



OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

Seminar Datenqualität & Seminar Optimierungs- und
Modellierungstechniken für Datenbanken

DQ - Tools

Martin Tobies - 10. Dezember 2009

Agenda

- Einführung
- MS SQL Server Integration Services (SSIS)
- Oracle Data Warehouse Builder (OWB)
- Zusammenfassung
- Quellen

Agenda

- Einführung
- MS SQL Server Integration Services (SSIS)
- Oracle Data Warehouse Builder (OWB)
- Zusammenfassung
- Quellen

Einführung – Motivation

- Datenqualität (DQ) immens wichtig
 - DQ → Qualität Warehouse (WH)
- ⇒ Wo setzen DQ-Tools an ?
- ⇒ Wie sieht die konkrete Umsetzung aus ?

Einführung – Grundlagen(1)

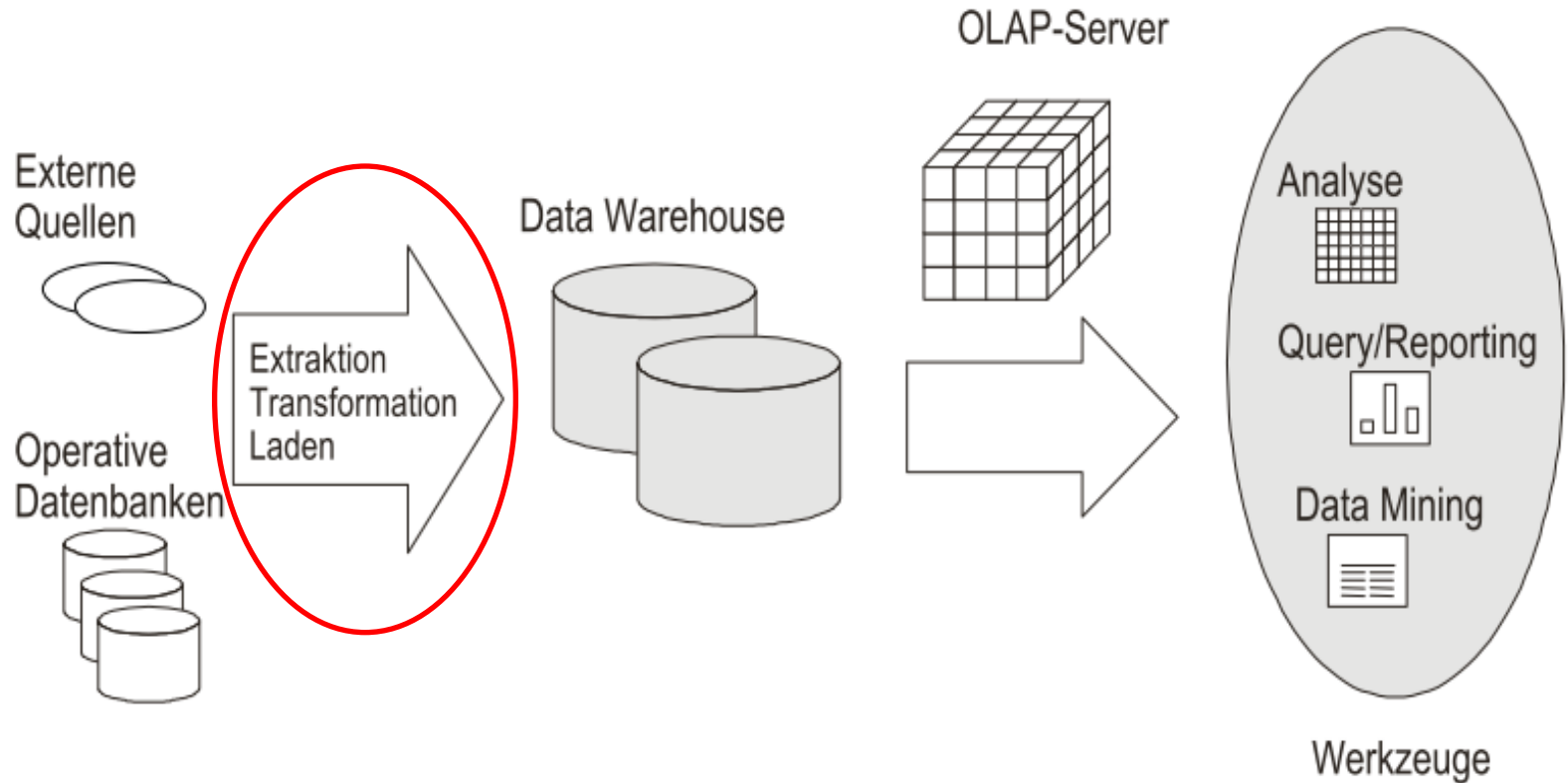


Abb. 1: Data Warehouse. Quelle [4]. Folie 8.

Einführung – Grundlagen(2)

- Profiling
 - ⇒ Analyse / Mess- und Prüfverfahren
- Cleansing
 - ⇒ Datenbereinigung / -aufbereitung
 - ⇒ Größter Aufwand
- Auditing
 - ⇒ Bewertung / Untersuchung des Prozesses

Agenda

- Einführung
- MS SQL Server Integration Services (SSIS)
- Oracle Data Warehouse Builder (OWB)
- Zusammenfassung
- Quellen

SSIS – Profiling(1)

- Vorgefertigte Profile (Nullwerte, Schlüsseleindeutigkeit...)
- Anwendbar auf einzelne Tabellen
- Aufruf der einzelnen Ergebnisse

⇒ Erstellung „Profiling Data Flow“

SSIS – Profiling(2)

