Working with Kafka and Spark streaming (in Windows)

Kafka installation steps

In this method, kafka is installed using a docker image.

1. Open new cmd.
   1. Go to a folder where you want to keep Kafka.

$ cd <folder\_path>

* 1. $ git clone <https://github.com/darshilparmar/kafka-in-10min-video-code.git>

This will install necessary dependencies from github page into the selected folder.

You can see the Kafka (2.7.0) – Scala (2.13) and zookeeper versions in docker-compose.yaml file.

1. Open the kafka folder in vscode.
2. Start docker server.
3. Start kafka project using vscode terminal.

$ docker-compose up

This will start the kafka broker and zookeeper servers.

\* Code already contains ipynb files for producer and consumer. So we can start our streaming right away.

1. Create a conda venv and select the new venv in the vscode to run the program in suitable kernel.

To test the kafka streaming:

1. To run kafka producer:
   1. Open new vscode terminal.
      1. $ docker exec -it <kafka\_docker\_id> /bin/bash

(to get docker\_id : $ docker ps)

* + 1. $ cd opt/kafka/bin
    2. Create a kafka topic

$ ./kafka-topics.sh --create --topic <myFirstTopic> --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1

* + 1. Create the producer:

$ ./kafka-console-producer.sh --topic <myFirstTopic> --bootstrap-server localhost:9092

1. To run kafka consumer:
   1. Open new vscode terminal.
      1. $ docker exec -it <kafka\_docker\_id> /bin/bash

(to get docker\_id : $ docker ps)

* + 1. $ cd opt/kafka/bin
    2. Create the consumer:

$ ./kafka-console-consumer.sh --topic <myFirstTopic> --bootstrap-server localhost:9092

1. Start typing in producer terminal.

Now, you should be able to see the content written in producer terminal in the consumer terminal also.

To connect Kafka with Pyspark

1. Install Pyspark (Here, installed pyspark version = 3.4.3).
2. Download required jar files from Maven repo. Make sure the versions are correct and compatible.

Kafka version=2.7

Commons-pool2 version=2.6.2

* 1. Spark sql kafka integration jar:

<https://mvnrepository.com/artifact/org.apache.spark/spark-sql-kafka-0-10_2.13/3.4.3>

* + 1. Click on the [jar] button in ‘Files’ menu to download.
  1. Kafka client jar:

<https://mvnrepository.com/artifact/org.apache.kafka/kafka-clients/2.7.0>

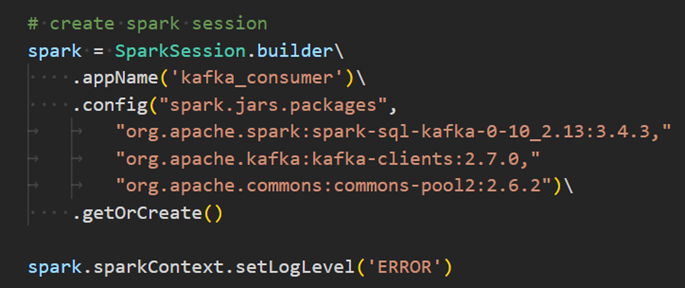
* 1. Commons-pool2 jar:

<https://mvnrepository.com/artifact/org.apache.commons/commons-pool2/2.6.2>

1. Copy the jar files from downloaded location to jar folder of the spark folder.

C:\apps\spark-3.4.3-bin-hadoop3-scala2.13\jars (<Pyspark\_path>\jar).

1. To start the connection, in a new ipynb file:
   1. Create Spark Session
      1. Add kafka package to spark session config.



* 1. Connect spark streaming with Kafka topic to read data streams



1. Extract Topic information and apply suitable schema
2. Analyse the data using structured streaming SQL queries
3. Store the results to a required sink.