Data Warehouse

Martin Clement
Head of Analytics
Martin.Clement@atvantage.com

DHBW
Duale Hochschule
Baden-Württemberg

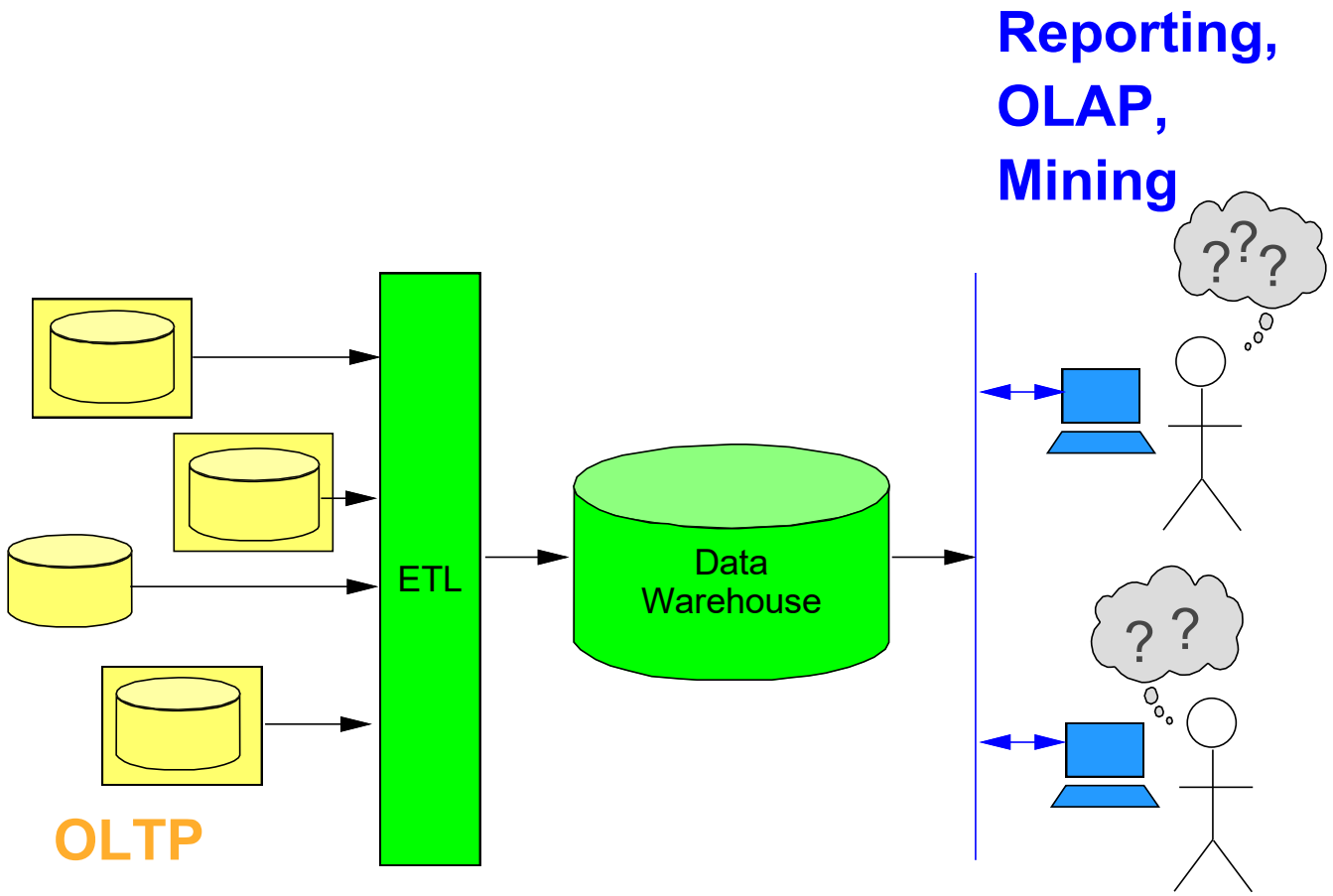
Data Warehouse, 2025

Extract –Transform - Load

ATVANTAGE

TIMETOACT GROUP

Basic Data Warehouse architecture

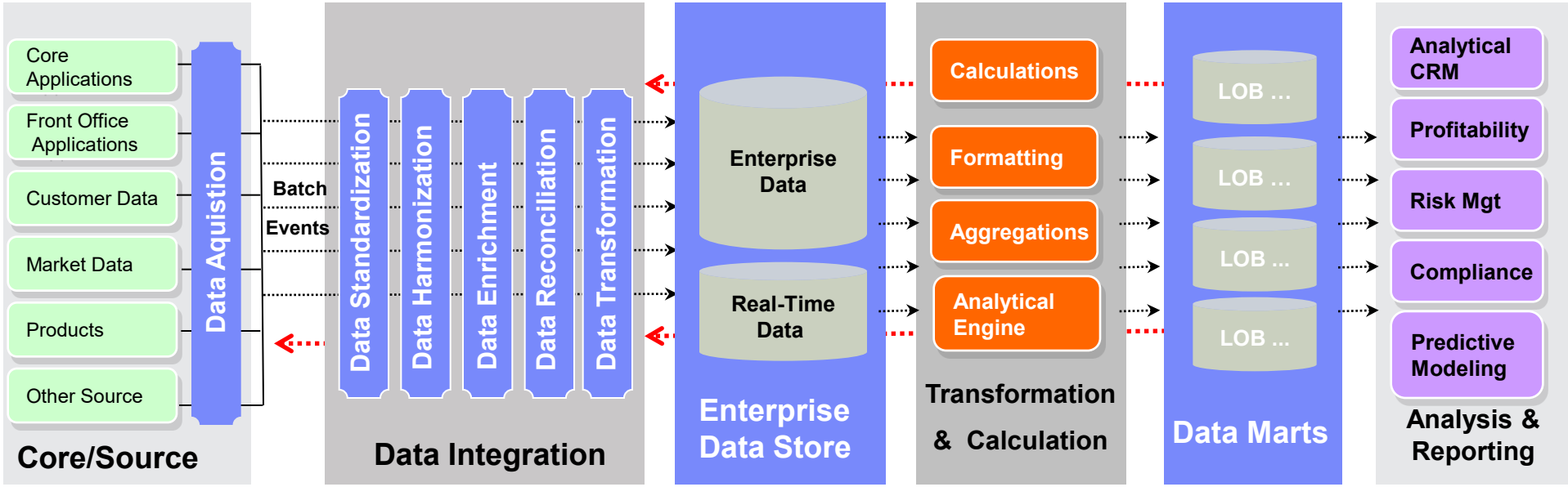


Tasks of the ETL Process

- Monitor changes in source systems
- Extract
  - capture and copy data from source systems (e.g. operational systems)
  - many different types of sources
    - Relational, hierarchical DBMSs
    - Flat files
    - Legacy systems
    - ERP systems
    - Other internal/external sources
- Transform
  - Filter data
  - Check and cleanse data
  - Transform data to CDW data model
- Load
  - fast load into staging area/CDW tables

Data Integration Architecture

Data-Driven Reference Architecture



Master Data Management - Data Quality Management - MetaData Management

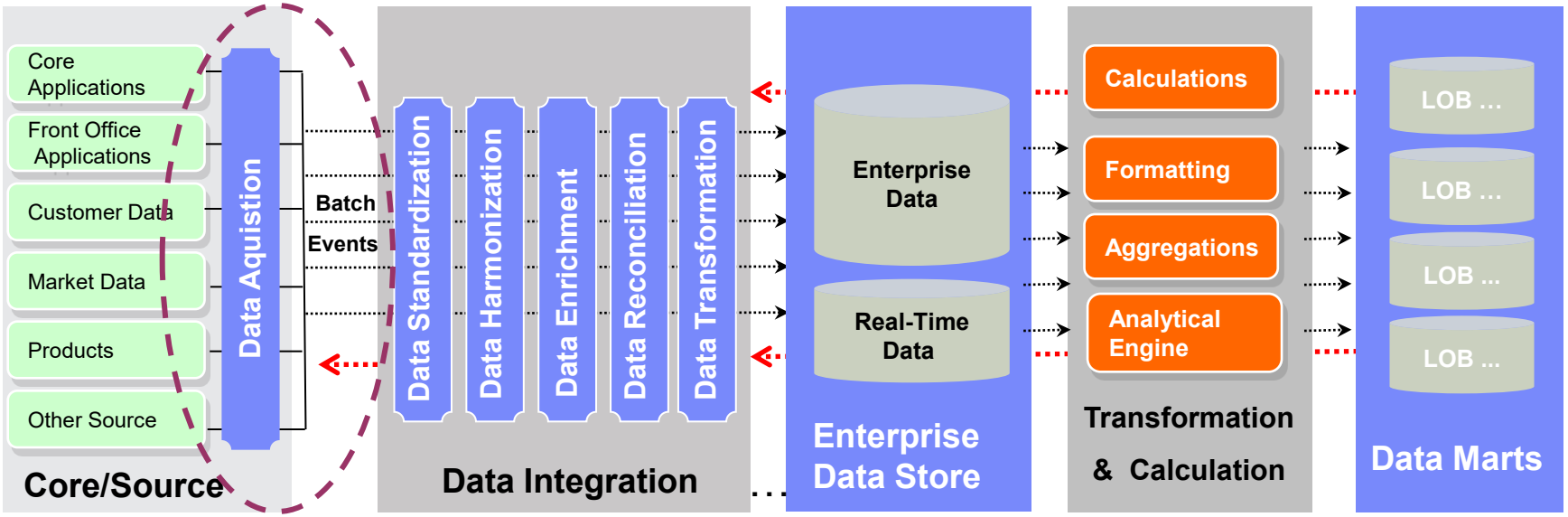
Governance (Process, Resources, Assets)

