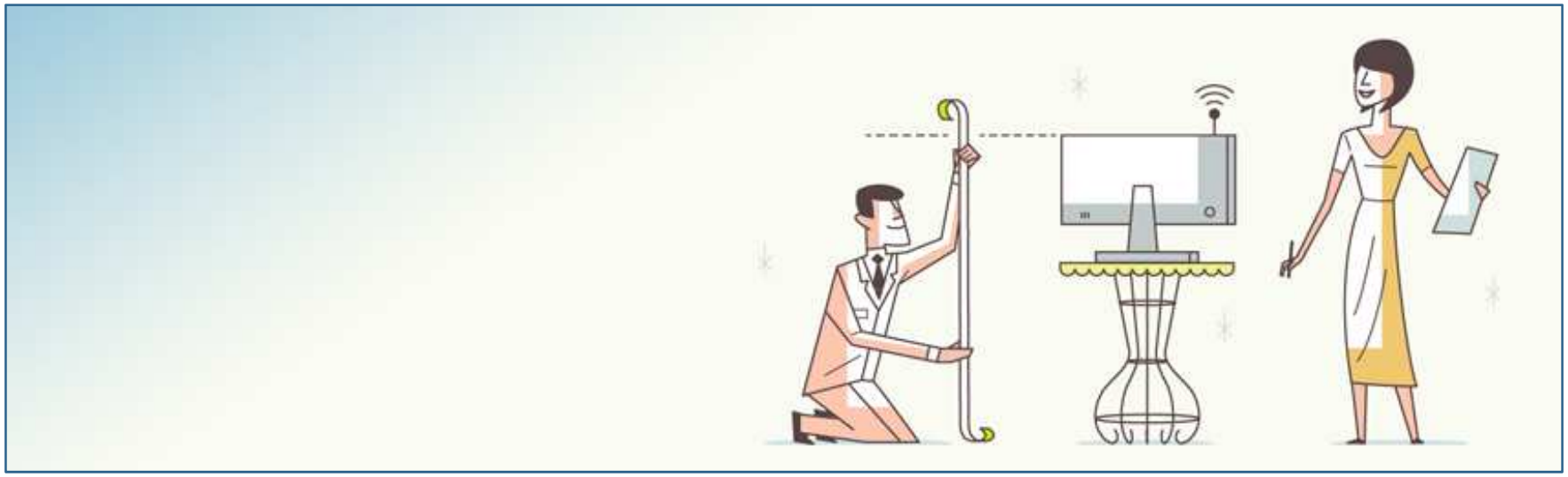
DATENQUALITÄT MIT ORACLE-PRODUKTEN



- gezielter Einsatz von ETL- und ELT-Tools im DQ-Bereich

IN EIGENER SACHE – DAS ORBIT ORACLE-PORTFOLIO

ORACLE BUSINESS INTELLIGENCE

ORACLE BI Enterprise Edition

ORACLE Application Express

ORACLE APEX

DWH/OLAP

ORACLE WAREHOUSE BUILDER

ORACLE DATA INTEGRATOR

PL/SQL

Oracle 11g Administration

OLTP

11g ORACLE DATABASE

ORACLE Datenbank

LICENSING

ORACLE

ORACLE Gold Partner

Specialized
Oracle Business Intelligence Foundation

ORACLE

SOLUTION PARTNER COMMUNITY
BUSINESS INTELLIGENCE
EPM – BI – DWH

Database Appliance

T-Series

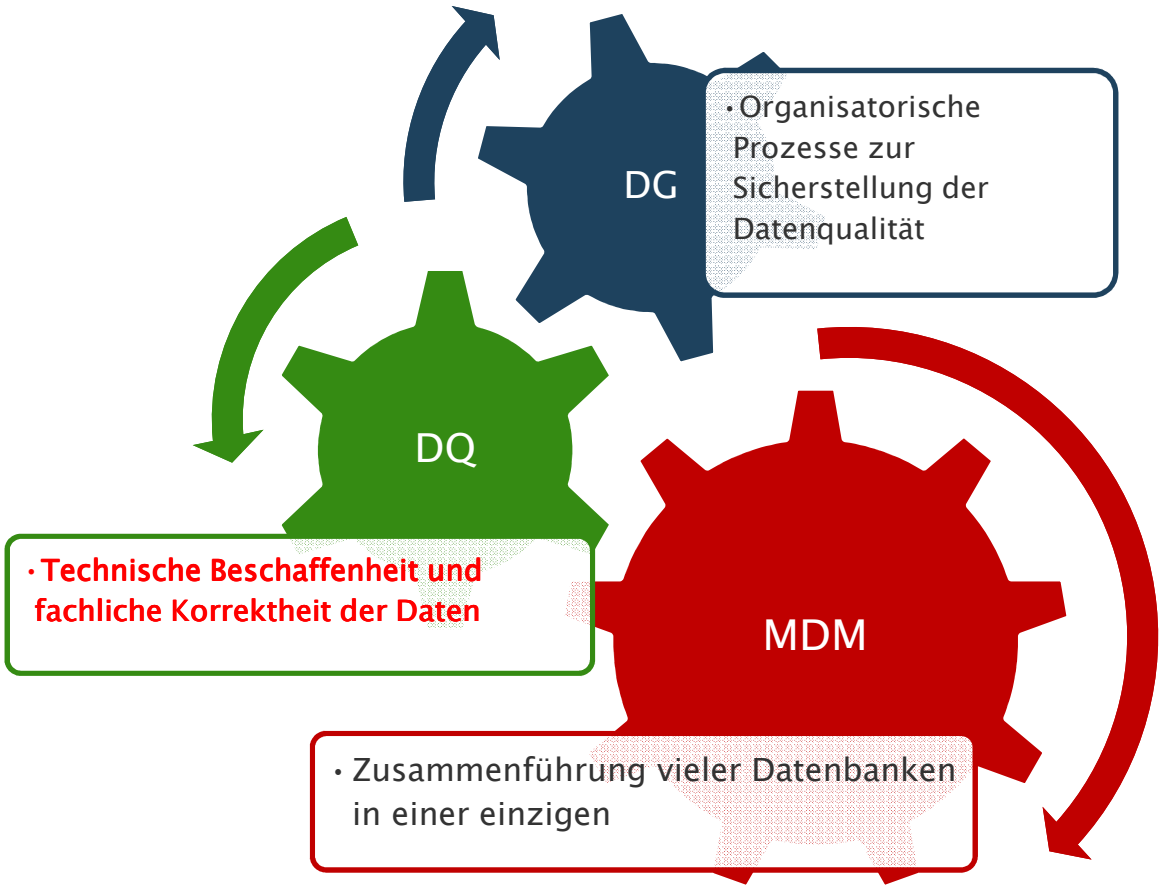
M-Series

Sun ORACLE

ORACLE Systems Familie

Storage-, Backup- und Virtualisierungs-lösungen

WORUM ES HIER GEHT



Motivation

MOTIVATION – WO IST DATENQUALITÄT WICHTIG?

- Kundendaten
  - Kontaktdaten
  - Zahlungsinformationen
- Materialdaten
  - Bestellnummern
  - Produktinformationen
- Bewegungsdaten
  - Umsätze
  - Lagerbestände
- Und vieles mehr!

MOTIVATION – WAS IST SCHLECHTE DATENQUALITÄT ?

- Unvollständige oder falsche Kontaktdaten
  - **Ursache**: Keine Plausibilitätsprüfung bei der Eingabe
  - Bis zu 30% Fehlerquote sind durchaus üblich
  - Allein 25% aller Adressen sind postalisch falsch
- Dubletten
  - **Ursachen**:
    - Falsche Eingaben
    - Zusammenführen von Datenbeständen
- Veraltete Daten
  - **Ursachen**:
    - Heirat, Scheidung und Umzüge der Kunden sowie
    - Schlechte interne Kommunikation (unklare Datenhoheit)

MOTIVATION – WARUM DATENQUALITÄT ?

