Data Warehouses - Architecture

Databases and Information Systems

Fabian Panse

panse@informatik.uni-hamburg.de

University of Hamburg

Acknowledgements

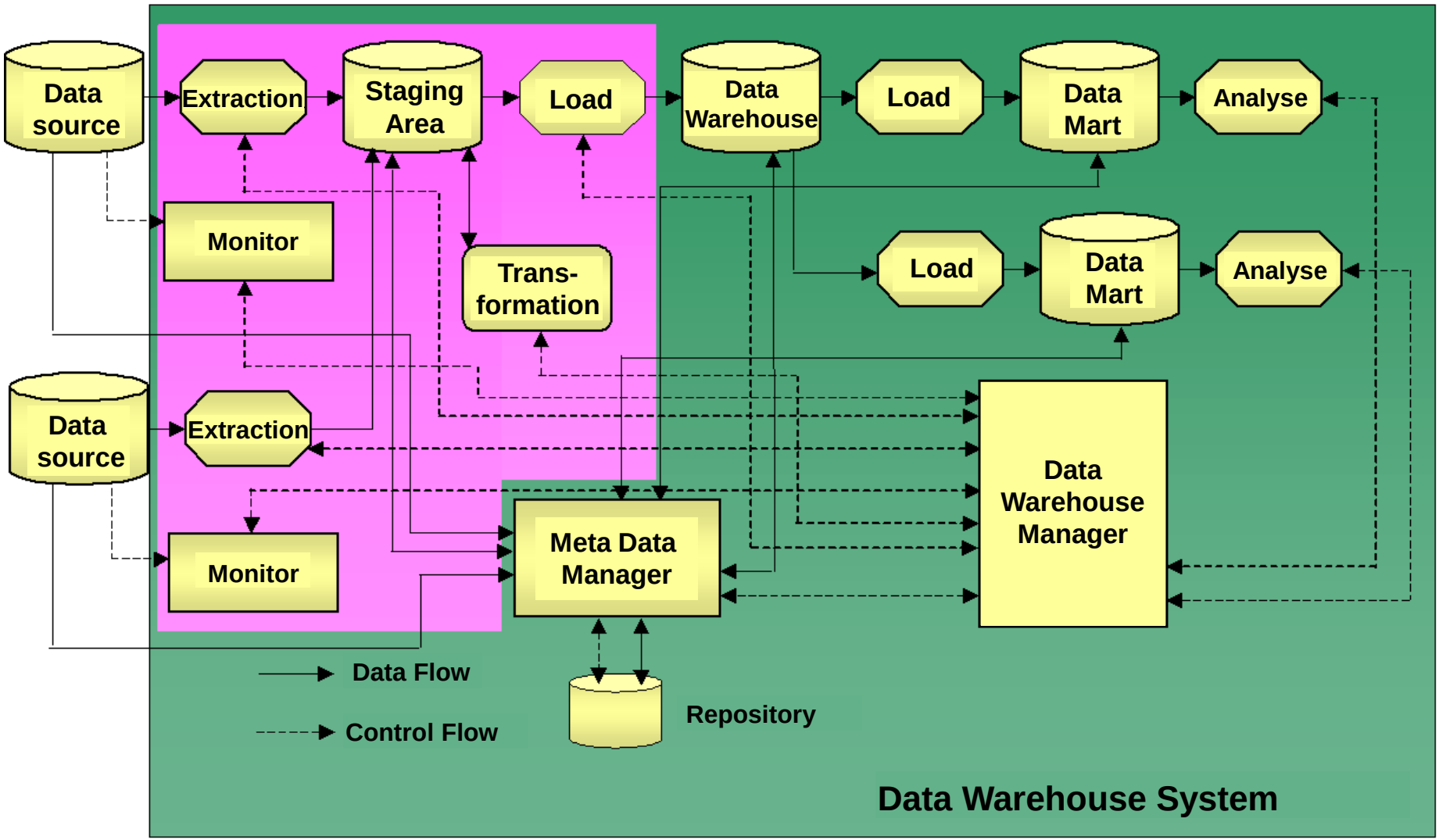
These slides are based on slides provided by

- Prof. Dr. Erhard Rahm
  University of Leipzig
  `http://dbs.uni-leipzig.de/`

Architecture of Data Warehouse Systems

- **Reference architecture**
  - Scheduler
  - Data sources
  - Data extraction
  - Transformation and load
- **Dependent vs. independent Data Marts**
- **Meta data management**
  - Classification of meta data (technical vs. subject-specific)
  - Architectures, mechanisms for interoperability
- **Operational Data Store (ODS)**
  - Real-time processing of integrated data
- **Master Data Management (MDM)**
  - Central management of operational data used by different departments of the same company

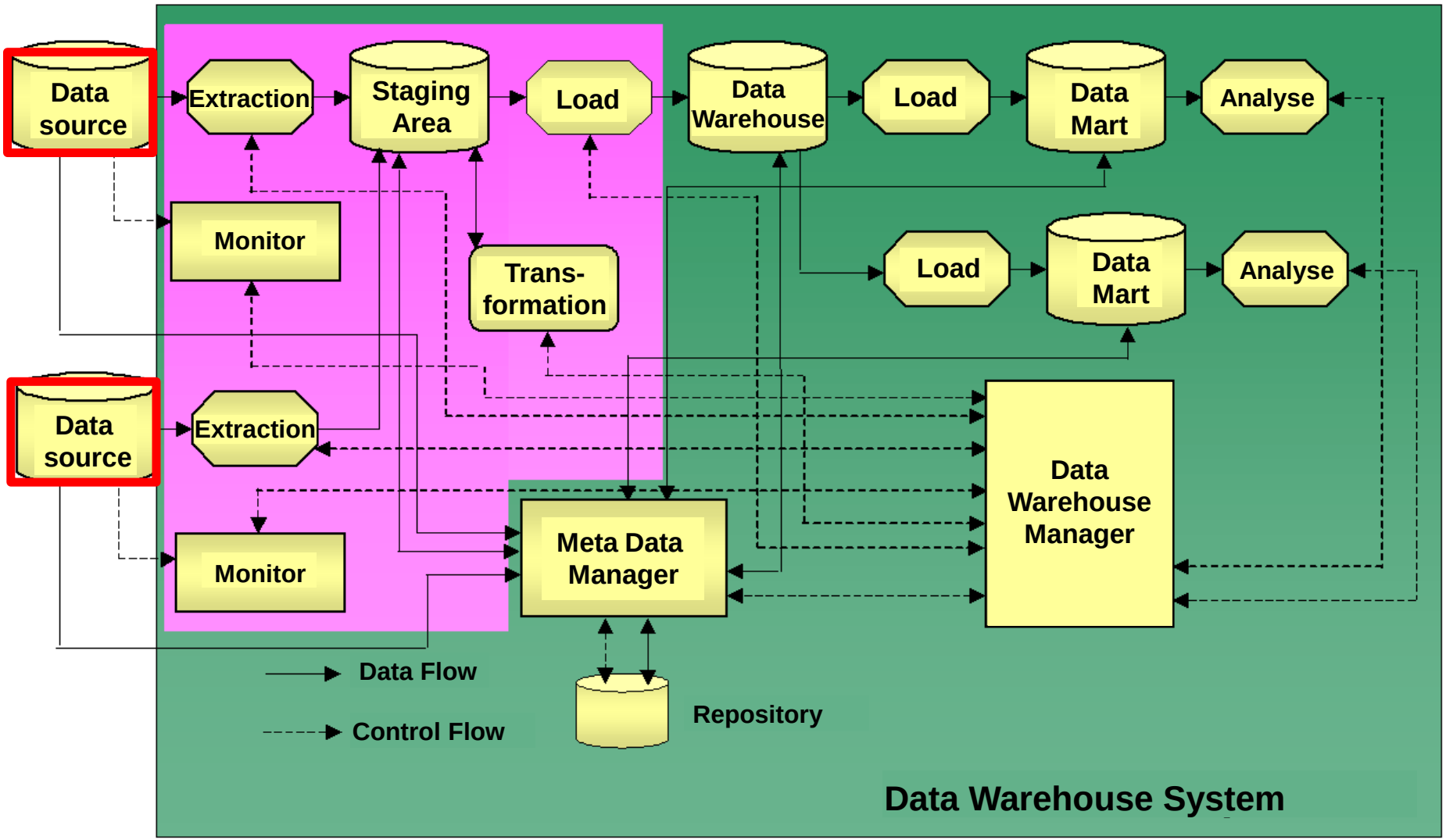
DW Reference Architecture



Phases of Data Warehousing

1. Monitoring of sources for changes in relevant data values
2. Copying of relevant data by means of extraction into a temporal staging area
3. Transformation of the extracted data in the staging area (cleaning, preparation, aggregation)
4. Copying the transformed data in the Data Warehouse (DW) so that it can serve as basis for different kinds of analyses
5. Load of data in Data Marts (DM)
6. Analyses: Operations on the data of the DW or the DMs