Infrastructure

Prerequisite of ETL - Understanding The Data

- Profile Existing Data Sources, Extracted Data
  - Analyze data structure, content, and quality
  - Find data relationships across systems
  - Uncover data issues that can affect subsequent transformation steps
    - missing values
    - Duplicates
    - inconsistencies

Data Integration Process



- **Data Acquisition**
  - Support data acquisition in the source systems with minimal impact into transaction systems
  - Capture of deltas for batch and real-time processing
  - Depends on data latency requirements

Monitoring

- Extracts from source systems
  - Intial extract for setting up the data warehouse
  - Periodical extracts for adding new/changed information to the data warehouse

- Question: How to determine what is new or what has changed in the source systems?

→ Task of „monitoring“

Monitor types

- The montitor determines all changes itself
  - the result of monitoring is a delta-file containg all changes
- The monitor determines only if something has changed (not what)
  - The changes have to be determined by the extraction-component

Aspects of monitoring

- Discovery of all changes vs. determining the net effect at extract/load time only
  - Example: an attribute value can be changed in two ways:
    - by one update operation
    - by one delete and one insert operation
  - The net effect of both is the same
  - However, history information is lost if the net effect is recorded only.
- Notification vs. Polling
  - Data sources notify monitor about changes (e.g. through triggers) or the monitor determines the changes itself
  - Frequency of polling
    - High → high load on source systems
    - Low → history information may be lost

Aspects of monitoring

- Internal vs. External monitoring
  - Internal:
    the source sytems offers sufficient hooks for determining the changes
    - E.g. Notification about changes
  - External:
    the changes have to determined from outside the source systems
    - E.g.; by comparinig consequtive extracts

Monitoring techniques

- Depend on characteristics of the data sources
  - based on modern relational DBMS
  - ...
  - legacy systems with proprietory data storage
- Types of techniques
  - Based on DBMS
    - Active mechanisms
    - Replication techniques
    - Protocol-based discovery
  - Controlled by application
    - Timestamp-based discovery
    - Snapshot-based discovery

Monitoring based on DBMS

- Active monitoring mechanisms
  - Based on (database) triggers
    - Example:
      - If new record is inserted in sales transaction table then insert transaction id and timestamp in change table
  - Advantage:
    - Triggers do not change operational applications
  - Disadvantage:
    - Performance impact on operation systems if triggers are used extensively

Monitoring based on DBMS

- Replication techniques
  - Snapshot
    - Read-only, local copies of one or multiple tables of the source systems
    - Incremental update through snapshot-log
      - Snapshot log
        - contains information of changes wrt to last snapshot operation
        - implemented through internal triggers
    - The monitor can use snapshot log to determine what has changed
  - Data replication
    - Target tables (like snapshot tables), not necessarily on local system
    - Creates delta-tables with the all (committed) changes, not only net effect

Monitoring based on DBMS

- Protocol-based discovery
  - Usage of database logs to determine changes
    - DBMSs write transaction logs in order to be able to undo partially executed transactions
    - This information can be used to determine all changes
  - Requires exact knowledge of the protocol mechanisms of the DBMS
  - Protocol files can be transferred to other systems to avoid additional load on source systems

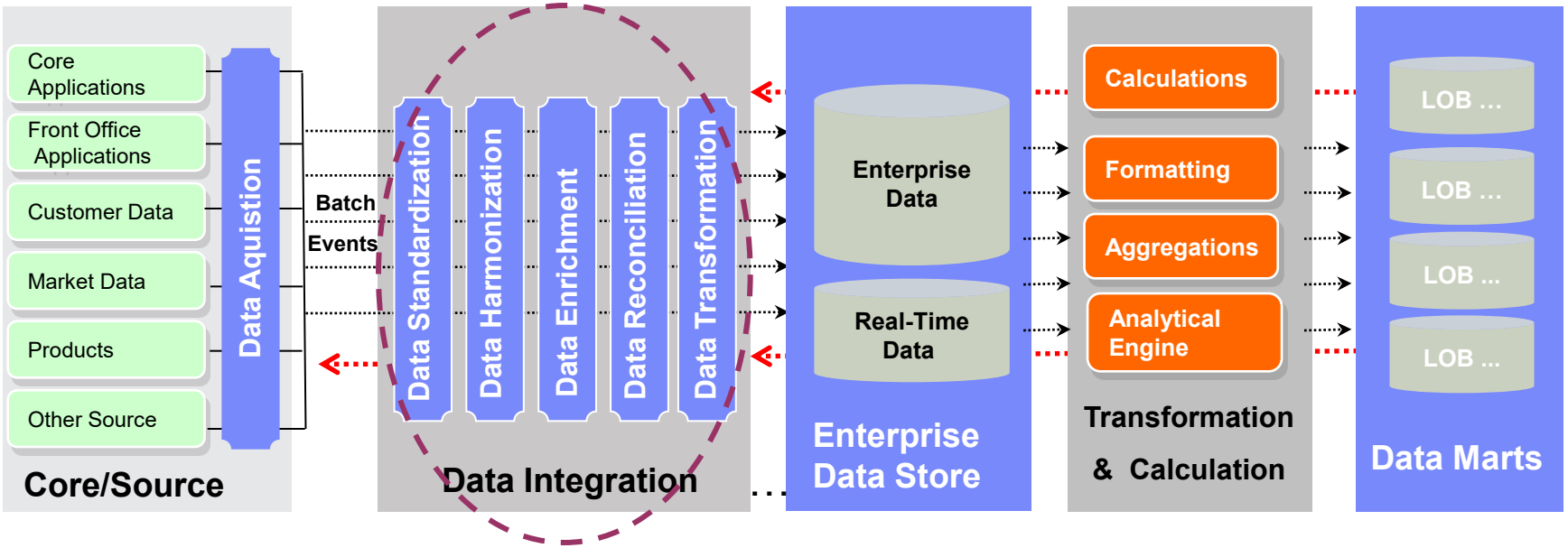
Monitoring controlled by the applications

- Approaches
  - Operational applications log themselves the changes in the operational data
    - Can require changes in the source code of the applications
      - Very difficult for legacy applications
    - Can have an impact on the performance
  - Independant monitoring applications that discover the changes in the operational applications

Application based monitoring techniques

- Timestamp based discovery
  - Every data item is associated with timestamp information about its validity period
  - Changed data can be determined from this timestamp information
  - Operational applications have to keep a limited change history
- Data comparison
  - Comparison of shapshots of the operational data at different points in time
  - Can be very complex
  - Sometimes the only possibility, for instance for legacy applications

