Best Practices in Data Warehousing Design and Implementation

Rein Mertens
SAS Netherlands
Best Practices…

- A best practice is a technique or methodology that, through experience and research, has proven to reliably lead to a desired result. A commitment to using the best practices in any field is a commitment to using all the knowledge and technology at one's disposal to ensure success. The term is used frequently in the fields of health care, government administration, the education system, project management, hardware and software product development, and elsewhere.

- Using past experiences to plan future developments
Data Warehousing approaches over time

- Use operational data direct? ("virtual dwh")
- Monolithic data warehouse? ("big bang")
- Data marts?
- Evolutionary approach? ("think big, start small")
Independent Data Marts

(b) An Information Architecture based on Data Marts without an Enterprise Layer
Data Warehousing in 2003 requires an Architected Environment

- Supports all forms of information use:
  - Suppliers/Organization/Customer/Enterprise
  - Query and reporting
  - Analytics / Data mining
  - Multi-dimensional analysis and OLAP

- Ensures end-to-end consistency:
  - Co-ordinates access to operational data
  - Foundation layer
  - Metadata support
    - Maintenance
    - User view
# Data Warehouses vs. Data Marts

<table>
<thead>
<tr>
<th>Data Warehouses</th>
<th>Data Marts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scope</strong></td>
<td></td>
</tr>
<tr>
<td>• Application-neutral</td>
<td>• Specific application requirement</td>
</tr>
<tr>
<td>• Centralized, shared</td>
<td>• LOB, department or user area</td>
</tr>
<tr>
<td>• Cross-LOB/enterprise</td>
<td>• Business process-oriented</td>
</tr>
<tr>
<td>• Architected</td>
<td>• Multiple databases with redundant data</td>
</tr>
<tr>
<td><strong>Data Perspective</strong></td>
<td></td>
</tr>
<tr>
<td>• Historical detailed data</td>
<td>• Detailed (some history)</td>
</tr>
<tr>
<td>• Some summary</td>
<td>• Summarized</td>
</tr>
<tr>
<td>• Lightly denormalized</td>
<td>• Highly denormalized</td>
</tr>
<tr>
<td><strong>Subjects</strong></td>
<td></td>
</tr>
<tr>
<td>• Multiple subject areas</td>
<td>• Single partial-subject area</td>
</tr>
<tr>
<td><strong>Data Sources</strong></td>
<td>• Multiple partial-subject areas</td>
</tr>
<tr>
<td>• Many</td>
<td>• Operational source snapshot</td>
</tr>
<tr>
<td>• Operational, external data</td>
<td></td>
</tr>
<tr>
<td><strong>Implementation Time Frame</strong></td>
<td></td>
</tr>
<tr>
<td>• Nine to 18 months for first stage (two or three subject areas)</td>
<td>• Four to 12 months</td>
</tr>
<tr>
<td>• Multiple-stage implementation</td>
<td></td>
</tr>
<tr>
<td><strong>Characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>• Flexible, extensible</td>
<td>• Restrictive, nonextensible</td>
</tr>
<tr>
<td>• Durable/strategic</td>
<td>• Short life/tactical</td>
</tr>
<tr>
<td>• Data orientation</td>
<td>• Project orientation</td>
</tr>
</tbody>
</table>
Data Warehousing in 2003

**Step #1 - Extract and Transform from source data into Staging Area**

- Source
  - Excel
  - SAS
  - SAP
  - Oracle
  - PeopleSoft

**Step #2 - Data Quality**

- Data Validation
- Slowly changing dimensions

**Step #3 - Transform into dimensional model**

- Central Warehouse
  - SAS
  - SAS
  - SAS

**ETL Studio Data Surveyors**

- More ETL flows
- ETL Studio transformations
- Data Quality plugins

- Customers
- Suppliers
- Finance
- HCM
- HCM

**Source**

- Excel
- SAS
- SAP
- Oracle
- PeopleSoft

**Central Warehouse**

- SAS
- SAS
- SAS

**Step #2 - Data Quality**

- Data Validation
- Slowly changing dimensions

**More ETL flows**

- ETL Studio
- Data Surveyors
- ETL Studio transformations
- Data Quality plugins
Data Warehousing in 2003

Step #1 - Extract and Transform from source data into Staging Area

ETL Studio Data Surveyors

Step #2 - Data Quality Data Validation Slowly changing dimensions

ETL Studio transformations Data Quality plugins

Step #3 - Transform into dimensional model

More ETL flows

Central Warehouse
Working with operational data

- How do you know about your operational world?
- External metadata
  - Operational databases
  - Operational systems
    - May need navigational assistance
- (Near)realtime?
- Data Profiling!
Data Warehousing in 2003

