Big Data Kafka Spark Streaming

FINISHED

```
%sh
pip install kafka-python
```

```
Requirement already satisfied (use --upgrade to upgrade): kafka-python in /opt/conda/lib/python2.7/sit
e-packages
You are using pip version 8.1.2, however version 22.1.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

Took 2 sec. Last updated by anonymous at May 26 2022, 4:03:39 PM.

FINISHED

```
%sh
mkdir /zeppelin/dep
cd /zeppelin/dep && wget https://repo1.maven.org/maven2/org/apache/spark/spark-streaming-kafka-0-8_2.1
```

```
mkdir: cannot create directory ‘/zeppelin/dep’: File exists
--2022-05-26 07:28:39--  https://repo1.maven.org/maven2/org/apache/spark/spark-streaming-kafka-0-8_
2.11/2.0.2/spark-streaming-kafka-0-8_2.11-2.0.2.jar
Resolving repo1.maven.org (repo1.maven.org)... 199.232.192.209, 199.232.196.209
Connecting to repo1.maven.org (repo1.maven.org)|199.232.192.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 294131 (287K) [application/java-archive]
Saving to: ‘spark-streaming-kafka-0-8_2.11-2.0.2.jar.6’

     0K .......... .......... .......... .......... .......... 17%  409K 1s
    50K .......... .......... .......... .......... .......... 34% 1.04M 0s
   100K .......... .......... .......... .......... .......... 52%  703K 0s
   150K .......... .......... .......... .......... .......... 69% 1.29M 0s
   200K .......... .......... .......... .......... .......... 87% 1.55M 0s
   250K .......... .......... .......... .......             100%  891K=0.4s

2022-05-26 07:28:40 (817 KB/s) - ‘spark-streaming-kafka-0-8_2.11-2.0.2.jar.6’ saved [294131/294131]
```

Took 1 sec. Last updated by anonymous at May 26 2022, 12:28:40 PM.

FINISHED

```
%sh
cd /zeppelin/dep && wget https://repo1.maven.org/maven2/org/apache/spark/spark-streaming-kafka-0-8-ass
```

```
  2800K .......... .......... .......... .......... .......... 25% 3.42M 3s
  2850K .......... .......... .......... .......... .......... 26% 1.66M 3s
  2900K .......... .......... .......... .......... .......... 26% 3.00M 3s
  2950K .......... .......... .......... .......... .......... 27% 2.41M 3s
  3000K .......... .......... .......... .......... .......... 27% 3.40M 3s
  3050K .......... .......... .......... .......... .......... 28% 3.18M 3s
  3100K .......... .......... .......... .......... .......... 28% 3.65M 3s
  3150K .......... .......... .......... .......... .......... 29% 3.42M 3s
  3200K .......... .......... .......... .......... .......... 29% 1.67M 3s
  3250K .......... .......... .......... .......... .......... 29% 3.29M 3s
  3300K .......... .......... .......... .......... .......... 30% 3.05M 3s
  3350K .......... .......... .......... .......... .......... 30% 2.41M 3s
  3400K .......... .......... .......... .......... .......... 31% 3.44M 3s
  3450K .......... .......... .......... .......... .......... 31% 3.18M 3s
  3500K .......... .......... .......... .......... .......... 32% 3.16M 3s
  3550K .......... .......... .......... .......... .......... 32% 1.95M 3s
  3600K .......... .......... .......... .......... .......... 33% 3.65M 3s
```

Took 4 sec. Last updated by anonymous at May 26 2022, 12:28:47 PM.

ERROR

```
%consumer.dep
z.reset()
z.load("/zeppelin/dep/spark-streaming-kafka-0-8-assembly_2.11-2.0.2.jar")
```

```
Must be used before SparkInterpreter (%spark) initialized
Hint: put this paragraph before any Spark code and restart Zeppelin/Interpreter
```

Took 1 sec. Last updated by anonymous at May 26 2022, 12:28:49 PM.

ERROR

```
%consumer.dep
z.reset()
z.load("/zeppelin/dep/spark-streaming-kafka-0-8_2.11-2.0.2.jar")
```

```
Must be used before SparkInterpreter (%spark) initialized
Hint: put this paragraph before any Spark code and restart Zeppelin/Interpreter
```

Took 0 sec. Last updated by anonymous at May 26 2022, 12:28:52 PM.

SPARK JOBS FINISHED

