INFORMATIONSINTEGRATION

PROBLEMSTELLUNG

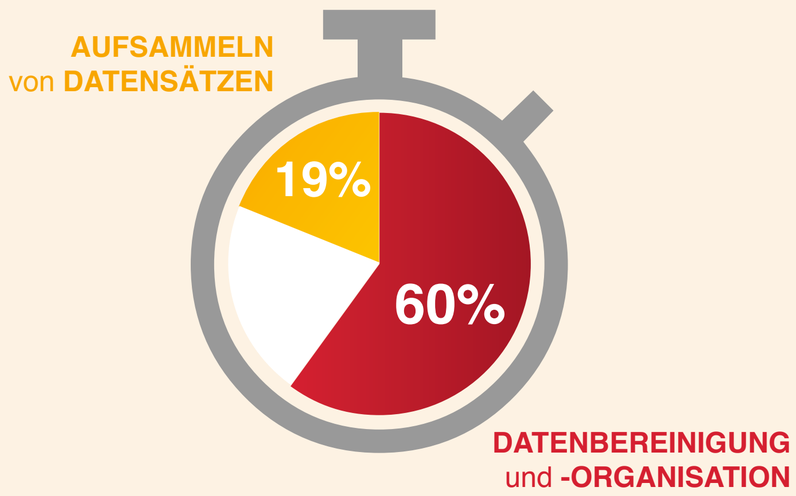
Ein gewöhnliches Großunternehmen verfügt über **5.000 verschiedene Datenbestände**. Diese wurden oft unabhängig voneinander entwickelt und sind auf vielfältige Weise heterogen.

Der Umgang mit dieser **Heterogenität** ist ein komplexes Problem. Insbesondere das Aufspüren semantischer Unterschiede ist aufwändig und **kaum automatisierbar**.

Aus diesem Grund wird nur ein Bruchteil der zur Verfügung stehenden Datenbestände in ein **Data Warehouse** integriert und damit der Analyse zugänglich gemacht. *Der Bedarf nach weiterer Integration bleibt enorm*.

Informationsintegration ist der Prozess, Daten aus **heterogenen Datenbeständen** in eine **homogene Struktur** zu überführen. Aspekte dieses Prozesses sollen im Folgenden dargelegt werden.

Wofür Data Scientists Zeit aufwenden



Quelle: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#46a3d8f06f63

„the most important way to support a data scientist is to help him find and clean data“

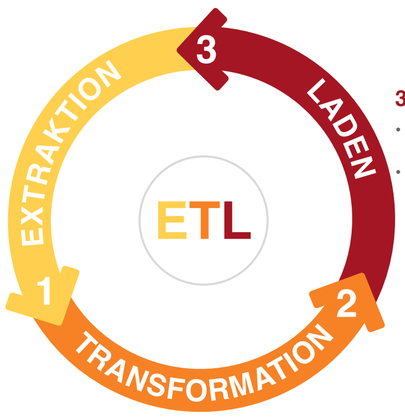
- Michael Stonebraker

„Data Science is 99% preparation, 1% misinterpretation“

- Big Data Borat; Twitter-Nutzer

DATA WAREHOUSE

- periodische Befüllung eines Data Warehouses
- hoher Aufwand pro Quellsystem
- Problem: erfordert globales Schema



1. EXTRAKTION

- Extraktion gewünschter Daten aus Quellsystem
- Problem: Verschiedene Dateiformate und Zugriffs-möglichkeiten (z.B. www)

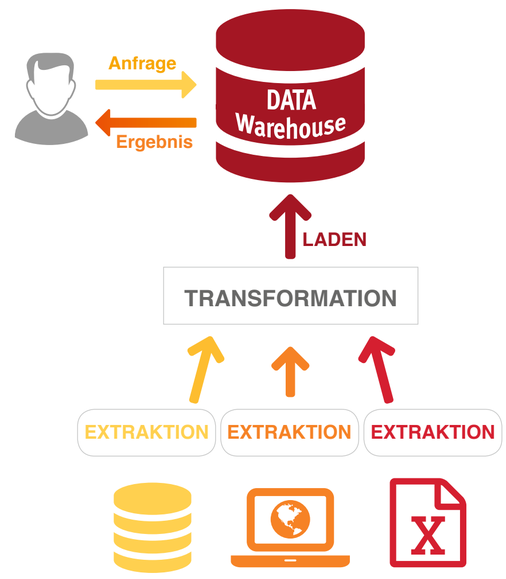
2. TRANSFORMATION

- Transformation der Daten in homogene Struktur
- Problem: ca. 10% der Daten typischerweise inkorrekt
- veraltet, redundant, inkonsistent, falsch formatiert, …

