Data Preprocessing 1

Thema: Bussiness Intelligence
Teil 1: OLAP & Data Warehousing

von Christian Merker

Gliederung

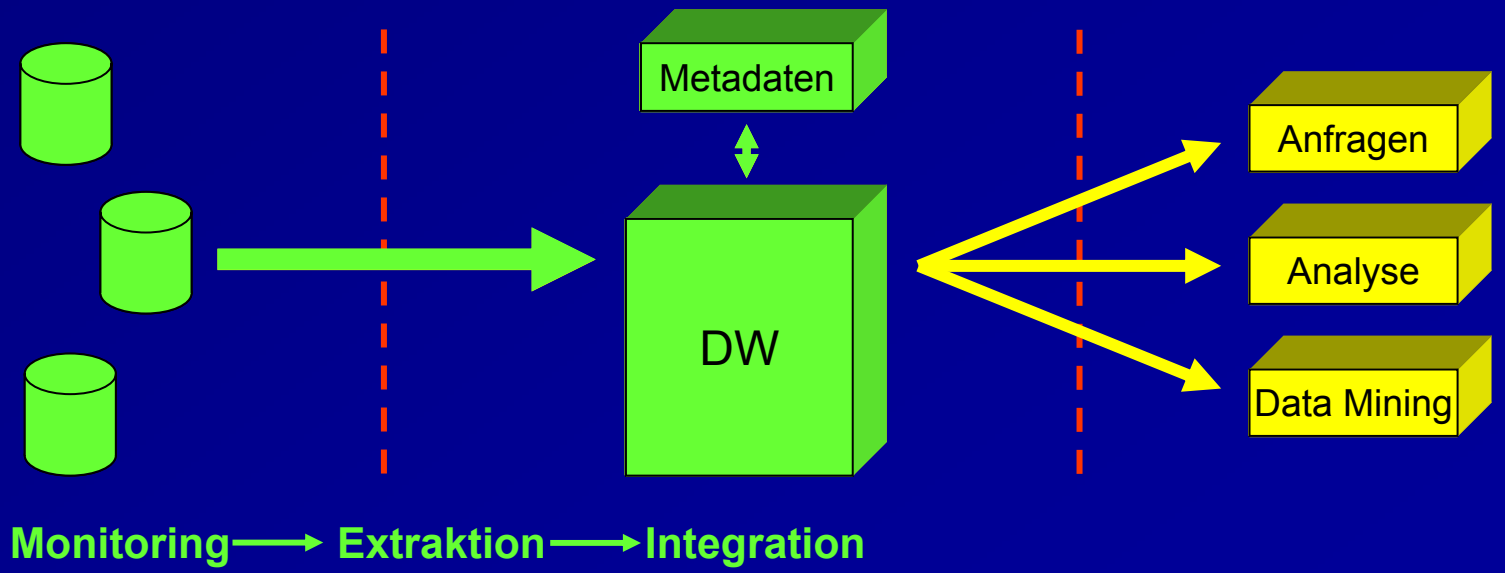
- Motivation
- Monitore
- Datenextraktion
- Schema- und Datenintegration
- Datenbereinigung
- Zusammenfassung

Motivation

- Warum der Aufwand ?

  600 Milliarden US-Dollar an Einnahmen entgingen amerikanischen Firmen 2001 durch „schmutzige“ Daten
  (Schätzung des Data Warehouse Instituts)

- Wo findet das Preprocessing statt ?



Monitore

Aufgaben:

- Daten in der Datenbank auf Änderungen überwachen
- Änderungen von relevanten Daten protokollieren
- Datenbeschaffungsprozess anstoßen

→ Monitor ist das Bindeglied zwischen Datenquellen und dem Data Warehouse

Monitore – Realisierung (1)

Grundlegende Ansätze

- Ansatz 1:
  - Monitor vermerkt alle Änderungen in einer Delta-Datei
  - Auf Anfrage wird die Delta-Datei an das Data Warehouse übertragen
- Ansatz 2:
  - Monitor vermerkt nur Hinweis, welche Datenquellen verändert wurden
  - Auf Anfrage werden die Hinweise an Data Warehouse übertragen
  - Data Warehouse ruft Extraktionskomponente auf, die alle geänderten relevanten Daten sucht