- Folgen schlechter Datenqualität
  - 30–50% höhere IT-Kosten
  - Dubletten verursachen z.B. unerwünschte Werbeanrufe und damit Abmahnungen
  - Längere Bearbeitungszeit → schlechte Kundenzufriedenheit
  - Veraltete Adressen verursachen Zahlungsausfälle und administrativen Zusatzaufwand
  - Angebote sind falsch
  - Reporting-Systeme produzieren falsche Berichte und Prognosen

- Qualitativ hochwertige Daten sind Voraussetzung für
  - Den wirtschaftlichen Erfolg einer Firma
  - Den Erfolg von nachfolgenden BI-Projekten, falls sich das DWH aus vielen Quellen speist

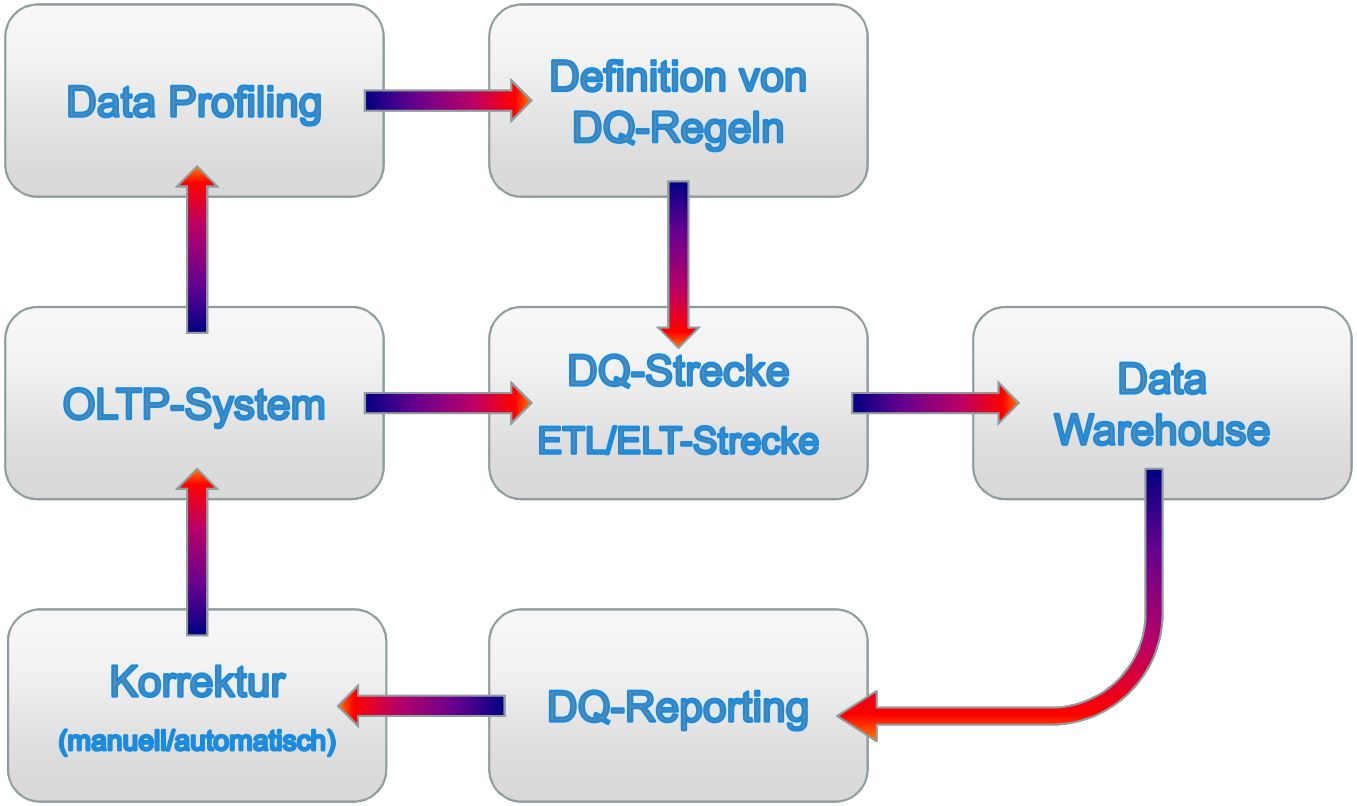
MOTIVATION – ZEHN GEBOTE

Lieber Kunde, Du sollst ...

