Requirements of the Data Warehousing

- Independence between data sources and analytical systems (w.r.t. availability, load, ongoing changes)
- Continuous provision of integrated and derived data (Persistence)
- Reusability of provided data
- Possibility to conduct arbitrary evaluations
Requirements of the Data Warehousing

- Support of custom views (e.g., w.r.t. time horizon, domain and structure)
- Erweiterbarkeit (z.B. Integration neuer Quellen)
- Automation of processes
- Uniqueness of data structures, access rights and processes
- Orientation on the main purpose: data analysis
12 OLAP rules by Codd

- Multidimensional, conceptual view
- Transparency
- Accessibility
- Performance
- Scalability
- Generic Dimensionality
- Dynamic handling of sparse multidimensional structures
- Multi-user mode
- Unrestricted operations
- Intuitive user interface
- Flexible reporting
- Any number of dimension and aggregation levels
FASMI

Fast Analysis on Shared Multidimensional Information

- Short response times (on average less than five seconds)
- Simple and flexible ways of evaluation
- Heterogeneous users with different rights
- Multidimensionality is an important criterion
- Questions on the number of required dimensions and ranges of values of associated attributes
Phases of the Data Warehousing

1. Controlling the sources for changes by using monitors
2. Copying relevant data by extraction into the temporary data cleaning area
3. Transformation of the data in the data cleaning area (cleansing, integration)
4. Copying of the data in an integrated database as the foundation base for various analyzes
5. Filling the data cubes (databases for analysis purposes)
6. Analysis: operations on data of the DW

Basic database and data cubes represent the data warehouse.
Data Warehouse Manager

- Central component of a DW system
- Initiation, control and monitoring of the individual processes (flow control).
- Initiation of the data collection process
  - At regular time intervals (every night, on weekends, etc.): Starting of the data extraction from sources and transferring to the data cleaning area
  - For a change in a source: starting of the respective extraction component
  - On explicit request of the administrator
  - Push vs. Pull strategy
  - Timeliness is a requirement for analytical tasks
Data Warehouse Manager

- After triggering the loading process:
  - Monitoring of the further steps (cleansing, integration, etc.)
  - Coordination of the processing order

- Event of a fault
  - Documentation of errors
  - Restart mechanisms

- Access to metadata from the repository
  - Controlling the flow
  - Parameters of the components
Data sources

- Suppliers of data for the data warehouse
  - Do not belong directly to the DW
  - Can be internal (company) or external (e.g., state institution)
  - Heterogeneous with respect to the structure, content and interfaces (databases, files)
  - Selection of sources and quality of the data of particular importance

- Factors for selection
  - Purpose of the DW
  - Quality of the source data
  - Availability (legal, social, technical)
  - Price for data acquisition (especially for external sources)
Data sources: classification

- Origin: internal, external
- Time: current, historical
- Use level: primary data, meta data
- Content: number, string, graphic, reference, document
- Display: numeric, alpha-numeric, BLOB
- Language and character set
- Degree of confidentiality
Data sources: quality requirements

- Consistency (absence of contradictions),
- Correctness (matching reality),
- Completeness (e.g., the absence of missing values or attributes),
- Reliability (e.g., confidence in the data sources),
- Accuracy (e.g., number of decimal places),
- Granularity (e.g., daily or monthly data),
- Timeliness (When was the last change performed vs. the occurrence of the data change),
- Relevance (How important is the data?),
- ...
Components

Data sources: quality requirements (2)

- Reliability (traceability of the origin, trustworthiness of the supplier),
- Understandability (content-wise and technical / structural for the respective target group),
- Usability (suitable format, expedience),
- Uniformity (data format),
- Intellegibility (interpretability) and
- Key integrity (keys and references)
Monitors

- **Task:**
  - Discovery of data manipulations in a data source

- **Strategies:**
  - **Trigger-based**
    - Active data base mechanisms
      - Activation of triggers due to data changes
      - Copy the modified tuple in another area
  - **Replication-based**
    - Use of replication mechanisms for the transmission of changed data
Monitors (2)

Strategies (contd.):

- **Log-based**
  - Analysis of transaction log data of the DBMS to detect changes

- **Timestamp-based**
  - Assigning a timestamp to tuples
  - Update in case of changes
  - Identification of chances since the last extraction by time comparison

