Hochschule Karlsruhe
University of
Applied Sciences

Data Engineering (DSCB330)
Wintersemester 2024/2025
Prof. Dr. Jannik Strötgen
jannik.stroetgen@h-ka.de

5. Dezember 2024

Übungsblatt 6

Aufgabe 6.1 – Data Lakes und Data Warehouses

a) Vergleichen Sie Data Lakes und Data Warehouses anhand von drei Charakteristika Ihrer Wahl.
b) Nennen Sie jeweils mindestens zwei Vorteile und Nachteile von ETL und ELT Prozessen.
c) Nennen Sie zwei typische Anwendungsfälle für einen Data Lake und zwei für ein Data Warehouse.

Aufgabe 6.2 – Data Lake und Metadaten

a) Welche Herausforderungen können bei der Nutzung eines Data Lakes auftreten, und wie können diese behoben werden?
b) Was ist der Unterschied zwischen schema-on-read und schema-on-write, und wie hängt dies mit Data Lakes und Metadaten zusammen?

Aufgabe 6.3 – Datenbase vs. Data Lake vs. Data Warehouse vs. Knowledge Graph

Ziel dieser Übungsaufgabe ist es, eine Recherche durchzuführen zu verschiedenen Systemen, die im Bereich Data Science und Data Engineering eine wichtige Rolle spielen, und die Sie als Data Scientist oder Data Engineer kennen und voneinander abgrenzen können sollten.
Wann immer Sie bei Ihrer Recherche einzelne Punkte sammeln, geben Sie bitte Quellen an, wo Sie Informationen gefunden haben. Sie können auch Vorlesungsmaterialien als Quellen nutzen.

- Recherchieren Sie im Internet und suchen Sie nach Gemeinsamkeiten und Unterschieden zwischen Databases, Data Warehouses, Data Lakes und Knowledge Graphs.

Aufgabe 6.4 – Metadaten

a) Welche Rolle spielt Wikidata als zentrale Metadatenquelle für Wikipedia? (Falls nötig, recherchieren Sie bitte)
b) Wie hilft die Versionshistorie als Metadatenquelle in Wikipedia, die Qualität eines Artikels zu bewerten?

Aufgabe 6.5 – Metadaten & Wikipedia API

Nutzen Sie die **Wikipedia-API**, um die folgenden Metadaten zu einem beliebigen Wikipedia-Artikel abzurufen:

1. Erstellungsdatum des Artikels
2. Datum der letzten Bearbeitung
3. Anzahl der Bearbeitungen
4. Kategorien des Artikels

Implementieren Sie ein Python-Skript, das den Benutzer nach einem Artikel-Titel fragt, die Wikipedia-API nutzt, um die genannten Metadaten abzurufen, und die Ergebnisse in einer übersichtlichen Form ausgibt.