1. Erkennen, dass du betroffen bist!
2. Verantwortliche für Datenqualität benennen!
3. Deinen Datenschatz hüten und anreichern!
4. Deine Daten zugänglich und leicht auffindbar machen!
5. Datenqualitätsprozesse automatisieren!
6. Datenqualität als internationale Aufgabe begreifen!
7. Dich auf Expertenwissen stützen!
8. Die Qualität deiner Daten schrittweise verbessern!
9. Die Ziele deiner Datenqualitäts-Aktivitäten immer vor Augen haben!
10. Die Früchte hoher Datenqualität ernten!

DQ-Vorgehensweise - allgemein

DQ-PROZESS – SCHEMATISCHE ABFOLGE



TECHNISCHE VORGEHENSWEISE – ÜBERSICHT

- Voranalysen (fakultativ): Profiling der Daten
  - Oracle Data Profiling, in OWB und ODI integrierbar
  - Informatica Analyst (eigenständiges Programm)
- Feste Implementierung von DQ-Regeln
  - Formulierung von SQL-Statements mit einheitlicher Signatur
    - Identifizierung der fehlerhaften Datensätze
    - Speicherung der maßgeblichen Metadaten (Schlüssel, zuständiger Sachbearbeiter, Datum etc.) in Sondertabellen
  - Integration in die DQ-Strecke je nach Technologie
    - PL/SQL-Prozeduren (mit ODI)
    - Generische Mappings, denen über eine Parameterdatei das SQL mitgegeben wird (OWB ab 11gR2, Informatica PowerStation durchgängig)

BESCHREIBUNG – SQL-BEISPIEL (ANREDE FEHLT)

```
SELECT

s.row_id, s.created, s.created_by, s.last_upd, s.last_upd_by,
'S_CONTACT s, CX_PARTY, S_ORG_EXT, S_ORG_EXT_X, CX_DIVISION', 'Ansprechpartner', 1,
S_CONTACT.fst_name || ' ' || S_CONTACT.last_name, x_address_salut, x_letter_salut,
attrib_46, sysdate, 0, 999, 'Sonderregel Anrede', 'X_ADDRESS_SALUT', X_ADDRESS_SALUT

FROM

S_CONTACT, CX_PARTY, S_ORG_EXT, S_ORG_EXT_X, CX_DIVISION

WHERE

S_CONTACT.PR_DEPT_OU_ID = S_ORG_EXT.ROW_ID AND S_ORG_EXT.ROW_ID =
CX_PARTY.ACCOUNT_ID AND CX_DIVISION.ROW_ID = CX_PARTY.ORGANIZATION_ID AND
S_ORG_EXT.ROW_ID = S_ORG_EXT_X.PAR_ROW_ID (+) AND CX_PARTY.TYPE = 'Account
Unit' AND emp_flg = 'N'

AND

(instr(lower(x_address_salut), 'herr ') = 0 and instr(lower(x_address_salut), 'frau ') = 0 or
instr(lower(x_letter_salut), 'herr ') = 0 and  instr(lower(x_letter_salut), 'frau ') = 0)
```

TECHNISCHE VORGEHENSWEISE – INFORMATICA

- Mit Informatica PowerStation
  - Sind mehrmalige Aufrufe generischer Mappings mit wechselnden Parameterdateien möglich, die das für die DQ-Prüfung relevante SQL enthalten
  - Kann die ETL-Strecke flexibel an sich ändernde DQ-Anforderungen angepasst werden, ohne die Vorgaben der Best Practice formal zu verletzen

TECHNISCHE VORGEHENSWEISE – OWB

- Mit OWB
  - Sind generische Mappings ab der Version 11gR2 möglich, denn es kann auch hier ein SQL Override definiert werden.
  - Kann die ETL-Strecke flexibel an sich ändernde DQ-Anforderungen angepasst werden, ohne die Vorgaben der Best Practice formal zu verletzen.