```
%producer.pyspark

df = (spark.read.format("com.databricks.spark.csv")
        .option("header", "true")
        .option("inferSchema","true")
        .load("/datadrive/census_1000.csv"))

df_list = df.collect()
df.show()
```

```
+---+---+-----------------+------------+------------+--------------------+-----------------+-----
---------+-------------------+-------+------------+------------+--------------+------+
|_c0|age|        workclass|   education|education-num|      marital-status|       occupation|  rel
ationship|          ethnicity| gender|capital-gain|capital-loss|hours-per-week|  loan|
+---+---+-----------------+------------+------------+--------------------+-----------------+-----
---------+-------------------+-------+------------+------------+--------------+------+
|  0| 39|        State-gov|   Bachelors|           13|       Never-married|      Adm-clerical| Not-
in-family|              White|   Male|        2174|           0|            40| <=50K|
|  1| 50| Self-emp-not-inc|   Bachelors|           13|  Married-civ-spouse|   Exec-managerial|
Husband|              White|   Male|           0|           0|            13| <=50K|
|  2| 38|          Private|     HS-grad|            9|            Divorced| Handlers-cleaners| Not-
in-family|              White|   Male|           0|           0|            40| <=50K|
|  3| 53|          Private|        11th|            7|  Married-civ-spouse| Handlers-cleaners|
Husband|              Black|   Male|           0|           0|            40| <=50K|
|  4| 28|          Private|   Bachelors|           13|  Married-civ-spouse|    Prof-specialty|
Wife|              Black| Female|           0|           0|            40| <=50K|
|  5| 37|          Private|     Masters|           14|  Married-civ-spouse|   Exec-managerial|
```

Took 2 sec. Last updated by anonymous at May 26 2022, 4:03:44 PM.

FINISHED

```
%sh
mkdir /zeppelin/dep
cd /zeppelin/dep && wget https://repo1.maven.org/maven2/org/apache/spark/spark-streaming-kafka-0-8-ass
cd /zeppelin/dep && wget https://repo1.maven.org/maven2/org/apache/spark/spark-streaming-kafka-0-8_2.1
cp /zeppelin/dep/spark-streaming-kafka-0-8_2.11-2.0.2.jar /zeppelin/interpreter/spark/dep
cp /zeppelin/dep/spark-streaming-kafka-0-8-assembly_2.11-2.0.2.jar /zeppelin/interpreter/spark/dep
```

```
  2100K .......... .......... .......... .......... .......... 19% 4.89M 4s
  2150K .......... .......... .......... .......... .......... 19% 4.20M 4s
  2200K .......... .......... .......... .......... .......... 20% 5.01M 4s
  2250K .......... .......... .......... .......... .......... 20% 4.85M 4s
  2300K .......... .......... .......... .......... .......... 21% 4.89M 3s
  2350K .......... .......... .......... .......... .......... 21% 3.84M 3s
  2400K .......... .......... .......... .......... .......... 22% 4.47M 3s
  2450K .......... .......... .......... .......... .......... 22% 4.95M 3s
  2500K .......... .......... .......... .......... .......... 23% 4.80M 3s
  2550K .......... .......... .......... .......... .......... 23% 3.48M 3s
  2600K .......... .......... .......... .......... .......... 24% 7.85M 3s
  2650K .......... .......... .......... .......... .......... 24% 4.00M 3s
  2700K .......... .......... .......... .......... .......... 24% 4.36M 3s
  2750K .......... .......... .......... .......... .......... 25% 4.37M 3s
  2800K .......... .......... .......... .......... .......... 25% 4.79M 3s
  2850K .......... .......... .......... .......... .......... 26% 2.50M 3s
  2900K .......... .......... .......... .......... .......... 26% 87.2M 3s
  2950K .......... .......... .......... .......... .......... 27% 4.63M 3s
```

Took 5 sec. Last updated by anonymous at May 26 2022, 12:29:06 PM.

ERROR

```
%producer.dep
z.reset()
z.load("/zeppelin/dep/spark-streaming-kafka-0-8_2.11-2.0.2.jar")

%consumer.dep
z.reset()
z.load("/zeppelin/dep/spark-streaming-kafka-0-8-assembly_2.11-2.0.2.jar")
```