Data Integration Process

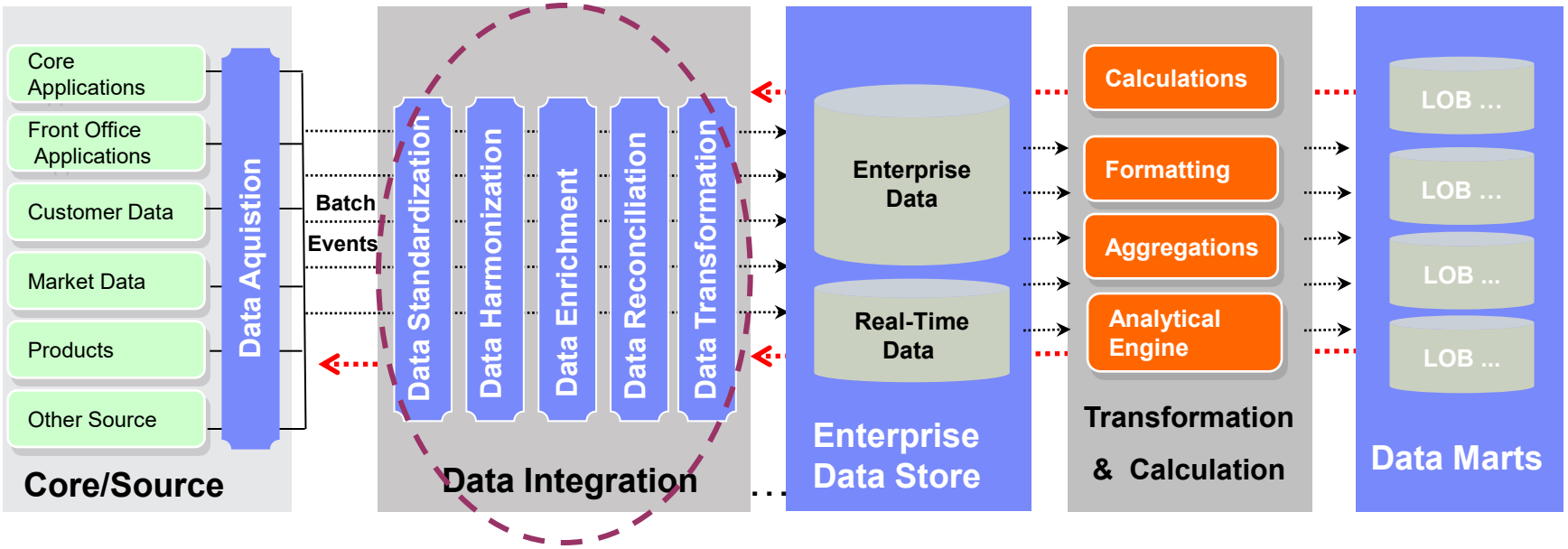


- **Data Integration**
  - Integrate source data into the consolidated Enterprise Data Warehouse
  - Implement complex integration processes
  - Building the „Single Version of Truth“

Extract

- Extraction intervals
  - Periodically – in regular intervals
    - Every day, week, etc.
    - Depends on the requirements on timeliness of the data warehouse data
  - Triggered by a specfic request
    - Addition of a new product
    - Query which involves more recent data
  - Triggered by specific events
    - Number of changes in operational data exceeds threshold
  - Instantly
    - Every change is directly propagated into the data warehouse
    - „real time data warehouse“

Data Integration Process



- **Data Standardization**
  - Match and reconcile related data elements
  - Remove duplicate, redundant data
  - Reengineer data to match single corporate standard
  - Quality processing to ensure data integrity

Transform - Integration of data

- Different keys for same entities in source data

AD_SYS	...	Kunde_Text	LOADTIME
AD-FX8257		Müller	31DEC1999:23:03:08
AD-FH2454		Meier	31DEC1999:23:03:08
AD-FX7059		Schulz	31DEC1999:23:03:08
AD-FT2567		Schmitz	31DEC1999:23:03:08
...	...	...	...

AC_SYS	Kunde_Text	Kunde_Status
3857_ACC	Müller	A
3525_ACC	Meier	A
3635_ACC	Schulz	A
3566_ACC	Schmitz	B
...	...	...

CC_SYS	Kunde_Grp	Kunde_Text	LOADTIME
59235395	Handel	Müller	31DEC1999:23:03:08
08485356	Industrie	Meier	31DEC1999:23:03:08
08555698	Industrie	Schulz	31DEC1999:23:03:08
85385386	Handel	Schmitz	31DEC1999:23:03:08
...	...	...	...

Kunde_ID	Kunde_Text	...	AD_SYS	CC_SYS	AC_SYS	...	LOADTIME
0001	Müller		AD-FX8257	59235395	3857_ACC		31DEC1999:23:03:08
0002	Meier		AD-FH2454	08485356	3525_ACC		31DEC1999:23:03:08
0003	Schulz		AD-FX7059	08555698	3635_ACC		31DEC1999:23:03:08
0004	Schmitz		AD-FT2567	85385386	3566_ACC		31DEC1999:23:03:08
...	...	...	...	...	...	...	...

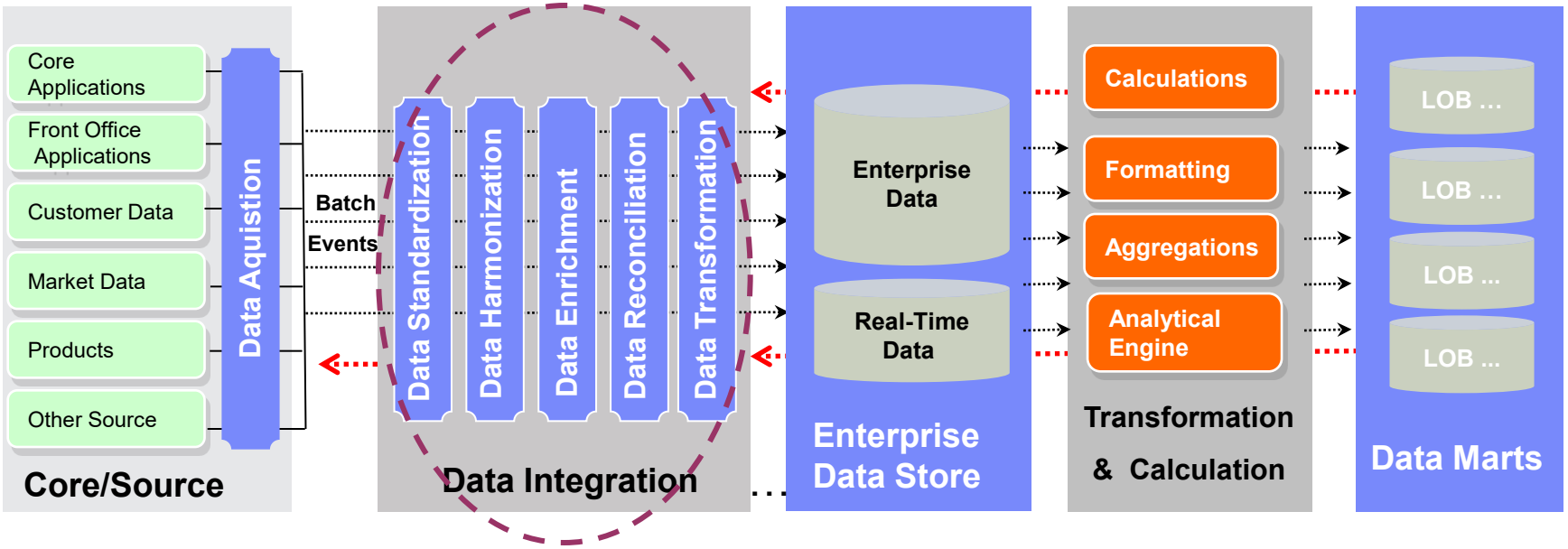
Legende: AD – Außendienstsystem, CC – Call-Center-Anwendung, AC – Abrechnungs-/Accounting-System

Quelle: Finger, R. (2002), Historisierungskonzepte, Vortrag im Rahmen der Seminarreihe „Data Warehouses und Data Marts – Effizienter Einsatz für das Controlling", Frankfurt am Main 2002.