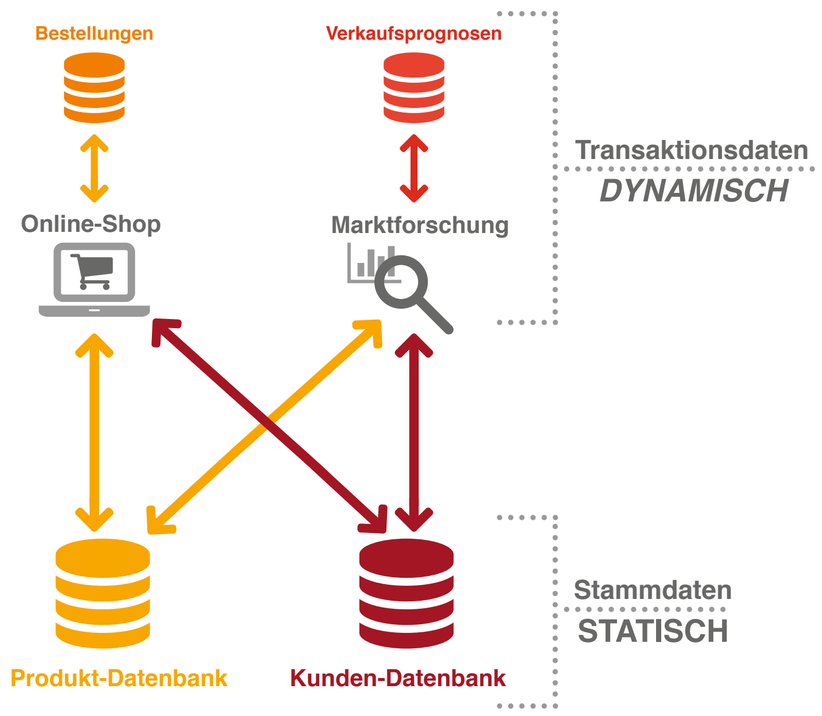
3. LADEN

- Bulk-Load der Daten ins Data Warehouse
- möglichst schnell

Architektur



MASTER DATA MANAGEMENT



Stammdaten-verwaltung

- Vereinheitlichung der Stammdaten eines Unternehmens:
  - Daten, die mit dem operativen Kerngeschäft verbunden sind (z.B. Kunden, Produkte)
- Unterschied zu Data Warehouse:
  - kein globales Schema notwendig

DATENBEREINIGUNG

Wichtiger Vorbereitungsschritt für Analyse

➡**Garbage-in-Garbage-out-Prinzip**

verlangt meist detailliertes Domänenwissen

Normalisierung:
- erleichtert weitere Verarbeitung der Daten
- z.B. Stammformreduktion: lesen, lesbar, Leser -> les

Konvertierung:
- z.B. $ -> €

Fehlende Werte:
- verursachen Probleme bei Statistiken
- Ersetzung von numerischen Werten mit Durchschnittswert
- Anwendung bekannter funktionaler Abhängigkeiten, um Fehler zu reparieren

Beseitigung von Ausreißern:
- z.B. Z-Score: Entfernung zum Erwartungswert abhängig von der Standardabweichung

Referenztabellen:
- für einheitliche Schreibweisen und Prüfung auf Konsistenz

Duplikaterkennung:
- z.B. Edit-Distanz: minimale Anzahl an Edit-Operationen, um den einen in den anderen String zu überführen
- Beachtung phonetischer Ähnlichkeit, Nähe der Buchstaben auf einer Tastatur, …

Datenfusion:
- Kombination von Duplikaten zu einem Datensatz
- Anreicherung von Daten mit zusätzlichen Informationen

Leonard von Merzljak
Bachelor Student
leonard.vonmerzljak@student.hpi.de

Quellen: http://www.redbook.io/ch12-dataintegration.html; Database Systems - The Complete Book, Hector Garcia-Molina et al.; Informationsintegration - Ulf Leser, Felix Naumann; https://www.youtube.com/watch?v=KRcecxdGxvQ&t=646s

Hasso Plattner Institut