Advanced Database Systems

Course Outline

- Introduction to Data Warehousing
- Meeting Business Needs
- Data Warehouse Concepts and Terminology
- Driving Implementation Through a Methodology
- Planning for a Successful Warehouse
- Analyzing User Query Needs
- Modeling the Data Warehouse
- Planning Warehouse Storage
- Building the Warehouse
- ETL
- Business Intelligence
- Web-enabling Warehouse

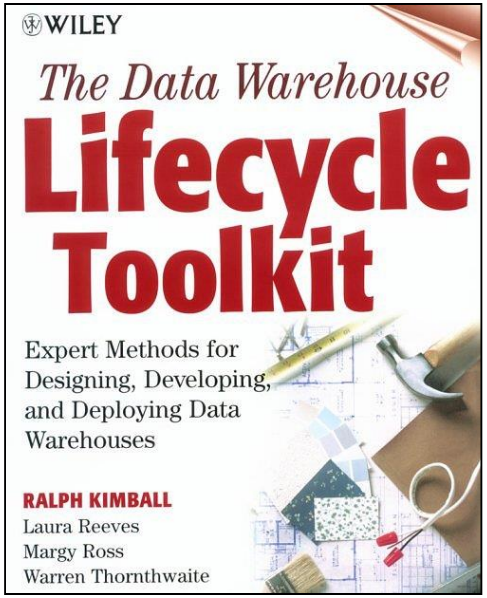
Introduction to Data Warehouse

Objective

- To provide basic understanding about data warehouse concepts
- In a way that everyone involved in data warehouse project have common understanding about data warehouse concepts
- So that the data warehouse project team can effectively communicate under the same understanding

Acknowledgement

This presentation is summarized from the first chapter of ‘The data warehouse lifecycle toolkit : expert methods for designing, developing, and deploying data warehouses’ by Ralph Kimball and others.