- **Snapshot-based**
  - Periodic copy of the dataset in a file (snapshot)
  - Comparison of snapshots to identify changes
Data cleaning area

- **Task:**
  - Central data management component of the data staging area
  - Temporary buffer for integration

- **Use:**
  - Execution of transformations (cleaning, integration, etc.) directly in the intermediate storage
  - Loading of transformed data into DW or base database only after successful completion of the transformation

- **Advantages:**
  - No influence on the sources or the DW
  - No acceptance of erroneous data
Extraction component

- Task: Transfer data from sources in the data staging area
- Function: dependent on the monitoring strategy
  - Periodically
  - On request
  - Event-controlled (e.g., when reaching a defined number of changed)
  - Immediate extraction

- Implementation:
  - Use of standard interfaces (e.g., ODBC, JDBC)
  - Exception handling in case of an error
Transformation component

- Preparation and adjustment of the data to load
  - Content-wise: data / instance integration and cleaning
  - Structural: schema integration
- Transferring all data in a uniform format
  - Data types,
  - Dates,
  - Units,
  - Encodings, etc.
- Removal of impurities (Data Cleaning or Data Cleansing)
  - Incorrect or missing values,
  - Redundancies,
  - Outdated values.
Transformations component (2)

- Data Scrubbing:
  - Utilization of domain-specific knowledge (e.g. business rules) to detect impurities
  - Example: detection of redundancies

- Data Auditing:
  - Application of data mining methods to uncover rules
  - Detection of deviations
Loading component

- **Task:**
  - Transfer of the adjusted and processed (e.g., aggregated) data to the base database or the DW

- **Features:**
  - Use of special loading tools (e.g., SQL*Loader by Oracle)
    - Bulk loading
  - Historicization: Changes in sources may not overwrite DW data, instead they are stored in addition

- **Loading process:**
  - Online: Base database or DW is still available
  - Offline: not available (time window: at night, during weekends)
Base database

- Task:
  - Integrated database for various analyzes
    → independent of specific analyzes, i.e., no aggregations yet
  - Supply of the DW with adjusted data (possibly by compression)

- Notes:
  - Often ommitted in practice
  - Equivalent to Operational Data Store (ODS) by Inmon
Data Cube

- Task: databases for analysis purposes (relational or multi-dimensional)
- Structure based on analysis needs
- Base: DBMS
- Features:
  - Support the load process
    - Fast loading of large amounts of data
      → Bulk loader, bypassing multi-user coordination and consistency checks
  - Support of the analysis process
    - Efficient query processing (index structures, caching)
    - Multidimensional data model (e.g., via OLE DB for OLAP)
Data Warehouse

In a narrower sense:
Base database and data cubes represent the data warehouse.

- In a broader sense, the data marts also provide components of the data warehouse.
Data Marts

- **Task:**
  - Providing a content-restricted view of the DW (e.g., Department)

- **Reasons:**
  - Autonomy, privacy, load balancing, data volume, etc.

- **Implementation:**
  - Distribution of the DW data

- **Formen:**
  - Dependent data marts
  - Independent data marts
Dependent Data Marts

- Distribution of the data set after
  - Integration and cleanup (based database) and
  - Organization in accordance with the analysis needs (data cube)

- "Hub and Spoke" architecture

- Data Mart:
  - Only extract (including aggregation) of the Data Warehouse
  - No adjustment or normalization

- Analyzes on data mart consistent with analyzes on DW

- Simple implementation:
  - Replication or view mechanisms of DBMS
"Hub and Spoke" architecture

- Analyse
- Analyse
- Analyse
- Analyse

Data Warehouses

Data Marts

Befüllen
Dependent Data Marts: Extraction process

- Structural extracts
  - Limited to parts of the schema
  - Example: only certain metrics or dimensions

- Content-based extracts
  - Content-based restriction
  - Example: only certain branches or the last year’s result

- Aggregated extracts
  - Reducing the granularity
  - Example: restriction on monthly results
Independent Data Marts

- Independently created "small" data warehouses (e.g., of individual organizations).
- Subsequent integration and transformation
- Problems:
  - Different analysis views (Data Mart, global Data Warehouse)
  - Consistency of the analysis due to additional transformation
Independent Data Marts
Analysis tools