Monitore – Realisierung (2)

Wichtige Aspekte bei der Realisierung

- Entdeckung aller Änderungen vs. Nettoeffekt
  - Entdeckung aller Änderungen
    - Alle Änderungen durch Transaktionen werden protokolliert
  - Nettoeffekt
    - Nur die Unterschiede zum letzten Ladevorgang werden protokolliert

Beispiel:

Person

ID	Name	Beruf
4711	Müller	Programmierer

Operationen: Einfügen einer Person und anschließendes Löschen der selben Person

Alle Änderungen:

INSERT auf Person (4711, Müller, Programmierer)
DELETE auf Person (4711, Müller, Programmierer)

Nettoeffekt:

Kein Eintrag, da kein Unterschied zum letztem Ladevorgang erkennbar ist

Monitore – Realisierung (3)

Weitere wichtige Aspekte

- Benachrichtigung vs. Polling
  - Polling
    - Monitor überprüft periodisch, ob Änderungen vorliegen
    - Kurze Periodendauer → belastet System unnötig
    - Lange Periodendauer → Änderungen werden evtl. nicht erkannt
  - Benachrichtigung
    - System erkennt Änderungen und informiert Monitor mittels Trigger
    - System wird nicht unnötig belastet

- Internes vs. Externes Monitoring
  - Intern
    - Quellsystem stellt Monitoring Möglichkeiten zur Verfügung
  - Extern
    - Quellsystem stellt keine Monitoring Möglichkeiten zur Verfügung
    - Änderungen müssen von außen entdeckt werden

Monitore – Techniken (1)

- Aktive Mechanismen
- Replikationsmechanismen
  - Snapshot
  - Datenreplikation
- Protokollbasierte Entdeckung
- Anwendungsunterstütztes Monitoring

Monitore – Techniken (2)

- Aktive Mechanismen
  - Erkennen vordefinierte Situationen bzw. Ereignisse
  - Führen beim Eintritt eines Ereignisses festgelegte Aktionen durch
  - z.B. Trigger, ECA-Regeln

Beispiel für ECA-Regeln

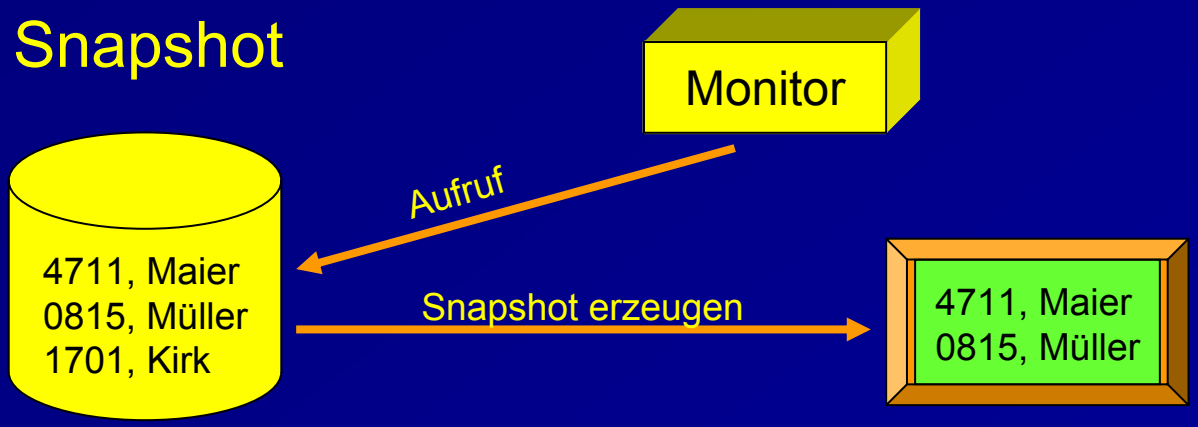
```
Event      Kontoeinzahlung
Condition  Kontostand > 5000 €
Action     Kontozinsen auf 3 % erhöhen
```

Monitore – Techniken (3)

