Data Engineering Prozesse

Cold Warm Hot ist eine Strategie um Data Engineering Prozesse anschaulich darzustellen zu können.
Es gibt kalte Daten mit denen noch nichts gemacht wurde, durch die Aufbereitung werden sie zu warmen Daten und schließlich nach der Analyse liegen heiße Daten vor.
Zuerst müssen die Rohdaten gesammelt werden, im nächsten Schritt werden sie vereinheitlicht, gefiltert und Duplikate werden eliminiert, das nennt man Aufbereitung. Im dritten Schritt werden die Daten mittels einer entsprechenden Technologie gespeichert, damit im nächsten Schritt die Analyse erleichtert wird. Hier sind kreative Lösungen gefragt. Zuletzt werden die Analyseergebnisse visualisiert wobei auf eine zielgruppengerechte Darstellungsart geachtet werden muss.

keine Fragen keine Antworten

Fragen stellen

Fragen beantworten

Rohdaten

„Data Lake“
Daten sammeln und so billig wie möglich speichern

Rohdaten werden z.B durch **Manual Logging** gesammelt

Sensordaten
Analysen auf Grundlage des Standortes
(Bsp.: Krankenhausbehandlungen)

verschiedene Datenquellen

z.B.: PDF, XLS, Web
→ nicht einheitlich
„Euros für Ärzte“: Daten aus 40 verschiedene Quellen

Aufbereitung

mit z.B. jupyter ← **Data Cleaning** → evtl. filtern

gemeinsames Format
mit verschiedenen Tools und händisch

Duplikaterkennung
Z.B.: Window Comparison über Datensatz laufen lassen („Euros für Ärzte“)

Duplikat-erkennung

false negative: Duplikat-Paar wird fälschlicherweise nicht erkannt → ist z.B. im Datenjournalismus die besser Alternative

Speicherung

Technologie aufgrund von Menge und Komplexität wählen

Hadoop
ist eine verteilt arbeitende Software, die auf dem **MapReduce** Algorithmus basiert

„Data Fusion and load“
Zusammenfassen und in Datenbank laden

Amazon S3
ist ein Cloud-Objektspeicher in dem große Datenmengen gespeichert werden können

Was ist MapReduce?

Programmiermodell für nebenläufige Berechnungen über große Datenmengen auf Computerclustern

Analyse

Machine Learning Tools

Interaction Mining
und Event Query Mining

Berechnungen
Z.B. mit Python/Pandas

Black Box Model
vs.
actual process models
„What if“ Analyse

Vereinigung von Wissen

Es muss Informatiker und z.B Journalistenwissen vereinigt werden um sinnvolle Antworten zu finden/erkennen

Visualisierung

→ mit z.B Tableau

Real-Time-Location-Map
Unterschiedliche Darstellung für verschiedene Personen

Petri Netze → Folding

generalisieren/abstrahieren

Statische vs. Dynamische
Überblick über Zeit/Dauer und Raum

Warum Daten-visualisierung?

Entscheidungsträger besitzen weniger Wissen als Entwickler
→ Mehrwert in möglichst kurzer Zeit darstellen

Carolin Goerke
(3. Semester Bachelor)

Hasso Plattner Institute, Potsdam, Germany

E-Mail: carolin.goerke@student.hpi.de