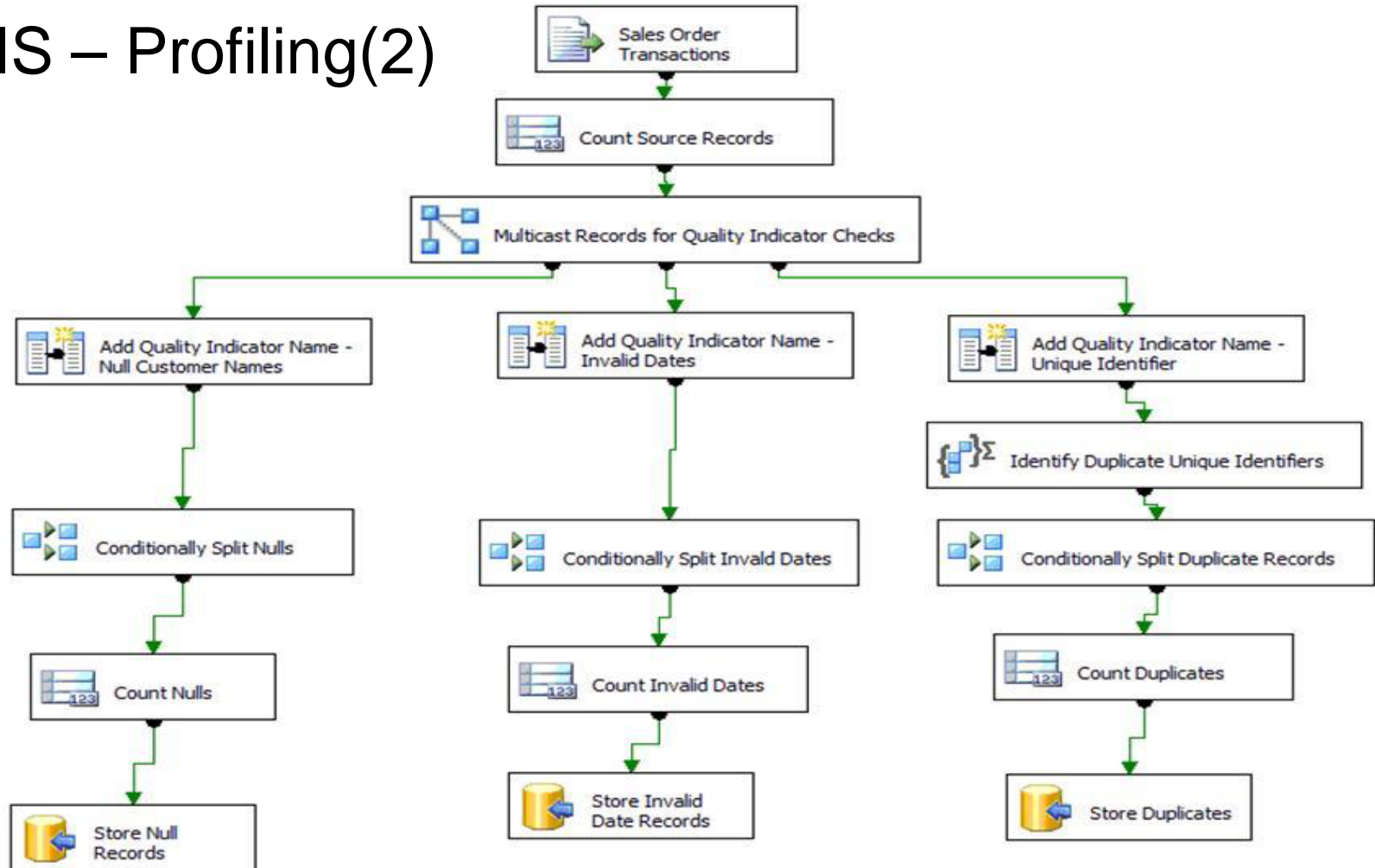


Abb. 2: Profiling Data Flow. Quelle [2]. Figure 1.

SSIS – Cleansing

- Vielfältige Methoden („Transformations“)
 - Duplicate → Fuzzy Lookup / Fuzzy Grouping
 - Validation von Werten durch Referenzdaten
 - Zusätzliche Quellen bei fehlenden Daten
 - CharacterMap
 - ...

⇒ Erstellung „Cleansing Data Flow“

SSIS – Auditing

- Aufzeichnung des Prozesses (History)
 - Vergleich Quelldaten mit aufbereiteten Daten
 - Statistikerstellung (Erfolge, Abbrüche, Dauer...)
-
- ⇒ Detailtabelle
 - ⇒ Fehler-Record-Tabelle
 - ⇒ Komponententabelle
 - ⇒ Batchtabelle
 - ⇒ Fehlertabelle

Agenda

- Einführung
- MS SQL Server Integration Services (SSIS)
- Oracle Data Warehouse Builder (OWB)
- Zusammenfassung
- Quellen

OWB – Profiling(1)

- Data Profiler
 - Quelldatenanalyse
 - Oracle DB's / Oracle gateways / ODBC / Flat file / SAP R/3 und andere ERP- Quellen

