Übung Big-Data-Technologien
Prof. Dr. Christoph Schmitz

Übungsblatt 10: Kafka

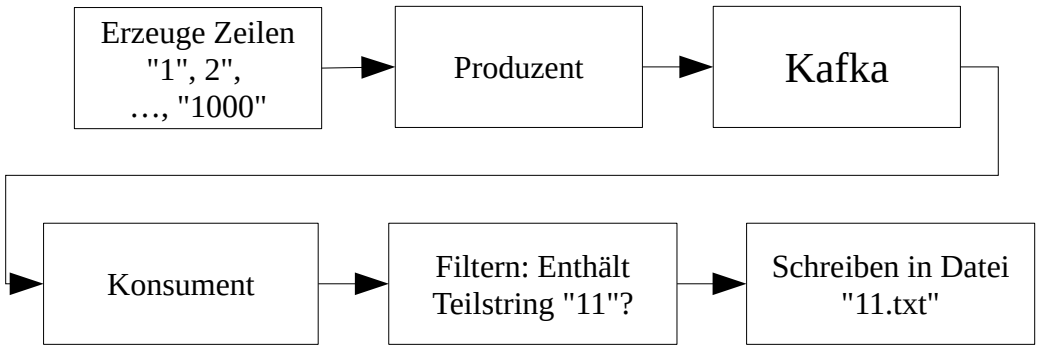
Hinweise

- Kafka läuft auf den Rechnern `infbdt{06-11}` jeweils auf Port 6667. Die zugehörigen Tools finden Sie unter `/usr/hdp/current/kafka-broker/bin`.
- ZooKeeper läuft auf den Rechnern `infbdt{09,10,11}` jeweils auf Port 2181.
- Die Authentifizierung erfolgt über Kerberos-Tickets (→ Prüfen mit `klist`).
- Für jeden Benutzer `bigdataXYZ` sind drei Topics bis `bigdataXYZ-1` bis `bigdataXYZ-3` angelegt, also z. B. bigdata042-1, bigdata042-2, bigdata042-3.
- Um die Authentifizierung zu konfigurieren, wird eine Datei `kafka.jaas` benötigt. Diese finden Sie in Stud.IP. Sie müssen diese Konfiguration anpassen (Benutzername!), auf den infbdtXYZ-Rechnern ablegen und den Kafka-Clients bekanntmachen:
  `export KAFKA_OPTS="-Djava.security.auth.login.config=$PWD/kafka.jaas`
- Für die Authentifizierung wird außerdem für Producer und Consumer die Option `security.protocol=SASL_PLAINTEXT` benötigt. Sie können diese setzen mit:

  `--producer-property security.protocol=SASL_PLAINTEXT` beim Producer bzw.
  `--consumer-property security.protocol=SASL_PLAINTEXT` beim Consumer.

Aufgabe 1

Implementieren Sie mit dem `kafka-console-producer` und dem `kafka-console-consumer` sowie Shell-Tools (`seq`, `grep` usw.) oder Skripten die folgende Topologie:



Hinweise:

- Der Produzent produziert selbst keine Daten, sondern nimmt Daten von der Standardeingabe entgegen. Sie müssen die Daten also anderweitig produzieren und mit Shell-Pipelines („|“) in den Produzenten eingeben.

- Der Konsument schreibt alle empfangenen Daten auf die Standardausgabe. Auch hier müssen Sie wieder mit Shell-Pipelines („|“, „>“) die Daten filtern und umleiten.
- `man seq`, `man grep` usw.

Aufgabe 2

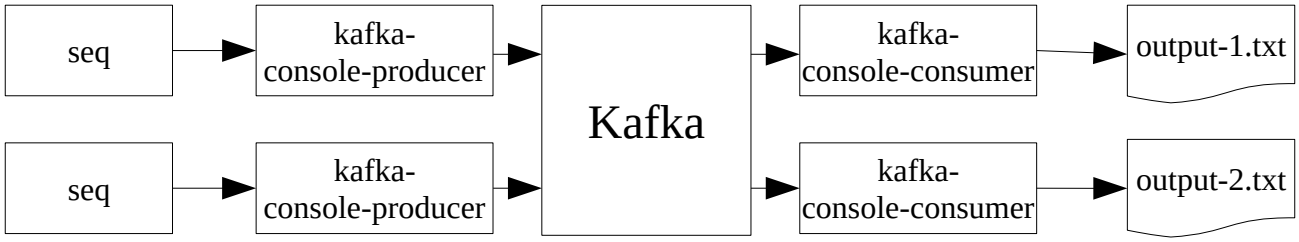
Lassen Sie mit zwei Instanzen des `kafka-console-producer` jeweils fortlaufende Zahlen 1, …, 1000 produzieren. Konsumieren Sie die Daten mit zwei Instanzen des `kafka-console-consumer`, so dass

(a) jede Nachricht an jeden Konsumenten geht (Broadcast)

(b) jede Nachricht an genau einen Konsumenten geht (Round Robin)

Leiten Sie die Ausgaben in Dateien um und zählen Sie nach, dass Broadcast und Round Robin jeweils so funktioniert hat wie gewünscht. Warum sind in der Variante (b) die Daten so ungleich über die beiden Consumer verteilt?

Insgesamt sollte Ihre Topologie so aussehen:



Lassen Sie die vier Producer und Consumer jeweils auf unterschiedlichen Rechnern laufen!