Transform - Integration of data

- Different keys for same entities in source data
  - Solution
    - Creation of a unique surrogate key
    - Storing the association of surrogate keys to the keys in the operational system in a mapping table
  - Determining if different keys point to the same entity is a difficult problem
    - Different ways of how a name can be written
      - Peter Maier
      - Maier, Peter
      - P. Maier

Data Integration Process



- **Data Harmonization**
  - Standardization of data so that they can be matched with other data and information regardless of the format.
  - Consolidate, Integrate, Cleanse, Normalize & Harmonize

Transform - Integration of data

- Unification of data
  - Unification of data types
    - Character string → date
      „20.01.2006“ → 20.01.2006
    - Character string → number
      „12345“ → 12345
  - Unification of encodings
    - For instance for gender F and M
    - Lookup-tables contain the mapping from old to new encodings
  - Unification of character strings
    - Unique way of naming and writing certain concepts
      - Küchengerät → Kuechengeraet
      - customer → client
    - Names:
      - „last name“, „first name“ like „Maier, Peter“

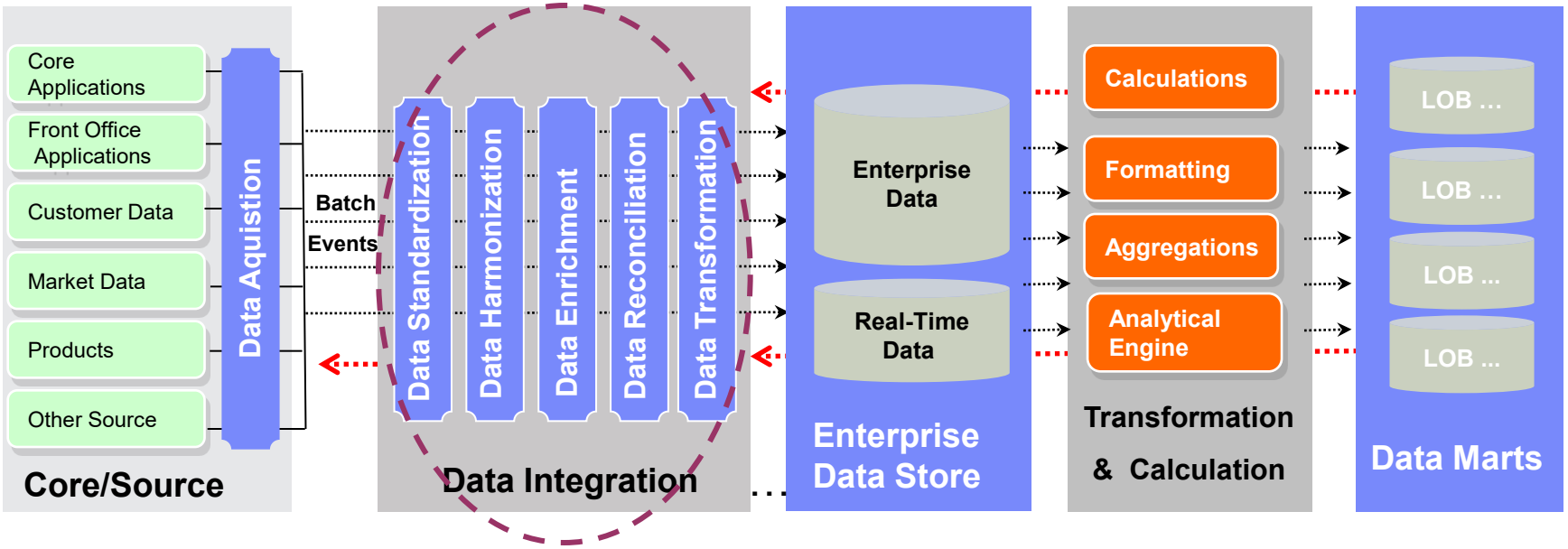
Transform - Integration of data

- Unification of data (cont.)
  - Unification of dates and timestamps
    - Rules for representing incomplete date information
      - If only month and year are known
    - Dates and timestamps wrt one specific timezone
      - Important for multi-national organizations
  - Combination of diffferent attributes to one attributes
    - day, month, year → date
  - Split of one attribute into two or more
    - Name → first name, last name
    - Product name - „Cola, 0.33 l“ →
      Product short name - „Cola“, size in liters - 0.33

Transform - Integration of data

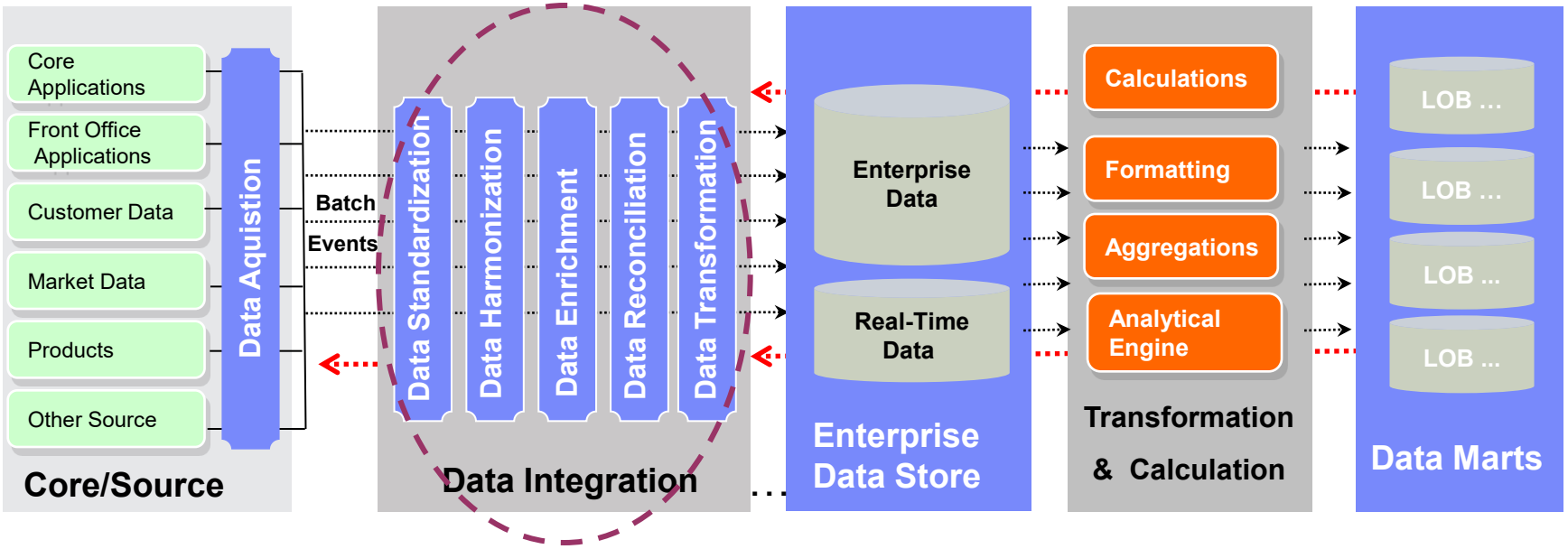
- Unification of data (cont.)
  - Computation of derived values
    - Proftit = sales price – purchase price
  - Aggregations
    - Revenue of the month/week computed from revenues of the day

Data Integration Process



- **Data Enrichment**
  - Supplementing internal data with data from external sources
  - Personal data such as date-of-birth and gender codes
  - Geographical data
  - Postal Data, Demographic information, Economic data , etc

Data Integration Process



- **Data Reconciliation**
  - Process to integrate and reconcile a view of data of the organization
    - Comparing Information from Multiple Sources
    - Correct one or both sources
    - Create a single source

Transform – Data cleansing

- Types of data quality problems
  - Wrong values of attributes
    - Phone number as address
    - Outliers for numeric attributes
      - Unlikely high price for a product
    - Can be discovered by statistical methods
  - Inconsistencies
    - Order date after payment date
    - Can be discovered by statistical and data mining methods
  - Multiple representations for one entity
    - Keys
    - Encodings

