Data Warehouse: Methodology and Tools

Concepts, Architectures and Products

Overview

- The Process of Planning and Building a Warehouse
- Data Warehouse Architecture (revisited)
- Classification of Tools
- Focus: OLAP Tools
  - Multidimensional Data Modeling
  - OLAP Architectures
  - OLAP query languages
- Tool Demonstration
- Summary
OLAP Design Cycle

Using the Data Warehouse

Requirement Analysis

Conceptual Design (Implementation Independent)

Logical + Physical Design (e.g. Product specific)

Implementation

Data Warehouse Architecture

Data Analysis

Reporting, OLAP, Data Mining

Data Storage

Repository

Data-Migration

Middleware (Populations-Tools)

Operational Data Sources
Classification of Tools

- Frontend Tools
- Data Storage Tools (Databases)
- ETL Tools
  - Extraction
  - Transformation
  - Loading
- Repository Systems
  - Metadata Storage

Repository Systems

- Manage Different Kinds of Metadata
  - Business Metadata
    - E.g. How is revenue computed
  - Technical Metadata
    - When was data last loaded from which system
    - Data model for OLTP and OLAP databases
- Functionality
  - Communication 'hub' for different tools
  - Guides user exploration
  - Guides development process
  - Impact analysis
- E.g. Viasoft Rochade, Softlab Enabler,...
ETL Tools

- **Extraction**: Range of Supported Data Sources
  - Mainframe legacy databases
  - COBOL Files
  - Relational Databases
  - Filebased data storage (Excel, Word, XML, ...)
- **Transformation**
  - (Graphical) Specification of Transformation Rules (Expressive Power)
- **Loading**
  - Ability to use database features (e.g., bulk loading)
- **Process Management**
  - Scheduling, Monitoring, Error Handling
- **Informatica PowerMart, Hummingbird Genio, Acta...**

Databases for DW

- **Special Indexing Techniques**
  - Multidimensional Indexes
  - Bitmap Indexes
  - Foreign Column Indexes
- **Support for Materialized Views (Preaggregation)**
- **Special Analytical Capabilities (e.g., SQL Extensions)**
  - Top N
  - Ranking
- **Bulk Loading Capabilities**
  - Offline, No concurrency control
Frontend Tools

Why did it happen?
What happened?
What will happen?

Additional Benefit

Reporting
Interactive OLAP
Ad hoc-Queries
Data Mining

Number of Users

The User’s view (OLAP Tool)
Multidimensional OLAP (MOLAP)

- specialized database technology
- multidimensional storage structures
- E.g. Hyperion Essbase, Oracle Express, Cognos PowerPlay (Server)
- Query Performance
- Powerful MD Model
- write access
- Database Features
  - multiuser access/backup and recovery
  - Sparsity Handling -> DB Explosion

Relational OLAP (ROLAP)

- idea: use relational data storage
- star (snowflake) schema
- E.g. Microstrategy, SAP BW
- advantages of RDBMS
  - scalability, reliability, security etc.
- Sparsity handling
  - Query Performance
  - Data Model Complexity
  - no write access
**Client (Desktop) OLAP**

- proprietary data structure on the client
- data stored as file
- mostly RAM based architectures
- E.g. Business Objects, Cognos PowerPlay

+ mobile user
+ ease of installation and use
- data volume
- no multiuser capabilities

---

**DW Integration**

- MOLAP
- ROLAP
- Client-OLAP
- Multidim. Database
- ROLAP-Engine
- DW-DB (mostly relational)
Combining Architectures I

Drill through

- highly aggregated data
- dense data
- 95% of the analysis requirements

Multidim. Database

- detailed data (sparse)
- 5% of the requirements

Relational Database

Combining Architectures II

Hybrid OLAP (HOLAP)

- equal treatment of MD and Rel Data
- Storage type at the discretion of the administrator
- Cube Partitioning

HOLAP System

Meta Data

Multidim. Storage

Relational Storage
OLAP Standards

- Idea: define interface between client and server
- Benefit: Component oriented architectures
- Proposal 1: OLAP Council
  - union of OLAP Tool producers
  - not implemented so far (even by the council members)
- Proposal 2: Microsoft - OLEDB for OLAP (shot ODBO)
  - standardizes a data model and an MD query language (MDX)
  - specification contains lots of optional functionality
  - all major vendors committed themselves to the standard
  - will be the de facto standard