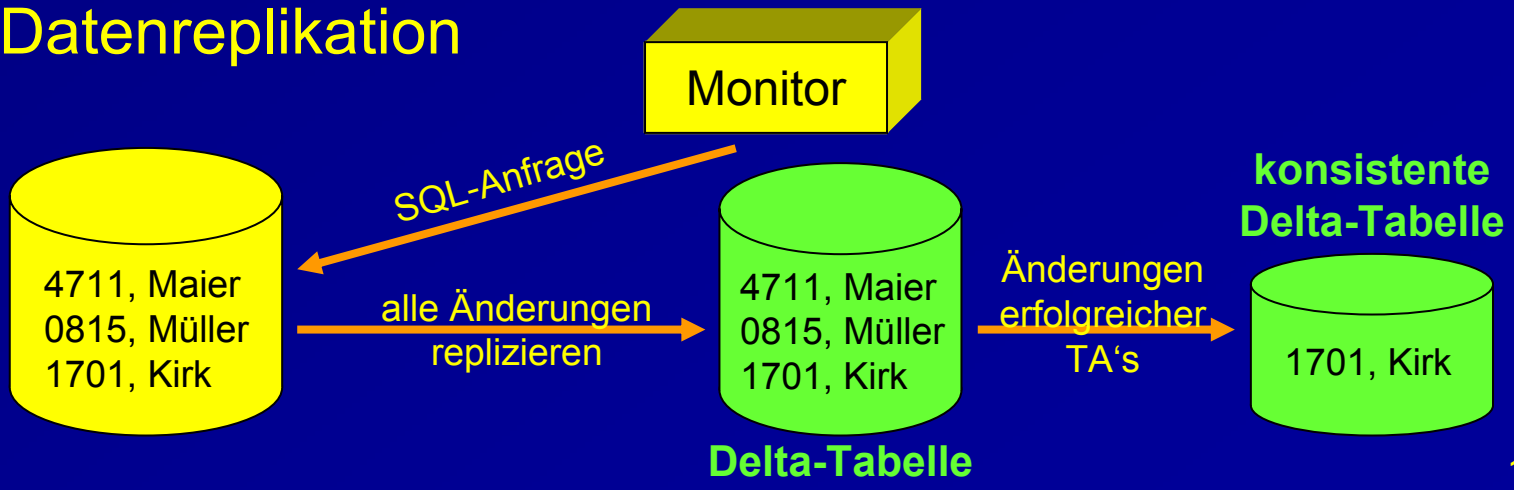
- Replikationsmechanismen
  - Snapshot
    - Lokale Kopie von Daten aus ein oder mehreren Quellen
    - Nur lesender Zugriff
    - Aktualisierung durch erneute Anfrage oder durch Snapshot-Logs
  - Datenreplikation
    - Geänderte Daten werden in Delta-Tabellen repliziert
    - Inkrementelle Aktualisierung der Delta-Tabellen
    - Einsatz im IBM DataPropagator:
      - Monitor merkt sich alle Änderungen (auch von Rollback Transaktionen)
      - Änderungen von erfolgreichen TA‘s kommen in konsistente Delta-Tabellen

Monitore – Techniken (4)

- Beispiele
  - Snapshot



  - Datenreplikation



Monitore – Techniken (5)

- Protokollbasierte Entdeckung
  - Protokolldatei wird benutzt, um Änderungen zu entdecken
  - Nur möglich, wenn Log-Datei genug Informationen liefert und von außen zugänglich ist
  - Übertragung der Log-Datei an Data Warehouse

- Anwendungsunterstütztes Monitoring
  - Letzte Möglichkeit, falls Monitor nicht anders realisierbar
  - Anwendungsprogramme müssen Delta-Datei erstellen
  - Problem: alte Programme sind meist schlecht entworfen und dokumentier

    → Alte Programme durch neue Programme ersetzen ☺

Extraktionsphase

Aufgabe:

Übertragung der geänderten relevanten Daten ins Data-Warehouse

Extraktionszeitpunkte:

