Zusammenfassung zur Vorlesung Data Warehousing

gehalten im WS 2003

von O. Univ. Prof. Dr. A Min Tjoa

rn

Februar 2004

Contents

1 The Compelling Need for Data Warehousing

Enterprises do need strategic information for many reasons. They often do have mountains of data which is spread across many types of incompatible structures and systems. Hence the available data is not readily usable for strategic decision making. OLTP (online transaction processing), the operational data and the informational data needed for analysis can be compared as follows:

- Operational systems
  - Take an order
  - Process a claim
  - Make a shipment
- Informational (decision support) systems
  - Show me the top selling products
  - Show me the problem regions
  - Show the highest margins
  - Tell me why (drill down)
  - Let me see other data (drill across)
  - Alert me when a district sells below target

Table (1) on page (4) lists the most important differences between transaction and analysis data. Processing requirements in the data warehouse environment:

	operational	*informational*
Data Content	Current Values	Archived, derived, summarized
Data Structure	Optimized for transactions	Optimized for complex queries
Access Frequency	High	Medium to low
Access Type	Read, update, delete	Read
Usage	Predictable, repetitive	Ad hoc, random, heuristic
Response Time	Sub-seconds	Several seconds to minutes
Users	Large number	Relatively small number

Table 1: Differences between operational and informational data

1. Running of simple queries and reports against current and historical data
2. Ability to perform 'what if' analysis in many different ways
3. Ability to query, step back, analyze, and then continue the process to any desired length
4. Spot historical trends and apply them for future results

1.1 Data Warehouse Defined

The data warehouse is an informational environment that

- Provides an integrated and total view of the enterprise
- Makes the enterprise's current and historical information easily available for decision making
- Makes decision-support transactions possible without hindering operational systems
- Presents a flexible and interactive source of strategic information

Data warehouse is really a simple concept:

- Take all the data you already have in the organization (e.g. from the operational systems)
- Include data from outside where necessary, such as industry benchmark indicators
- Clean, transform and store it, and then
- Provide useful strategic information
- a data warehouse is an environment, NOT a product (rather a blend of many technologies)

2 Data Warehouse: The Building Blocks

A data warehouse is a

- subject oriented,
- integrated,
- nonvolatile, and
- time variant

collection of data in support of management's decisions. - Bill Inmon
The data in the data warehouse is:

- Separate
- Available
- Integrated
- Time stamped
- Subject oriented
- Nonvolatile
- Accessible

2.1 Data Warehouses vs. Data Marts

Data warehouses contain an enterprise-wide view of the data, whereas data marts contain the department specific views of the processes.

- Top down approach
    - Build the data warehouse for the whole enterprise
        - Advantages
            - Enterprise view of data
            - Single, central storage of data
            - Iterations may lead to quicker results for the top down approach
        - Disadvantages
            - Takes longer to build
            - High risk to failure
            - Needs high-level of cross-functional skills
- Bottom-up approach
    - Start to build data marts
        - Advantages
            - Faster and easier implementation
            - Less risk of failure
            - Can schedule important data marts first
        - Disadvantages
            - Each department has his own narrow view of the data
            - Redundant data in every mart
            - Unmanageable interfaces

2.2 Components of a Data Warehouse

- Source data component
    - Choose segments of data from the different operational systems
    - Internal data (data held in ’private’ files)
    - Archived data (historical snapshots of data from operational systems)
    - External data (data from external sources)
- Data staging component (ETL)
    - Data extraction
    - Data transformation
        - Cleaning

        - Standardization
        - Combining pieces of data from different sources
    - Data loading

- Data storage component
    - A separate repository (from operational data)
    - Large volumes of historical data for analysis (read only)

- Information delivery component
    - Deliver data from the warehouse to different users

- Metadata component
    - Data about the data in the warehouse

- Management and control component
    - On top of other components
    - Coordinates services and activities within the data warehouse
    - Moderates the information delivery to the users
    - Management and control functions
    - ...

3 Trends in Data Warehousing

Data Warehousing is becoming mainstream and revolutionizes the way people perform business analysis and how people make strategic decisions in every industry. The growing market is flooded with numerous products for data modeling, data acquisition, data quality, data analysis, ...

3.1 Significant Trends

- Data visualisation
    - More chart types
    - Interactive visualisation

- Use of parallel processing (hardware or software)

- Query tools
    - Flexible presentation (easy to use and able to present results online and on reports in many different formats)

    - Aggregate awareness (automatically route queries to the aggregate tables when aggregate results are desired)
    - Crossing subject areas (automatically cross over from one subject data mart to another)

- Data fusion (merging of data from disparate sources)

- Data warehousing and ERP

    - Data in ERP packages
    - Integrating ERP and data warehouse
    - Integration Options (how to combine e.g. first into ERP then into warehouse et vice versa)

- Integration of knowledge management

    - involves integration of unstructured data (BLOB et cetera)

- Data Warehousing and customer relationship management

- Active data warehousing (one-on-one service)