WILEY

The Data Warehouse

Lifecycle Toolkit

Expert Methods for Designing, Developing, and Deploying Data Warehouses

RALPH KIMBALL
Laura Reeves
Margy Ross
Warren Thornthwaite

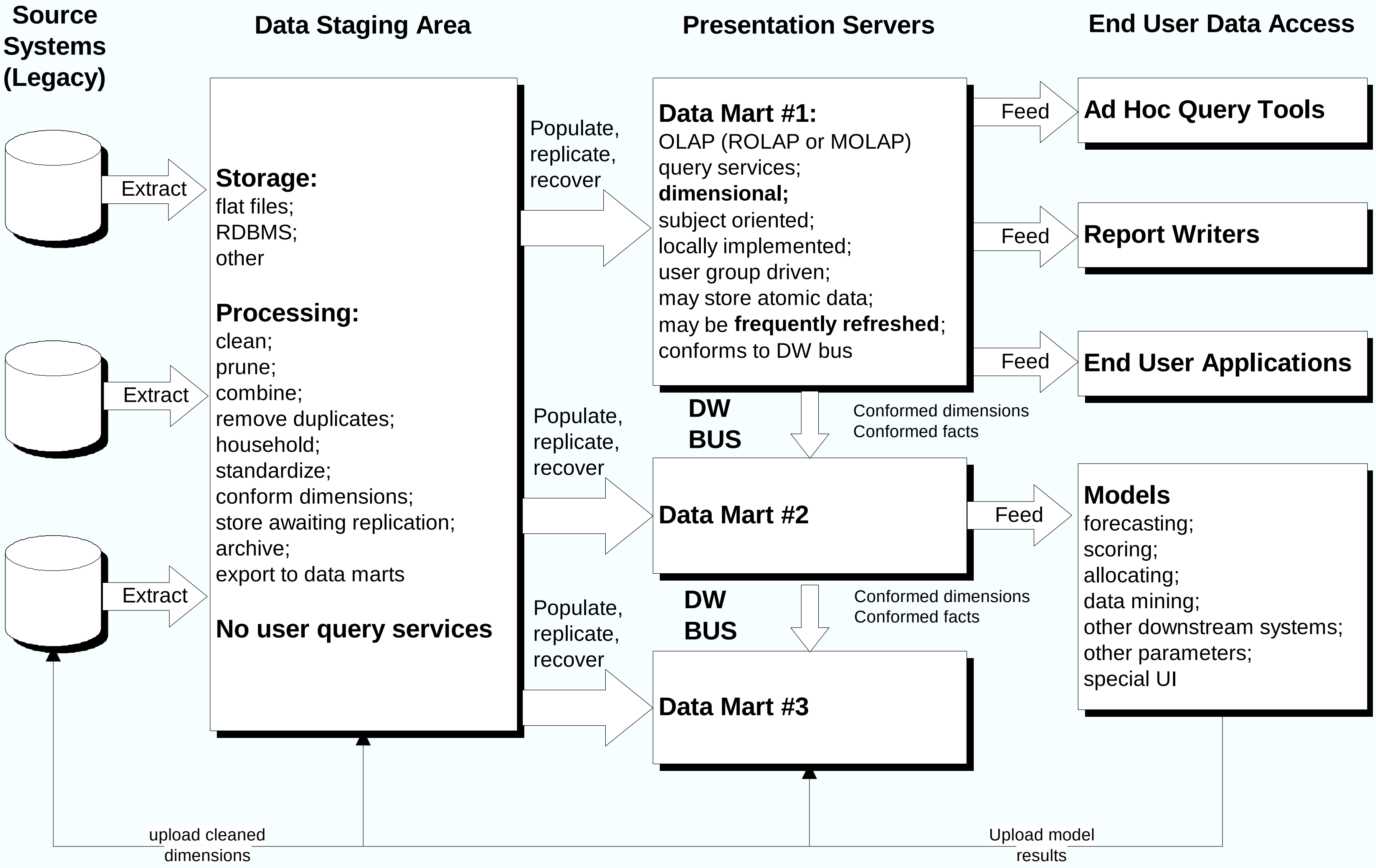
Agenda

- The goals of a data warehouse
- Basic elements of the data warehouse
- Basic processes of the data warehouse

The goals of a data warehouse

- Makes an organization’s information accessible
- Makes the organization’s information consistent
- Is an adaptive and resilient source of information
- Is a secure bastion that protect our information asset
- Is the foundation for decision making

Basic elements of the data warehouse



Source Systems (Legacy)

Extract

Extract

Extract

Data Staging Area

Storage:
flat files;
RDBMS;
other

Processing:
clean;
prune;
combine;
remove duplicates;
household;
standardize;
conform dimensions;
store awaiting replication;
archive;
export to data marts

No user query services

Populate, replicate, recover

Populate, replicate, recover

Populate, replicate, recover

Presentation Servers

Data Mart #1:
OLAP (ROLAP or MOLAP)
query services;
dimensional;
subject oriented;
locally implemented;
user group driven;
may store atomic data;
may be **frequently refreshed**;
conforms to DW bus

DW BUS

Conformed dimensions
Conformed facts

Data Mart #2

DW BUS

Conformed dimensions
Conformed facts

Data Mart #3

End User Data Access

Feed

Ad Hoc Query Tools

Feed

Report Writers

Feed

End User Applications

Feed

Models
forecasting;
scoring;
allocating;
data mining;
other downstream systems;
other parameters;
special UI

upload cleaned dimensions

Upload model results

Source System

- An operational system of record whose function it is to capture the transactions of the business
- Queries against source systems are narrow, account-based query that are part of the normal transaction flow and severally restricted
- Maintain little historical data

Data Staging Area

- A storage area and set of processes that clean, transform, combine, deduplicate, household, archive, and prepare source data for use in the data warehouse
- It is more likely to be spreaded over a number of machines
- It does not provide query and presentation services