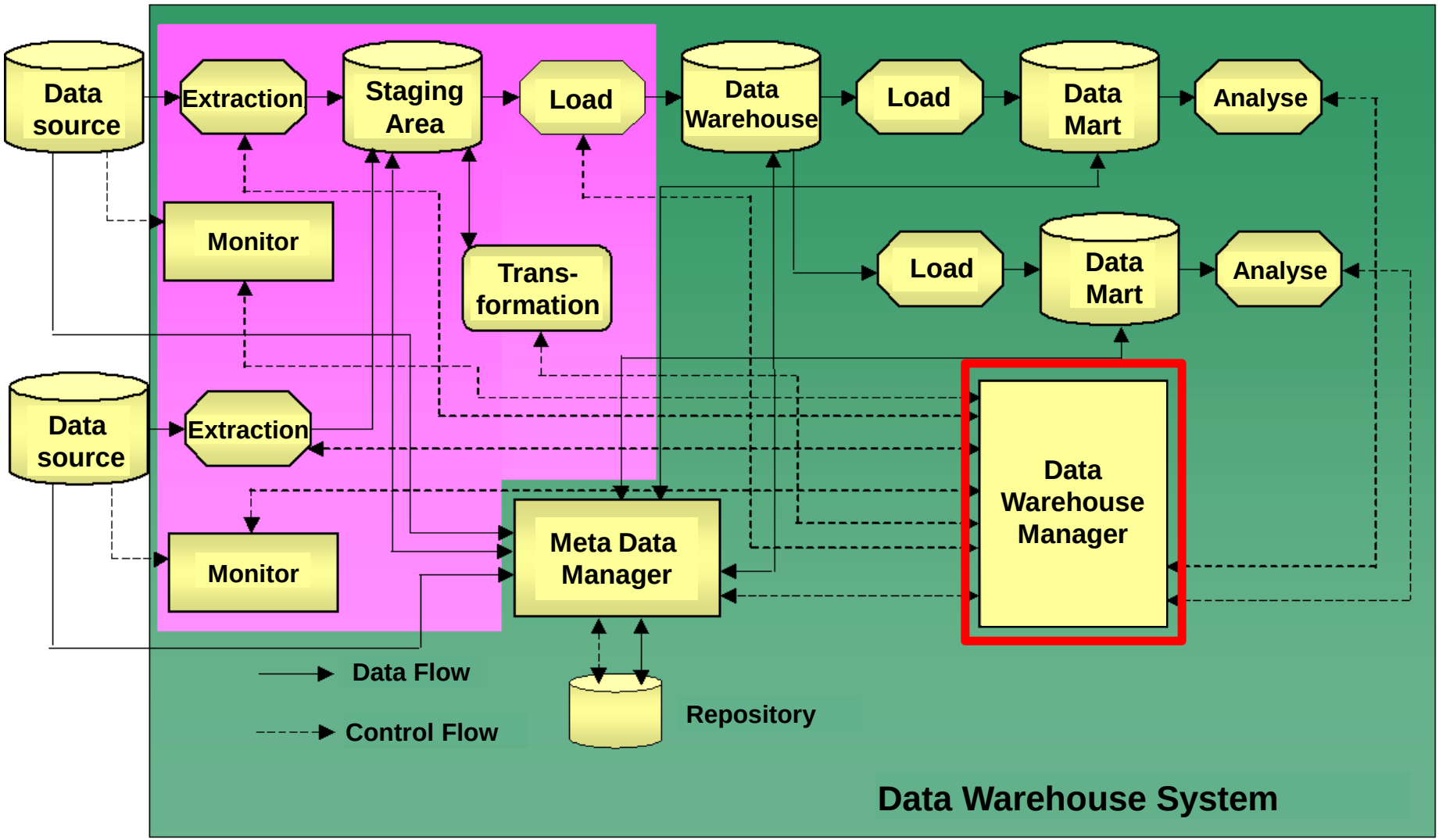
DW Reference Architecture - Data Sources



Data Sources

- **Provide the data for the DW** (usually do not belong to the DW)
- **Characteristics**
  - Internal (company) or external (e.g. Internet)
  - Maybe fee-based
  - In general autonomous
  - In general heterogeneous with respect to structure, content and interfaces (databases, data files)
- **Quality requirements**
  - Availability of meta data
  - Consistency (free from contradictions)
  - Correctness (agreement with reality)
  - Completeness (e.g. no missing values, attributes, or tuples/objects)
  - Up-to-dateness
  - Understandability
  - Usefulness
  - Relevance

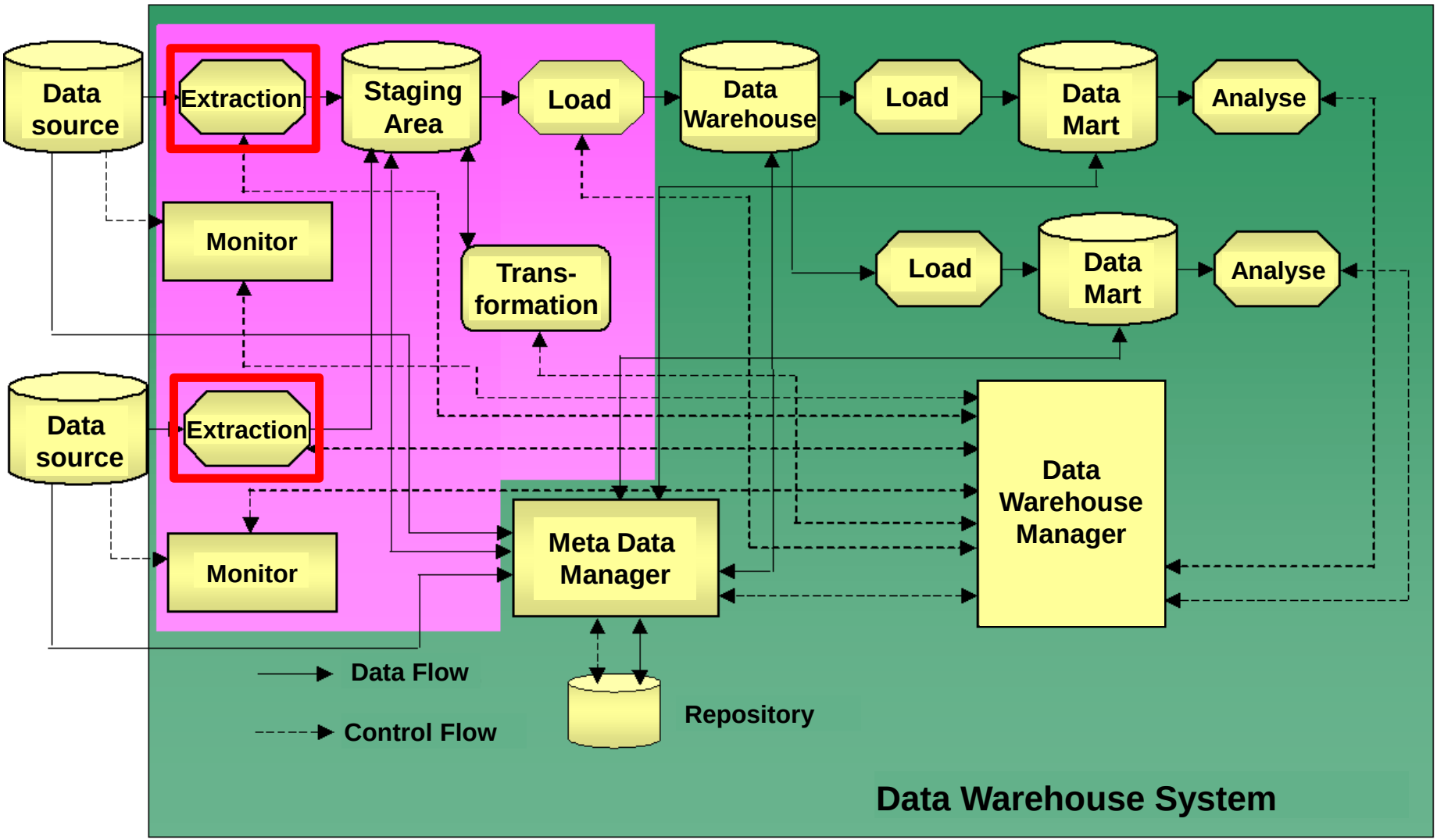
DW Reference Architecture - DW Manager



Data Warehouse Manager/Scheduler

- **Flow control:**
  - Initiation, controlling and monitoring of the individual processes
- **Initiation of data acquisition processes and transfer of acquired data to the staging area:**
  - Periodically (each night, weekend, end of month, etc.)
  - If the data of a source changes:
    Start of the corresponding extraction component
  - On explicit demand of the administrator
- **Case of error:**
  - Documentation of errors, restart mechanisms
- **Access to meta data from the repository:**
  - For flow control
  - Parameter of the individual components

