Triple-Driven Data Modeling Methodology in Data Warehousing: A Case Study

Authors: Yuhong Guo, Shiwei Tang, Yunhai Tong, and Dongqing Yang (Peking Univ., China)

Friday, Nov. 10, 2006
Triple-driven: why and how?

- **Motivation**
  - Existing methods are used in isolation
  - Data models from single principle are incomplete, which cannot obtain satisfaction

- **Solution**
  Goal-driven
  Data-driven
  User-driven
  \[\rightarrow\] DW data model
Outline

- **Background**
  - Motivation & related work
  - Objectives
  - CLIC DW case study
- **Proposed Methodology**
- **Discussion**
- **Conclusions**
Motivation

- **Open Problems & Challenges**
  - DW conceptual modeling is still under user’s dissatisfactions (*DMDW’03*)
  - Lack of comprehensive documentation and dissemination of requirement engineering methods (*DaWaK’05*)

- **What leads to this?**
Background

Related work

- **Existing Data-driven**
  - Emphasis: integrate, reorganize source schemas
  - Lack: match data sources with information requirements

- **Existing Goal-driven**
  - Emphasis: decompose business process
  - Lack: embody business goals into design elements

- **Existing User-driven**
  - Emphasis: facilitate user participations
  - Lack: translate user requirements into design elements

The three methods are *complementary* and should be used in parallel to achieve optimal design
Our objectives

- Tackle four research questions
  - **Triple-driven**: How to integrate the three existing approaches to warehouse design
  - **Data-driven**: How to identify warehouse elements from operational data sources
  - **Goal-driven**: How to embody corporate strategy and business objectives
  - **User-driven**: How to translate user requirements into appropriate design elements
The CLIC DW Planning Project

- **Objective**
  - Develop data model for a central DW

- **Diversity needs for the DW**
  - Centralize the data scattered
  - User querying, reporting, analysis, decision

- **2 core application systems**
  - Cbps system: core business process system
  - Callcenter system: customer consultation, complaint, inquiry...
Outline

- Background
- **Proposed Methodology**
  - Framework
  - Goal-Driven
  - Data-Driven
  - User-Driven
  - Combine triple-driven
- Discussion
- Conclusions
Proposed Methodology

Framework

Goal Driven

1.1 Develop corporate strategy
1.2 Identify main business fields
1.3 Define KPIs of each business field
1.4 Identify target users
1.5 Identify subject areas
1.6 Key performance indicators

Data Driven

2.1 Identify data source systems
2.2 Classify data tables of each source system
2.3 Delete pure operational tables and columns
2.4 Map the remainder tables into the subject areas
2.5 Integrate the tables in the same subject area to form each subject’s schema

User Driven

3.1 User interview
3.2 Report collection and analysis
3.3 Business questions
3.4 Analytical requirements represented by measures and dimensions

2006-11-10
Goal-Driven

Main tasks

- Identify main business fields
  - CRM, RM, ALM, F&PM
- Identify subjects
- Define KPIs (Key Performance Indicators)
- Identify users
  - Query users
  - report users
  - analytical users
  - data miners
**Goal-Driven: Identify subjects**

- **Subject**
  - Object that will be analyzed in each business field
  - High information class

- **Subject Level**
  - subject->sub-subject...

- **Guideline**
  - number of the subjects in each level is about 10, not more than 20 (manageable for human)
Goal-Driven: Define KPIs

**KPIs of Customer Relationship Management**

<table>
<thead>
<tr>
<th><strong>KPIs</strong></th>
<th><strong>Definition</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Satisfaction Index</td>
<td>The quality of the services given by a department from the view of customers in the targeted segments.</td>
</tr>
<tr>
<td>Customer Retention Rate</td>
<td>The ability of a company’s department to retain customers in the targeted measurement segments.</td>
</tr>
<tr>
<td>Revenue Per Customer</td>
<td>The profitability on target customer segments.</td>
</tr>
<tr>
<td>Customer Acquisition Rate</td>
<td>The ability of a company’s center/department to acquire customers in the target segments.</td>
</tr>
</tbody>
</table>
Data-Driven

- **Main tasks**
  - Identify candidate data sources
  - Classify data source tables
    - Transaction tables
    - Component tables
    - Report tables
    - Classification tables
    - Control tables
  - Delete pure operational tables & columns
  - Map the remainder tables into the subjects
  - Integrate the tables in the same subject
## Proposed Methodology

### Data-Driven: Map tables into subjects

<table>
<thead>
<tr>
<th>Candidate Systems</th>
<th>Remainder Tables</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>...</th>
<th>Subject N</th>
</tr>
</thead>
<tbody>
<tr>
<td>System1</td>
<td>Transaction Table 1</td>
<td>[ ]</td>
<td>[ ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Component Table 1</td>
<td>[ ]</td>
<td>[ ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System2</td>
<td>Transaction Table 1</td>
<td>[ ]</td>
<td>[ ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Component Table 1</td>
<td>[ ]</td>
<td>[ ]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data-Driven: Integrate tables in the same subject

Proposed Methodology

Source tables mapped into Subject 1

Integrated Schema of Subject 1

Sys1      Sys2
T■...     T■...
C■...     C■...
T□... + T□...
C□...     C□...

Subject M  Subject N

Sys1_T  Sys1Sys2_C  Sys2_T
Star schema VS. anti-star schema

Anti-star schema shifts attention to a component object, not an event, which leads to delicate analysis of an object according to its behaviors.
User-Driven

Main tasks

- User interview -> Business questions
  - Which customers are most profitable based upon premium revenue?
  - Which channels customers like most?
  - What are the top five reasons that customers return products?

- Reports collection and analysis
  - Analytical requirements represented by measures and dimensions

Proposed Methodology
User-Driven: Analytical requirements

- **Measures:**
  - count, premium, insured amount, policy numbers, net profit, loss ratio...

- **Dimensions:**
  - sex (male, female, unknown)
  - income (<1000, 1000-5000, 5000-10000, 10000-20000, >20000)
  - marriage status (married, unmarried, divorced)
  - education degree (<primary school, high school, >undergraduate)
  - age (< 20, 21-25, 26-30, 30-35, 36-40, 40-50, 50-65, >65)
Proposed Methodology

Combine the triple-driven results

atomic data

Goal-driven

Data-driven

aggregate data

User-driven

2006-11-10

DOLAP’06
Outline

- Background
- Proposed Methodology
- Discussion
- Conclusions
Discussion

- Advantages
  - It ensures actual business value of the DW and stability of the data model
  - It raises acceptance of users towards the DW
  - It ensures the DW is flexible to support the widest range of analysis
  - It leads to a design capturing all specifications

- Impact & Experience: encouraging

Scattered operational databases

Ambiguous business and user needs

effective & comprehensive design
Conclusions

- Main contributions
  - **Integrate** three existing single-driven by “subject”
  - **Identify** valuable elements from data sources by mapping tables into “subjects”
  - **Embody** corporate strategy and business objectives into data model by KPIs
  - **Translate** user needs into design elements by report analysis & business questions
Thanks for your attention

If you have any question, please contact the first author:

yhgguo@pku.edu.cn