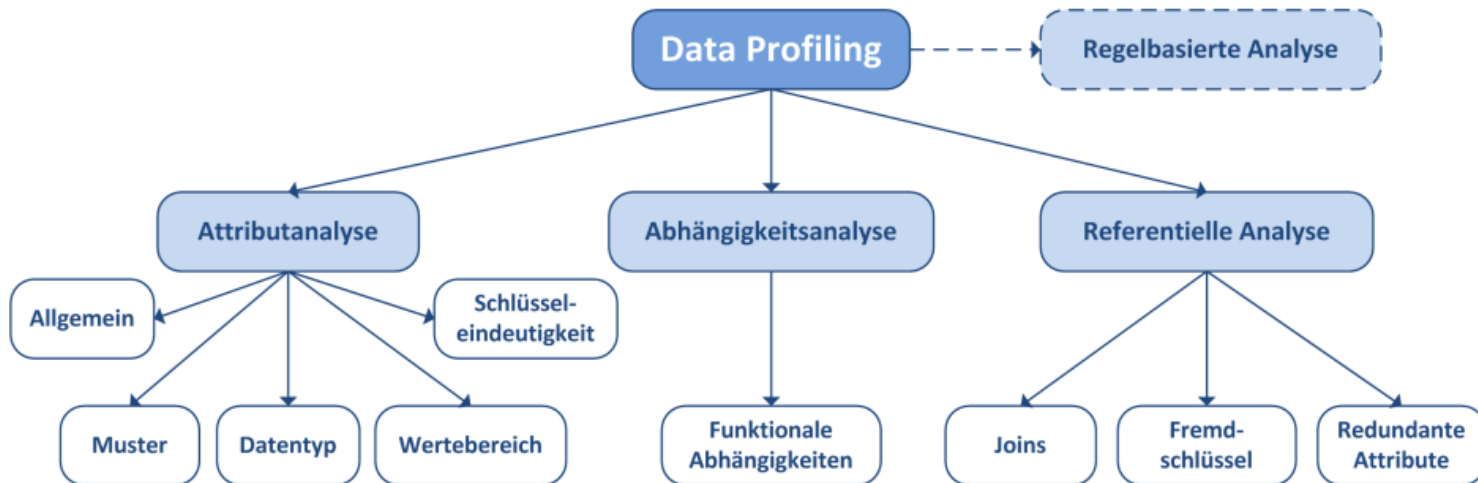


Abb. 3: Data Profiler. Quelle [1]. Abb. 5.9.

OWB – Profiling(2)

Profile Results Canvas

Unique Key Functional Dependency Referential Data Rule
Data Profile Profile Object Aggregation Data Type Pattern Domain

Here are the aggregation analysis results for EMPLOYEES, which has 12 columns and 109 rows.

Columns	Minimum	Maximum	# Distinct	% Distinct	NOT NULL	Disco
LAST_NAME	Abel	Zlotkey	102	93.6%	Yes	Yes
MANAGER_ID	100	205	18	16.5%	No	Yes
PHONE_NUMB...	011.44....	650.50...	107	98.2%	No	Yes
SALARY	2100	24000	57	52.3%	No	Yes

Derive Data Rule Remove Data Rule

Tabular Graphical

Data Drill Panel

Here are drill results on EMPLOYEES column SALARY related to Maximum value.

Distinct values: All

	SALARY	# Rows	% of 109
1	24000	1	.9%
2	17000	2	1.8%
3	14000	1	.9%

Displaying 57 Rows out of 57 more

Rows for the selected distinct value:

	PHONE_NUM...	SALARY
1	515.123.4567	24000

Displaying 1 Rows out of 1 more

Abb. 4: Aggregation.
Quelle [5]. Figure 18-2.

OWB – Cleansing(1)

- Correction Mapping mithilfe von Regeln
- Match/Merge
 - Duplikate
- Transformation
 - Parsing
 - Standardization
 - Augmentation



OWB – Cleansing (2)

Row	First Name	Last Name	PHN	SSN
A	John	Doe	650-123-1111	NULL
B	Jonathan	Doe	650-123-1111	555-55-5555
C	John	Dough	650-123-1111	555-55-5555

Datensatz

Name	Position	Rule Type	Usage	Description
Rule_1	1	Conditional	Active	Match SSN
Rule_2	2	Conditional	Active	Match Last Name and PHN

Datenregeln

Attribute	Position	Algorithm	Similarity Score	Blank Matching
SSN	1	Exact	0	Do not match if either is blank

Attribute	Position	Algorithm	Similarity Score	Blank Matching
LastName	1	Exact	0	Do not match if either is blank
PHN	2	Exact	0	Do not match if either is blank

Abb. 5: Matching. Quelle [5]. Table 23-5 – 23-8.