- Periodisch
  - Aktualisierung nach Ende eines Zeitintervalls (z.B. alle 24 h)
  - Länge des Intervalls hängt von der Dynamik der Daten ab
- Anfragegesteuert
  - Anstoß durch expliziten Aufruf von außen (z.B. durch Administrator)
- Ereignisgesteuert
  - Anstoß durch bestimmtes Ereignis
    (z.B. bestimmte Anzahl an Änderungen)
- Sofort
  - Bei sehr hohen Aktualitätsanforderungen (z.B. Börse)
  - Data Warehouse ist immer so aktuell wie die Quellsysteme

Schemaintegration (1)

Ziel:

- Analyse der lokalen Schemata
- Integration des globalen Schemas im Data Warehouse

4 Phasen der Integration

- Vorintegrationsphase
- Vergleichphase
- Vereinheitlichungsphase
- Restrukturierungsphase

Schemaintegration (2)

Vorintegrationsphase

- Analyse der Quellschemata
- Auswahl der zu integrierenden Teile

Vergleichsphase

Ziel : Mögliche Konflikte entdecken

Konfliktklassen:

- Namenskonflikte
- Semantische Konflikte
- Strukturkonflikte
- Datenmodellkonflikte

Schemaintegration (3)

Vereinheitlichungsphase

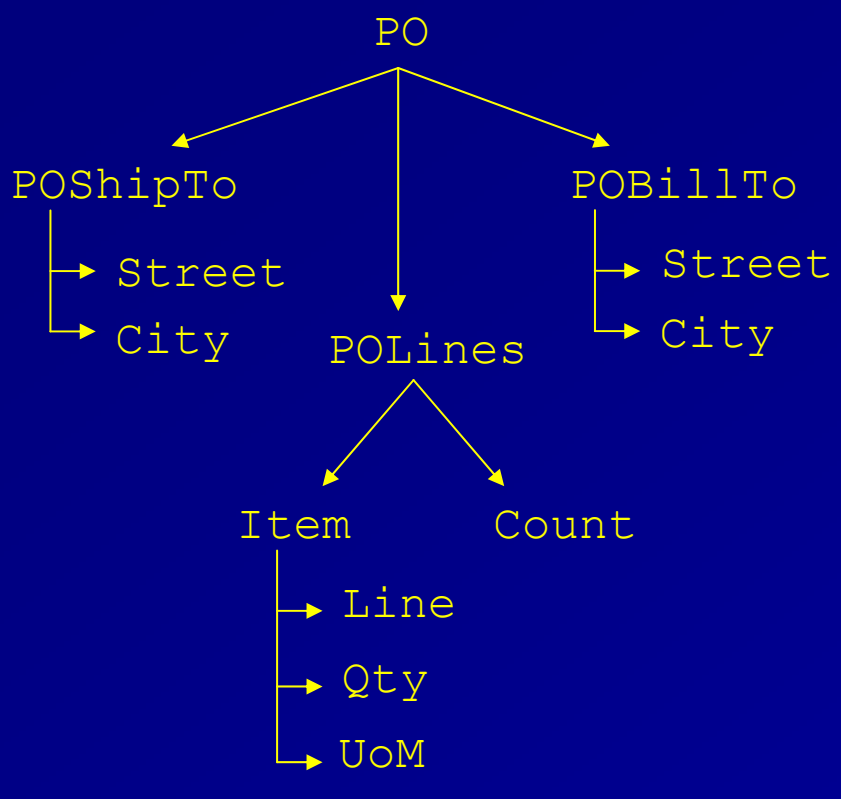
Ziele :

