BDA, Praktikumsbericht 4

Gruppe mi6xc: Alexander Kniesz, Maximilian Neudert, Oskar Rudolf

Quellen

Das PySpark Notebook findet man hier.

Aufgabe 1

a)

Zuerst verbinden wir uns auf eine Node, die als producer dienen soll, zum Beispiel `saltshore`:

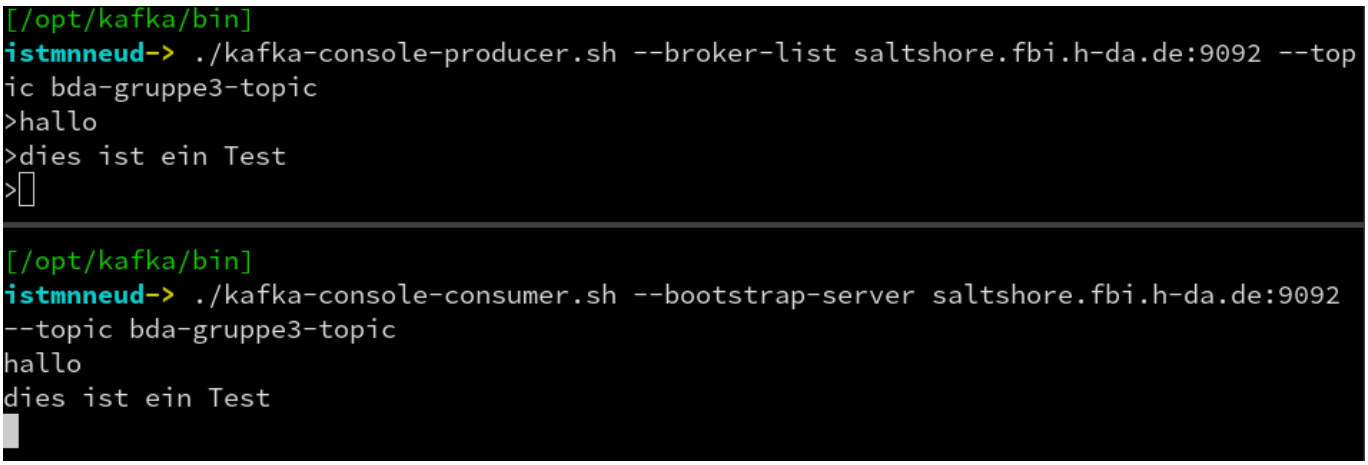
```
mosh istuser@saltshore.fbi.h-da.de
```

anschließend wechseln wir in das Verzeichnis mit den vorbereiteten Scripts und starten den producer:

```
cd /opt/kafka/bin
./kafka-console-producer.sh --broker-list saltshore.fbi.h-da.de:9092 --
topic bda-gruppe3-topic
```

Analog gehen wir vor, verbinden uns auf eine andere Node, wechseln in den Ordner und starten den consumer, der sich zum Producer zunächst ohne `--form-beginning` verbindet:

```
mosh istuser@sunspear.fbi.h-da.de
cd /opt/kafka/bin
./kafka-console-consumer.sh --bootstrap-server saltshore.fbi.h-da.de:9092 -
-topic bda-gruppe3-topic
```



Wenn wir beim Producer nun Nachrichten schreiben, dann werden diese mit kurzer Verzögerung vom Consumer empfangen und dort auf der Console ausgegeben. Fügen wir nun zusätzlich den Parameter `--form-beginning` hinzu, so erhalten wir erwartungsgemäß alle Nachrichten, die bisher auf dem angegebenen Topic gesendet wurden. Da wohl ein paar Spaßvögel mit Scripts das Topic geflutet haben dauert das sogar eine Weile auszugeben.

```
[/opt/kafka/bin]
istmnneud-> ./kafka-console-producer.sh --broker-list saltshore.fbi.h-da.de:9092 --top
ic bda-gruppe3-topic
>hallo
>dies ist ein Test
>neue Nachricht
>
```

```
Turns reality (turn reality)
They already know they can't fuck with Iggy
Switchin' up the game (switch it up, switch it up now)
It's Iggy Iggs!
Is Iggy the ziggy-iggy the baddest of 'em all?
Turns reality (turn reality)
They already know they can't fuck with Iggy
Switchin' up the game (switch it up, switch it up now)
It's Iggy Iggs!
Is Iggy the ziggy-iggy the baddest of 'em all?
Turns reality (turn reality)
They already know they can't fuck with Iggy
Switchin' up the game (switch it up, switch it up now)
It's Iggy Iggs!
Is Iggy the ziggy-iggy the baddest of 'em all?
Turns reality (turn reality)
They already know they can't fuck with Iggy
Switchin' up the game (switch it up, switch it up now)
dies ist ein Test
neue Nachricht
```

b)

Schauen wir uns die Spalten der Kafka Ausgabe per SQL an

so sehen wir, dass wir neben `key` und `value` auch eine Reihe an Metadaten wie `timestamp`, `topic` und `partition` erhalten.

Starten wir die WordCount Query

```
%pyspark
#aufgabe 1b): Deklaration des Kafka-Consumer-Streams und Start der query

from pyspark.sql.functions import explode
from pyspark.sql.functions import split

# read text file
lines = spark \
.readStream \
.format("kafka") \
.option("kafka.bootstrap.servers", "141.100.62.88:9092") \
.option("subscribe", "bda-gruppe3-topic") \
.option("startingOffsets", "earliest") \
.option("kafkaConsumer.pollTimeoutMs", "8192") \
.load()

# cast value object to string
lines = lines\
    .withColumn("value", lines.value.cast('string'))\
    .select('value')

# split lines into words
words = lines\
    .select(explode(split(lines.value, " "))\
    .alias("word"))

# count words
count_words  = words\
    .groupBy("word").count()\
    .orderBy("count", ascending=False)

# Start running the query that prints the running counts to memory sink
writer = count_words \
    .writeStream \
    .queryName("mi6xc_kafkawords") \
    .outputMode("complete") \
    .format("memory")

query = writer.start()
```

so erhalten wir folgendes Ergebnis:

```
%sql
select * from mi6xc_kafkawords
```

settings

word	count
¯_(ツ)_/¯	2710818
Robin	2701986
weiß	491688
was	459938
du	428187
RobinRobin	414451
gestern	396433
getan	364681