KAFKA CHALLENGE

Speicherung der aktuellen Trends auf Wikipedia und historischen Veränderungen von Themen mittels Apache Kafka.

Struktur

1. Einleitung
2. Methodik
3. Prototyp
4. Aspekte

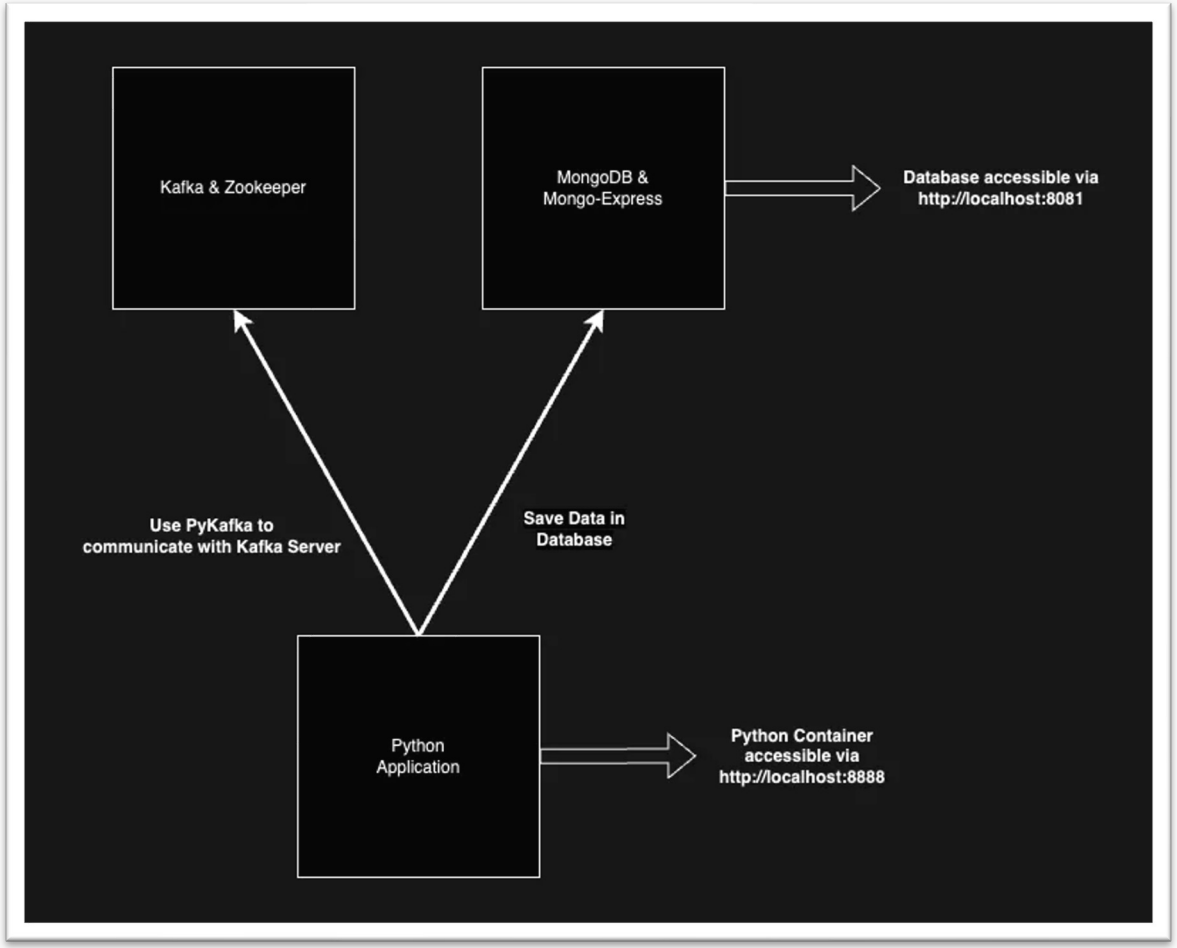
EINLEITUNG

Herangehensweise

- Einlesen in Kafka (Container, Topics, Consumer, Producer)
- Erstelle Repository
- Erstelle Kafka Container und debugge mittels KafkaCat
- Erstelle MongoDB Container für die Datenbank
- Erstelle Python Application um Logik zu implementieren (erst lokal, später im Container)