- Emergence of standards (multi-vendor products have to cooperate)

- Web enabled data warehouse (bring data warehouse to the web)

4 Planning and Project Management

4.1 Overview

- Many organizational changes for enterprise-wide information analysis

- Until now each department and each user 'owned' their own data

- You may uncover problems with the production system when building the data warehouse

- Planning is essential

4.2 Consideration of Key Issues

- Value and Expectations

    - Be sure that a data warehouse is the most viable solution
    - Make a list of realistic benefits and expectations

- Risk Assessment

    - How much money will go down the drain in case of failure?
    - What are the risks without having a data warehouse?
    - Risk assessment is company specific

- Top-down or Bottom-up
    - Is there a need to quickly install a data warehouse?
    - Do you have the large resources needed to build a corporate-wide data warehouse first

- Build or Buy
    - How much of the data marts should you build yourself?
    - How much of these may be composed of ready-made solutions

- Single Vendor or Best-of-Breed
    - A single vendor solution has advantages (integration, look and feel, . . .)
    - Only a few vendors offer fully integrated solutions (e.g. IBM, NCR)

Let business requirements drive your data warehouse, not technology!

4.3 Justifying Your Data Warehouse

Top management support is essential for introducing a data warehouse. Many companies are able to introduce data warehousing without a full cost justification analysis (e.g. in case of competitive pressures). However, not every company's top management is so easy to please. The typical approaches for justifying the introduction of a data warehouse are as follows:

- Calculate the current technology cost and compare to the costs for the data warehouse
- Calculate the business value of the proposed data warehouse and compare to the costs

4.4 The Initial Stage of the Data Warehouse Project

Data warehouse projects are different from projects building the transaction process systems, they tend to be more complex, have a broader scope, and involve many different technologies. Involve the users in every stage of the development. There should also be a stage of assessment of readiness for the data warehouse. A readiness assessment report for the data warehouse project includes:

- Lower the risks of big surprises
- Reassess corporate commitment

- Review the project scope and size
- Identify critical success factors
- Restate user expectations

4.4.1 The Project Team

- Several trained and specially skilled persons needed
- Two things can break a project
  - Complexity overload
  - Responsibility ambiguity
- Each person should be given specific responsibilities of a particular role based on her skill and experience level
- Identify roles needed
- Assign individual persons to the team roles (do not forget about users)

Data warehousing projects are not yet standardized as fas as the job titles go. Important members of the project team are:

- Executive sponsor
  - Direction, support
- Project manager
  - Assignments, monitoring, control
- Data modeler
  - Relational and dimensional modeling
- Data warehouse administrator
  - DBA functions
- Development programmer
  - In-house programs and scripts
- ...

As said before users should participate in any of the project stages.
Possible scenarios of failure are:

- Data basement(poor quality data without proper access)

- Data jail house (user can not reach the data)
- Data cottage (stand-alone, fragmented, island data mart)
- ......

Key success factors are:

- Ensure long-term support from the executive sponsors
- Get the users enthusiastically involved throughout the project
- Remember: architecture first, then tools, then products
- ...

5 Defining the Business Requirements

As a data warehouse is an information delivery system, it is important to concentrate on what information the users really need. Although users often can not precisely define which information they need, one can check the industry's best practices or try gather information in interviews (also on how the users think about the business, et cetera).

5.1 Dimensional Nature of Business Data

Users think in terms of business dimensions, hence those dimensions are important while collecting requirements. Different users have different views of the data:

- Marketing vice president
  - How much did my new product generate (by month, ...)
- Marketing manager
  - Give me sales statistics (by products, by month, ...)
- Financial controller
  - show me expenses (by month, ...)

5.2 Information Packages - a New Concept

- Information subject: Sales Analysis
- Dimensions:
  - Time
  - Locations
  - Products

  - ...

- Hierarchies
  - Year
  - Country
  - Product class
  - ...
- Measured facts
  - Forecast sales
  - Budget sales
  - ...

5.3 User Interviews

- Interviews
  - Two to three persons at a time
  - Also encourage users to prepare for the interview
  - Easy to schedule
- Group sessions
  - Up to twenty persons at a tame
  - Not good for initial data gathering
  - Need to be very well organized

Interviews should be based on user profiles and make use of interview techniques to get the information that is wanted. JAD centers are another possibility (several users meet for discussion workshops for several days) .

6 Requirements as the Driving Force for Data Warehousing

Planning for the architecture involves the determination of the size and content of each component.