OWB – Auditing

- Data Auditor
- Validierung der Daten gegen Regeln
- Ad hoc oder im Work Flow
- Ergebnisdarstellung in Fehlertabelle

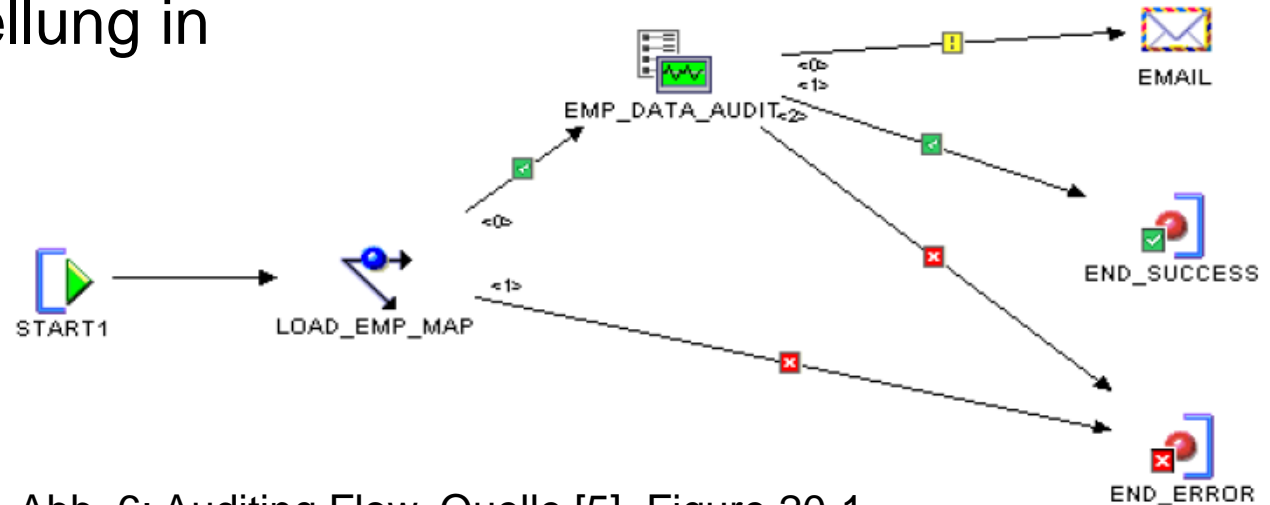


Abb. 6: Auditing Flow. Quelle [5]. Figure 20-1.

Agenda

- Einführung
- MS SQL Server Integration Services (SSIS)
- Oracle Data Warehouse Builder (OWB)
- **Zusammenfassung**
- Quellen

Zusammenfassung

- mächtige & umfangreiche Tools
 - Geringe Anschaffungskosten
 - Vorteile in jeweiliger Umgebung (OracleBD's, Windows)
- ⇒ (teurere) Alternativen z.b. „Power Center“(Informatica), „Data Stage“ (IBM)

Agenda

- Einführung
- MS SQL Server Integration Services (SSIS)
- Oracle Data Warehouse Builder (OWB)
- Zusammenfassung
- Quellen

Quellen:

- [1] Borowski, E.: „Entwicklung eines Vorgehensmodells zur Datenqualitätsanalyse mit dem Oracle Warehouse Builder“. Diplomarbeit. FU Berlin, 2008.
- [2] Vitt, E. (2006): „Data Quality Solutions“. Technical Report. URL: <http://msdn.microsoft.com/en-us/library/aa964137%28SQL.90%29.aspx> [Stand: 05.12.2009]
- [3] Knight, B./Veerman, E./Dickinson, G./Hinson, D./Herbold, D.: „Professional SQL Server® 2008 Integration Services“. Indianapolis: Wiley Publishing, Inc. 2008
- [4] Sattler, K.: „Datenqualität – eine Datenbankorientierte Sichtweise“. Vortrag. 10. Datenbank-Tutorientage. Karlsruhe, 01.03.2005
- [5] Oracle: „Oracle® Warehouse Builder Data Modeling, ETL, and Data Quality Guide 11g Release 2 (11.2)“. White Paper. 2009

Fragen?

???