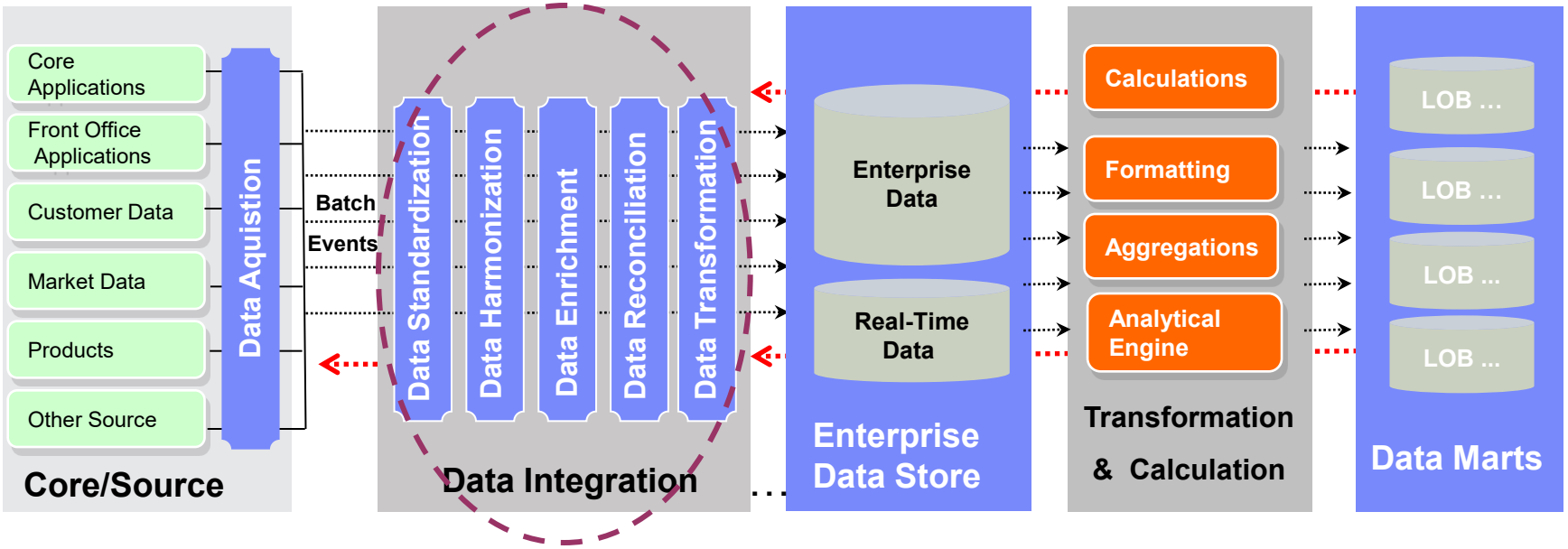
Transform – Data cleansing

- Types of data quality problems (cont.)
  - Missing values
    - In applications missing values are encoded by specific values like 01.01.1900 for unknown date
    - In databases NULL represents an unknown value
    - Missing values can represent
      - an unknown value
        - like date of birth of a customer
      - a value that does not exist
        - like „engine type“ for bicycle in a vehicles table
      - it is possible that both cases are possible
        - like e-mail address of a customer

Transform – Data cleansing

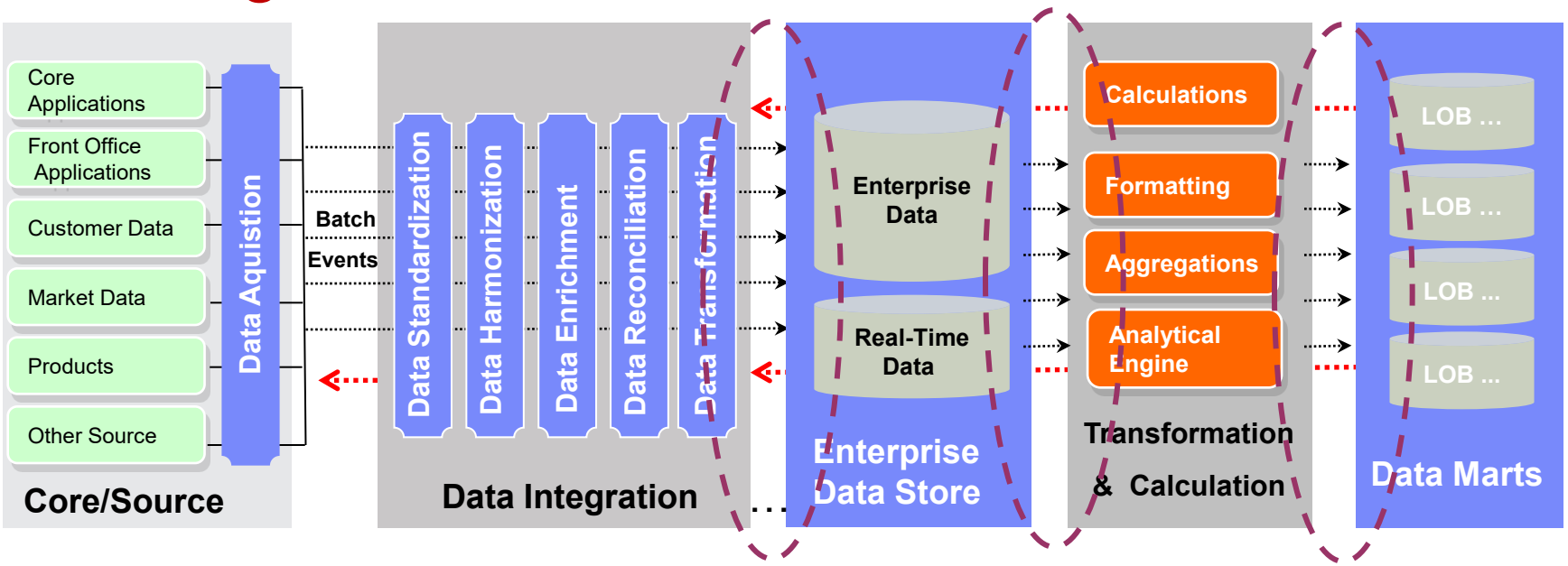
- Correcting the data
  - Automatically
    - E.g., address of a customer if a correct reference table exists
  - Manually
  - At the source systems
    - Common master data management across all operational applications

Data Integration Process



- **Data Transformation**
  - Apply transformations to source data to adequate to corporate standards (based on transformation rules), so that data can be leveraged for data analysis

Data Integration Process

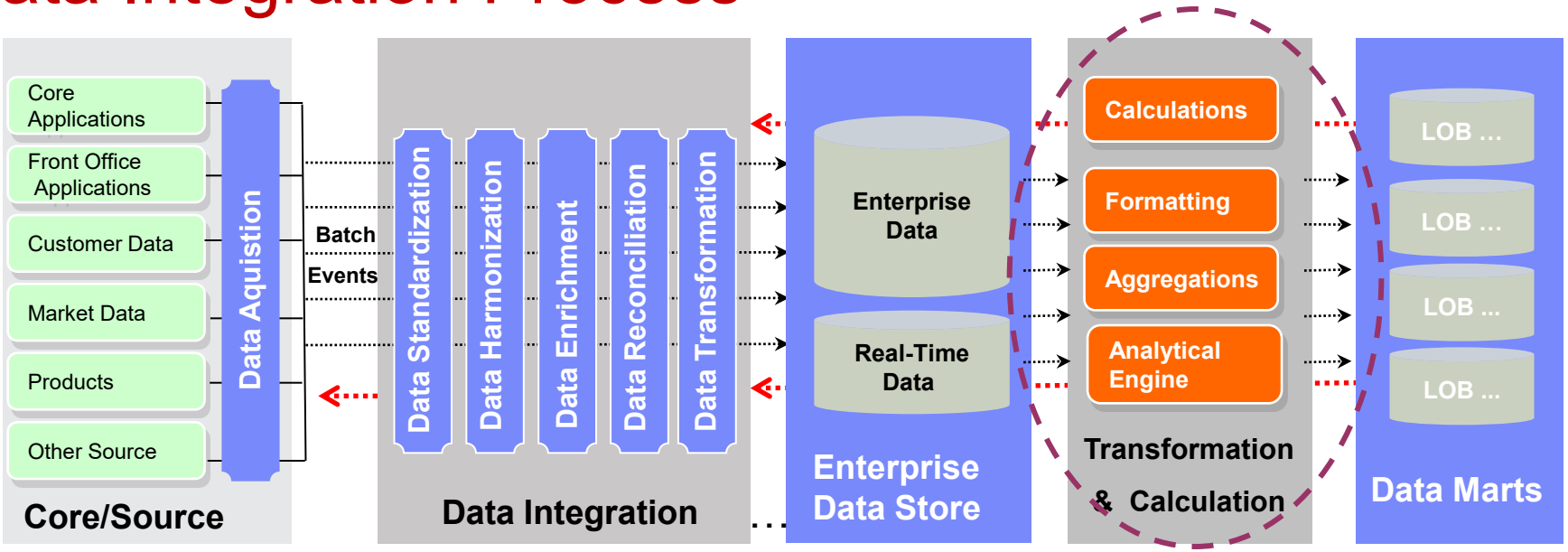


- **Data Publish and Delivery**
  - Guarantee delivery of the data with integrity
  - Populates targets structures in the EDW and DM
  - Publish information for close loop analysis (integration with operational systems)

Load

- Efficient load operations are important
  - „bulk load“
- Online load
  - Data warehouse is still accessible
  - For incremental updates
- Offline load
  - Data warehouse is offline
  - For updates that require the recomputation of a cube

Data Integration Process



- **Transformations and Calculations**
  - Calculations: To meet specific business requirements for LOB (data marts)
  - Formatting: Format data for easy understanding and analysis by end users
  - Aggregations: Summarize data for analysis
  - Analytical Engines: Data Mining and others analytical engines required for advanced analysis

Excercise

- For one of the following companies
  - Retail Bank
  - Telecommunication company
  - Online book store (like Amazon.com)
  - Discount furniture store (like IKEA)
  - Supermarket