DW Reference Architecture - Extraction



Data Extraction

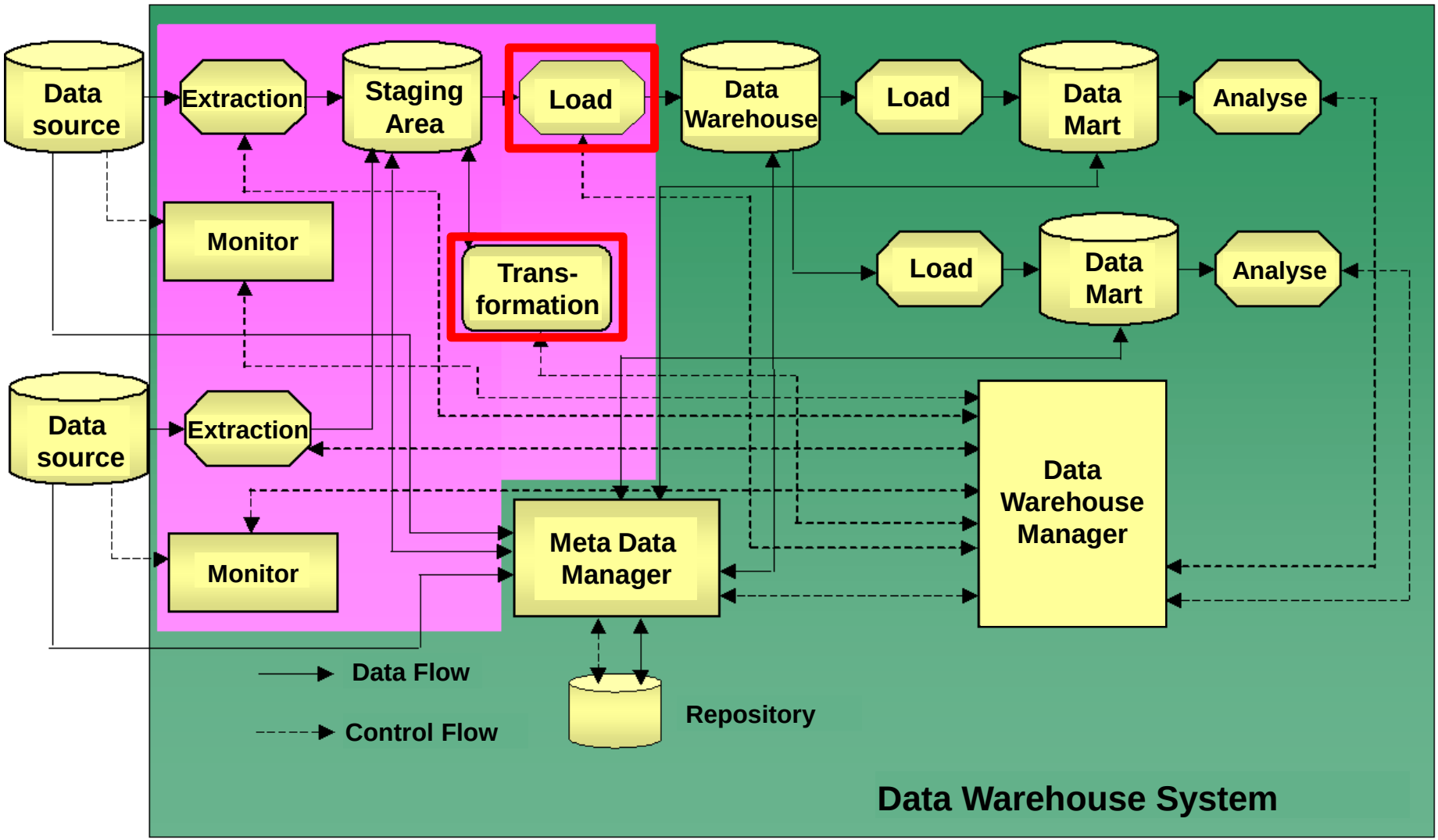
- **Monitoring:** Detection of data manipulations in the data source
  - Internal data source: active mechanisms
  - External data source: polling/periodic request
- **Extraction components:**
  Transfer of data from sources to staging area
  - Periodically
  - On request
  - Event-driven (e.g. if a predefined number of changes is reached)
  - Immediate extraction
- **Different functionalities of different source systems**
- **Usage of standard interfaces (e.g. ODBC) or self-developed**
- **Performance problems in cases of large data sets**
- **Autonomy of the source system has to be preserved**

Data Extraction: Strategies

- **Snapshots**
  - Periodic copying of the database into a data file
- **Log-based**
  - Analyses of transaction-log files of the DBMS to detect relevant changes
- **Trigger**
  - Activation of triggers in the case of data changes and copying of the changed tuples
- **Usage of DBMS-specific mechanisms for data replication**

	Autonomy	Performance	Usability
Snapshot	o/+	--	+
Log	o	-	o/-
Trigger	o	+	o
Replikation	o	o	o

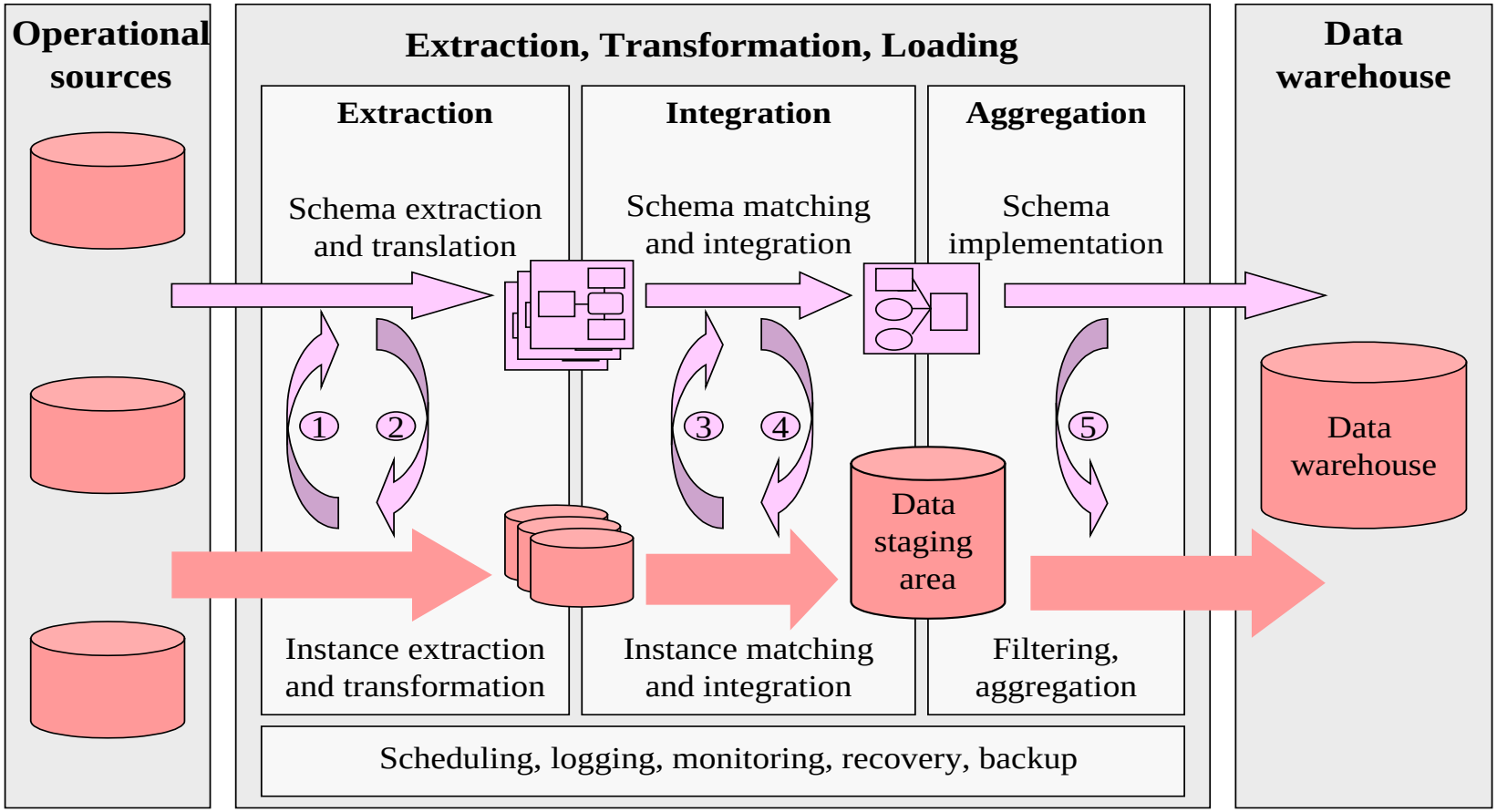
DW Reference Architecture - Transformation and Load



