Data Engineering in der Praxis
- Prozessschritte und Herausforderungen -

Die Data Science beschäftigt sich mit der Analyse umfangreicher Datensätze. Je nach Anwendungsbereich und Aufgabenstellung ist dabei ein unterschiedlicher Workflow erforderlich, um aus den Rohdaten die gewünschten Informationen extrahieren zu können.

Das Data Engineering hat das Ziel, den Data-Science-Workflow möglichst effizient und skalierbar zu gestalten. Die Prozessschritte und Herausforderungen, welche dabei zu berücksichtigen sind, werden in diesem Poster überblicksartig dargestellt.

Mögliche Aufgabenstellungen in der Data-Science

Optimierung von Betriebsabläufen

- Reduktion des Energieverbrauchs in Produktionsstätten
- Verkürzung der Wartezeit für Patienten in Kliniken

Maschinen-Sensoren
GPS-Tags von Mitarbeitern

Vorhersage von Kundenverhalten

- Generieren interessenabhängiger Produkt- bzw. Jobempfehlungen

www.

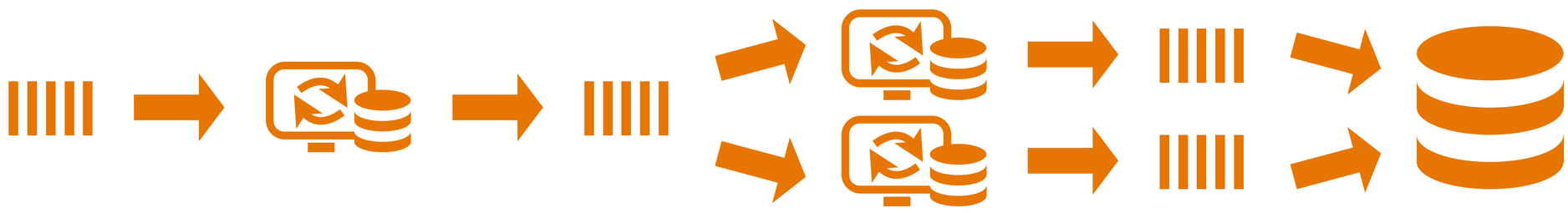
Aufzeichnung von Applikations- / Website-Nutzerdaten

Automatisierung von Unternehmensprozessen

- Automatisches Prüfen von Rechnungen
- Automatisches Filtern von Nachrichtenmeldungen

Text- und Bilddateien z.B. aus Emails und Websites

Die Rohdatenquellen generieren einen aus Ereignissen bestehenden Datenstrom



Echtzeitanalyse des Datenstroms

- Anwenden vorab definierter Abfragen auf den Datenstrom
- erfordert schnellen Datenzugriff

Aufteilung des Datenstroms zur Performancesteigerung

- Herausforderung: Management der Systemzustände zur Wahrung der Datenkonsistenz

Speicherung des Datenstroms im Data Lake

- Vorhalten der Daten für komplexere Aufbereitungs- und Analyseschritte im Batch-Processing

Erzeugen eines Data-Warehouse durch Aufbereitung der Rohdaten aus dem Data Lake

- Entfernen von **Duplikaten** und fehlerhaften Daten
- Erkennen von **Lücken** im Datensatz, ggf. Einfügen von Approximationen

- **Aggregation des Datensatzes**, Entfernen nicht benötigter Daten
- **Anpassen der Datenstruktur** an die geplante Analyse

Information Retrival & Predictive Analytics

Lineare Modelle
- z.B. Ermitteln der Schlüssel-faktoren des Energiebedarfs

Matrix Factorization
- z.B. zum Generieren von Produktempfehlungen

Supervised Machine Learning
- z.B. für Bild- und Textanalyse

Erzeugen / Training geeigneter Modelle → **Anwenden der Modelle und Bewerten der Modellperformance** → **Verbessern der Modelle**

Visualisierung und Anwendung der gewonnenen Information

Josafat-Mattias Burmeister

Bachelorstudent im Studiengang IT-Systems-Engineering
Hasso Plattner Institut Potsdam

E-Mail: josafat-mattias.burmeister@student.hpi.de

HPI Hasso Plattner Institut