TECHNISCHE VORGEHENSWEISE – ODI

- ODI bietet
  - Keine Möglichkeit, generische Mappings für DQ-Belange einzurichten.
  - Mit der Option „ODI Procedures“ ist dennoch eine flexible DQ-Modellierung möglich, denn das Prüf-SQL kann in Form einer gespeicherten Prozedur implementiert und mit „ODI Procedures“ aufgerufen werden.

Spezielle Optionen – ODI

SPEZIELLE DQ-OPTIONEN – ODI

- Die sog. „Data Quality Firewall“ bietet einen ganzheitlichen Ansatz, der aus folgenden Schritten besteht:
  - Valide Daten durchlaufen die ELT-Strecke ohne Eingriff.
  - Invalide Datensätze werden in gesonderte Tabellen umgeleitet.
  - Zurückgewiesene Daten können anschließend überprüft werden.
  - Reparierte Daten können nachträglich weitergeleitet oder recycled werden.
- Die für die Datenqualität benutzten Regeln werden im Metadaten Repository gespeichert. Sie können entweder durch Reverse Engineering aus der Datenbank ausgelesen werden (z.B. Constraints oder Fremdschlüssel) oder aber als frei definierte Regeln im Mapping Designer erstellt werden.

SPEZIELLE DQ-OPTIONEN – ODI

- Dublettenerkennung
  - Anhand von Merkmalen wie gleichen E-Mail-Adressen oder gleicher Produktbezeichnung bei verschiedenen Bestellnummern.
  - Implementierung mit Primärschlüsseln bzw. Unique Constraints.
- Referenzielle Integrität
  - Kunden ohne Account Manager oder Bestellungen invalider Kunden.
  - Implementierung durch Schlüsselvergleiche wie WHERE-Bedingungen bzw. komplexere Funktionen wie Spalte A = f(Spalte B, Spalte C).
- Wertevalidierung
  - Konsistenzprüfung auf Datensatzebene.
  - Beispiele: fehlende PLZ oder formal falsche E-Mail-Adresse.
  - Implementierung über SQL-Bedingungen.

SPEZIELLE DQ-OPTIONEN – ODI

- Mit ODI kann die Datenqualität an folgenden Stellen überwacht werden:
  - Innerhalb der Quellsysteme;
  - Auf der ELT-Strecke;
  - Innerhalb des Data Warehouse.
- Technischer Aufbau
  - Es werden ODI-Mappings verwendet, in denen die entsprechende Logik enthalten ist.
  - In den Quellsystemen bzw. im DWH sorgen getaktete Prozesse für den Transport fehlerhafter Daten in Sondertabellen.
  - Auf der ELT-Strecke werden fehlerhafte Datensätze nicht in das DWH geschrieben sondern ebenfalls in spezielle Tabellen umgeleitet.
  - Als vierte Möglichkeit können als fehlerhaft erkannte Datensätze aus dem DWH wieder in die ELT-Strecke eingebracht werden.

Spezielle Optionen – OWB

SPEZIELLE DQ-OPTIONEN – OWB

- Die Stärken des OWB liegen in der
  - Bereinigung von Adressen
  - Entfernung von Dubletten.
- Für jeden dieser beiden Schritte ist im OWB-Funktionsumfang ein Transformationsoperator enthalten.
- Beide Operatoren haben definierte Eigenschaften und können in Mappings eingebaut werden.

OWB – OPERATOR „NAME AND ADDRESS“

- Der Operator „Name and Address“ korrigiert inhaltliche Fehler durch
  - Parsen oder Trennen von Namens- und Adressdaten in einzelne Elemente;
  - Postalisch korrektes Standardisieren von Abkürzungen, Titeln und Adressen;
  - Validieren und Korrigieren von Adress- und Ortsinformationen;
  - Anreicherung mit Zusatzinformationen (Geschlecht, PLZ, Landesvorwahl etc.).
- Der zugrunde liegende „Name and Address Server“ wird bei Installieren der Oracle-Datenbank mit installiert.