Data Transformation and Load

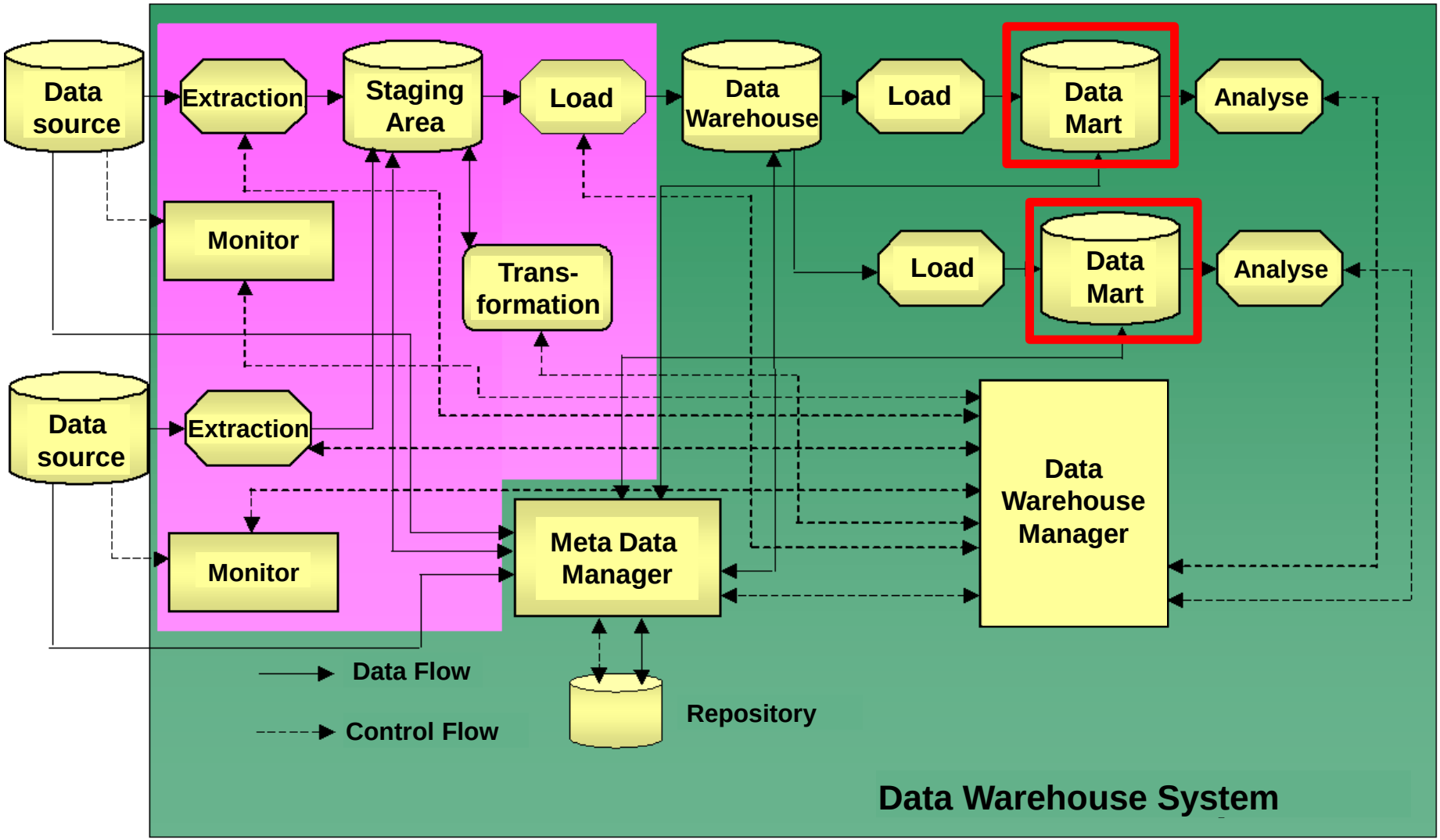
- **Staging area:**
  - Temporary storage for integration, cleaning and preparation
  - Loading of the data in the DW only after transformation was successful
  - No influences on data sources or the DW
  - No loading of error-prone data
- **Transformation component:**
  - Preparation of the data for the loading process
  - Adaptation to the DW schema (e.g. merge/split of attributes)
  - Unification and standardization of data types, date values, units, encodings
  - Data auditing: usage of data mining methods to detect rules and outliers
  - Data cleaning: handling of incorrect, missing, redundant and outdated values
- **Load component:**
  - Transfer of the cleaned, prepared and aggregated data in the DW
  - Usage of specific loading tools (e.g. Bulk Loader)
  - Historicization: additional storage of changed data instead of overwriting
  - Offline vs. Online load (availability of the DW during the loading process)

ETL Process: Overview



E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches.* IEEE Techn. Bulletin on Data Engineering, 2000

