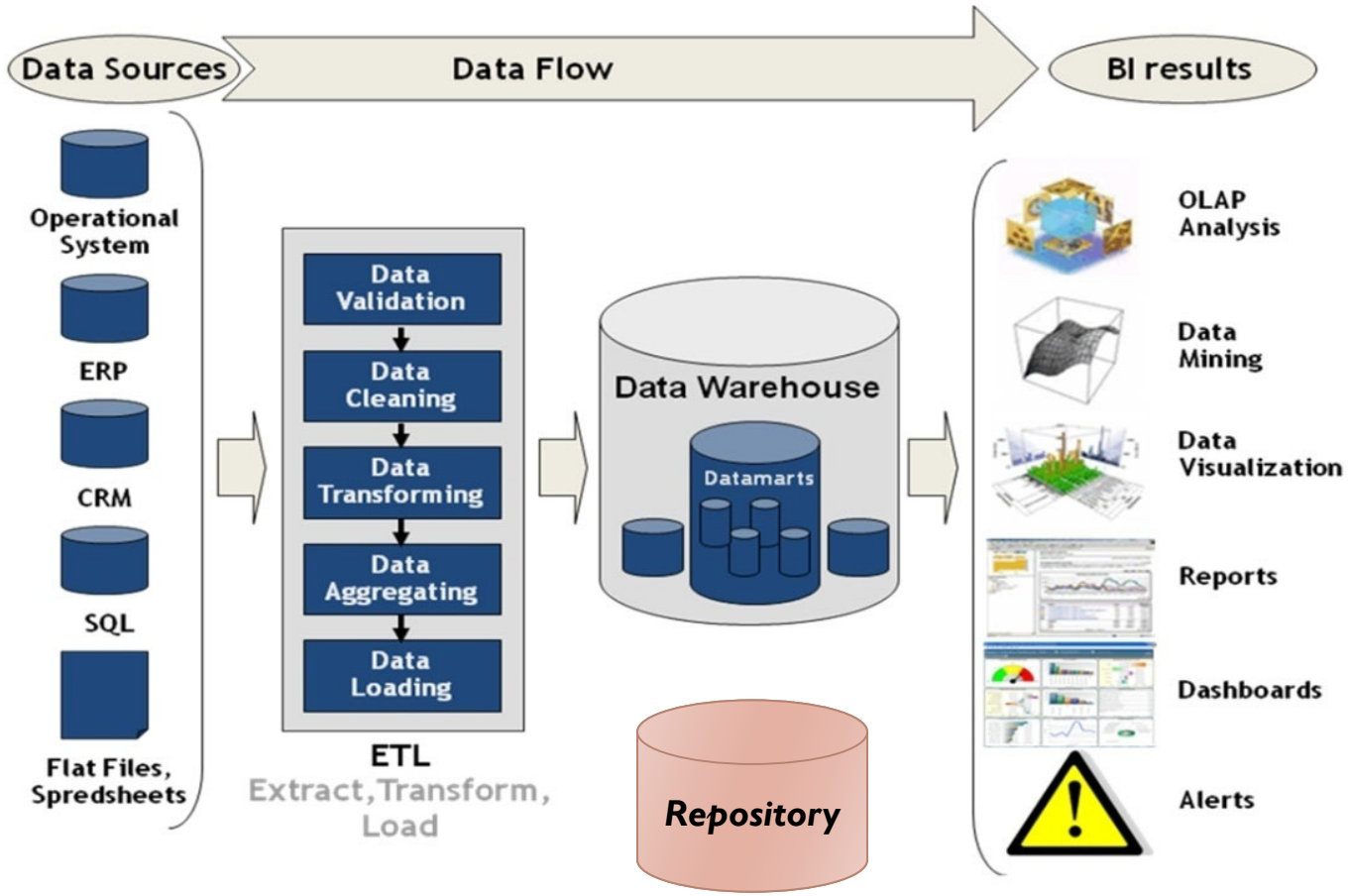
Lecture 10: Data quality

1. ***Data quality management in a Data Warehouse***

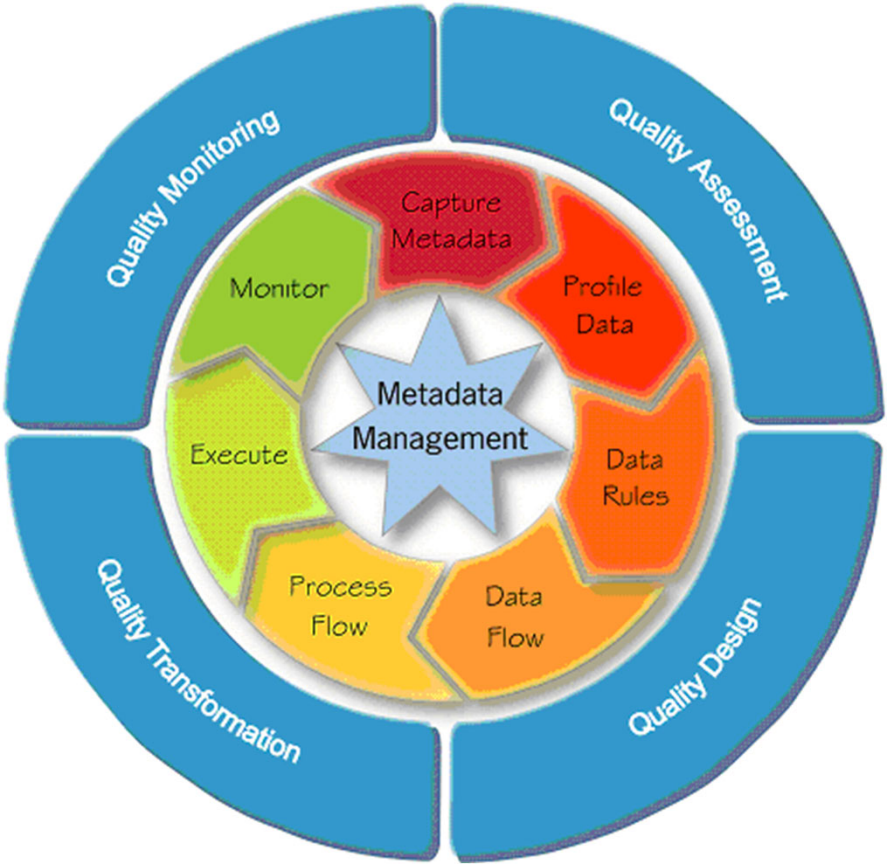
Pär Douhan, pdo@du.se

ETL


Data Sources
Data Flow
BI results
Operational System
ERP
CRM
SQL
Flat Files, Spredsheets
Data Validation
Data Cleaning
Data Transforming
Data Aggregating
Data Loading
ETL
Extract, Transform, Load
Data Warehouse
Datamarts
Repository
OLAP Analysis
Data Mining
Data Visualization
Reports
Dashboards
Alerts

Oracle Data Quality Management

Oracle Warehouse Builder



Data Quality Management

A quality process that *assesses*, *designs*, *transforms*, and *monitors* quality.

