FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Datenbereinigung und Data Provenance

Verfahren und Begriffsdefinitionen, Prof. Dr. Markus Grüne

Fachbereich 3 Wirtschaft und Recht

Inhaltsüberblick

Data Provenance

Datenbereinigung und Fehler

Data Provenance

Zentrale Fragen:

- Wie sind meine Daten entstanden?
- Sind diese plausibel?

Andere Begriffe: Data Lineage

Todo:

- Rückverfolgung der Datengenese zu den Quellen
- Sehr schwer, da sehr viele mögliche Architekturen mit vielen Transformationsschritten

Data Provenance

Why-Provenance:

- Zusammenhang zwischen Quell- und Zieltupeln
- Welche Tupel wurden verwendet?

How-Provenance:

- Wie wurden die Tupel kombiniert, um das Ziel zu produzieren?

Where-Provenance:

- Zusammenhang zwischen dem Quell- und Zielort an dem die Daten „residieren“
- Z.B. Herkunftszelle bestimmen

Data Provenance

Möglichkeiten zur Analyse:

- instanzbasiertes Vorgehen → bestimmte Datenkonstellationen untersuchen
- anfragebasiertes Vorgehen → Abfrageausdrücke untersuchen z.B. mittels Operatorbaum (Abfragepfade)

Ggf. dann Änderungen der Datenquellen

Daten(be)reinigung…
Zwecke

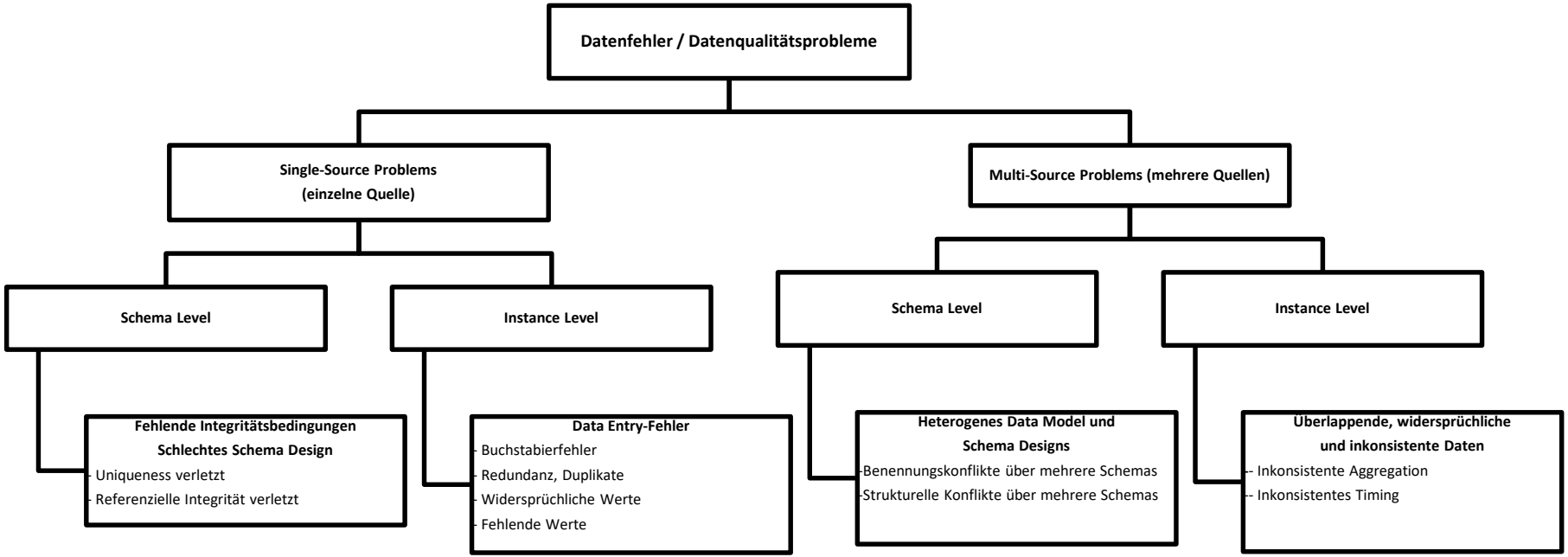
Behebung von fehlerhaften / verrauschten Daten

Fehler entstehen u.a. durch

- Falsche Verarbeitungsschritte (ETL)
  - Logische Fehler u.a.
- Zusammenspielen / Integration von Daten
  - unterschiedlicher Zeitscheiben
  - Overlapping Data Sources
- Falsche Schema Mappings / Schema Transformationen
- Duplikate und Konflikete → welche Daten soll ich dann nehmen?