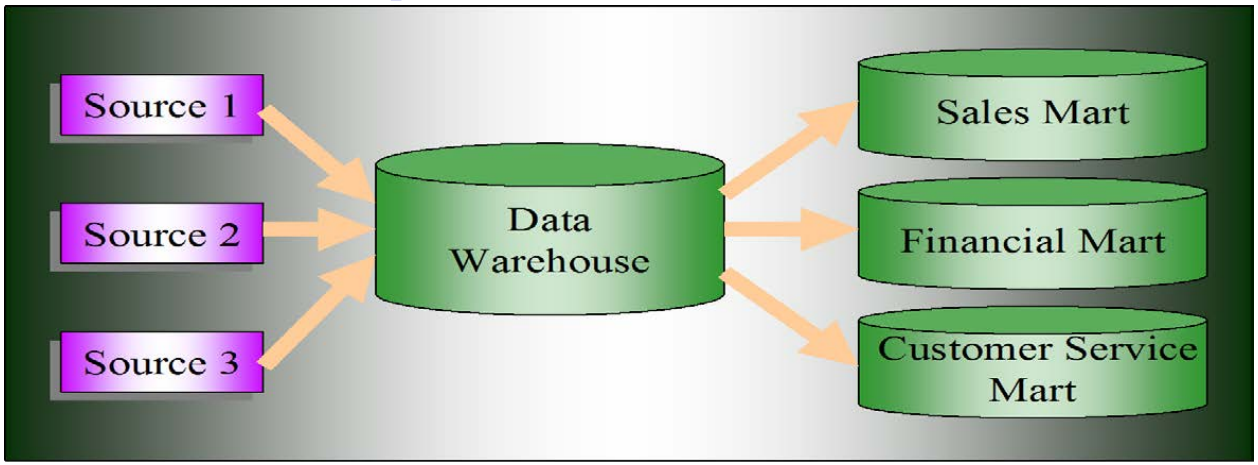
DW Reference Architecture - Data Marts



Data Marts

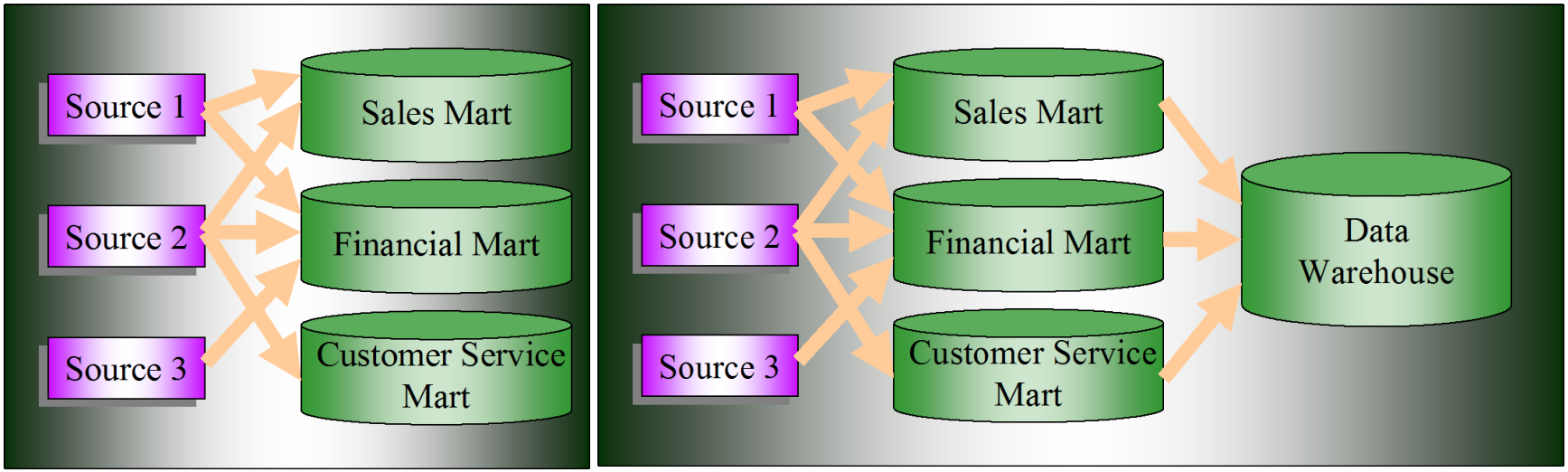
- **What is a Data Mart?**
  - A particular part of the Data Warehouse
  - Restriction to a particular topic or business area
  - Leads to distributed DW solutions
- **Reasons for Data Marts**
  - Performance: shorter response time, less users, better load balancing
  - Autonomy, data privacy
  - Maybe faster to realize
- **Problems**
  - Redundant data storage
  - Additional transformation costs
  - It is more difficult to ensure consistency
- **Variants**
  - Dependent Data Marts
  - Independent Data Marts

Dependent Data Marts



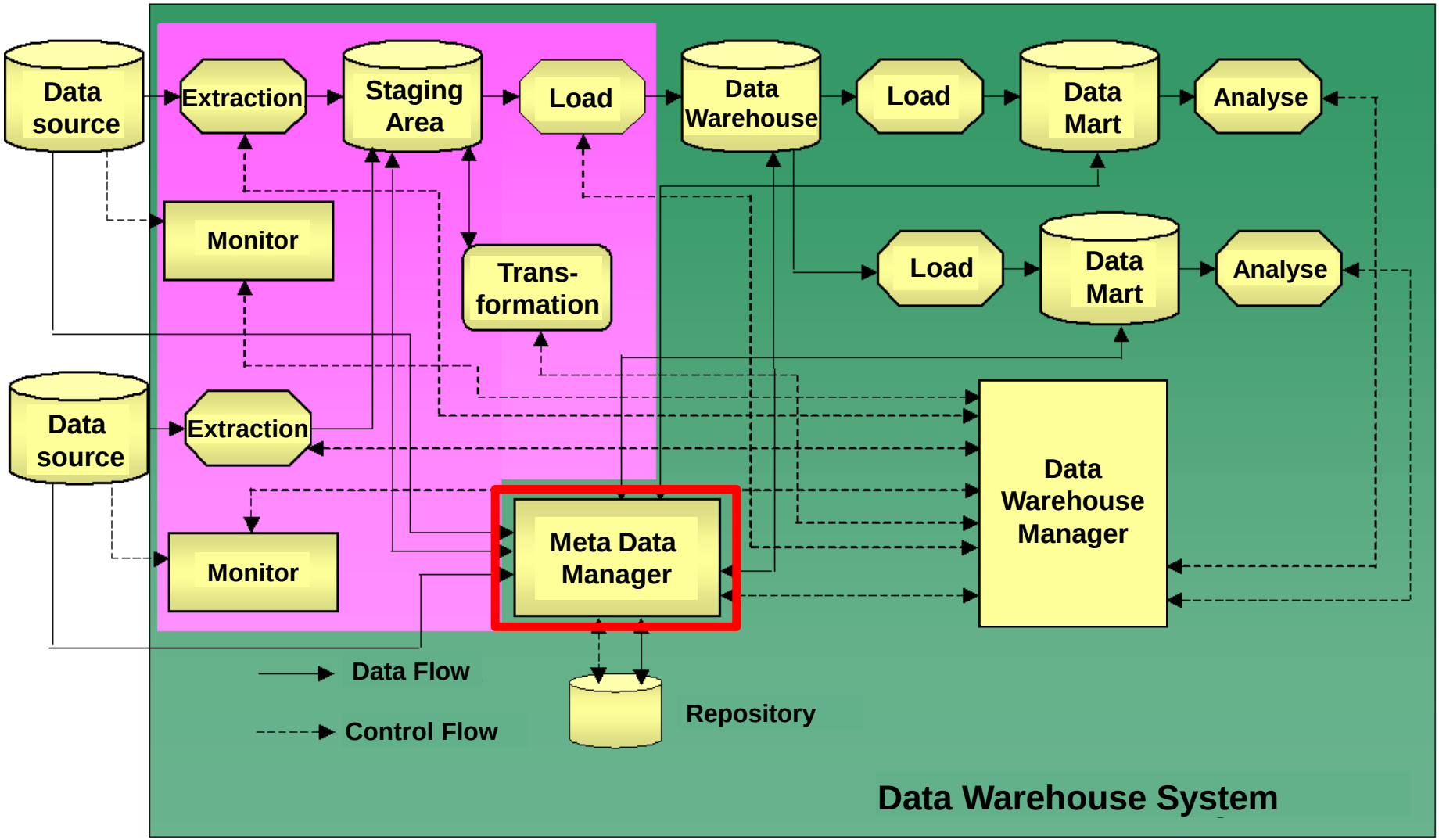
- **"hub and spoke" architecture**
- **Data Marts are extracted from the central Warehouse**
  - Structural extracts (part of the whole schema, e.g. particular facts)
  - Substantial extracts (particular time periods, branches, ...)
  - Aggregation (lower level of granularity, e.g. numbers per month)
- **Benefits:**
  - Simple to derive (replication mechanisms of the Warehouse-DBS)
  - Analyses on Data Marts are consistent with analyses on the Warehouse
- **Drawback:** Development time (company-DW needs to be developed first)

Independent Data Marts



- **Variant 1: no central, company-wide DW**
  - Simpler and faster to develop compared to a DW
  - Same data is stored in different Marts, risk of inconsistency
  - Costs increase proportional with the number of DMs
  - Difficult to extend (compared to dependent DMs)
  - No company-wide analyses possible
- **Variant 2: independent DMs + derivation of a DW from the DMs**
- **Variant 3: independent DMs + shared use of dimensions**

DW Reference Architecture - Meta Data Management



Meta Data Management

- **Requirements for meta data management/repository**
  - Provision of all relevant meta data (meta data should be up-to-date)
  - Flexible access based on powerful interfaces
  - Version and configuration management
  - Support of technical and subject-specific tasks
  - Active usage by DW processes (data transformation, analyses)
- **Forms of realization**
  - Tool-specific: inherent part of tools
  - General usable: generic and extendable repository schema (meta data model)
- **Large number of proprietary meta data models**
- **Efforts for standardization**
  - Open Information Model (OIM): Metadata Coalition (MDC) - 2000 abandoned
  - Common Warehouse Metamodel (CWM): Object Management Group (OMG)
- **Integration of or respectively exchanges between decentralized meta data management systems are often required**

