schemagenerator

August 6, 2024

[1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.avro.functions import from_avro, to_avro
import pandas as pd
import json
```

[2]:
```
# spark.sparkContext.stop()

#Spark Session creation configured to interact with Kafka
spark = SparkSession.builder.appName("pyspark-notebook").\
config("spark.jars.packages","org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.
 ↪0,org.apache.spark:spark-avro_2.12:3.0.0,org.mongodb.spark:
 ↪mongo-spark-connector_2.12:3.0.0").\
getOrCreate()
```

```
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
:: loading settings :: url = jar:file:/usr/local/lib/python3.7/dist-packages/pys
park/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
org.apache.spark#spark-avro_2.12 added as a dependency
org.mongodb.spark#mongo-spark-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-
parent-16f14b5a-8441-41ad-adcd-5cbe646c7eac;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.0 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.0 in
central
        found org.apache.kafka#kafka-clients;2.4.1 in central
        found com.github.luben#zstd-jni;1.4.4-3 in central
        found org.lz4#lz4-java;1.7.1 in central
        found org.xerial.snappy#snappy-java;1.1.7.5 in central
        found org.slf4j#slf4j-api;1.7.30 in central
        found org.spark-project.spark#unused;1.0.0 in central
        found org.apache.commons#commons-pool2;2.6.2 in central
        found org.apache.spark#spark-avro_2.12;3.0.0 in central
        found org.mongodb.spark#mongo-spark-connector_2.12;3.0.0 in central
        found org.mongodb#mongodb-driver-sync;4.0.5 in central
        found org.mongodb#bson;4.0.5 in central
```

```
        found org.mongodb#mongodb-driver-core;4.0.5 in central
:: resolution report :: resolve 514ms :: artifacts dl 9ms
        :: modules in use:
        com.github.luben#zstd-jni;1.4.4-3 from central in [default]
        org.apache.commons#commons-pool2;2.6.2 from central in [default]
        org.apache.kafka#kafka-clients;2.4.1 from central in [default]
        org.apache.spark#spark-avro_2.12;3.0.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.0 from central in
[default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.0 from central
in [default]
        org.lz4#lz4-java;1.7.1 from central in [default]
        org.mongodb#bson;4.0.5 from central in [default]
        org.mongodb#mongodb-driver-core;4.0.5 from central in [default]
        org.mongodb#mongodb-driver-sync;4.0.5 from central in [default]
        org.mongodb.spark#mongo-spark-connector_2.12;3.0.0 from central in
[default]
        org.slf4j#slf4j-api;1.7.30 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.7.5 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   14  |   0   |   0   |   0   ||   14  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-16f14b5a-8441-41ad-
adcd-5cbe646c7eac
        confs: [default]
        0 artifacts copied, 14 already retrieved (0kB/11ms)
24/07/22 06:02:24 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
24/07/22 06:02:25 WARN Utils: Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.
```

[3]:
```
print(spark.version)
```

```
3.0.0
```

[4]:
```
# spark = SparkSession.builder.getOrCreate()
# spark.sparkContext.stop()
```

[8]:
```
#Read data from Kafka
data = spark\
  .readStream\
  .format("kafka")\
  .option("kafka.bootstrap.servers", "ec2-3-107-14-79.ap-southeast-2.compute.
 ↪amazonaws.com:9092")\
  .option("subscribe", "tweets")\
  .option("startingOffsets", "earliest")\
  .load()\
  .selectExpr("CAST(key AS STRING)", "CAST(value AS STRING)")\
  .select("value")
```

[9]:
```
#write streaming data as a text file
data.\
writeStream.\
format("text").\
option("checkpointLocation", "checkpoint/schema").\
option("format", "text").\
option("path", "schema/in").\
outputMode("append").\
start()
```

[9]:
```
<pyspark.sql.streaming.StreamingQuery at 0x72699067f278>
```

[10]:
```
#extract schema by reading the file written above
smallBatchSchema = spark.read.json("schema/in/*.txt").schema
```

```
24/07/22 06:03:15 WARN package: Truncated the string representation of a plan
since it was too large. This behavior can be adjusted by setting
'spark.sql.debug.maxToStringFields'.
```

[11]:
```
import os
import json

# Đường dẫn tới thư mục và file
dir_path = "schema/out"
file_path = os.path.join(dir_path, "tweet_schema.json")

# Tạo thư mục nếu chưa tồn tại
if not os.path.exists(dir_path):
    os.makedirs(dir_path)

# Ghi schema vào file JSON
with open(file_path, "w") as f:
    json.dump(smallBatchSchema.jsonValue(), f)

print(f"Schema đã được ghi vào {file_path}")
```

```
Schema đã được ghi vào schema/out/tweet_schema.json
```

[ ]: