Week - 8: Assignment Handout:

Stream the data stored on the GCS bucket into Kafka by breaking the data into batches of 10 records that are written to Kafka separated by a sleep time of 10 seconds until 100 records are written. Use Spark Streaming to read from Kafka every 5 seconds and emit the count of rows seen in the last 10 seconds.

Links:

- https://spark.apache.org/docs/latest/structured-streaming-kafka-integration.html
- https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html
- https://github.com/apache/spark/blob/v3.3.2/examples/src/main/python/sql/streaming/structured_network_wordcount.py
- https://kafka.apache.org/quickstart
- https://cloud.google.com/pubsub/docs/migrating-from-kafka-to-pubsub
- https://tmuxcheatsheet.com/

Installing Dependencies

```
sudo apt update && sudo apt install -y tmux vim openjdk-8-jdk wget curl git python3-pip # python-is-python3 # (Optional)
pip install kafka-python pyspark pyspark[sql]
wget https://dlcdn.apache.org/kafka/3.4.0/kafka_2.13-3.4.0.tgz # Get Kafka Sources
```

Installing Kafka

New Terminal Session

```
tar -xzvf kafka_2.13-3.4.0.tgz
cd kafka_2.13-3.4.0
bin/zookeeper-server-start.sh config/zookeeper.properties
```

New Terminal Session

```
bin/kafka-server-start.sh config/server.properties
```

Create Events

New Terminal Session

```
bin/kafka-topics.sh --create --topic test-topic --bootstrap-server localhost:9092
bin/kafka-topics.sh --list --bootstrap-server localhost:9092
bin/kafka-topics.sh --describe --topic test-topic --bootstrap-server localhost:9092
```

Checking The created topic

```
bin/kafka-console-producer.sh --topic test-topic --bootstrap-server localhost:9092
bin/kafka-console-consumer.sh --topic test-topic --from-beginning --bootstrap-server localhost:9092
```

Start Producer-Consumer Demo from Python-to-Kafka Local Data

New Terminal Session

```
cd ~/demo
python3 kafka_consumer.py
```

New Terminal Session

```
cd ~/demo
python3 kafka_producer.py
```

Start Producer-Consumer Demo from Python-to-Kafka Local Data

Start Pyspark-Kafka Demo script (JSON)

New Terminal Session

```
cd ~/demo
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.3.2 kafka_streaming_json_demo.py
```

Start Kafka-Producer (JSON)

New Terminal Session

```
cd ~/demo
python3 kafka_producer.py
```

Clean Up

```
rm -rf /tmp/kafka-logs /tmp/zookeeper /tmp/kraft-combined-logs
```