Presentation Server

- The target physical machine on which the data warehouse data is organized and stored for direct querying by end users, report writers, and other applications
- Data should be presented and stored in a dimensional framework

Dimensional Model

- A specific discipline for modeling data that is an alternative to entity-relationship (E/R) model
- Main components are fact tables and dimension tables
- Better for decision support than E/R model

Data Mart

- A logical subset of the complete data warehouse
- A data warehouse is made up of the union of all its data marts
- Without conformed dimensions and conformed facts, a data mart is a stovepipe
- Data mart can contains not only the summary data but also atomic data

Operational Data Store (OSD)

- Served as the point of integration for operational systems
- Important for legacy systems that grew up independently
- More on real-time, account-based data query
- Should not be considered as part of the decision support system

OLAP, ROLAP, MOLAP

- OLAP – Online Analytic Processing
  - The general activity of querying and presenting text and number data from data warehouse in a dimensional style
- ROLAP – Relational OLAP
  - A set of user interface and application that give a relational database a dimensional flavor
- MOLAP – Multidimensional OLAP
  - A set of user interface, application and proprietary database technology that have a strong dimensional flavor

End users data access

- End user application/Report Writers
  - A collection of tools that query, analyze, and present information targeted to support a business need
- Ad hoc query tool
  - A data access tool that invite the user to form their own queries by directly manipulating relational tables and joints

End users data access (cont)

- Modeling Application
  - A sophisticated kind of data warehouse client with analytic capabilities that transform or digest the output from data warehouse
    - Forecasting models
    - Behavior scoring models
    - Allocation models
    - Data mining tools

Meta Data

- All of the information in the data warehouse that is not the actual data itself
- Could be spread across several machines
- Could be in various formats and diversity in the usage

Basic processes of the data warehouse

- Extracting
- Transforming
- Loading and indexing
- Quality assurance checking
- Release/publishing

Basic processes of the data warehouse (cont.)

- Updating
- Querying
- Data feedback
- Auditing
- Securing
- Backing up and recovering

Extracting

- First step of getting data into data warehouse environment
- Reading and understanding the source data
- Copying the parts that are needed from source system to the data staging area for future work

Transforming

- Many possible transforming steps includs
  - Cleaning : correct misspells, resolve domain conflict, deal with missing data elements
  - Purging : select fields from legacy data that are not useful for the data warehouse
  - Combining data sources : match the key
  - Creating surrogate keys : enforce referential integrity between dimension tables and fact tables
  - Building aggregates : for better performance

Loading and Indexing

- Usually replicating the dimension and fact tables from data staging area to each of the data mart
- Usually perfom using bulk loading facility of the data mart
- Indexing should be done for query performance

Quality Assurance Checking

- After data loading/indexing, last step before publishing
- Can be done by running a comprehensive exception report over the entire set of newly loaded data
- The exception report could be built using the report writer facility of the data mart

Release/Publishing

- When each data mart has been freshly loaded and quality assured, the user community must be notified that the new data is ready
- Also including the communication of the changes in underlying dimension and new assumptions introduced into the measure or calculated facts

Updating

- Modern data marts may be updated, sometimes frequently
- “Managed load updates”, not transactional updates
- Triggers of the update includes
  - Data correction
  - Changes in labels
  - Changes in hierarchies
  - Changes in status
  - Changes in corporate ownership

Querying

- Means all the activities of requesting data from a data mart including
    - Ad hoc query by end users
    - Requests from models
    - Data mining
- Take place in presentation server, never takes place in data staging area

Data Feedback

- May include the upload of cleaned dimension descriptions fom data staging area to legacy source systems
- May include the upload of the results of a complex query or a model run or a data mining analysis back into data mart

Auditing

- Critically important to know where the data come from and what were the calculation performed
- Special audit record should be created during the extraction and transformation processes

Securing

- Data warehouse security must be manage centrally
- Users must be able to access all authorized data marts with a single sign on
- Development of the Internet increases the need of data warehouse security architect role