Practical Case Study

Building a Warehouse
The Modeling Process

- Which business process is being modeled?
- What is the subject of analysis (fact) and what is being measured?
- On what granularity level is active analysis being done?
- Which properties (dimensions) determine the measures?
- Which different levels of aggregation are meaningful?
- What additional information is needed for the different levels?
- What is the variability and the cardinality of the dimensions?
Facts

- Fact = Subject of Analysis: Sales
- Measures = Attributes describing facts: Quantity, Price
- Derived Measures: Profit
- Additivity of Measures:
  - globally additiv: Quantity
  - additiv for some dimensions: Items in stock
  - additiv resp. to plants:
  - not additiv w.r.t. time: not additiv w.r.t. time
  - not additiv at all: profit margin

Dimensions

- Dimensions = static structure of business information
- Used for navigating the data space
- Choosing the necessary granularity
- Dimension Members = Instances of a dimension
  - e.g. 8.12.1997 and Juli 1997 are members of dimension “time”
- Structuring Dimension
  - using different dimension levels (hierarchies)
  - using descriptive attributes
**Simple Hierarchies**

- **Month:** Januar 99, Februar 99, März 99, April 99, Mai 99, Juni 99, Juli 99, August 99, Sept. 99
- **Quarter:** 1. Quartal 99, 2. Quartal 99, 3. Quartal 99
- **1/2 Year Period:** 1. Halbjahr 99, 2. Halbjahr 99
- **Year:** 1999

**Dimension Level**

**Unbalanced Hierarchies**

- **Plant/Site:** Plant 1, Plant 1, Plant 0815
- **Business Unit:** Bu 1, Bu 2
- **Business Division:** Div A, Div B
- **Great Outdoors:**

© 2001 FORWISS, Carsten Sapia - sapia@forwiss.de
Criteria for a ‘good’ MD Design

- dimensions should be independent
- dimensionality of a cube should be max. 7-8 dimensions
  - interpretation of results is difficult for a large number of dimensions
- hierarchies should have a fan-out of max. 30
  - long drill-down times
  - large drill-down results
  - insert additional levels for structuring purposes (e.g. insert state between city and country)

Graphical Notation (ME/R)

A Fact and its measures

.. is characterized by dimensions

A Dimension Level with attributes

..can be classified according to...
Example Data Model

Cognos PowerPlay- Architecture
Logical+Physical Design

Using the Data Warehouse

Requirement Analysis

Conceptual Design (Implementation Independent)

Logical + Physical Design (tool specific)

Implementation

Practical Demonstration
Summary and Conclusions

- Multidimensional modeling is performed on different levels
  - conceptual model (tool independent level) following requirement analysis
  - logical and physical design before implementation
- Distinction between two types of data
  - quantifying data: measures, cells of the cube, fact table
  - qualifying data: properties, dimensions, dimension tables
- Hierarchical structures of dimensions can be complex
- ME/R notation can be used to document conceptual models
- Several ways to map an MD model to a relational DB

Canonical Query (I)

Restriction Element

Result Granularity

Query Result

Result Measures

m1 | m2
A
B
Canonical Query (II)

Canonical Query Definition

Result Measures

\[ m_1, \ldots, m_k \]

Restriction Elements

\[ r_1, r_2, \ldots, r_n \]

Result Granularity

\[ g_1, g_2, \ldots, g_n \]

SQL:

```
SELECT g_1, \ldots, g_n, \text{aggr}(m_1), \ldots, \text{aggr}(m_k)
FROM FactName, Dim_1, \ldots, Dim_n
WHERE Dim_1.level(r_1) = r_1 \text{ AND } \ldots \text{ AND } Dim_n.level(r_n) = r_n
\text{AND Dim}_1.d_1 = \text{FactName}.d_1 \text{ AND } \ldots \text{ AND } \text{Dim}_n.d_n = \text{FactName}.d_n
GROUP BY g_1, \ldots, g_n
```