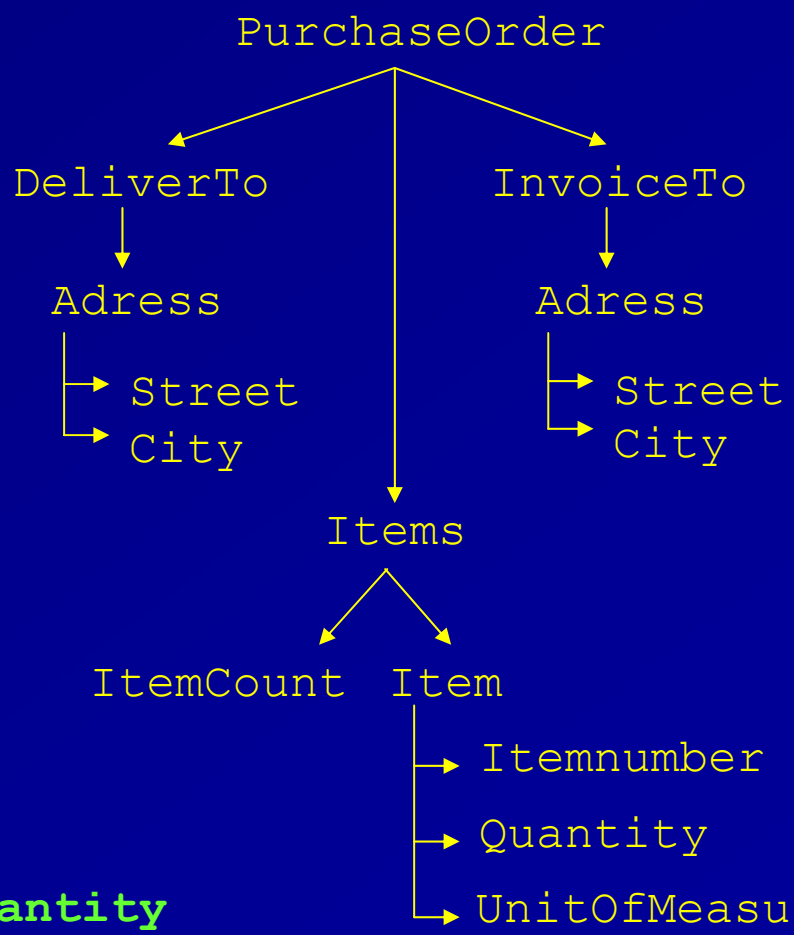
- Lokale Schemata für die Zusammenführung zu modifizieren
- Konflikte aus der Vergleichsphase beseitigen
- Oftmals manuelle Eingriffe nötig → sehr zeitaufwendig

Restrukturierungsphase

- Erstellen des globalen Schemas im Data Warehouse
- Überprüfung der Schema Qualität:
  - Vollständigkeit
  - Minimalität
  - Verständlichkeit
  - Korrektheit