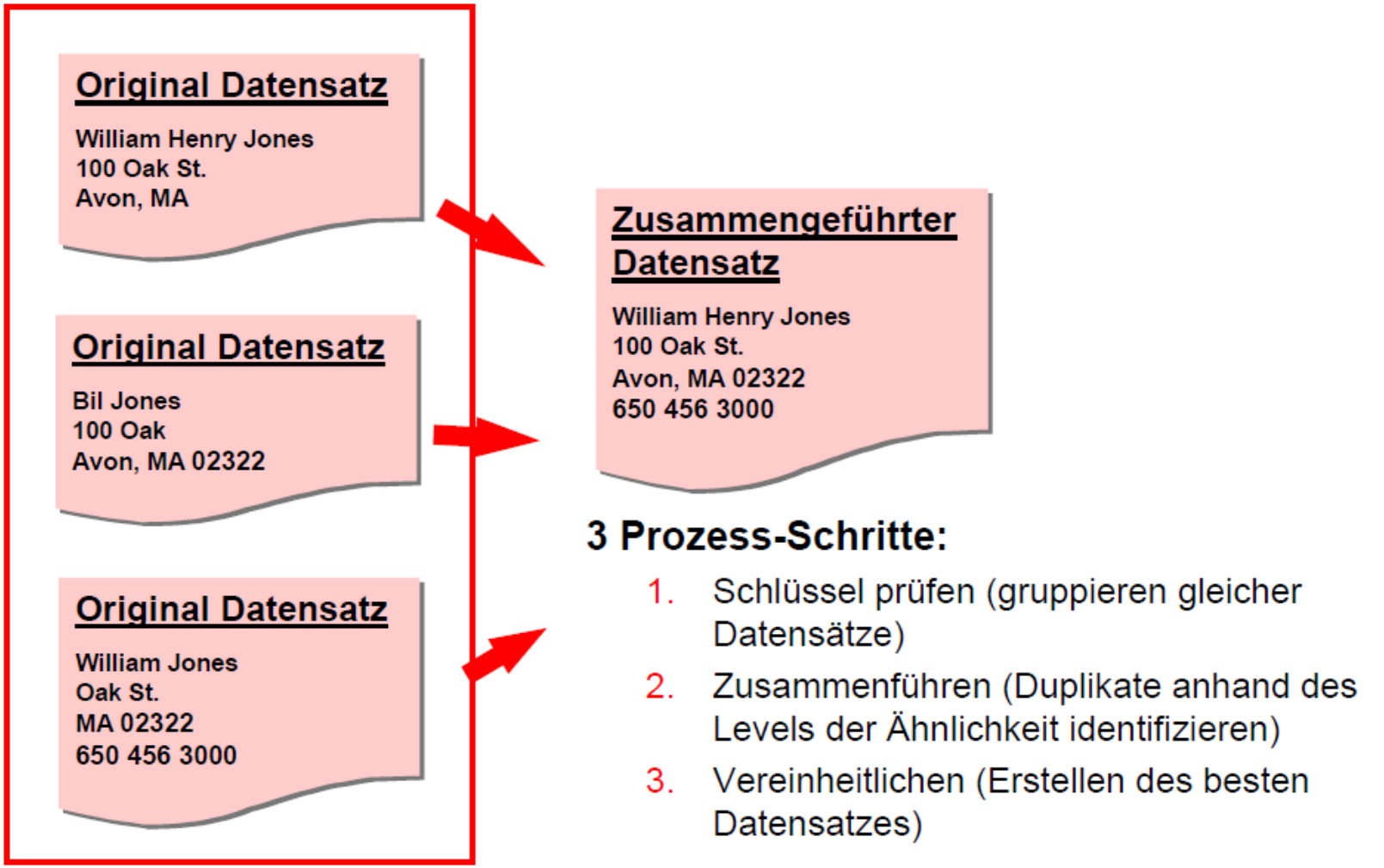
OWB – OPERATOR „NAME AND ADDRESS“

- Es ist die Einbindung von Fremdsoftware und Adressbibliotheken erforderlich, die für die Verwendung mit OWB zertifiziert sein müssen.
  - Als Beispiele seien Trillium oder First Logic genannt.
  - Hoher Bedarf an Speicherplatz (mehrere GB).
- Eine Übersicht zu allen Anbietern lässt sich über folgende Seite abrufen: oracle.com/technology/products/warehouse/htdocs/OTN_Partners.html
- Die Adressbibliotheken
  - Enthalten aktuelle in- und ausländische Adressen sowie korrekte Schreibweisen und Abkürzungen.
  - Ergänzen zur Laufzeit fehlende Angaben und korrigieren veraltete oder falsche Daten (z.B. fehlerhafte Zuordnung einer Postleitzahl).
- Beispiel:
  - „Fritz Schmitz|Fliegenbusch 4711|Krefeld-Hüls“
  - „Fritz|Schmitz|Am Fliegenbusch|4711|Krefeld|Deutschland|47839“.

OWB – OPERATOR „MATCH AND MERGE“

- Daten sollten zuvor mit „Name and Address“ bereinigt werden
- Die Adaptierung der Match- und Merge-Regeln erfolgt mit Hilfe eines intuitiv zu bedienenden Assistenten. Dabei stehen vordefinierte elementare Match-Regeln zur Verfügung, z.B.:
  - Strip Noise Words
  - Match on abbreviations
  - Match on acronyms
  - Detect switched name order
  - Match on initials
  - Detect compound name.
- Um die Dublettenerkennung möglichst sicher zu machen, sollten die Daten zuvor bereits mit dem Operator „Name and Address“ bereinigt worden sein.

OWB – BEISPIEL ADRESSENBEREINIGUNG



Quelle: Oracle

OWB – OPERATOR „MATCH AND MERGE“

- Der strukturell unveränderliche Operator „Match Merge“ besteht aus einer Eingabegruppe und den beiden Ausgabegruppen
  - MERGE (mit den konsolidierten Daten)
  - XREF (mit der Dokumentation des Konsolidierungsprozesses)