- Source data
  - Operational source systems
  - Computing platforms, operating systems, databases, files
  - Departmental data such as spreadsheets or other files
  - External data sources

- Data staging
  - Data mapping between data sources and staging area data structures
  - Data transformations
  - Data cleansing
  - Data integration
- Data storage
  - Size of extracted and integrated data
  - DBMS features
  - Growth potential
  - Centralized or distributed
- Information delivery
  - Types and numbers of users
  - Types of queries and reports
  - Front-end DSS applications
- Metadata
  - Operational Metadata
  - ETL metadata
  - End-user metadata
  - Metadata storage
- Management and control
  - Data loading
  - External sources
  - Alert Systems
  - End-user information delivery

6.1 Special Considerations

In the requirements definition phase there are some factors you need to pay special attention:

- ETL
  - Data extraction
    - Clearly identify all the internal data sources
  - Data transformation

        - Examine each data element to be stored in the data warehouse against the source data elements
    - Data loading
        - Determine how often each major group of data must be kept up-to-date in the data warehouse

- Data Quality
    - Identify potential sources of data pollution particularly in the early phase (see section (13) on page (31))

- Metadata
    - Identify needed Metadata at the beginning of the project (see section (9) on page (22))

6.2 Tools and Products

The data warehouse architecture should be based on requirements NOT on tools. Select the architecture based on requirements and THEN look for proper tools. Tools are available for the following functions:

- ETL
    - Middleware
    - Data extraction
    - Data transformation
    - ...
- Warehouse storage
    - Data marts
    - Metadata
- Information access/delivery
    - Report writers
    - Query processors
    - OLAP
    - Alert Systems
    - Data mining
    - ...

6.3 Data Storage Specifications and DBMS Selection

The data storage specification can be found by bottom-up or top-down approach:

- Top-down approach - define the storage specifications for
  - The data staging area
  - The overall corporate data warehouse
  - Any multidimensional databases for OLAP
- Bottom-up approach - define the storage specifications for
  - The data staging area
  - Each of the confirmed data marts, beginning with the first
  - Any multidimensional databases for OLAP

Choose any DBMS that fills your needs regarding data load, metadata management, openness, type of queries, . . . Think about the storage size the data warehouse will need.

6.4 Tips for Requirements Definition

Another endless list of words.

7 The Architectural Components

The three major areas of the data warehouse architecture are shown in figure (1) on page (16). The data warehouse architecture consists of following distinct components:

- Different objectives and scope
  - Scope
    - The number and extent of the data sources
    - The data transformations and integration functions
    - Data granularity and data volumes
  - The impact of the data warehouse on the existing operational systems
- Data content
  - The 'read only ' data in the data warehouse is the primary component in the architecture
  - Keep data integrated from various sources (transformation, cleansing, integration)
  - Very high volumes of historical data
- Complex analysis and quick response

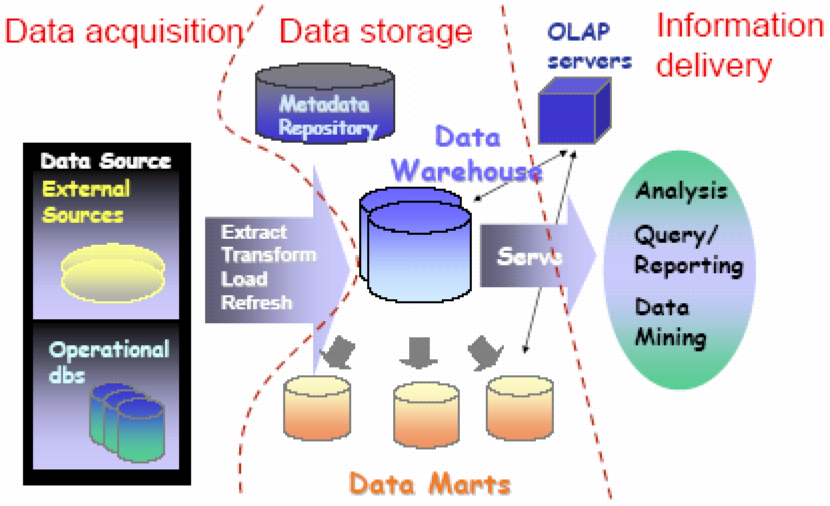

Figure 1: Data Warehouse Architecture

  - Support complex analysis of the strategic information by the users

- Flexible and dynamic
  - Business conditions keep on changing which calls for additional business requirements to be included in the data warehouse

- Metadata-driven
  - The metadata component interleaves with and connects the other components

7.1 Technical Architecture

The technical architecture is the complete set of functions and services provided within its components and NOT the set of tools needed to perform the functions and provide the services. The overall architecture includes the data acquisition technical architecture seen in figure (2) on page (17), the data storage technical architecture seen in figure (3) on page (18), and the information delivery technical architecture in figure (4) on page (19). Those architectures are going to be described in detail.