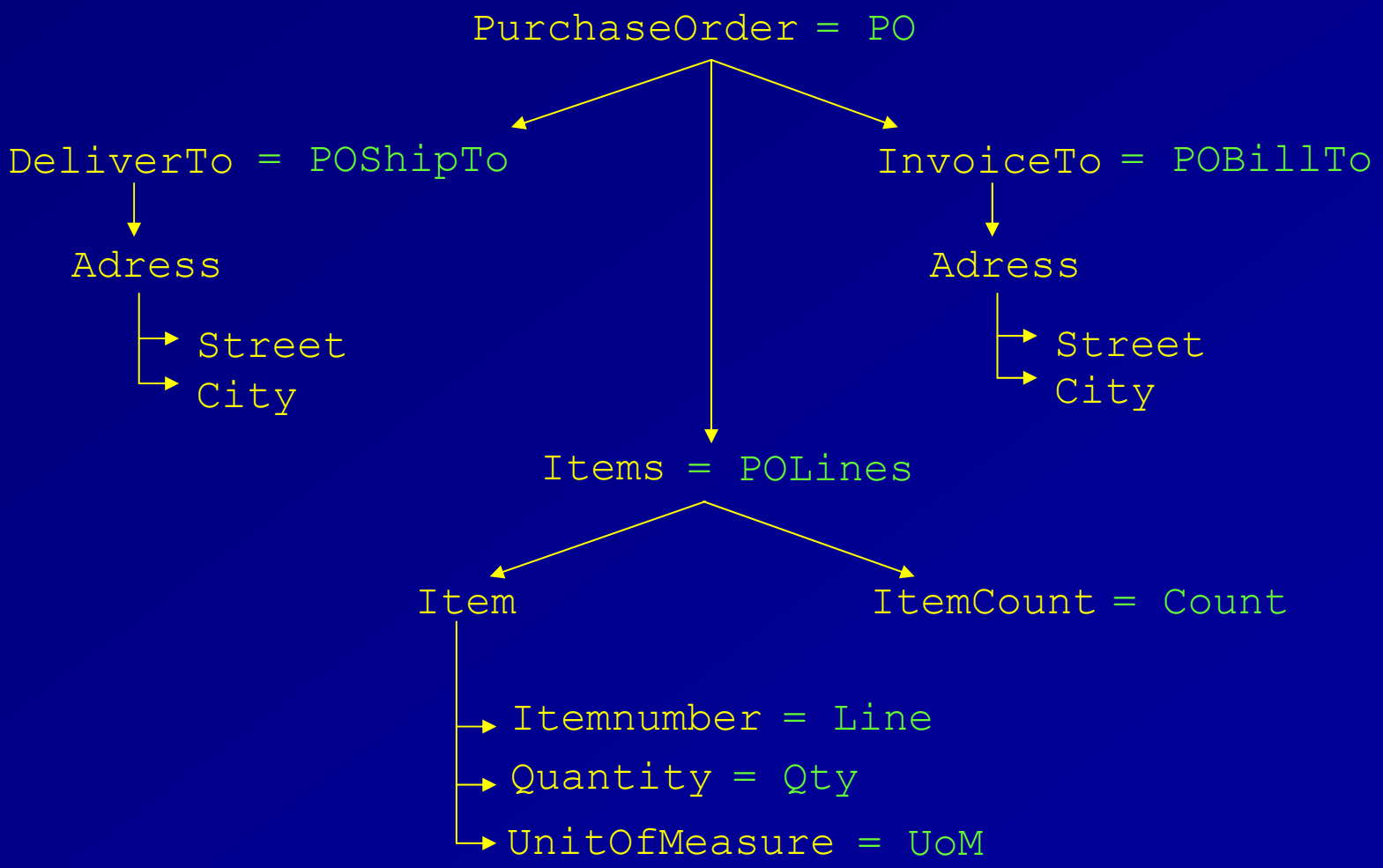
Schemaintegration (4) - Beispiel





Qty → Quantity
UoM → UnitOfMeasure
Bill → Invoice

Schemaintegration (5) - Beispiel



Datenintegration (1)

Schlüsselbehandlung

- Probleme
  - Schlüssel sind lokal eindeutig, aber global nicht mehr
  - Schlüssel können eine implizite Semantik besitzen
  - Zwei Objekte aus verschiedenen Quellsystemen können sich auf das selbe reale Objekt beziehen
- Lösungen
  - Einführung eines zusätzlichen Schlüsselattributs
  - Wenn die Semantik bekannt ist, kann der Schlüsselinhalt aufgespaltet und in mehrere Attribute gespeichert werden
  - Einsatz statischer Verfahren um Ähnlichkeiten zu erkennen

Datenintegration (2)

Vereinheitlichung von Zeichenketten

- Alle Zeichen auf Groß- bzw. Kleinbuchstaben umwandeln
- Umlaute in mehrbuchstabige Zeichenketten transformieren
- Einsatz von Verfahren um ähnlich klingende Wörter zu finden (z.B. SoundEx)

→ sorgt für einheitliche Basis um Vergleiche durchzuführen

Beispiel:

```
Küchengerät     →   KUECHENGERAET
VCR             →   Videorekorder

SCHMIDT → SCMDT → SMD → S530
SMITH   → SMT   → SMT → S530
```

Datenintegration (3)

Vereinheitlichung von Datumswerten

- Transformation in landesspezifisches Format (z.B. TT.MM.JJJJ)
- Moderne Datenbanken bieten interne und externe Darstellung an
- Altsysteme: oft Datum als Zeichenkette oder nur mit zweistelliger Jahreszahl

Umrechnung von Maßeinheiten

- Verwendung verschiedener Maßeinheiten auf der Welt

→ Umrechnung mittels Umrechnungstabellen

10 inch = 25,4 cm = 0.254 m

- Eintrag muss Vermerk über die verwendete Maßeinheit besitzen
- Bei gleicher Maßeinheit eventuell Skalierung notwendig