- Größten Probleme:
  - *Wie könnten Beispieldaten aussehen?*
  - *Historische Artikel auch speichern oder nur Anzahl Edits?*

METHODIK


Kafka & Zookeeper
MongoDB &
Mongo-Express
Database accessible via
http://localhost:8081
Use PyKafka to
communicate with Kafka Server
Save Data in
Database
Python
Application
Python Container
accessible via
http://localhost:8888

IDEE

Idee

- Producer
  - *Prototyp: producer.ipynb um Beispiel-Datenpunkte zu erstellen*

- Consumer
  - *Prozessiere Daten alle 60 Sekunden*
  - *Speichere historische Daten (neue und alte Artikel) in Datenbank Wikipedia_Historic*
  - *Zähle globale und deutsche Edits und speiche in Datenbank Wikipedia_Trend*
  - *Nutze MongoDB Datenbank und speichere Daten als JSON*

PROTOTYP

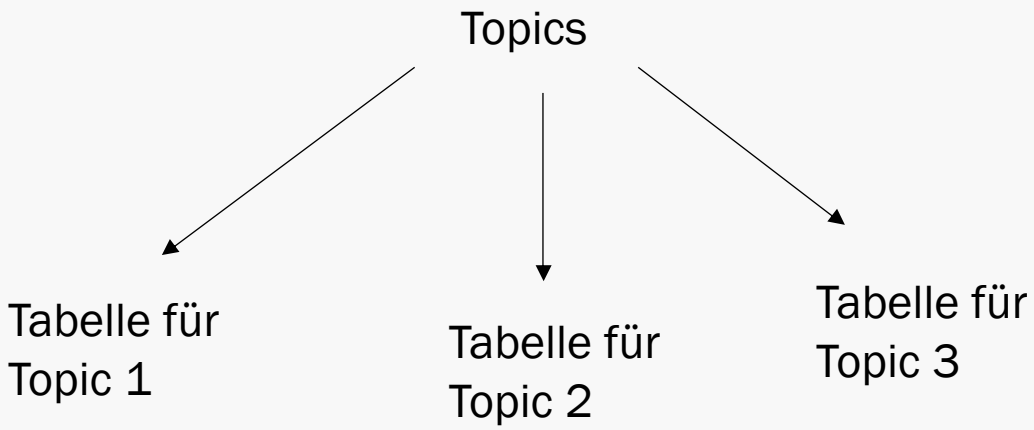
ASPEKTE

Was wäre eine mögliche Datenbank zur Speicherung der Daten?

- MongoDB war eine schlechte Entscheidung:
  - *Table-Based Struktur sinnvoller, da ich strukturelle Daten habe*
    → Strings, Integerer
  - *Würde im Endeffekt einfach eine PostgreSQL Database nutzen*
    - Möglicherweise Cassandra, wenn es NoSQL sein soll
  - *Ein Data Warehouse (OLAP) anstelle eines DBMS (OLTP) könnte auch sinnvoll sein: Amazon Redshift, Bigquery, Microsoft Synapse*

Welches Datenmodell wäre deiner Meinung nach sinnvoll zur Ablage der Events?

- Relationales Datenmodell für PostgreSQL:



Welche Topics wären sinnvoll?

Multi-Topic Struktur	Single-Topic Struktur
Mehrere Topics können einen höheren Overhead verursachen → Mehrere Tabellen verwalten	Consumer muss zwischen den Wikipedia-Themen differenzieren
Wikipedia Artikel müssen bestimmten Topics zugeordnet werden	Hohes Datenaufkommen → Engpässe bei Verarbeitung wenn Partitionierung nicht richtig konfiguriert
Bessere Skalierung über Topics	Eine einzige Topic-Struktur vereinfacht die Konfiguration


VIELEN DANK FÜR DIE
AUFMERKSAMKEIT!