Meta Data: Example

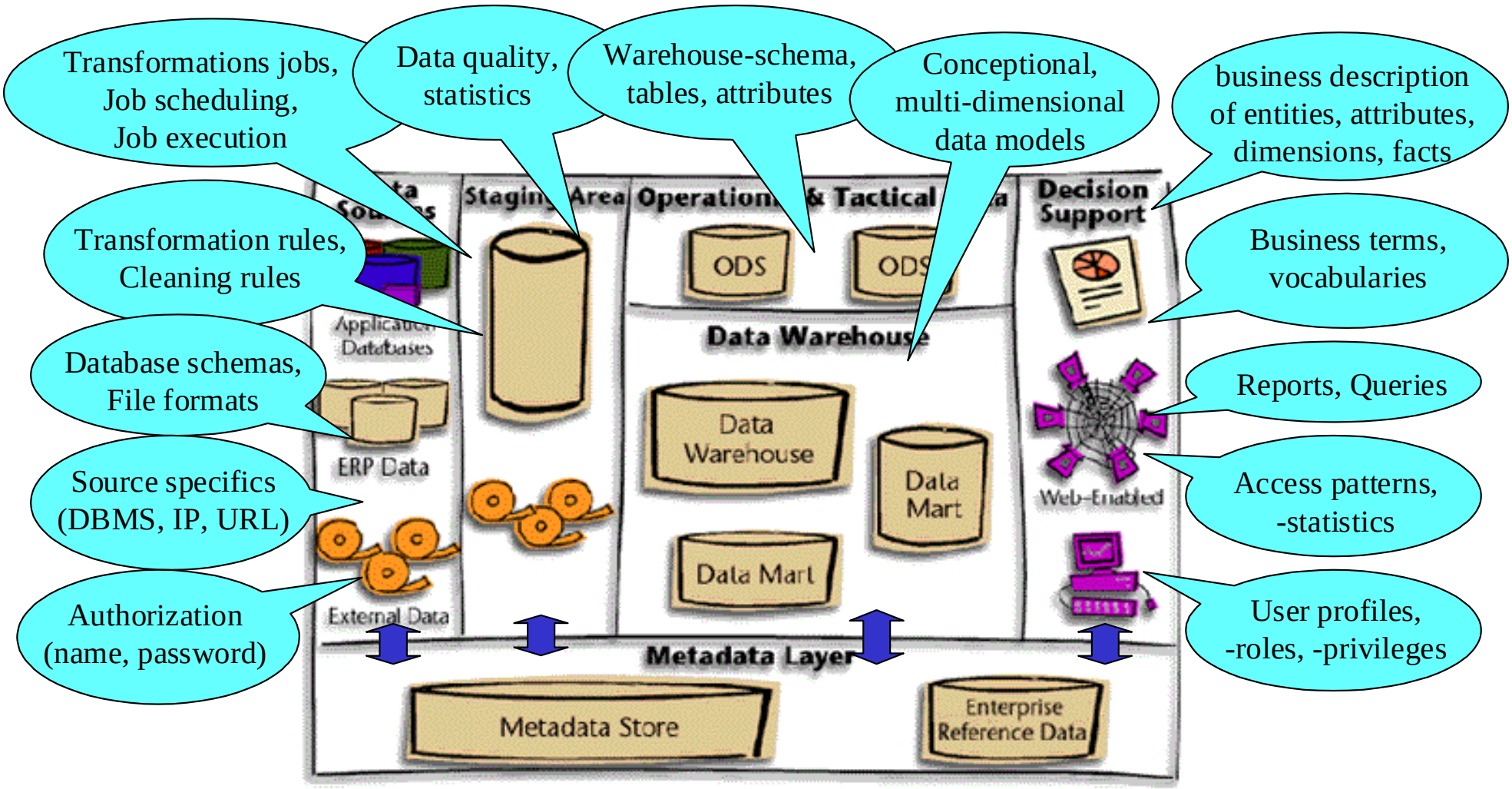
Personal : Tabelle

Feldname	Felddatentyp	Beschreibung
Personal-Nr	AutoWert	Nummer, die einem neuen Angestellten automatisch zugewiesen wird.
Nachname	Text	
Vorname	Text	
Position	Text	Position des Angestellten.
Anrede	Text	In Begrüßungen verwendete Anrede.
Geburtsdatum	Datum/Uhrzeit	
Einstellung	Datum/Uhrzeit	
Straße	Text	Straße oder Postfach.
Ort	Text	
Region	Text	Bundesland oder Provinz.
PLZ	Text	
Land	Text	
Telefon privat	Text	Telefonnummer mit (internationaler) Vorwahl.
Durchwahl Büro	Text	Interne Durchwahlnummer zum Büro
Foto	O	
Bemerkungen	M	
Vorgesetzte(r)	Z	

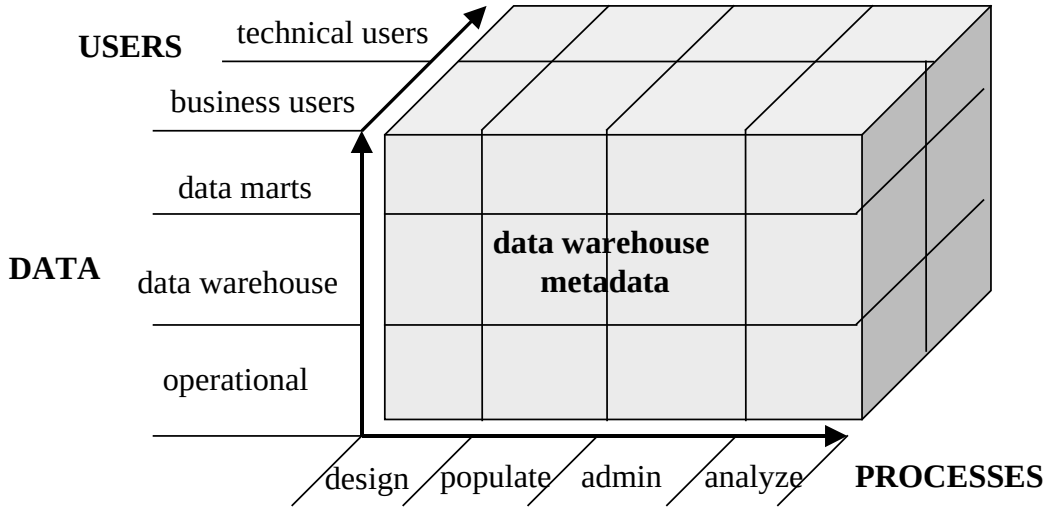
Personal : Tabelle

Personal-Nr	Nachname	Vorname	Position	Anrede	Geburtsdatum
1	Davolio	Nancy	Vertriebsmitarbeiterin	Frau	08. Dez. 48
2	Fuller	Andrew	Geschäftsführer	Herr	19. Feb. 52
3	Leverling	Janet	Vertriebsmitarbeiterin	Frau	30. Aug. 63
4	Peacock	Margaret	Vertriebsmitarbeiterin	Frau	19. Sep. 37
5	Buchanan	Steven	Vertriebsmanager	Herr	04. Mrz. 55
6	Suyama	Michael	Vertriebsmitarbeiter	Herr	02. Jul. 63
7	King	Robert	Vertriebsmitarbeiter	Dr.	29. Mai. 60
8	Callahan	Laura	Vertriebskoordinatorin	Frau	09. Jan. 58
9	Dodsworth	Anne	Vertriebsmitarbeiterin	Frau	27. Jan. 66