```
Must be used before SparkInterpreter (%spark) initialized
Hint: put this paragraph before any Spark code and restart Zeppelin/Interpreter
```

Took 2 sec. Last updated by anonymous at May 26 2022, 12:29:37 PM.

ERROR

```
%producer.dep
z.reset()
z.load("/zeppelin/dep/spark-streaming-kafka-0-8_2.11-2.0.2.jar")
```

```
Must be used before SparkInterpreter (%spark) initialized
Hint: put this paragraph before any Spark code and restart Zeppelin/Interpreter
```

Took 1 sec. Last updated by anonymous at May 26 2022, 12:29:37 PM.

RUNNING 0%

```
%producer.pyspark
import time
import json
import random
import logging

from kafka import KafkaProducer
from kafka.errors import KafkaError

KAFKA_BROKER = "172.25.0.12:9092"
KAFKA_TOPIC = "default_topic"

producer = KafkaProducer(bootstrap_servers=[KAFKA_BROKER])
index = 0

while True:

    row_dict = df_list[index].asDict()

    future = producer.send(
        topic=KAFKA_TOPIC,
        key=str(row_dict["_c0"]).encode("utf-8"),
        value=json.dumps(row_dict).encode("utf-8"))

    try:
        record_metadata = future.get(timeout=10)
    except KafkaError:
        # Decide what to do if produce request failed...
        logging.exception("Error")
        pass

    producer.flush()

    index += 1
    time.sleep(random.uniform(0.1,3.0))
```

Started 23 minutes ago.

RUNNING 0%

```
%consumer.pyspark
import json
from pyspark.streaming.kafka import KafkaUtils
from pyspark.streaming import StreamingContext

import os
os.environ['PYSPARK_SUBMIT_ARGS'] = '--packages org.apache.spark:spark-streaming-kafka-0-8_2.11:2.0.2


try:
    # Reset streaming context if exists
    ssc.stop(stopSparkContext=False, stopGraceFully=False)
except:
    pass

ssc = StreamingContext(sc, batchDuration=2)

REDDIT_TOPIC = "default_topic"
KAFKA_BROKERS = "172.25.0.12:9092,172.25.0.13:9092"

stream = KafkaUtils.createDirectStream(
                            ssc,
                            [REDDIT_TOPIC],
                            {"metadata.broker.list": KAFKA_BROKERS})

stream = stream.map(lambda x: json.loads(x[1]))
stream = stream.map(lambda x: (x["_c0"], x["loan"]))
```

```
stream.pprint()

#ssc.start()
ssc.awaitTermination()
```

```
-------------------------------------------
Time: 2022-05-26 11:24:46
-------------------------------------------
(829, u' <=50K')

Time: 2022-05-26 11:24:48
-------------------------------------------
(830, u' <=50K')


-------------------------------------------
Time: 2022-05-26 11:24:50
-------------------------------------------
(831, u' >50K')


-------------------------------------------
Time: 2022-05-26 11:24:52
-------------------------------------------
(832, u' <=50K')
```

Started 23 minutes ago.

ABORT

```
%consumer.pyspark
import sys

from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from pyspark.streaming.kafka import KafkaUtils


ssc = StreamingContext(sc,5)

print "Connected to spark streaming"

def process(time, rdd):
    print("========= %s =========" % str(time))
    if not rdd.isEmpty():
        rdd.count()
        rdd.first()

ssc = StreamingContext(sc, 5)
kafkaStream = KafkaUtils.createStream(ssc, "server:2181", "pysparkclient1", {"smartPlug": 1})
kafkaStream.pprint()
kafkaStream.foreachRDD(process)

#ssc.start()
ssc.awaitTermination()
```

```
Connected to spark streaming
-------------------------------------------
Time: 2022-05-26 11:03:54
-------------------------------------------
(1, u' <=50K')


-------------------------------------------
Time: 2022-05-26 11:03:56
-------------------------------------------
(2, u' <=50K')
```

Big Data Kafka Spark Streaming

```
(3, u' <=50K')

-------------------------------------------
Time: 2022-05-26 11:03:58
-------------------------------------------

-------------------------------------------
```

Took 17 min 37 sec. Last updated by anonymous at May 26 2022, 4:04:03 PM.

```
%consumer.pyspark
```

READY