Data Warehouse and Data Mining

Lecture No. 02
Lifecycle of Data warehouse

Naeem Ahmed
Email: naeemmahoto@gmail.com
Department of Software Engineering
Mehran University of Engineering and Technology Jamshoro
Outline

• Lifecycle of DW

• Classical SDLC vs. DW SDLC

• Operating DW

Acknowledgements: Wolf-Tilo Balke and Silviu Homoceanu
http://www.ifis.cs.tu-bs.de
Lifecycle of DW

**Data Warehouse System Development Life Cycle (SDLC)**

- **Design**
  - End-user interview cycles
  - Source system cataloging
  - Definition of key performance indicators
  - Mapping of decision-making processes underlying information needs
  - Logical and physical schema design
Lifecycle of DW

• Prototype
  – Objective is to **constrain** and in some cases **reframe** end-user requirements

• Deployment
  – Development of documentation
  – Training
  – Operations and management processes

• Operation
  – Day-to-day maintenance of the DW needs a good management of ongoing **Extraction**, **Transformation** and **Loading** (ETL) process
Lifecycle of DW

- **Enhancement** needs the modification of
  - HW - physical components
  - Operations and management processes
  - Logical schema designs
Lifecycle of DW

• Classical SDLC vs. DW SDLC

• DW SDLC is almost the *opposite* of classical SDLC
Lifecycle of DW

- Classical SDLC vs. DW SDLC

<table>
<thead>
<tr>
<th>Classical SDLC</th>
<th>DW SDLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements gathering</td>
<td>Implement warehouse</td>
</tr>
<tr>
<td>Analysis</td>
<td>Integrate data</td>
</tr>
<tr>
<td>Design</td>
<td>Test for bias</td>
</tr>
<tr>
<td>Programming</td>
<td>Program against data</td>
</tr>
<tr>
<td>Testing</td>
<td>Design DSS system</td>
</tr>
<tr>
<td>Integration</td>
<td>Analyze results</td>
</tr>
<tr>
<td>Implementation</td>
<td>Understand requirements</td>
</tr>
</tbody>
</table>

- Because it is the opposite of SDLC, DW SDLC is also called CLDS
Lifecycle of DW

• CLDS is a **data driven** development life cycle
• It starts with data
  – Once data is at hand it is integrated and tested against bias
  – Programs are written against the data and the results are analyzed and finally the requirements of the system are understood
  – Once requirements are understood, adjustments are made to the design and the cycle starts all over
• “**spiral development methodology**”
Operating a DW

• In Operating a DW the following phases can be identified
  – Monitoring
  – Extraction
  – Transforming
  – Loading
  – Analyzing
Operating a DW: Monitoring

• Monitoring
  – Surveillance of the data sources
  – Identification of data modification which is relevant to the DW
  – Monitoring has an important role over the whole process deciding on which data the next steps will be applied on

• Monitoring techniques
  – Active mechanisms - Event Condition Action (ECA) rules:
    
    | EVENT | Payment                  |
    |-------|--------------------------|
    | CONDITION | Account sum > 10 000 € |
    | ACTION   | Transfer to economy account |
Operating a DW: Monitoring

• Monitoring techniques
  – Replication mechanisms
    • Snapshot:
      – Local copy of data, similar to a View
      – Used by Oracle 9i
    • Data replication
      – Replicates and maintains data in destination tables through data propagation processes
      – Used by IBM
Operating a DW: Monitoring

• Monitoring techniques
  – Protocol based mechanisms
    • Since DBMS write protocol data for transaction management, the protocol can be used also for monitoring
    • Difficult due to the fact that the protocol format is proprietary and subject to change
  – Application managed mechanisms
    • Hard to implement for legacy systems
    • Based on time stamping or data comparison
Operating a DW: Extraction

- **Extraction**
  - Reads the data which was selected throughout the monitoring phase and inserts it in the data structures of the workplace
  - Due to large data volume, compression can be used
  - The time-point for performing extraction can be:
    - Periodical:
      - Weather or stock market information can be actualized more times in a day, while product specification can be actualized in a longer period of time
    - On request:
      - For example when a new item is added to a product group
Operating a DW: Extraction

• Extraction
  – The time-point for performing extraction can be:
    • Event driven:
      – Event driven extraction can be helpful in scenarios where time, or the number of modifications over passing a specified threshold triggers the extraction. For example each night at 03:00 or each time 50 new modifications took place, an extraction is performed
    • Immediate:
      – In some special cases like the stock market it can be necessary that the changes propagate immediately to the warehouse
  – The extraction largely depends on hardware and the software used for the DW and the data source
Operating a DW: Transforming

- Transforming
  - Implies adapting data, schema as well as data quality to the application requirements
- Data integration:
  - Transformation in de-normalized data structures
  - Handling of key attributes
  - Adaptation of different types of the same data
  - Conversion of encoding:
    - “Buy”, “Sell” → 1,2 vs. B,S → 1,2
  - Normalization:
    - “Michael Schumacher” → “Michael, Schumacher” vs. “Schumacher Michael” → “Michael, Schumacher”
Operating a DW: Transforming

- Transforming
  - Data integration:
    - Date handling:
      - “MM-DD-YYYY” $\Rightarrow$ “MM.DD.YYYY”
    - Measurement units and scaling:
      - 10 inch $\Rightarrow$ 25.4 cm
      - 30 mph $\Rightarrow$ 48.279 km/h
  - Save calculated values
    - Price_incl_VAT = Price_excl_VAT * 1.19
  - Aggregation
    - Daily sums can be added into weekly ones
    - Different levels of granularity can be used
Operating a DW: Transforming

• Transforming
  – Data cleaning:
    • Consistency check
      – Delivery_date < Order_date
    • Completeness
      – Management of missing values as well as NULL values
Operating a DW: Loading

• Loading
  – Loading usually takes place during weekends or nights when the system is not under user stress
  – Split between initial load to initialize the DW and the periodical load to keep the DW updated
  – Initial loading
    • Implies big volumes of data and for this reason a bulk loader is used
    – Usually performed by partitioning, parallelization and incremental actualization
Operating a DW: Analyzing

• Analyze
  – Data access
    • Useful for extracting goal oriented information:
      – How many iPhones 3G were sold in the Braunschweig stores of T-Mobile in the last 3 calendar weeks of 2008?
      – Although it is a common OLTP query, it might be too complex for the operational environment to handle
  – OLAP
    • Falsely used as representing DW because it is used to analyze data contained in DW
    • Used to answer requests like:
      – In which district does a product group register the highest profit
      – How did the profit change in comparison to the previous month?
Operating a DW: Analyzing

• Analyze
  – OLAP
  • Used to answer requests like:
    – Mostly known as organized on a multidimensional data model
    – Common operations for analyze are:
      » Pivoting/Rotation
      » Roll-up, Drill-down and Drill-across
      » Slice and Dice
  – Data mining
    • Useful for identifying hidden patterns
    • Refers to two separate processes:
      – KDD (Knowledge Discovery in Databases)
      – Prediction
Operating a DW: Analyzing

• Analyze
  – Data mining
    • Useful for answering questions like:
      – How did the sales of this product group evolve?
    • Methods and procedures for data mining
      – Clustering, Classification, Regression, Association rule learning