data services (stock data),
  - and other documents (.doc, XML, WWW).
Data Sources

- Data sources are often operational systems, providing the lowest level of data.
- Data sources are designed for operational use, not for decision support, and the data reflect this fact.
- Multiple data sources are often from different systems, run on a wide range of hardware and much of the software is built in-house or highly customized.
- Data sources can be designed using different logical structures.
Data Extraction

- Identification of concepts and objects does not have to be easy.
- **Example**: Extract information about sales from the source system.
  - What is meant by the term **sale**? A sale has occurred when
    - the order has been received by a customer,
    - the order is sent to the customer,
    - the invoice has been raised against the order.
  - It is a common problem that there is no table **SALES** in the operational databases; some other tables can exist like **ORDER** with an attribute **ORDER_STATUS**.
Change Monitoring

• Change monitoring is directly connected with data warehouse refreshment.
• Detect changes to an information source.
• Different monitoring techniques: external and intrusive techniques.
• Monitoring Techniques
  ▶ Snapshot vs. timestamped sources
  ▶ Queryable, logged, and replicated sources
  ▶ Callback and internal action sources
Outline

1 Motivation

2 Data Extraction

3 Transformation and Integration of Data

4 Load of Data

5 Data Warehouse Refreshment

6 Summary
Transformation and Integration of Data

• Transformation and integration of data is the most important part of data warehousing.
• This phase consists in removing all inconsistencies and redundancies of data coming to the data warehouse from operational data sources.
• The data are conform to the conceptual schema used by the warehouse.
• Integration concerns data and data schemas.
• Different levels of integration: schema, table, tuple, attribute values.
Conflicts and Dirty Data

- Different logical models of operational sources,
- Different data types (account number stored as **String** or **Numeric**),
- Different data domains (gender: **M, F, male, female, 1, 0**),
- Different date formats (dd-mm-yyyy or mm-dd-yyyy),
- Different field lengths (address stored by using 20 or 50 chars),
- Different naming conventions: homonyms and synonyms,
- Semantic conflicts, when the same objects are modeled on different logical levels,
- Structural conflicts, when the same concepts are modeled using different structures.
Conflicts and Dirty Data

- Different entries for the same attribute (state name or abbreviation),
- Text fields can possess hidden information (contact person name can be given in the company name field),
- Wrong attribute entries (attribute name contains company name or contact person name),
- Inconsistent information concerning the same object,
- Information concerning the same object, but indicated by different keys,
- Missing values,
- ...
Data Cleansing

- We would like to analyze high-quality data, since our goal is to support decision making.
Data Cleansing Techniques

- Conversion and normalization methods (date formats “dd/mm/rrrr”, names conventions: Jan Kowalski),
- Parsing text fields in order to identify and isolate data elements:
  - Transformation (splitting the text into records \{title = mgr, first name = Jan, last name = Kowalski\}),
  - Standardization (Jan Kowalski, magister ⇒ mgr Jan Kowalski),
- Dictionary-based methods (database of names, geographical places, pharmaceutical data),
- Domain-specific knowledge methods to complete data (postal codes),
- Rationalization of data (PHX323RFD110A4 ⇒ Print paper, format A4),
- Rule-based cleansing (replace gender by sex)
- Cleansing by using data mining.
Data Cleansing Techniques

• Deduplication ensures that one accurate record exists for each business entity represented in a database,

• Householding is the technique of grouping individual customers by the household or organization of which they are a member; this technique has some interesting marketing implications, and can also support cost-saving measures of direct advertising.

• Example:
  ▶ Consider the following rows in a database:

<table>
<thead>
<tr>
<th>Name</th>
<th>Street</th>
<th>City</th>
<th>State</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tim Jones</td>
<td>123</td>
<td>Main Street</td>
<td>Marlboro</td>
<td>MA 12234</td>
</tr>
<tr>
<td>T. Jones</td>
<td>123</td>
<td>Main St.</td>
<td>Marlborough</td>
<td>MA 12234</td>
</tr>
<tr>
<td>Timothy Jones</td>
<td>321</td>
<td>Maine Street</td>
<td>Marlborog</td>
<td>AM 12234</td>
</tr>
<tr>
<td>Jones, Timothy</td>
<td>123</td>
<td>Maine Ave</td>
<td>Marlborough</td>
<td>MA 13324</td>
</tr>
</tbody>
</table>

  ▶ The sales for around $500 are counted for each tuple.
  ▶ Is it the same person?
Outline

1. Motivation
2. Data Extraction
3. Transformation and Integration of Data
4. Load of Data
5. Data Warehouse Refreshment
6. Summary
Load of Data

- After extracting, cleaning and transforming, data must be loaded into the warehouse.
- Loading the warehouse includes some other processing tasks: checking integrity constraints, sorting, summarizing, creating indexes, etc.
- Batch (bulk) load utilities are used for loading.
- A load utility must allow the administrator to monitor status, to cancel, suspend, and resume a load, and to restart after failure with no loss of data integrity.
Load of Data

• Concern very large data volumes.
• Sequential loads can take a very long time.
• Full load can be treated as a single long batch transaction that builds up a new database.
• Using checkpoints ensures that if a failure occurs during the load, the process can restart from the last checkpoint.
Outline

1. Motivation
2. Data Extraction
3. Transformation and Integration of Data
4. Load of Data
5. Data Warehouse Refreshment
6. Summary
• Refreshing a warehouse means propagating updates on source data to the data stored in the warehouse.
• Follows the same structure as ETL process.
• Several constraints: accessibility of data sources, size of data, size of data warehouse, frequency of data refreshing, degradation of performance of operational systems.
Data Warehouse Refreshment

- Detect changes in external data sources.
- Extract the changes and integrate into the warehouse.
- Update indexes, subaggregates and materialized views.
Data Warehouse Refreshment

• Periodical refreshment (daily or weekly).
• Immediate refreshment.
• Determined by usage, types of data source, etc.
Refreshment and Load

- Refreshment can be asynchronous.
- Load of data requires a long-term access to source data.
- Refreshment concerns less data.
- Refreshment is faster.
1. Motivation
2. Data Extraction
3. Transformation and Integration of Data
4. Load of Data
5. Data Warehouse Refreshment
6. Summary
Summary

• ETL process is a strategic element of data warehousing.
• Main concepts: extraction, transformation and integration, load, data warehouse refreshment and metadata.
• New emerging technology . . .
J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan-Kaufmann 2000


Ch. Todman, *Projektowanie hurtowni danych. Zarzadzanie kontaktami z klientami (CRM)*, Wydawnictwa Naukowo-Techniczne 2003