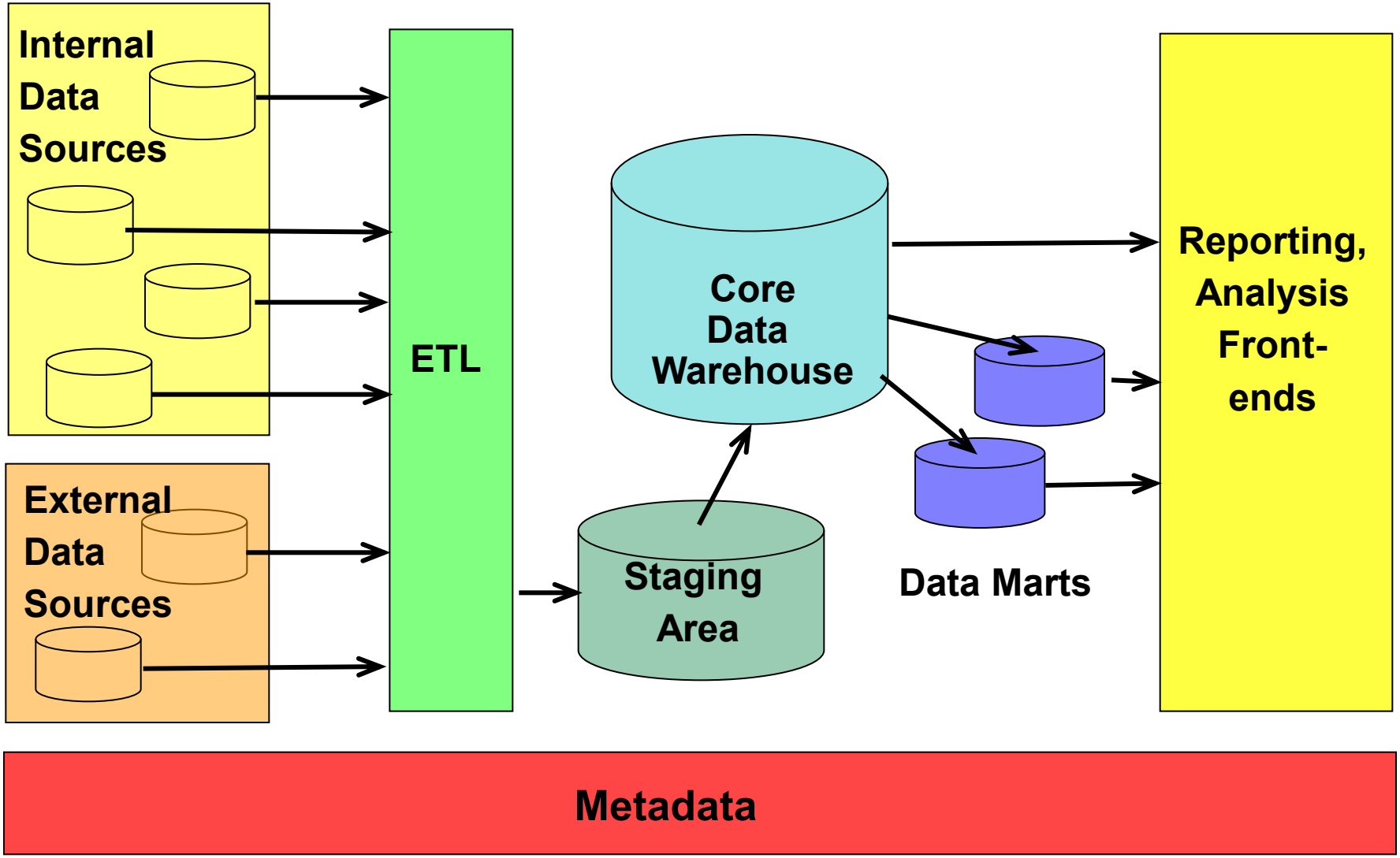
  describe 5 potential data quality problems.
- Which impact might these problems have on its business?

Metadata

ATVANTAGE

TIMETOACT GROUP

Standard Data Warehouse Architecture



What Is Metadata, anyway?

Types of metadata

- Business Metadata
  - Definition of business vocabulatory and relationships
  - Definition of the value range
  - Linkage to physical representation

GL Account Number

The ten digit account number for general ledger. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

Business

Technical

Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = ACCT_NO

data type = char(11)

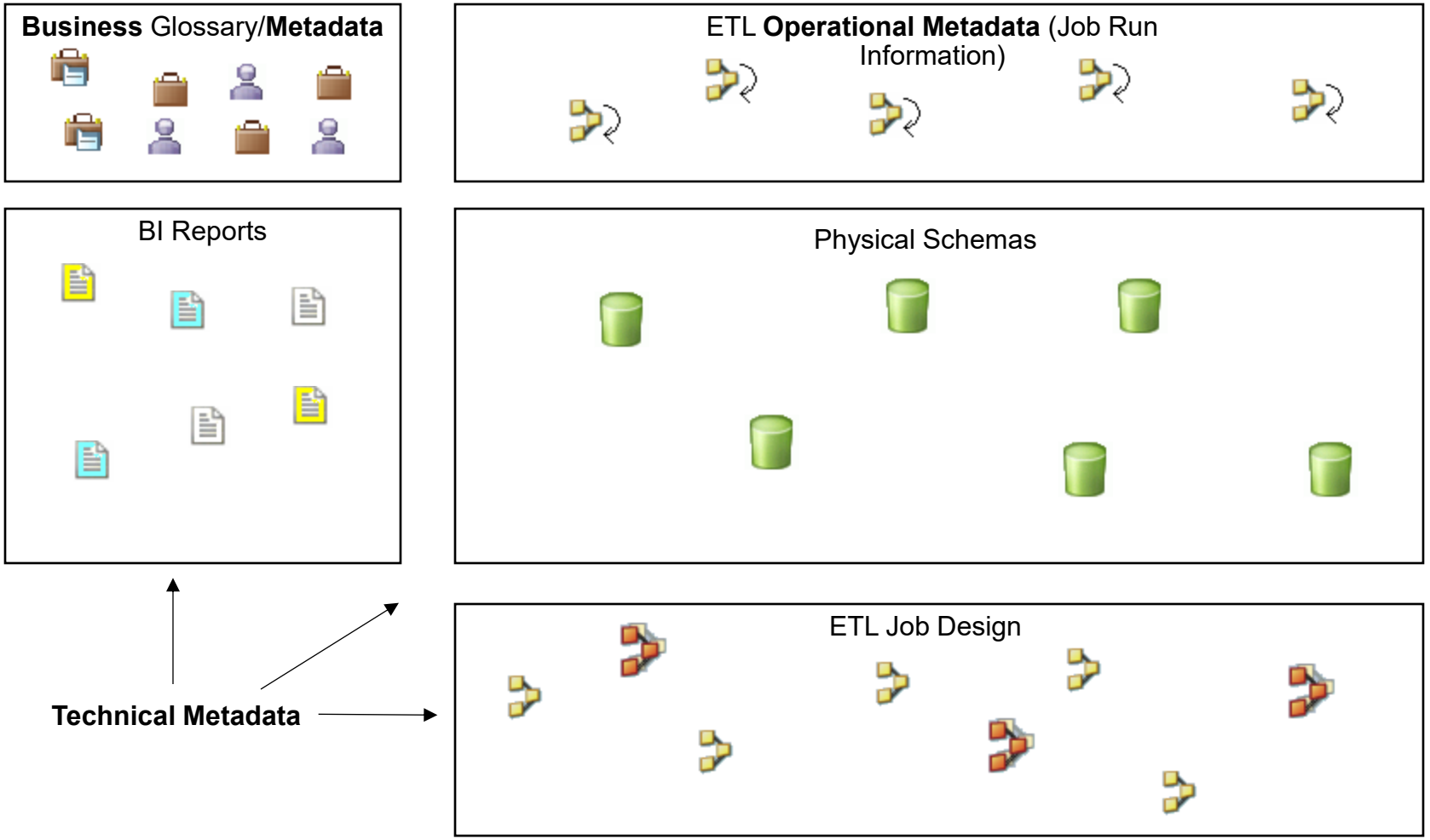
Types of metadata (cont.)