Phases in the Data Quality Life Cycle

Ensuring data quality involves the following phases:

1. Quality Assessment
2. Quality Design
3. Quality Transformation
4. Quality Monitoring

The Quality Management Process

The Data Quality Management Process

Quality data is *crucial* to decision-making and planning.

The aim of building a data warehouse is to have an *integrated*, *single* source of data

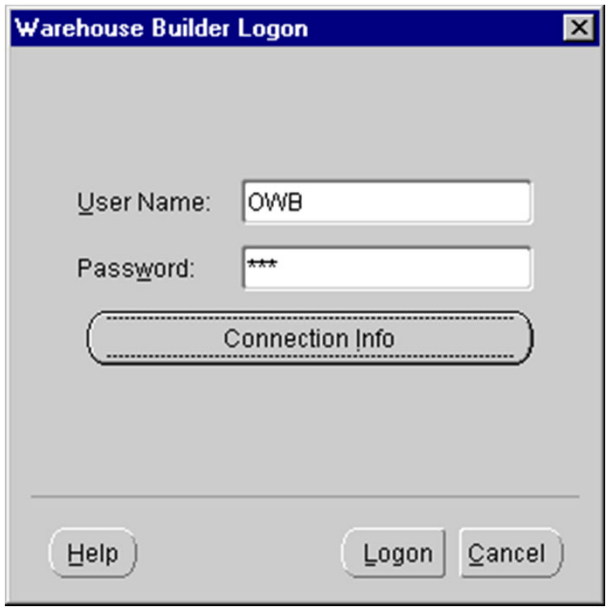
Since the data is usually sourced from a number of *disparate* systems, it is important to ensure that the data is *standardized* and *cleansed* before loading into the data warehouse.