Step #1 - Extract and Transform from source data into Staging Area

Source
- Excel
- SAS
- SAP
- Oracle
- PeopleSoft

Step #2 - Data Quality
- Data Validation
- Slowly changing dimensions

More ETL flows

Step #3 - Transform into dimensional model

Central Warehouse
- SAS
- SAS
- SAS

Finance

Customers

Suppliers

Step #1 - Extract and Transform from source data into Staging Area

ETL Studio
- Data Surveyors

Staging
- SAS
- SAS
- SAS

ETL Studio transformations
- Data Quality plugins

Finance

Customers

Suppliers

Step #3 - Transform into dimensional model

HCM
The importance of staging areas

- Support an incremental approach
- A place for various data validation
  - Data cleansing
  - Data content validation
  - Data integrity
  - Key validation
- Slowly changing dimension management
### Load Summary

<table>
<thead>
<tr>
<th>Table</th>
<th>Last Update</th>
<th>Load No.</th>
<th>Read</th>
<th>Unchanged</th>
<th>Changed</th>
<th>New</th>
<th>I.Del</th>
<th>Created</th>
<th>I.Del Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETAIL</td>
<td>01/01/1997</td>
<td>1</td>
<td>647</td>
<td>0</td>
<td>0</td>
<td>647</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DIMAIRCRAFT</td>
<td>01/01/1997</td>
<td>1</td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DIMROUTE</td>
<td>01/01/1997</td>
<td>1</td>
<td>12,213</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Enterprise Table Record Summary

<table>
<thead>
<tr>
<th>Table</th>
<th>Total</th>
<th>Open</th>
<th>Closed</th>
<th>I.Deletes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETAIL</td>
<td>647</td>
<td>647</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DIMAIRCRAFT</td>
<td>112</td>
<td>112</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FACTSALES</td>
<td>12,213</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Enterprise Table Size (Bytes)

<table>
<thead>
<tr>
<th>Table</th>
<th>01/01/1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETAIL</td>
<td>57,344</td>
</tr>
<tr>
<td>DIMAIRCRAFT</td>
<td>193,536</td>
</tr>
<tr>
<td>FACTSALES</td>
<td>1,722,368</td>
</tr>
<tr>
<td>Total</td>
<td>1,973,248</td>
</tr>
</tbody>
</table>
Data Warehousing in 2003

Step #1 - Extract and Transform from source data into Staging Area

ETL Studio
Data Surveyors

Staging

Source
Excel
SAS
SAP
Oracle
PeopleSoft

Finance

More ETL flows

Central Warehouse

Step #2 - Data Quality
Data Validation
Slowly changing dimensions

ETL Studio transformations
Data Quality plugins

Step #3 - Transform into dimensional model

HCM

Suppliers

Finance

Customers

SAS

SAS

SAS

Finance

Customers

SAS

SAS

SAS

SAS

HCM
What is a central data model?

- Normalized relational information
- Foundation Layer
- Example: Orion Star sample data model
Metadata import

- Common Warehouse Metamodel (CWM)
- Meta Integration Model Bridge (MIMB)
- Registers lots of metadata at once
Data Warehousing in 2003

Step #1 - Extract and Transform from source data into Staging Area

ETL Studio Data Surveyors

Step #2 - Data Quality Data Validation Slowly changing dimensions

ETL Studio transformations Data Quality plugins

Step #3 - Transform into dimensional model

More ETL flows

Finance

Central Warehouse

SAS

SAS

SAS

HCM

Suppliers

Customers

SAP

Oracle

PeopleSoft

Step #1 - Extract and Transform from source data into Staging Area

ETL Studio Data Surveyors

Step #2 - Data Quality Data Validation Slowly changing dimensions

ETL Studio transformations Data Quality plugins

Step #3 - Transform into dimensional model
Working with dimensional models

- Purpose-oriented data mart
- Dimension tables
- Fact tables with measures
Physical Data Model of Order Star Schema

**Customer_Dim**
- **Customer_Id**: INTEGER
  - **Customer_Country**: CHARACTER(2)
  - **Gender**: CHARACTER(1)
  - **Customer_Name**: CHARACTER(40)
  - **Customer_Firstname**: CHARACTER(20)
  - **Customer_Lastname**: CHARACTER(30)
  - **Birthday**: DATE
  - **Customer_Type**: CHARACTER(40)
  - **Customer_Group**: CHARACTER(40)
  - **Age**: DECIMAL(3)

**Organization_Dim**
- **Employee_Id**: INTEGER
  - **Country**: CHARACTER(2)
  - **Employee_Name**: CHARACTER(40)
  - **Group**: CHARACTER(40)
  - **Section**: CHARACTER(40)
  - **Department**: CHARACTER(40)
  - **Company**: CHARACTER(40)
  - **Job_Title**: CHARACTER(25)
  - **Gender**: CHARACTER(1)
  - **Salary**: DECIMAL(12,2)
  - **Birthday**: DATE
  - **Emp_Hire_Date**: DATE
  - **Emp_Term_Date**: DATE
  - **Start_Date**: DATE
  - **End_Date**: DATE