- Technical metadata
  - Report metadata
    - Report definitions
    - Data sources
    - Colum definitions
  - Logical and physical metadata of data model
    - Table structure
    - Definition of columns
    - Relationships between tables and columns
  - ETL metadata
    - Job design
    - Input-/output tables
    - computations
- Operational metadata (of ETL jobs)
  - Start time and duration
  - Return code

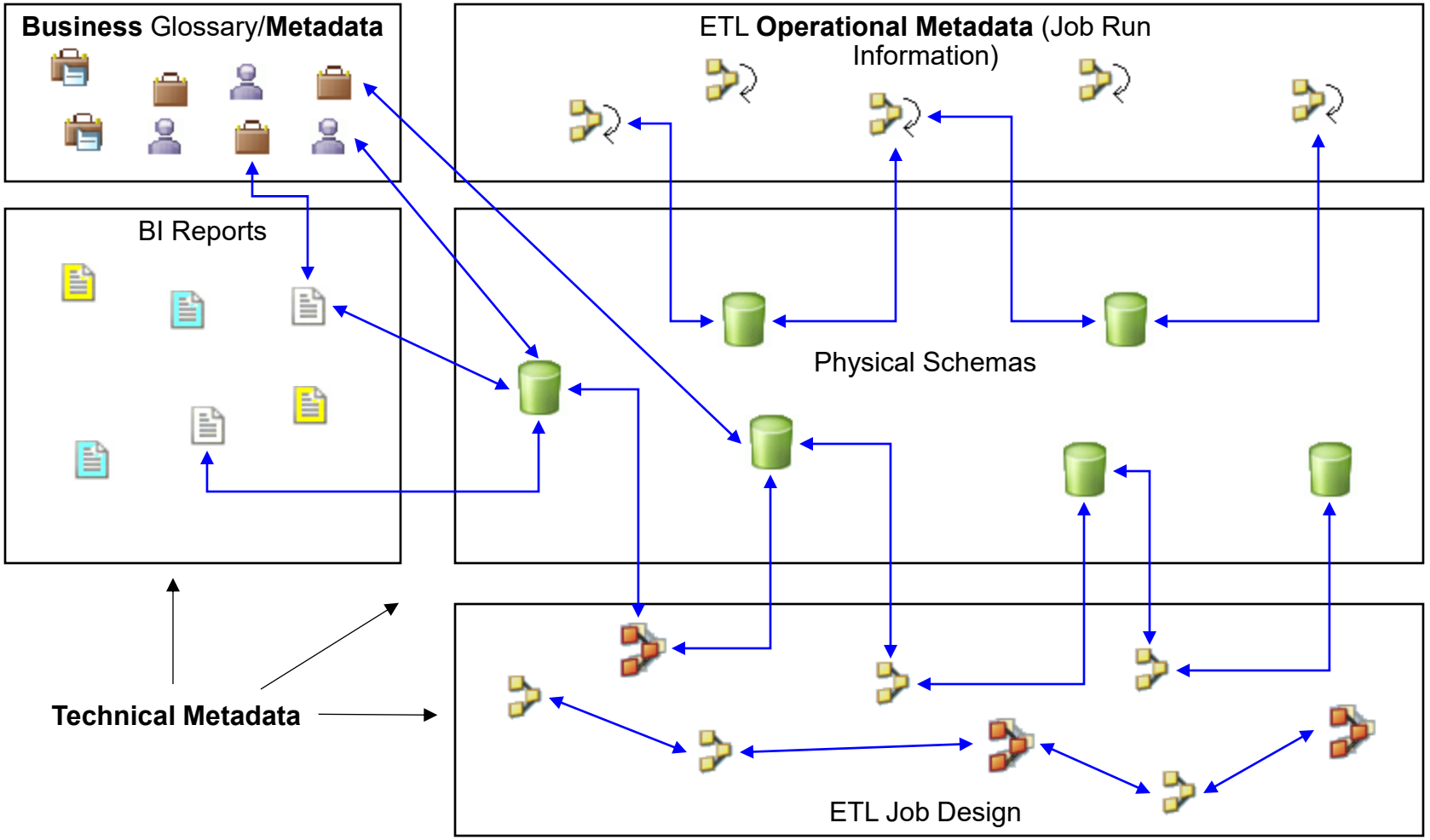
Why to use a common metadata repository

- Components of a data warehouse system are interconnected
  - BI report designer has to know
    - the table definitions
    - the meaning of the column values
  - ETL job designer has to know
    - the table definitions
    - the exact definition of the measures
  - ...
- Common metadata repository ensures consistency accross all components
- Enables cross component metadata analyisis
  - Data Lineage
  - Impact Analysis

The Areas of Metadata

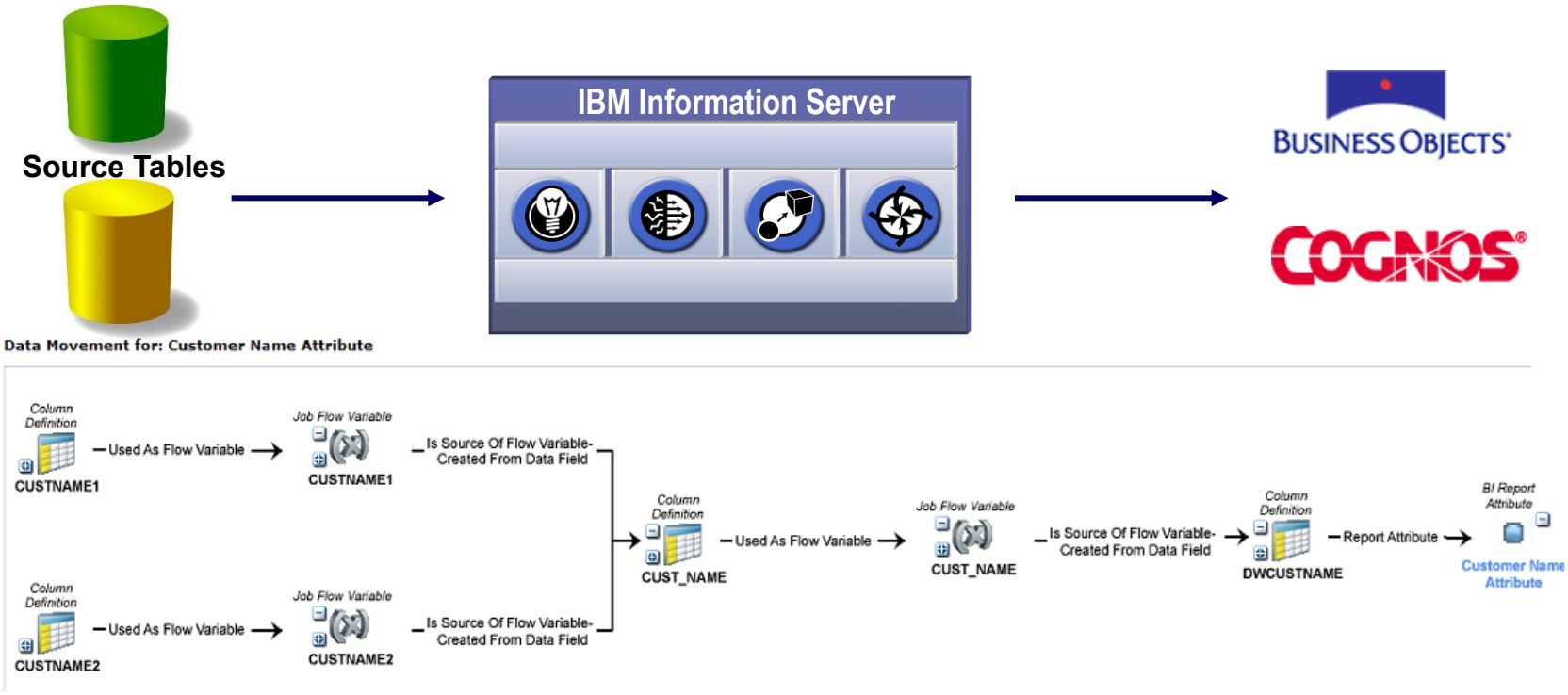


The Areas of Metadata Connected



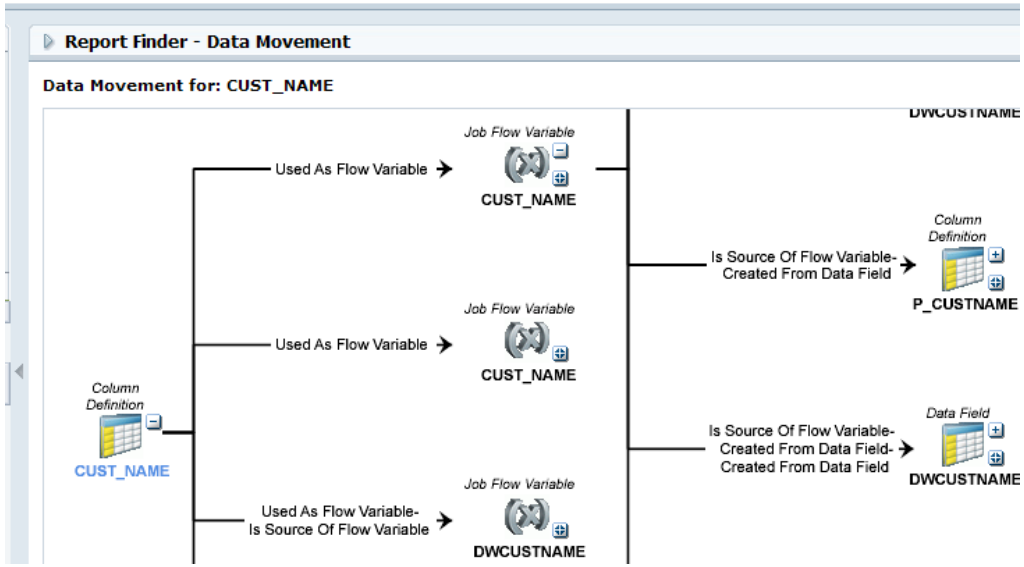
Where does a field of data in this report come from?