7.1.1 Data Acquisition

- Data extraction
  - Select the data sources
  - Generate automatic extract files from operational systems using replication and other techniques
  - Transport extracted files from multiple platforms

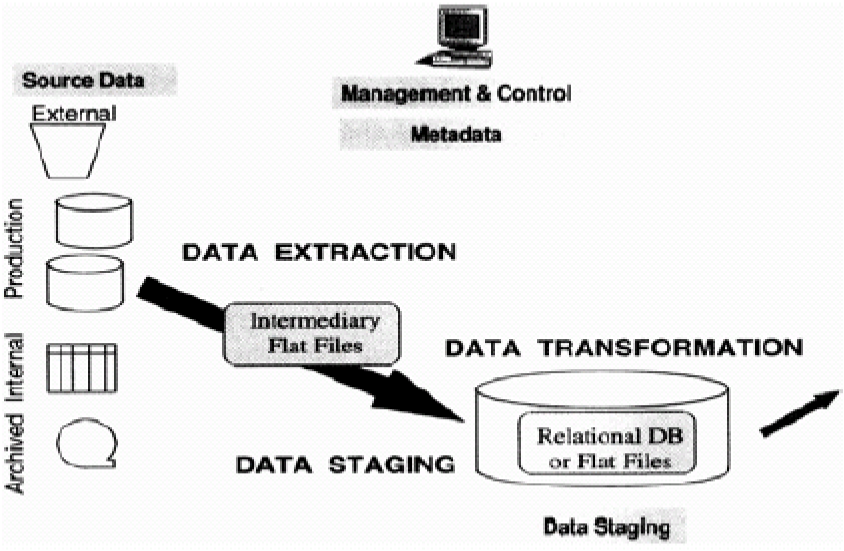

Figure 2: Data Acquisition: Technical Architecture

  - Reformat input from outside sources
  - Resolve inconsistencies for common data elements from multiple sources

- Data transformation
  - Map input data to data for the data warehouse repository
  - Clean data
  - Denormalize extracted data sources in order to fit the dimensional data model
  - Convert data types
  - Calculate and derive attribute values
  - Check for referential integrity
  - Aggregate data is needed
- Data staging
  - Provide backup and recovery for staging area repositories
  - Sort and merge files
  - Create and populate database
  - Consolidate datasets and create flat files for loading through DBMS utilities

7.1.2 Data Storage

- Load data for full refreshes of data warehouse tables

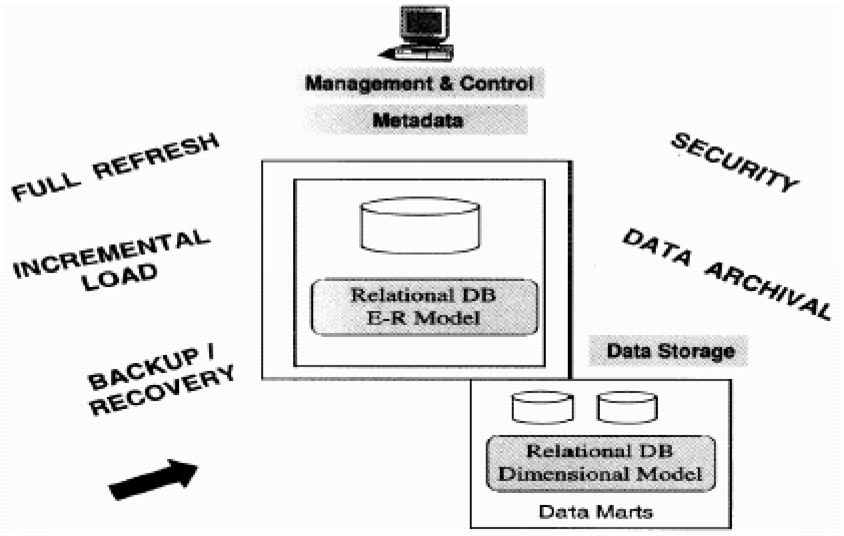

Figure 3: Data Storage: Technical Architecture

- Optimize the loading process
- Provide backup and recovery for the data warehouse database
- Provide security

7.1.3 Information Delivery

- Monitor user access to improve service and for future enhancements
- Allow users to browse data warehouse content
- Enable queries to be aware of aggregate tables for faster results
- Provide report generation
- Store results sets of queries and reports for future use
- Provide multiple levels of data granularity
- OLAP again