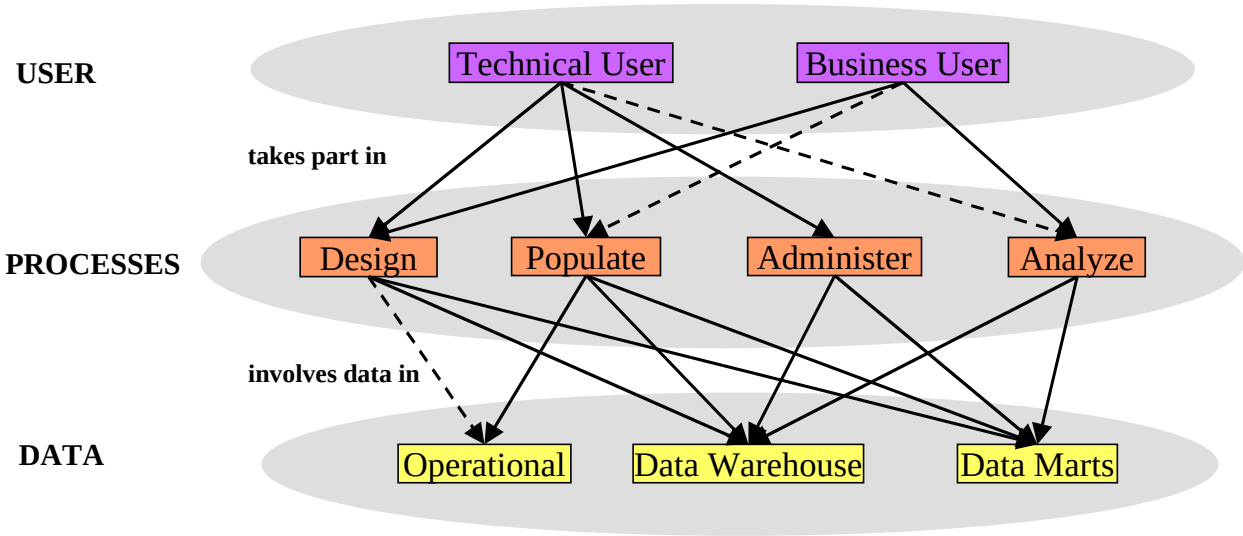
Meta Data in the Data Warehouse Context



Classification of Data Warehouse Meta Data



Not all combinations are relevant



Technical Meta Data

- **Schemas**
  - Database schemas, data file formats
- **Source-, Target-system**
  - Technical details for access (IP, protocol, username and password, etc.)
- **Data dependencies: (technical) Mappings**
  - Operational Systems $\leftrightarrow$ Data Warehouse, Data Marts: Data transformation rules
  - Data Warehouse, Data Marts $\leftrightarrow$ tools for data access: Technical description of queries, reports, cubes (SQL, Aggregation, Filters, etc.)
- **Warehouse-Administration (data updating, archiving, optimization)**
  - System statistics (e.g. usage patterns, user-/group-specific CPU/IO-usage) for resource planning and optimization
  - Frequency (scheduling), Logging information, execution status of jobs
  - Rules, functions for data selection for archiving

Business Meta Data

- **Information models, conceptual data models**
- **Company-/sector-specific business terms, vocabulary, terminologies, ontologies**
- **Mapping between business terms and Warehouse/Data Mart-elements (dimensions, attributes, facts)**
- **Description of queries, reports, cubes, operating numbers (from the business perspective)**
- **Data quality**
  - Lineage: from which sources originate the individual data values? owner?
  - Accuracy: which transformations have been applied to the data values?
  - Timeliness: when was the last update?
- **User information**
  - Relationships between users, roles of users, information objects, fields of interest and activities
  - Assignment of users to roles, of roles to activities or fields of interest, and of activities to information objects and fields of interest

Business Meta Data: Example

Business terms in the sector of insurances

Liability Insurance:

Insurance covering the legal liability of the insured resulting from injuries to a third party to their body or damage to their property.

Life Insurance:

Insurance providing payment of a specified amount on the insured's death, either to his or her estate or to a designated beneficiary.

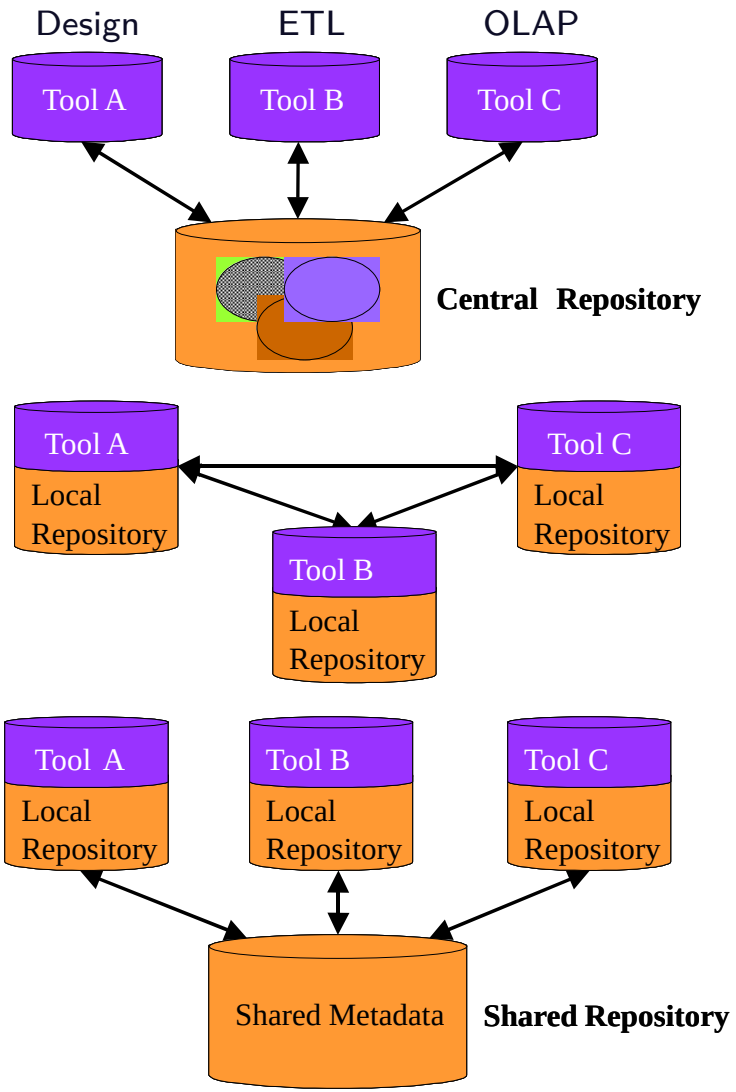
Liquor Liability Insurance:

Provides protection for the owners of an establishment that sells alcoholic beverages against liability arising out of accidents caused by intoxicated customers.

Long-Term Disability Insurance:

Insurance to provide a reasonable replacement of a portion of an employee's earned income lost through serious illness or injury during the normal work career.

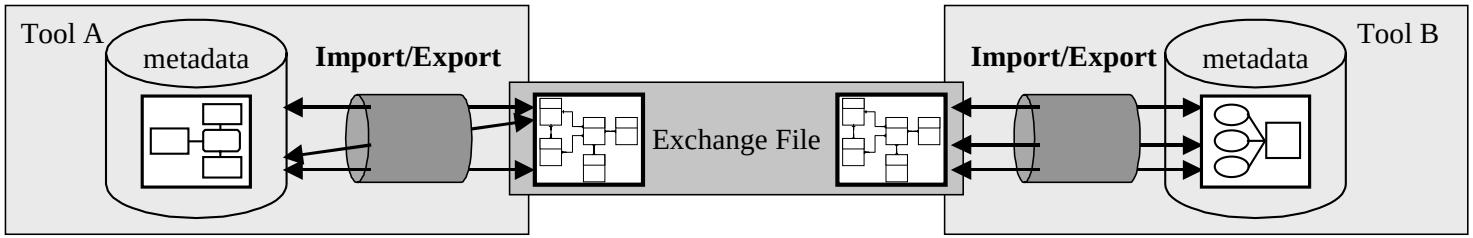
Meta Data: Alternative Architectures



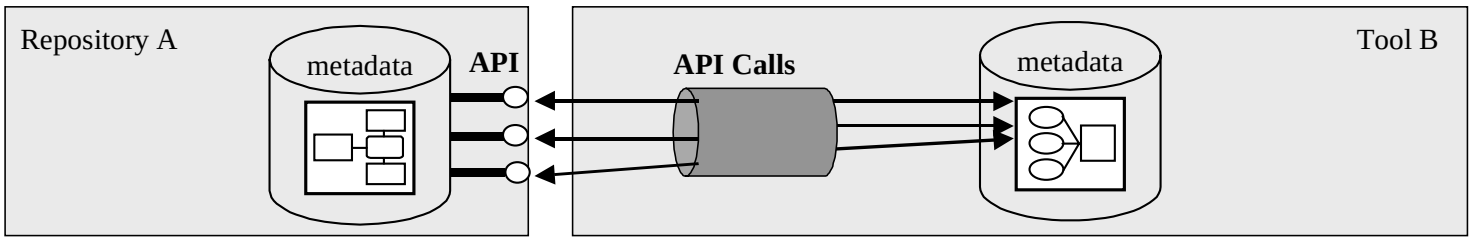
- **Central repository**
  - No replication of meta data
  - Dependency to central repository also for local meta data
  - Low autonomy, slow meta data access
- **Distributed repositories**
  - Maximal independence
  - Fast access to local meta data
  - Several connections to exchange meta data
  - High degree of meta data replication
  - Synchronization is complicated
- **Federated (shared repository)**
  - Uniform representation of shared meta data
  - Local autonomy
  - Limited amount of meta data exchange
  - Controlled redundancy

Mechanisms for Interoperability (1)

- **Exchange of data files**
  - No direct access to repository, asynchronous
  - Platform-independent and simple to realize
  - Standard formats: MDIS, CDIF, XML