Quality Assessment

1. Quality Assessment

determine the quality of the source data

1. ***Load* the source data**, which could be stored in *different* sources, into Warehouse Builder.
2. **Import *metadata and data*** (from both Oracle and non-Oracle sources)
3. **Use *data profiling*** to *assess* the data quality:
4. **Data profiling uncovers** data *anomalies*, *inconsistencies*, and *redundancies* by analyzing the *content*, *structure*, and *relationships* within the data.



Quality Assessment

Data Profiling

Data profiling, also called *data archeology*, is the statistical analysis and assessment of the quality of data values within a data set for *consistency*, *uniqueness* and *logic*.

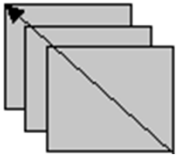
We can use to data profiling to *discover* and *measure defects* in our data before we start working with it.

Types of Data Profiling

Attribute Analysis

Functional Dependency

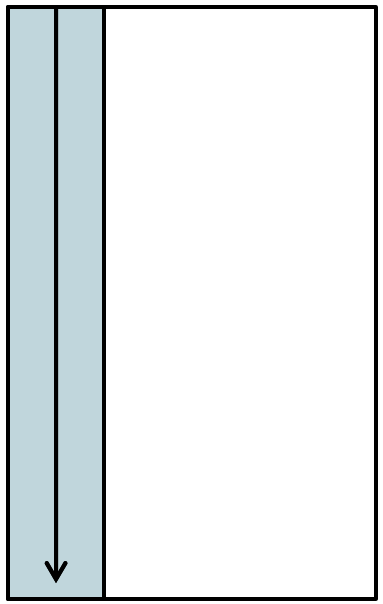
Referential Analysis



Quality Assessment

Data Profiling - Attribute Analysis

Attribute analysis looks for information about *patterns*, *domains*, *data types*, and *unique values* in a given column.



1. **Identified patterns:** *dates*, *e-mail addresses*, *phone numbers*, and *social security numbers*.

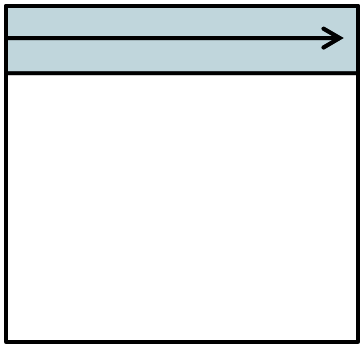
2. **Domain analysis** identifies a domain or set of *commonly used values* within the attribute by capturing the most frequently occurring values;

3. **For example**, the Status column in the Customers table is profiled and the results reveal that 90% of the values are among the following: "MARRIED", "SINGLE", "DIVORCED". Further analysis and drilling down into the data reveal that the other 10% contains misspelled versions of these words with few exceptions.

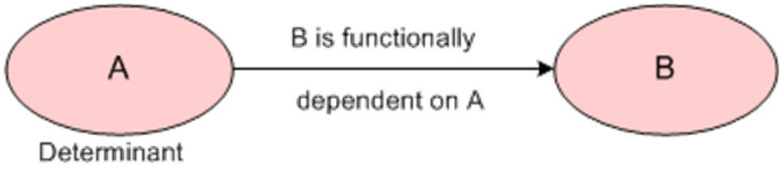
Quality Assessment

Data Profiling - Functional Dependency

Functional dependency analysis reveals information about column relationships.



