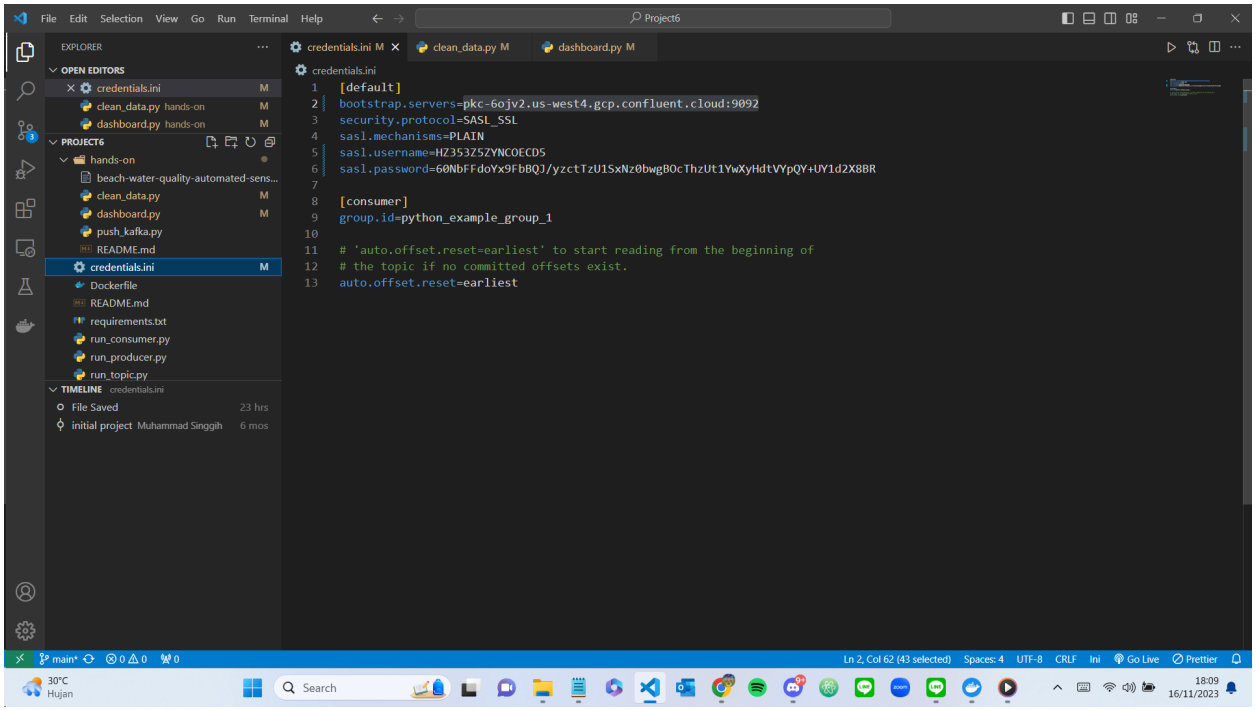
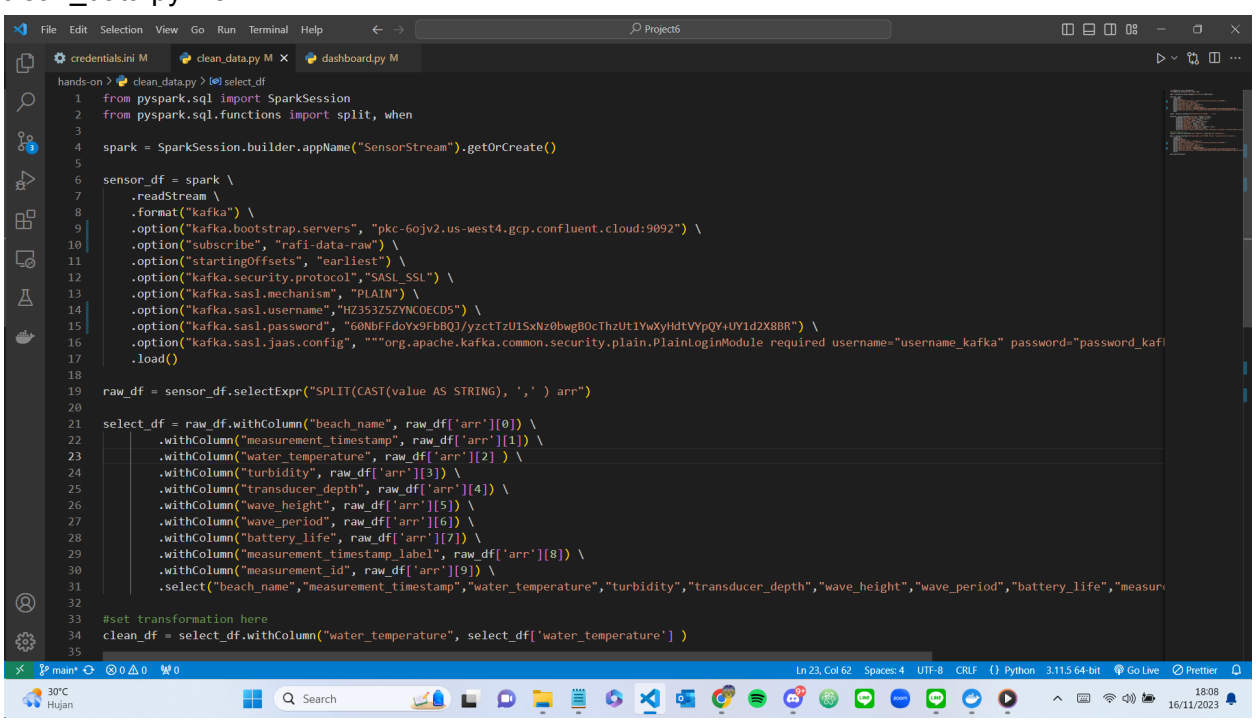
Project Walkthrough Documentation