**Time_Dim**
- **Date**: INTEGER
  - **Weekday_Name**: SMALLINT
  - **Weekday_Eu**: SMALLINT
  - **Week_Name**: CHARACTER(40)
  - **Month_No**: SMALLINT
  - **Month_Name**: CHARACTER(7)
  - **Quarter**: CHARACTER(6)
  - **Year**: CHARACTER(4)

**Geography_Dim**
- **Street_Id**: INTEGER
  - **Street_Name**: CHARACTER(40)
  - **Zipcode**: CHARACTER(10)
  - **City**: CHARACTER(40)
  - **County**: CHARACTER(40)
  - **Province**: CHARACTER(40)
  - **Region**: CHARACTER(40)
  - **State**: CHARACT(30)
  - **Country**: CHARACTER(2)
  - **Continent**: CHARACTER(40)

**Product_Dim**
- **Product_Id**: INTEGER
  - **Product_Name**: CHARACTER(45)
  - **Supplier_Id**: INTEGER
  - **Product_Group**: CHARACTER(40)
  - **Product_Category**: CHARACTER(40)
  - **Product_Line**: CHARACTER(40)

**Order_Fact**
- **Customer_Id**: INTEGER (FK)
- **Employee_Id**: INTEGER (FK)
- **Street_Id**: INTEGER (FK)
- **Product_Id**: INTEGER (FK)
- **Order_Date**: DATE (FK)
- **Order_Id**: INTEGER
  - **Order_Type**: SMALLINT
  - **Delivery_Date**: DATE
  - **Amount**: DECIMAL(5)
  - **Price**: NUMERIC(8,2)
  - **Cost_Price_Pr_Unit**: DECIMAL(12,2)
  - **Promotion**: DECIMAL(5,2)
Next Steps

- **OLAP**
  - Design cubes for common queries/reports
  - Star schema is good basis for this

- **Semantic layer for end user reporting**
  - SAS Information Map Studio
Cube Designer - Finish

Review the cube definition below and select Finish to save the cube definition in the metadata.
Select whether the cube is to be created or not. You may also choose to save the full set of PROC CLAP code to a file.

General:
- Cube Name: OrionStar
- Cube Description: Orion Star Cube
- Repository: Enterprise
- Olap Schema: SASApplicationServer3 - OLAP Schema
- Path: C:/projects/seugi21/level/SharedDataStorage/olap
- Work Path:

Input:
- Repository: Enterprise
- Star Schema Fact Table: ORDER_FACT

- Save the metadata, delete the existing cube, and recreate the cube.
- Save the metadata, delete the existing cube but do not recreate the cube.

Save PROC CLAP Code
SAS® Information Map Studio

HR Data

HR Information Map

- User's View of Map
- Employee ID
- Reason Code
- Duration
- Start Date
- End Date
- Avg. Duration
- Reason Code
- Short Absences
- Long Absences
- Reason Codes
  - FMLA
  - Vacation
  - Health
  - FMLA (Paid)
  - FMLA (Unpaid)
  - Military
  - Personal (Unpaid)
  - Short Term Disability
  - Long Term Disability

Copyright © 2003, SAS Institute Inc. All rights reserved.
How to develop

- Document it!
- Multi-developer projects
- Check-in, check-out
- Audit history
- Dev, Test, Prod environments
- Reusable added code
How to deploy

- Execute ETL in a production environment
- Easily managed deployment
- Schedule for periodic runs
  - Uses integrated LSF Scheduler from Platform Computing
  - Ready for use with standard schedulers too
    - cron (unix), at (Windows)
The visual below shows the data flow from the selected item, both from the sources that feed into it and the targets that are fed from it.
Successful Data Warehouse implementations

- Shift corporate culture to rely upon DWH
- Align planning with corporate strategy
- Provide adequate resources to maintain/expand
- Demonstrate return-on-investment
- Build executive support
- Maintain complete and up-to-date content
- Employ correct technology/Upgrade at best freq.
- Plan sufficient training
Successful Data Warehouse implementations

- Shift corporate culture to rely upon DWH 92%
- Align planning with corporate strategy 75%
- Provide adequate resources to maintain/expand 75%
- Demonstrate return-on-investment 50%
- Build executive support 42%
- Maintain complete and up-to-date content 42%
- Employ correct technology/Upgrade at best freq. 42%
- Plan sufficient training 33%
Summary

- Planning is key
- Architected environment
- Don’t be afraid of ERP data
- Integrate cleansing and validation
- Don’t underestimate the value of data modeling
- Metadata
- Be sure the DWH is used