8 Infrastructure as the Foundation for Data Warehouses

8.1 Operational Infrastructure

- People
- Procedures

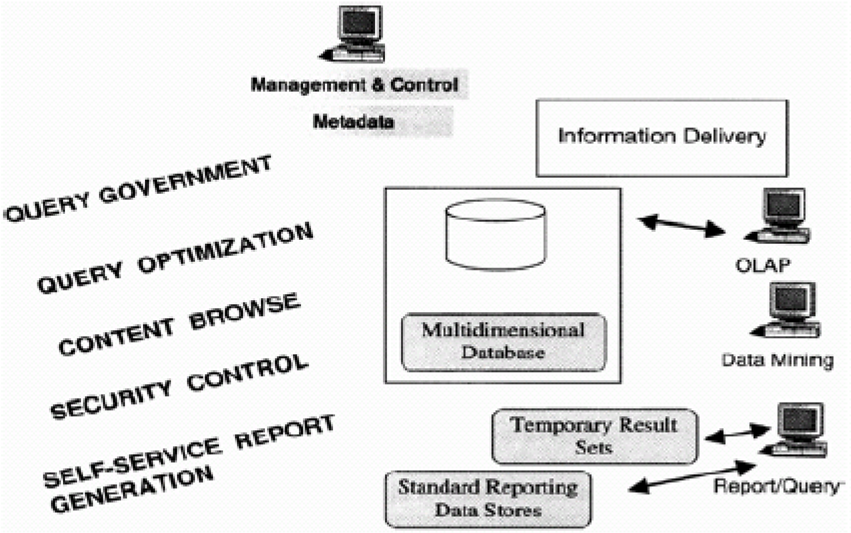

Figure 4: Information delivery: Technical Architecture

- Training
- Management software
- These are not the people and procedures needed to keep the data warehouse going NOT to develop it

8.2 Physical Infrastructure

- Hardware and operating system
  - Two principles:
    - Leverage as much of the existing physical infrastructure as possible
    - Keep the infrastructure as modular as possible
  - Criteria for hardware selection
    - Scalability
    - Support
    - ...
  - Criteria for operating system selection
    - Scalability
    - Security
    - ...

Using a single platform is not always an option, because of general legacy systems problems. Commonly one faces many different technologies to support the different needs for data warehousing.

8.3 Server hardware options

- Symmetric multi processing (SMP)
- Massively parallel processing (MMP)
- Clusters
- Non-uniform memory architecture (NUMA)

8.4 Selection of the DBMS

- Load balancing
- Parallel processing options
- Query governor
  - to anticipate and abort runaway queries
- Query optimizer
  - to parse and optimize user queries
- Query management
  - to balance the execution of different types of queries
- Load utility
  - for high-performance data loading, recovery, and restart
- Metadata management
  - with an active data catalog or dictionary
- Scalability
  - in terms of both number of users and data volume
- Extensibility
  - having hybrid extensions to OLAP databases
- Portability
  - across platforms
- Query tool APIs
  - for tools from leading vendors
- Administration
  - providing support for all DBA functions

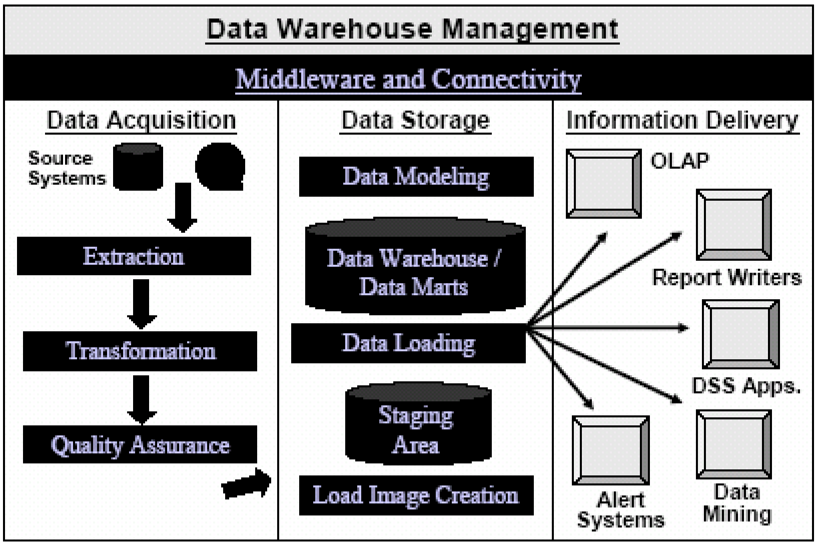

Figure 5: Tools for Your Data Warehouse

8.5 Architecture First, Then Tools

Figure (5) on page (21) shows which types of tools are used across the data warehouse architecture.

- Data modeling
  - Model both source and target systems
  - Provide forward/reverse engineering capabilities
  - Provide dimensional modeling capabilities
- Data extraction
  - Bulk extraction for full refreshes
  - Or change-based replication for incremental loads
- Data transformation
- Data loading
- Data quality
  - Assist in locating and correcting data errors
- Queries and reports
  - Allow users to produce graphic-intensive, sophisticated reports
  - Support users to formulate and run queries

- Online analytical processing
  - Allow users to run complex dimensional queries
- Alert systems
  - Get user's attention based on defined exceptions
- Middleware and connectivity
  - Transparent access to source systems and databases in different environments and on multiple platforms
- Data warehouse management
  - Assist data warehouse administrators in day-to-day management
  - Tools to track user queries

9 The Significant Role of Metadata

9.1 Why Metadata is Important

Three types of metadata:

- Operational metadata
  - Contains all of the information about the operational data sources
- Extraction and transformation metadata
  - Data about extraction from source systems
  - Data about transformation of data in the staging area
- End-user metadata
  - Metadata for end-users

Important questions about the data in the data warehouse:

- Are there any predefined queries?
- Is there information about unit sales and unit costs by product?
- How old is the data in the warehouse?
- From where did they get the data in the warehouse (which sources)?
- When was the last time fresh data was brought in?

