-Data Engineering: Prozesse und Techniken-

Wie werden immer größer werdende Datenmengen gesammelt, aufbereitet, parallelisiert, schließlich zum Nutzer gebracht und visualisiert? Wie durchläuft ein Rohdatenstrom diesen Prozess in Echtzeit? Dieser Kernfrage des **Data Engineerings** wird hier auf den Grund gegangen.

Durch welche smarten Algorithmen und Techniken können aus den Datenmengen neue Informationen bereitgestellt werden? Mit dieser Frage beschäftigt sich die **Data Science**.

Aktuelle Aufgabenstellungen

Optimieren von Geschäftsprozessen
- Kürzere Wartezeit für Patienten
- Prüfen, speichern und begleichen von Rechnungen
- Automatisches Scannen von Artikeln

Verbesserung von Betriebsaufgaben
- Energiesparen in Produktionsstätten und Städten
- Genaue Auswertung kostenintensiver Crashtests
- Komplexe Suche nach Patienten
- Suchen/Recherchieren relevanter Nachrichten

Analyse der Kunden
- Finden abgestimmter Jobangebote
- Zeigen interessanter Shoppingvorschläge
- Filtern von wichtigen Nachrichten

Beginn der Data Pipeline

Eine Pipeline verwirklicht den Datenfluss von den Rohdaten bis zur Visualisierung. Sie wird so konstruiert, dass neue Daten in Echtzeit den Prozess durchlaufen.

Suchen und Finden der Daten

Woher kommen die Daten?

- Manuelles Generieren:
- Automatisches Generieren:
- Nicht eigene Quellen:

Beispiele aus der Praxis:
- Arzt gibt Patientendaten und Diagnose ein
- RTLS: GPS-basiertes Tracking von Patienten und Mitarbeitern
- Daten aus Abteilungen oder Unternehmen

Vorbehandlung der Daten

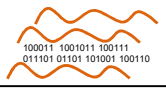
Entfernen fehlerhafter Daten und Berichtigung

Beispiele aus der Praxis:
- Patient oder Arzt lässt Trackingsensor liegen, wodurch falsche Daten produziert werden
- Personaler benennen Jobs unpräzise, es ist nicht ersichtlich, was gesucht wird

Speicherung der Rohdaten

Speichern der Daten im Data Lake

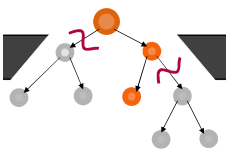
Alle Rohdaten werden gespeichert, damit diese für die komplexe Analyse zur Verfügung stehen.

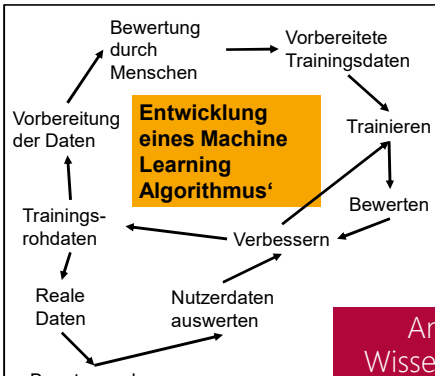


Auswählen relevanter Daten

Welche Daten sind kennzeichnend für das Ergebnis? Es kann keine gesamte Datenbank pro Anfrage analysiert werden.

Static / Dynamic Partition Elemination (s.A. Redshift) Entfernen von Zeilenpartitionen, sodass große Blöcke nicht durchsucht werden müssen





Durch welche Verfahren werden Muster und passende Lösungen gefunden?

Interaction Mining: Kombination von Interaktionen, Aktivitäten und deren Attribute, um z.B. komplexe Arbeitsprozesse mit vielen Personen optimal zu planen.

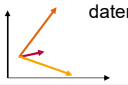
Event Query Mining: Formalisieren und Vorhersagen von Interaktionen und Zeitfenster, um optimale Lösung zu finden.

II-Miner: Kombination mehrerer Techniken, um korrekte und minimale Event-Patterns zu finden.

Energy Efficient Coal Mining (o. GRIDS):Wissensgewinnung aus verschiedensten Energiedaten zur Energieeinsparung

Prediction Analysis: schätzen und relevante Ergebnisse **filtern**, z.B. Suchen und Filtern von Jobangeboten

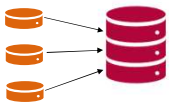
Analyse und Wissensgewinnung



Organisieren im Data Warehouse

Ungeordnete Rohdaten werden in strukturierten Datenbanken organisiert.

- Erkennen von Datenlücken und Approximieren
- Anpassen der Datenstruktur auf die kommende Analyse
- Entfernen von Duplikaten und nicht benötigten Daten



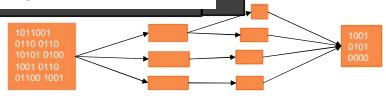
Beispiele aus der Praxis:
- Amazon Redshift Spectrum: Data Warehousing Tool, das Anfragen schnell bearbeiten kann (>1 Exabyte)

Parallelisierung der Datenverarbeitung

Wie werden Daten geteilt, sodass große Datenmengen parallel bearbeitet werden können?

Amazon Redshift Spectrum: Aufteilung der Anfrage auf lokalen Rechner und mehrere tausend Nodes, die S3 durchsuchen

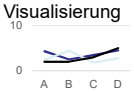
Stream Processing: Parallelisierung mit Apache Flink und relationalen APIs zur parallelen Analyse



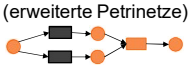
Visualisierung der Ergebnisse

Was soll der Nutzer sehen?

- **Weglassen** irrelevanter Informationen
- Wahl der Diagramme / Visualisierung



Folding: Abstrahieren der komplexen Datenmodelle auf für Menschen zu verstehende **GSPN** (erweiterte Petrinetze)



Wie werden Informationen veranschaulicht?

Ermöglichung datengetriebener Entscheidungen

Wie werden verschiedene Datensätze kombiniert?

- Mensch erlangt Wissen durch Kombinierung
- VizQL: Umwandlung von Benutzereingaben in SQL

Wie können Querys in Echtzeit berechnet werden?
>1TB Rohdaten ⟷ Echtzeitanwendung

Cold	Warm	Hot
Rohdaten-sammlung im Data Lake	Data Warehouse mit vorbehandelten Daten	Hochentwickelte Datenbanken mit analysierten Daten, die in Echtzeit Anwenderfragen beantworten

Erkenntnisse aus den Daten beim Anwender

Georg Lange
Bachelorstudent am Hasso-Plattner-Institut Potsdam

E-Mail: Georg.Lange@student.hpi.de

HPI Hasso Plattner Institut