

Übung 5 - Apache Kafka mit Movie-Datensatz

Ziel der Übung

In dieser Übung lernen Sie, ein Apache Kafka Deployment einzurichten und mit Kafka-Datenströmen zu arbeiten. Sie erstellen einen Producer, zwei Consumer und skalieren das Kafka-Cluster und die Topic-Partitionen.

1. Kafka Deployment mit Docker Compose

1. Setup:

- Erstellen Sie eine `docker-compose.yml`
- Nutzen Sie das offizielle Kafka Docker-Image (ohne Zookeeper)

2. Testen Sie das Deployment:

- Starten Sie das Cluster und prüfen Sie die Logs, um sicherzustellen, dass Kafka korrekt läuft.
-

2. Python Producer: Film-Datensätze in ein Topic schreiben

1. Setup des Producers:

- Verwenden Sie die Python-Bibliothek `kafka-python`.
- Laden Sie den Movie-Datensatz aus der CSV-Datei.
- Schreiben Sie jedes Datensatz-Objekt einzeln in ein Kafka-Topic `movie-data`.
- Serialisieren Sie die Daten als JSON, bevor Sie sie ins Topic senden.

2. Testen Sie den Producer:

- Starten Sie den Producer und stellen Sie sicher, dass alle Datensätze korrekt ins Topic geschrieben werden.
-

3. Zwei Python Consumer mit unterschiedlichen Consumer Groups

1. Setup der Consumer:

- Erstellen Sie zwei Consumer in Python, die das Kafka-Topic `movie-data` lesen:
 - **Consumer A** gehört zur Consumer Group `group-a`.
 - **Consumer B** gehört zur Consumer Group `group-b`.

2. Funktionalität:

- Lassen Sie beide Consumer parallel laufen.
 - Verifizieren Sie, dass jeder Consumer alle Nachrichten aus dem Topic unabhängig voneinander verarbeitet.
-

4. Skalierung des Kafka-Clusters auf 3 Nodes

1. Cluster erweitern:

- Passen Sie die `docker-compose.yml` Datei an, um einen drei Kafka-Broker hinzuzufügen.
- Verknüpfen Sie den neuen Broker mit Zookeeper.

2. Verifizieren:

- Prüfen Sie, ob die Broker im Cluster verfügbar sind, indem Sie die Kafka-CLI verwenden (`kafka-topics.sh --describe`).
-

5. Partitionen des Topics auf 2 erhöhen

1. Partitionierung:

- Erhöhen Sie die Anzahl der Partitionen des Topics `movie-data` von 1 auf 2 mit der Kafka-CLI (`kafka-topics.sh --alter`).
- Beachten Sie, dass bereits existierende Nachrichten nicht neu verteilt werden.

2. Testen Sie die neue Partitionierung:

- Starten Sie den Producer und Consumer erneut.
- Verifizieren Sie, dass Nachrichten jetzt auf beide Partitionen verteilt werden.