Entity Name:	Customer
Alias Names:	Account, Client
Definitions:	A person that purchases goods from the company.
Remarks:	Customer entity includes regular, current and past customers.
Create Date:	January 15, 1999
Last Update Date:	January 17, 2002
Update Cycle:	weekly
Latest (full) refresh:	February 1, 2001
...	...

Table 2: Metadata Example for Customer Entity

Metadata describes all the relevant aspects of the data in the data warehouse fully and precisely. The answers to those questions are kept in a place called the metadata repository. Table (2) shows an example of metadata for the entity *customer*. Metadata is a critical need for

- using the data warehouse, users need information about the data
- building the data warehouse, metadata about source systems, target mappings, ...
- administering the data warehouse, impossible without metadata (*metadata is like a nerve center*)

9.2 Questions and Issues about Data Warehouse Administration

- Date Extraction / Transformation / Loading
  - How to handle data changes?
  - How to include new sources?
  - Where to cleanse the data?
  - How to switch to new data transformation techniques?
- Data from External Sources
  - How to add new external data sources?
  - How to drop some external data sources?
  - When mergers and acquisitions happen, how to bring in new data to the warehouse?
  - How to verify all external data on ongoing basis?
- Data Warehouse

  - How to add a new summary table?
  - How to expand storage?
  - How to continue ongoing training?
  - How to perform disaster recovery drills?
  - How to maintain the security system?
  - When to schedule backups?
  - . . . ?

9.3 Who Needs Metadata?

- Casual users:
  - Information discovery
    - List of predefined queries and reports
    - Business views
  - Meaning of data
    - Business terms
    - Data definitions
    - Filters
    - Data sources
    - Conversion
    - Data owners
  - Information access
    - Authorization request
    - Information retrieval into desktop applications such as spreadsheets
- Power users:
  - Information discovery
    - Databases
    - Tables
    - Columns
  - Meaning of data
    - Business terms
    - Data definitions
    - Data mapping
    - Cleansing functions
    - Transformation rules
  - Information access

        - Query tool sets
        - Database access for complex analysis

- IT professionals:
    - Information discovery
        - Databases
        - Tables
        - Columns
        - Server platforms
    - Meaning of data
        - Data structures
        - Data definitions
        - Data mapping
        - Cleansing functions
        - Transformation rules
    - Information access
        - Program code in SQL, 3GL, 4GL
        - Front-end applications
        - Security

9.4 Metadata Is Essential for Every Tier of the Data Warehouse Architecture

- Data warehouse architecture
    - Data acquisition
    - Data storage
    - Information delivery

See (07.19 - 07.21)[1].

9.5 However, even More Aspects of Metadata

- The metadata repository
    - Business metadata
        - Connection between business users and the data warehouse
    - Technical metadata

[1] Those numbers relate to the overhead transparencies of the book (the ones used in the lecture).

        - Meant for IT staff for development and administration of the data warehouse

- Metadata requirements
    - Capturing and storing data
    - Metadata integration
    - Metadata standardization
    - Keeping Metadata synchronized
    - Metadata Exchange

10 Principles of Dimensional Modeling

The requirements definition completely drives the data design for the data warehouse.

- Overall decisions
    - Choosing the process (selecting the subjects from the information packages)
    - Determining the level of detail
    - Identifying the dimensions
    - Choosing the facts
    - Choosing the duration of the database

See example Automaker Sales (07.6).

- Criteria for combining the tables into a dimensional model
    - Model should provide the best data access
    - The whole model must be query-centric
    - It must be optimized for queries and analyses
    - Model must show that dimension tables interact with the fact table
    - The model should allow rolling up and drilling down along dimension hierarchies
- Differences between dimensional modeling for the data warehouse and er-modeling for OLTP systems

10.1 Star Schema Modeling

The star schema mainly consists of one fact table and several dimensional tables. The dimensional tables usually capture entities like *customer*, *salesperson* or *product*. The fact table is derived from the dimensional tables and consists of foreign keys to any of the dimensional tables and some aggregated attributes. (see 07.11 - 7.20 for further explanations)

OLTP	DW
details of events or transactions	overall processes
individual events	how managers view the business
window into micro-level transactions	reveals business trends
details necessary to run the business	centered around a business process
strictly normalized (consistency, . . .)	normalization issues not critical

Table 3: OLTP and DW differences

11 Dimensional Modeling: Advanced Topics

This chapter covers the problems of changing dimensions, large dimensions, the snowflake schema and aggregate tables.