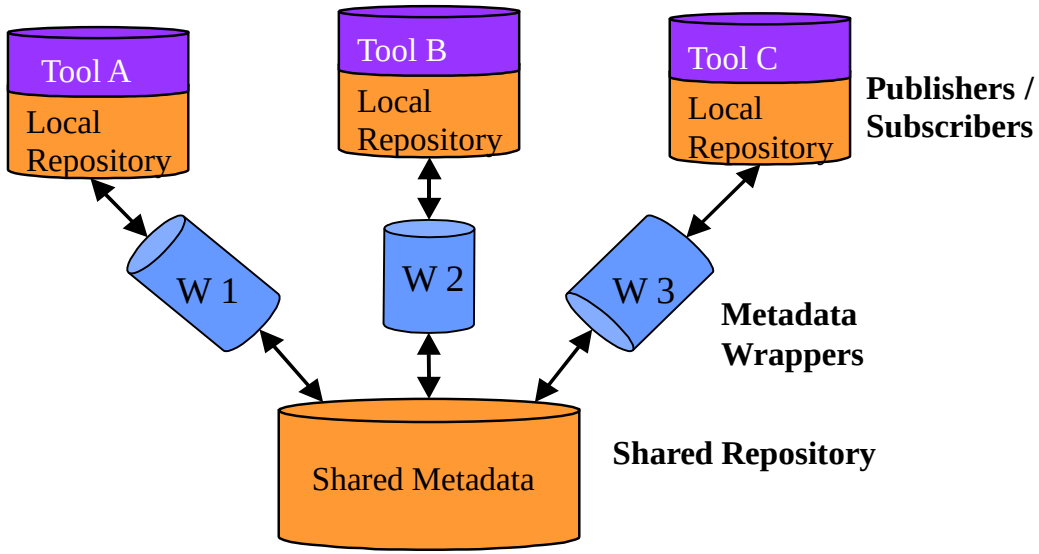


- **Application Programming Interface (API)**
  - Direct access to repository, synchronous
  - Most often proprietary
  - Standards for data and meta data access: ODBC, OLEDB for OLAP



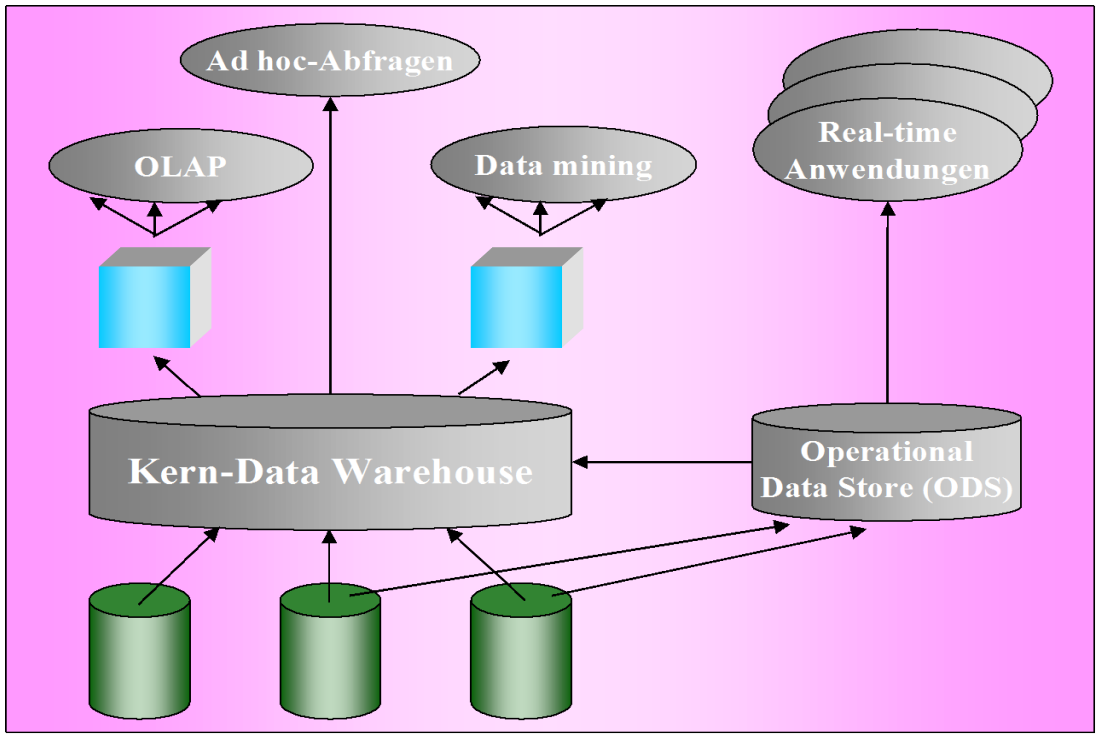
Mechanisms for Interoperability (2)

- **Meta data wrapper**
  - Mapping between different representations of meta data
  - Exchange between wrapper and repositories (local/shared) either asynchronous (file exchange) or synchronous (API-based)
  - Enables usage of a common API for shared repository
  - Adding a new tool becomes simpler
  - If local repositories have copies of shared data, centralized replication control by shared repository



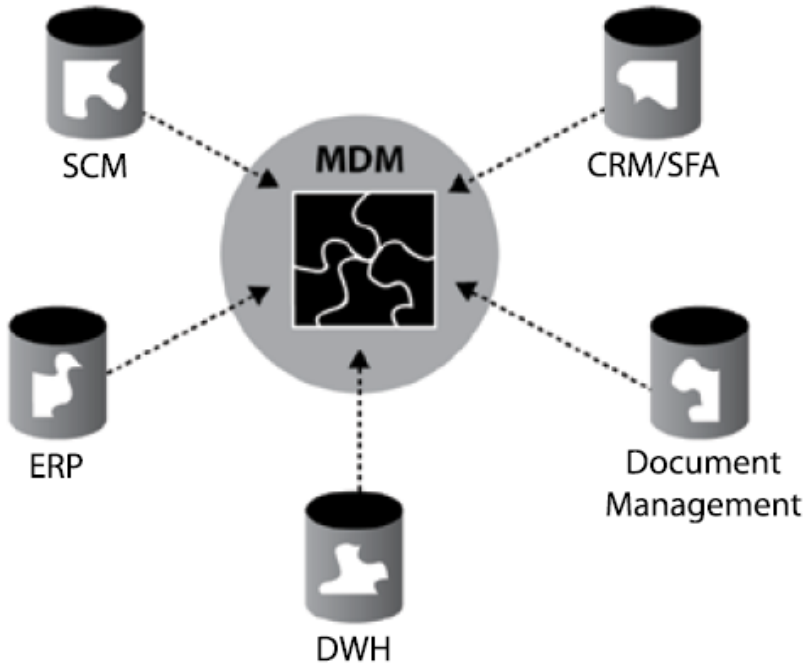
Operational Data Store (ODS)

- **Optional component of a DW architecture to support real-time applications (e.g. stock market-based) on the integrated data**
  - More up-to-date than the Data Warehouse, more data than the sources
  - Lower level of aggregation, not for analysis purposes (no historical data)
  - Integrated data can change, but is not passed back to operational DBs
  - Can serve as DW source
  - Example: preparation of an offer based on real-time market prices
- **Problems**
  - More redundant data
  - ODS contains changed data values
    $\Rightarrow$ Inconsistency between ODS and operational DBs



Master Data Management (MDM)

- **Usage of integrated master data (reference data) not only for analysis purposes, but also for operative applications and business processes**
  - CDI: Customer Data Integration
  - Data of products, accounts, employees, customers, suppliers, ...
- **MDM construction is similar to DWH construction, but has different purposes**
  - Replication (Caching) of master data in applications with the possibility to change
  - Master data and application data are (eventually) consistent
- **MDM has to be scalable and extendable**



Quelle: IBM

SCM = Supply Chain Management
ERP = Enterprise Resource Planning
CRM = Customer Relationship Management
SFA = Sales Force Automation

Summary

- **Components of the reference architecture**
  - Data sources
  - ETL components including Monitoring and Scheduling
  - Staging area
  - Data Warehouse and Data Marts
  - Analysis tools
  - Meta data management
- **Extraction strategies:** Snapshot, Trigger, Log-Transfer, Replication
- **Dependent vs. independent Data Marts**
- **Systematic management of DW-Meta data is necessary**
  - Technical meta data vs. business meta data, ...
  - Often: Coexistence of local repositories with proprietary meta data models
  - Interoperability of meta data (file exchange vs. Low-Level Repository APIs)
- **Support of operative applications on integrated data**
  - ODS: Online Data Store
  - MDM: Master Data Management