- Beide Ausgabegruppen, werden im PL/SQL-Format generiert. Letztere kann den Anforderungen angepasst werden.
- Ausführung von OWB-Mappings mit Match-Merge-Operator
  - Nur im zeilenbasierten Modus möglich;
  - Regeln werden nach der Oder-Logik ausgewertet;
  - Beschleunigung durch sog. „Match bins“, bei denen begrenzte Bereiche durchsucht werden und nicht die ganze Tabelle.
  - **Beispiel**: Gruppierung von Kundenadressen nach Stadt oder PLZ.

OWB – OPERATOR „MATCH AND MERGE“

- Sobald die Dubletten herausgefunden sind, kann mit Hilfe von insgesamt elf Merge-Regeln eine Konsolidierung erfolgen.

- Beispiele:

  - Soll etwa der längste Vorname als konsolidierter Datensatz „überleben“, so wird die Regel „Min Max Record“ verwendet.
  - Mit TAKE_VERIFIED_ADDRESS wird festgelegt, dass der jenige Datensatz verwendet werden soll, der auch in der Postdatenbank (Drittanbieter) zu finden ist.

Weitere nützliche Tools

WEITERE NÜTZLICHE TOOLS

- Informatica Analyst: Data Profiling
  - Java-Anwendung
  - Wertespektrum & Häufigkeit bestimmter Ausprägungen;
  - Erkennung von Mustern, z.B. nn.a-an.
- Informatica Developer
  - Eclipse-basierter graphischer Editor
  - Entwicklung von IPC-kompatiblen Mappings
  - zusätzliche, über IPC hinausgehende DQ- und statistische Funktionen
- Oracle Application Express
  - Generierung von HTML-Seiten (Reports, Eingabemasken)
  - Validierung möglich (not NULL, reguläre Ausdrücke, String-Vergleiche, Werteprüfung von Zahlen)
  - Beschränkungen des Wertebereichs bei Zahlen werden über SQL oder PL/SQL definiert, ebenso die Validierungen ganzer Seiten.

VIELEN DANK

- für Ihre Aufmerksamkeit

➔ Falls Sie Fragen zu dieser Präsentation haben, sprechen Sie uns einfach an.

➔ Ihr Ansprechpartner

Dr. Götz Gleitsmann

+49 228-95693-673

goetz.gleitsmann@orbit.de

www.orbit.de