11.1 Changing Dimensions

- Most dimensions are constant over time
- Many dimensions only change slowly
- The product key of the source record does not change
- The descriptions and other attributes change slowly over time
- In the source OLTP, the new values overwrite the old ones
- Overwriting of dimension table attributes is not always the appropriate option
- How changes are made depends on the types of changes and what information must be preserved in the data warehouse
- Type 1 changes: correction of errors
  - No need to preserve history in the data warehouse
  - Overwrite the attribute value in the dimension table row with the new value
- Type 2 changes: preservation of history
  - Need to preserve history in the data warehouse
  - Add a new dimension table row with the new value of the changed attribute
  - No changes to the original row
  - The new row is inserted with a new surrogate key
- Type 3 changes: tentative soft revisions
  - Relate to 'soft' changes in the source systems

  - Need to keep track of history with old and new values of the changed attribute
  - Used to compare performance across the transition (tracking forward and backward)
  - Add an 'old' field in the dimension table for the affected attribute (e.g. old_territory_name)
  - Keep the new value in the 'current' field (e.g. current_territory_name)
  - You may add a effective date (when the change happened)

11.2 Large Dimensions

For instance the customer dimension usually is in the range of 20 million rows with up to 150 dimension attributes.

- Multiple hierarchies, e.g. product dimension
  - Hierarchy for finance (description, division, stackable, ...)
  - Hierarchy for marketing (description, brand, category, ...)
- Divide large, rapidly changing dimensions, e.g. customer dimension
  - Customer dimension (customer_key, name, address, ...)
  - Behavior dimension (behavior_key, type, returns, ... split from former customer dimension)

11.3 Snowflake schema

- Migrating from star to snowflake
  - Partially or fully normalize (a few) dimension tables
- Advantages
  - Small savings in storage space
  - Normalized structures are easier to update and maintain
- Disadvantages
  - Complexity (especially for users)
  - Degraded query performance (more joins are needed)

12 Data Extraction, Transformation, and Loading

12.1 Major Tasks in the ETL Process

- Combine several source data structures into a single row in the target database
- Split one source data structure into several structures
- Read different data formats from different source systems and different data structures
- Aggregate for populating aggregate fact tables
- Derive target values for input fields (e.g. age from birth date)
- Change cryptic values to meaningful values (e.g. 1 and 0 to male and female)

12.2 Data Extraction

Following issues should be considerated:

- Source identification (identify source applications and source structures)
- Method of extraction (manual or tool-based, decision for each data source)
- Extraction frequency (weekly, daily, monthly, ...)
- Time window (for extraction process)
- Job sequencing (e.g. one job needed to be finished for another one to start)
- Exception handling (decide how to handle input records that can not be extracted)

Decide between the following types of data extraction:

- ’as is’ (static) data
  - Capture data at a given point in time
  - Taking a snapshot of the relevant source data
  - Primarily used for the initial load of the data warehouse
- Data of revisions / Incremental data capture
  - Immediate
    1. Capture through transaction logs
    2. Capture through database triggers
    3. Capture in source applications
  - Deferred
    1. Capture based on date and timestamps
    2. Capture by comparing files

12.3 Data Transformation

The basic tasks in the field of data transformation are:

- Selection of data from the source systems
- Splitting/Joining
- Conversion
- Summarization/Aggregates
- Enrichment (rearrangement and simplifications of simple fields)

The major types of transformation are:

- Format revisions (changes to the data types and lengths of individual fields)
- Decoding of fields (change values to make sense to the user)
- Calculated and derived values
- Splitting of single fields or aggregating to composite data fields
- Merging of information
- Character set conversion
- Conversion of units of measurements
- Date/time conversion
- Key restructuring (e.g. product number and country code to new key)
- Duplication

12.4 Data Loading

- Initial load (populating warehouse for first time)
- Incremental load (applying ongoing changes in periodic manner)
- Full refresh (completely erasing the contents of one or more tables and reloading with fresh data)

Data can be applied by the following modes:

- Load
    - Replace existing data by fresh one
    - Erase old values and load new ones

- Append
  - Appending one or more new rows to the data in the warehouse
  - Those new rows were not existent in the warehouse before
- Destructive Merge
  - Replace one or more rows by new ones from the source systems
  - Old rows will be deleted
- Constructive Merge
  - Add one or more rows by new ones from the source systems
  - Old rows will not be deleted
  - Use of surrogate keys (see type 2 changes in section (11.1) on page (27))

As the key of the fact table is the concatenation of the keys of the dimension tables, it is loaded after the dimension tables.

13 Data Quality: A Key to Success

The main characteristics of high-quality data are:

- Accuracy
- Domain integrity
- Data type
- Consistency
- Redundancy
- Completeness
- Usefulness
- Duplication (see data marts)
- Clarity
- ...