- Business Intelligence Tools
- Task:
  - Presentation of the data collected
  - Interactive navigation
  - Analysis options
- Analysis:
  - Simple arithm. operations (e.g., aggregation) ... complex statistical analysis (e.g., data mining)
  - Preparation of the results for further processing or forwarding
Analysis tools: Representation I

- **Tables**
  - **Pivot Tables:** = crosstabs
    - Feature values in the row and column header
  - Analyzing by interchanging rows and columns
  - Change of table dimensions
  - Nesting of table dimensions

<table>
<thead>
<tr>
<th></th>
<th>Beer</th>
<th>Red wine</th>
<th>Summe</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sachsen-Anhalt</td>
<td>45</td>
<td>32</td>
<td>77</td>
</tr>
<tr>
<td>Thüringen</td>
<td>52</td>
<td>21</td>
<td>73</td>
</tr>
<tr>
<td><strong>Summe</strong></td>
<td>97</td>
<td>53</td>
<td>150</td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sachsen-Anhalt</td>
<td>60</td>
<td>37</td>
<td>97</td>
</tr>
<tr>
<td>Thüringen</td>
<td>58</td>
<td>20</td>
<td>78</td>
</tr>
<tr>
<td><strong>Summe</strong></td>
<td>118</td>
<td>57</td>
<td>175</td>
</tr>
</tbody>
</table>
Analysis tools: Representation II

- Graphics
  - Visualization of large data sets
  - Net, dot, surface graphs

- Text and multimedia elements
  - Addition of audio or video data
  - Inclusion of document management systems
Components

Analysis tools: Functionality

- **Data Access**
  - Reporting Tools
  - Reading of data, changing Presentation in reports
  - Presentation in reports
  - "Traffic lights": rule-based formatting
  - Base: SQL
Components

Analysis tools: Example [Cognos, 2012]
Analysis tools: Functionality

OLAP
- Interactive data analysis, classification navigation
- Reports with aggregated values (metrics / indicators)
- Navigation operations:
  - Drill Down,
  - Roll Up,
  - Drill Across,
  - Dice und
  - Slice
- Aggregation and calculation functions (statistic, economic)
- Validating hypotheses, plausibility check
Analysis tools: Example [Cognos, 2012]
Analysis tools: Functionality

- **Data Mining**
  - Uncovering previously unknown relationships
    → Patterns, paths, rules
  - Methods (among others):
    * Classification: assignment of data to predefined classes
    * Association rules
    * Clustering: segmentation, i.e., grouping data regarding their characteristic values
    * Forecast
Data Mining: Examples

![Diagram showing data points for age (Alter) and income (Einkommen), with categories for creditworthiness (Kreditwürdigkeit) and employment status (Ang.verhältnis).]
Types of analyzes

- Komplexität
- Flexibilität

- Datenvisualisierung
- Charting
- Standard-Reporting
- Analytical-Reporting
- Ad-hoc-Reporting
- CRM-Analysen
- Forecasting/Budgeting
- OLAP
- Data Mining

© Sattler / Saake / Köppen
Data Warehouse Technologies
Last Change: 06.01.2019
Analysis tools: Implementation

- Standard Reporting:
  - Reporting tools of classical reporting
- Record books:
  - Graphical development environment for creating presentations of tables, graphs, etc.
- Ad-hoc Query & Reporting:
  - Tools for the creation and presentation of reports
  - Hide database connection and query languages
Analysis tools: Implementation

- **Analysis Clients:**
  - Tools for multidimensional analysis
  - Include navigation, manipulation (computing), advanced analysis and presentation functions

- **Spreadsheet add-ins:**
  - Extension of spreadsheets for data connection and navigation

- **Development Environments:**
  - Supporting the development of own analytic applications
  - Provision of operations on multidimensional data
Repository

Task:
- Storing the metadata of the DW system

Metadata:
- Information simplifying the construction, maintenance and administration of the DW system and enabling information retrieval
- Examples:
  - Database schemas,
  - Access rights,
  - Process information (processing steps and parameters), etc.
Metadata Manager

- **Tasks:**
  - Control of the metadata management
  - Access, query, navigation
  - Version and configuration management

- **Forms:**
  - General use: extensible base schema
  - Tool-specific: fixed part of tools

- Frequently integration and exchange between decentralized metadata management systems necessary
Summary

- Reference architecture for Data Warehouse systems
- Process of the Data Warehousing
- Roles of components
- Data Marts as extracts of the DW
- Analysis tools: Classification and examples