Datenfehler / Daten-Qualitätsprobleme

Quelle: Jahn



Datenbereinigung bei IIS

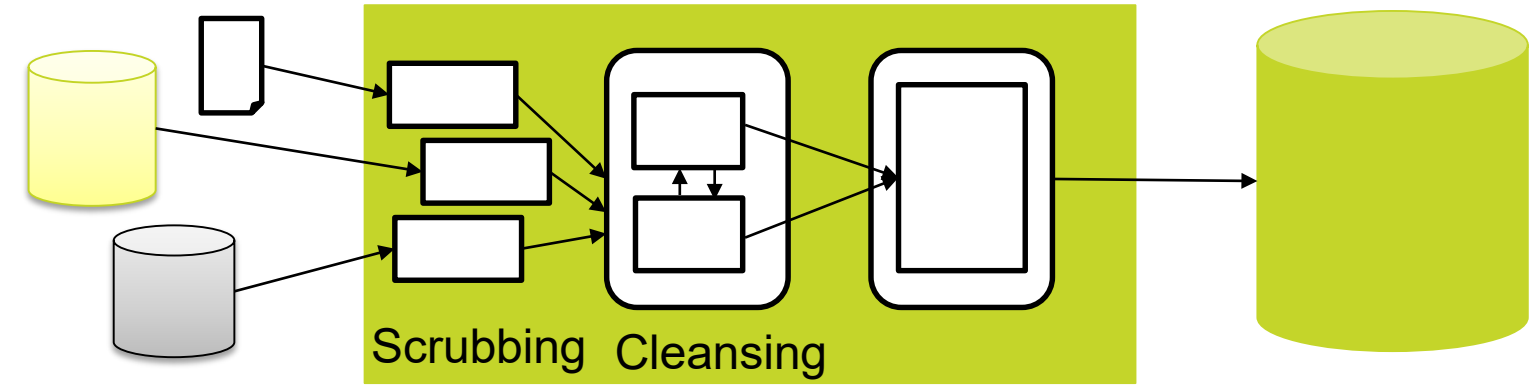
Wiederholung: IIS = Integriertes Informationssystem

Z.B. umgesetzt durch föderiertes DBMS, Multi-DBMS, DWH, Data Lake (?)

In föderierten Systemen muss die Bereinigung online stattfinden → Mediator

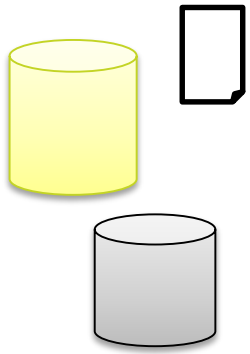
In DWH findet die Bereinigung vor dem Load in den Core statt.

Datenbereinigung im Data Warehouse Extract Transform Load (ETL)



In Anlehnung an: Skript Informationsintegration, SoSe2010, M. Herschel, Uni Tübingen

Fehler einer Datenquelle Scrubbing



Datenquellen

Unzulässige Attributwerte: z.B. Geburtsdatum 24.24.24

Fehlende Attributwerte/ Attributwertteile: z.B. unvollständige Telefonnummern

Wechselnde Eintragsformate bei Freiformfeldern: z.B. „Daimler“, „DB Stuttgart“, „Mercedes Benz AG, Stuttgart“ → häufige Fehlerquelle für Duplikate

Rechtschreibfehler

Kryptische Datenwerte: z.B. „Cert X-903“ für Qualifikation

Falsch erfasste Attribute: z.B. Straße = „Frankfurt“

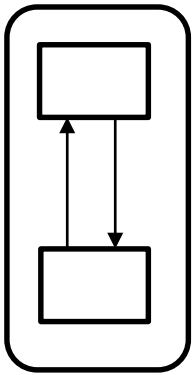
Inkonsistenzen auf Datensatzebene: z.B. Alter 50, aber Geburtsdatum 30.12.2010

Inkonsistenzen des Zustands von Tabellen: z.B. doppelte Verwendung desselben Keys

Fehler in Fremdschlüsselbeziehungen: Verweise auf fehlende Datensätze

In Anlehnung an: RDD10

Fehler bei der Integration mehrerer Datenquellen – Cleaning / Cleansing



Cleansing

Duplikate → z.B. mittels Heuristiken erkennen (Distanzmaße), Containment

Vernachlässigung unterschiedlicher Datenformate und Zeichensätze

Auswahl falscher Zeitfenster → Datenquellen haben andere Bezugszeiträume und „passen nicht zusammen“ → Bereinigung durch Neuladen

Fehlerhafte Einträge können teils durch Algorithmen bereinigt werden, wenn die Ähnlichkeit der Daten heuristisch bestätigt werden kann

Daten mit unterschiedlichen Standards: m, ft, .. → können normiert werden

Falsche Schema Mappings → Bereinigung des Schema Mappings

Ggf. manuelle Nachkorrektur und Protokollierung der Korrekturen

Produkte

Data Quality Services (SQL Server), Microsoft

Azure Data Factory, Microsoft

Data Cleanser, Oracle

…

Lessons Learned

Data Provenance versucht, die Entstehung von Daten zu erklären und nachvollziehbar zu machen.

Bereinigungen sollten nach Fehleranalyse nur in den Quelldaten erfolgen.

Datenfehler sind unvermeidbar in integrierten Informationssystemen.

Datenfehler können klassifiziert werden.