This are the main types of data quality problems:

- Dummy values in fields
- Absence of data values

- Unofficial use of fields
- Cryptic values
- Contradicting values
- Violation of business rules (e.g. order date after delivery date)
- Reused primary keys
- Non unique identifiers
- Inconsistent values
- Incorrect values
- Multipurpose fields

And many sources of data pollution:

- System conversions
- Data aging (e.g. new currency)
- Poor database design
- Input errors
- Fraud

Some discovery and correction features in addition to standard DBMS features like domain integrity and referential integrity checking are:

- Quickly identify duplicate records
- Find inconsistent data
- Check for correctness of domains and if all values are inside the range of the domains
- Normalize inconsistent data
- Group and relate customer records belonging to the same household
- Provide measurements of data quality
- Validate for allowable values

13.1 Data Cleansing

Data cleansing is a very time-consuming and not attractive work. Users often do not give data quality the attention they should. Furthermore companies are often terrified of launching a data cleansing initiative because of the high costs of both money and time. Basic data cleansing decisions are:

- Which data to cleanse
- Where to cleanse
- How to cleanse
- How to discover the extent of data pollution
- Setting up a data quality framework

Suppose any incoming data is wrong and start your analysis from that point of view. An overall data purification process (located at the data staging area) which takes polluted data as input from the source systems and produces cleansed data as output to the data warehouse, could consist of the following points:

- Prioritize data into high, medium, and low categories
- Prepare schedule for data purification beginning with the high priority data
- Ensure that techniques are available to correct duplicate records

Final tips on data quality:

- Identify high-impact pollution sources
- Do not try to do everything with in-house programs
- Select proper tools
- Agree on standards and reconfirm these
- Try to get users to do the unattractive data quality work
- Get the executive sponsor of the warehouse project to be actively involved
- Wherever needed, bring in outside experts

14 Matching Information to the Classes of Users

Users do have more possibilities when working with the data warehouse than working with OLTP. Data warehousing helps planning and assessing of results of campaigns. Information usage modes are:

- Verification
  - Find hypotheses
- Discovery
  - Let hypotheses be generated

14.1 Data Warehouse User Classes

- Tourists
  - Executive: interested in business indicators
  - Status of the indicators at routine intervals
  - Easy navigation from one indicator to the next
- Operators
  - Support staff: interested in current data
  - Immediate answers based on reliable current data
  - Current state of the performance metrics
  - Current data needed
  - Quick access to detailed information
  - Was the latest campaign for the new product successful?
- Farmers
  - Analysts: interested in routine analysis
  - Run predictable queries easily and quickly
  - Obtain same types of information at predictable intervals
  - Current data with simple comparisons to historical data
- Explorers
  - Skilled analysts: interested in highly ad hoc analysis
  - Totally unpredictable and intensely ad hoc queries
  - Complex analysis
  - Long analysis sessions
- Miners
  - Special purpose analysts: interested in knowledge discovery
  - Access to mountains of data to analyze and mine

  - What is the impact of the dollar exchange rate and of the campaign on the rise in exports?

In a managed reporting environment information is pushed to the user, not pulled by the user as in the case of queries. Predefined Reports are automatically generated and delivered to the user (by e-mail, automatic fax et cetera).

15 OLAP in the data warehouse

OLAP, online analytical processing, concerns the processing of data that is manipulated for analysis. One can not have a data warehouse without OLAP. Guidelines for an OLAP system are:

1. Multidimensional conceptual view
2. Transparency
3. Accessibility
4. Consistent reporting performance
5. Client/server architecture
6. Generic dimensionality
7. Dynamic sparse matrix handling
8. Multiuser support
9. Unrestricted cross-dimensional operations
10. Intuitive data manipulation
11. Flexible reporting
12. Unlimited dimensions and aggregation levels
13. OLAP analysis models
14. Treatment of nonnormalized data
15. Missing values
16. DBMS tools
17. 'drill through to detail level' functionality
18. Incremental database refresh
19. SQL interface

OLAP systems

- let business users have a multidimensional and logical view of the data in the data warehouse,
- facilitate interactive query and complex analysis for the users,
- allow users to drill down for greater details or roll up for aggregations of metrics along a single business dimension or across multiple dimensions,
- provide ability to perform intricate calculations and comparisons, and
- present results in a number of meaningful ways, including charts and graphs.

OLAP is much more than an information delivery system for the data warehouse. A data warehouse stores the data and provides simpler access to the data. An OLAP system complements the data warehouse by lifting the information delivery capabilities to new heights. General features of OLAP:

- Multidimensional analysis
- Consistent performance
- Fast response times for interactive queries
- Drill-down and roll-up
- Navigation in and out of details
- Slice-and-dice or rotation
- Multiple view models
- Easy scalability
- Time intelligence (year-to-date, fiscal periods)

The (simplified) sales example:

- Fact table
  - Sales facts
    - product_key
    - time_key
    - store_key
    - fixed_costs
    - variable_costs
    - direct_sales
- Dimension tables

  - Product
    - product_key
    - product_name
    - product_category
  - Store
    - store_key
    - store_name
    - store_territory
    - store_region
  - Time
    - time_key
    - date
    - month
    - year
    - quarter

A three-dimensional display: A table with the time dimension as rows, the product dimension as columns and the store dimension as pages (see 07.26 for roll up see 07.38). There are several models of OLAP in use:

- ROLAP: relational online analytical processing (star and snowflake schema)
- MOLAP: multidimensional online analytical processing
- DOLAP: desktop online analytical processing. A variation of ROLAP