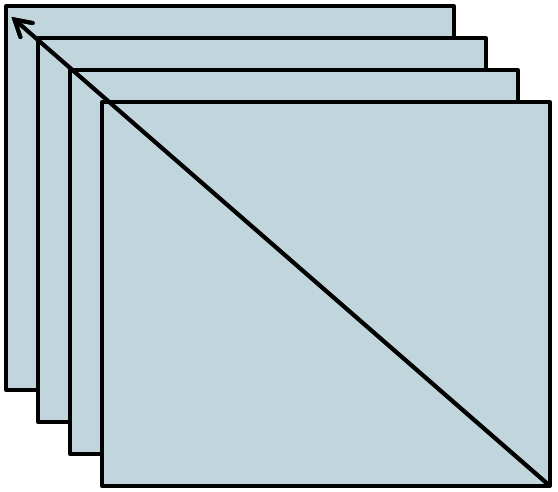
This enables us to search for things such as one attribute *determining* another attribute within a data object (table).



Quality Assessment

Data Profiling - Referential Analysis

Detect data objects that *refer to other objects*.



1. ***Parent* - and *child*** *objects (Master – Detail, PK – FK)*
2. **Things detected include** *orphans*, *childless objects* and *redundant objects*
3. ***Orphans*** are values that are found in the child object, but not found in the parent object
4. ***Childless objects*** are values that are found in the *parent object*, but not found in the *child object*
5. ***Redundant attributes*** are values that exist in both the parent and child objects

Quality Design

2. Quality Design

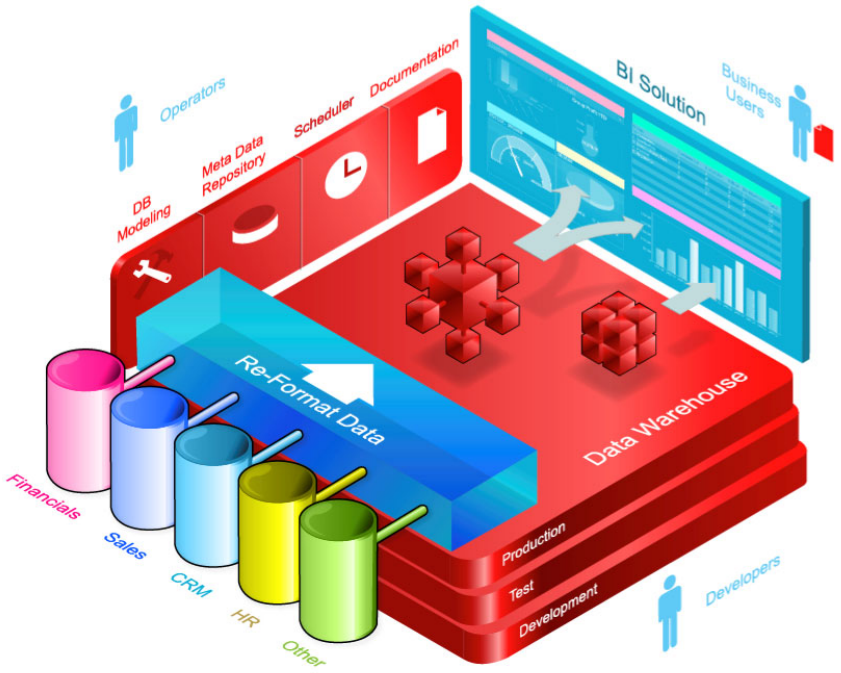
Designing the quality processes. Specify the legal data within a data object or legal relationships between data objects using **data rules**.

1. **Data rules** are definitions for *valid data values and relationships*
2. **The *metadata*** for a data rule is stored in the repository
3. **To use a data rule**, you apply the data rule to a data object
4. **We can create a data rule** called *gender_rule* that specifies that valid values are 'M' and 'F'. You can apply this data rule to the *emp_gender* column of the Employees table.

Quality Transformation

3. Quality Transformation

The quality transformation phase consists of *running* the correction mappings that are used to correct the source data.



Quality Monitoring

4. Quality Monitoring

Data monitoring is the process of examining your data over time and alerting you when the data violates any business rules that are set.

1. **Quality monitoring builds on your *initial data profiling* and data quality initiatives.** It enables you to monitor the quality of your data over time

2. **To monitor data** using Oracle Warehouse Builder you need to create *data auditors*. Data auditors ensure that your data complies with the business rules you defined