- "Data lineage"
  - Import & Browse Full BI Report Metadata
  - Navigate through report attributes
  - Visually navigate through data lineage across tools
  - Combines operational & design viewpoint



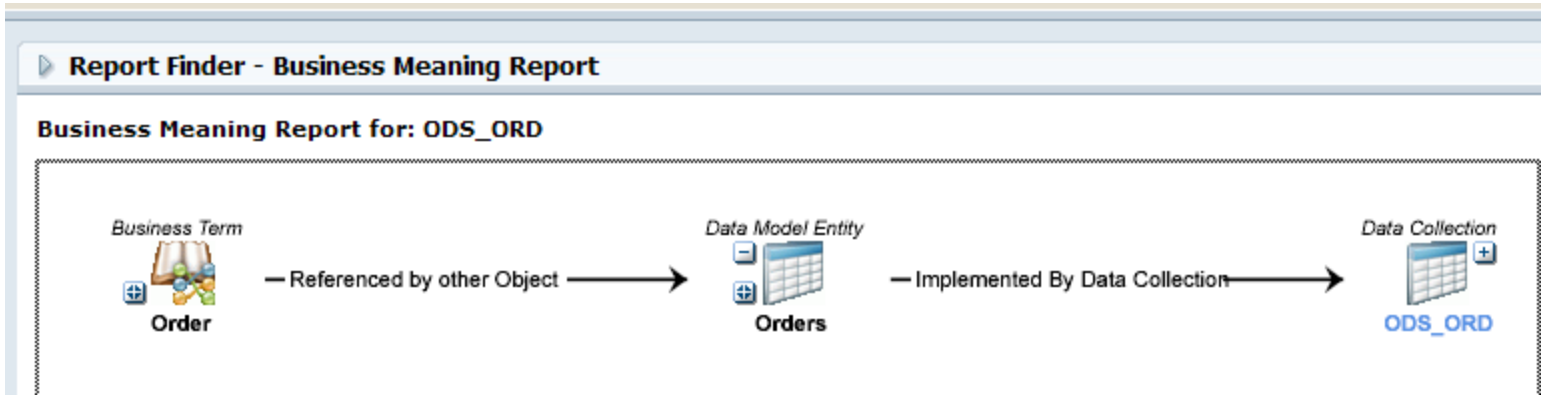
What happens if I change this column?

- "Impact Analysis"
  - Show complete change impact in graphical or list form
  - Includes impact on reports in BI tools
  - Visually navigate through impacted objects across tools
  - Allows impact analysis on any object type



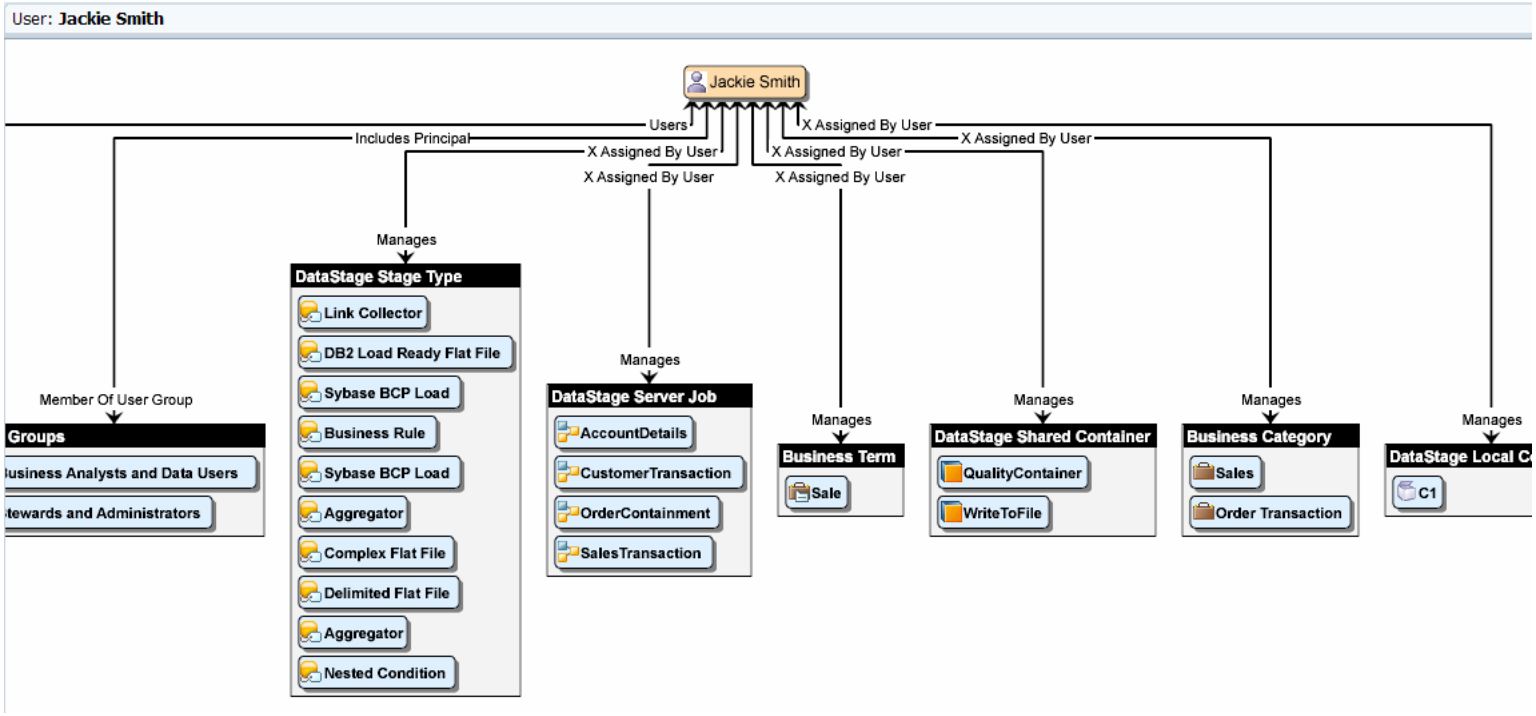
What does this field mean?

- Show relationships between business terms, data model entities, and technical and report fields
- Provides cross-tool mapping of business terms
- Allows field meaning to be understood
- Allows business term relationships to be understood



What objects does this user own?

- Shows objects that user manages
- Shows stewardship relationships on business terms
- Shows user group associations

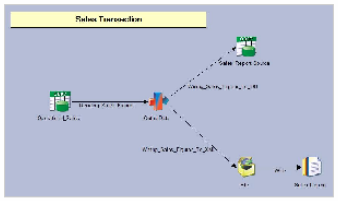


What happened on the last job run?

- Navigation through complete job details
- Navigation of complete operational metadata

DataStage Server Job: **SalesTransaction**

Image



Sales_TX.JPG

DataStage Job

Job	SalesTransaction
Type	SERVER
Project	LAURIE-DEMO1:TAB
Folder	\\Jobs\\ACME_Sales
Description	Reconcile all current transactional information into the operational data store, comprising activity of the past 12 months.
Data Steward	Jackie Smith
Contains DataStage Stages	File Operational_Sales OutputData Sales_Report Sales_Report_Source
Contains DataStage Links	Reading_Sales_Figures Write Writing_Sales_Figures_To_DB Writing_Sales_Figures_To_XML
Contains DataStage Local Containers	None

DataStage Job Usage

Operational Job Locator Information	SoftwareExecutable(Job)=";SalesTransaction";
Job Runs	SalesTransaction 2007-01-24 16:21:12 SalesTransaction 2007-01-24 18:26:44 SalesTransaction 2007-01-25 12:06:21
Previous DataStage Job	OrderContainment

Exercises

- List 3 reasons why common metadata is important in the context of warehousing.
- Define criteria for the evaluation of an ETL tool
- How does a relational DBMS (like Oracle, DB2, MS SQL Server) meet these requirements?