Datenbereinigung (1)

Aufgabe:

Inkorrekte, unvollständige und inkonsistente Daten entdecken und nach Möglichkeit behandeln

Häufig auftretender „Schmutz“

- Fehlerhafte Werte
  - Unvollständige oder fehlerhafte Daten (z.B. Adressangaben)
  - → Abgleich mit Daten aus anderen Quellen
  - Domänenunabhängige Validierung erfordert oft manuelle Eingriffe
- Redundanz
  - Daten, die mehrfach in einer oder verschiedenen Quellen gespeichert wurden
  - Entstehen häufig durch fehlende Normalisierung des Schemas
  - → Einsatz von Verfahren zur Duplikateleminierung

Datenbereinigung (2)

Häufig auftretender „Schmutz“ (Fortsetzung)

- Nullwerte
  - Es existiert kein Wert für das Attribut
    (z.B. Verfallsdatum bei nicht-verderblichen Lebensmitteln)
  - Wert des Attributs war zum Zeitpunkt des Eintragens nicht bekannt
- Behandlung von Nullwerten
  - Fehlende Werte manuell einfügen:
    Sehr zeitaufwendig und in großen DB‘s nicht realisierbar
  - Globale Konstante:
    Kann zu Verfälschungen von Statistiken führen
  - Mittelwert:
    Für numerische Werte gut, für Zeichenketten schlecht
  - Häufigster Wert:
    Ordnet Nullwert den häufigsten Wert der Spalte zu
    → keine Verfälschungen der Statistik

Datenbereinigung (3)

Sorted-Neighborhood-Methode

Arbeitet in 3 Phasen um Duplikate zu entdecken und zu eliminieren

- Schlüssel generieren
  - Jedes Tupel erhält einen Key
  - Key besteht aus den ersten 3 Ziffern oder Konsonanten jedes Feldes

  → Ähnlichkeitsfunktion
- Sortieren
  - Datensätze nach dem generierten Schlüssel sortieren
- Mischen
  - Mischen der Tupel, die (sehr) ähnliche Keys besitzen
  - Zusätzliche Ähnlichkeitsbedingungen können Mischvorgang noch effektiver machen

Datenbereinigung (4)

Sorted-Neighborhood-Methode (Beispiel)

ID	Name	Vorname	Adresse	Key
4711	Schmidt	Fritz	Hauptstr. 168	471SCHFRTHPT
0815	Maier	Kurt	Auf dem Acker 4	081MERKRTFDM
4711	Schmid	Fritzchen	Hauptstrasse 168	471SCHFRTHPT
0815	Mayer	Kurt W.	Auf dem Acker 4	081MYRKRTFDM

ID	Name	Vorname	Adresse	Key
4711	Schmidt	Fritz	Hauptstr. 168	471SCHFRTHPT
4711	Schmid	Fritzchen	Hauptstrasse 168	471SCHFRTHPT
0815	Maier	Kurt	Auf dem Acker 4	081MERKRTFDM
0815	Mayer	Kurt W.	Auf dem Acker 4	081MYRKRTFDM

Datenbereinigung (5)

Sorted-Neighborhood-Methode (Beispiel Fortsetzung)

ID	Name	Vorname	Adresse	Key
4711	Schmidt	Fritz	Hauptstr. 168	471SCHFRTHPT
4711	Schmid	Fritzchen	Hauptstrasse 168	471SCHFRTHPT
0815	Maier	Kurt	Auf dem Acker 4	081MERKRTFDM
0815	Mayer	Kurt W.	Auf dem Acker 4	081MYRKRTFDM

ID	Name	Vorname	Adresse	Key
4711	Schmidt	Fritz	Hauptstr. 168	471SCHFRTHPT
0815	Maier	Kurt	Auf dem Acker 4	081MERKRTFDM

Zusammenfassung

- Monitorrealisierungen und –techniken
- Extraktionskomponente
- Schemaintegration in 4 Phasen
- Integration von Daten aus verschiedenen Quellen
- Bereinigung der Daten von inkorrekten Werten und Duplikaten

Vielen Dank

Fragen ?