Credentials.ini file:

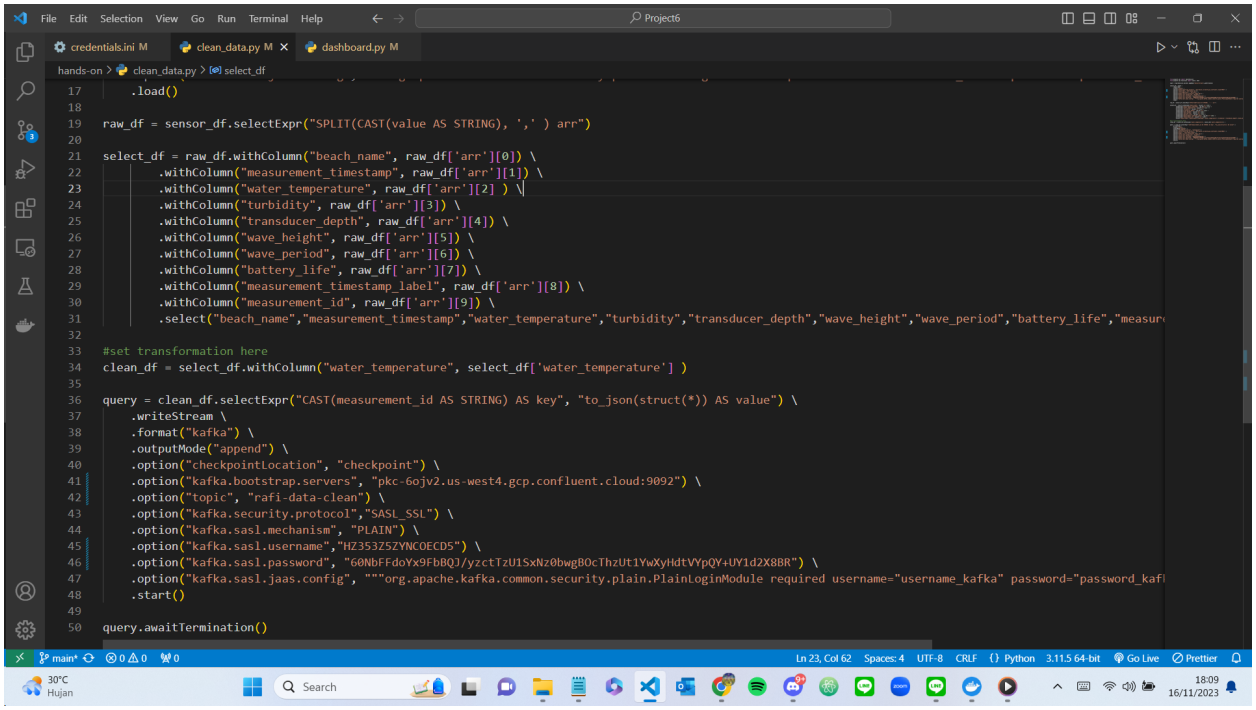


```ini
[default]
bootstrap.servers=pkc-6ojv2.us-west4.gcp.confluent.cloud:9092
security.protocol=SASL_SSL
sasl.mechanisms=PLAIN
sasl.username=HZ353Z5ZYNCOECD5
sasl.password=60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR

[consumer]
group.id=python_example_group_1

# 'auto.offset.reset=earliest' to start reading from the beginning of
# the topic if no committed offsets exist.
auto.offset.reset=earliest
```

clean_data.py file:



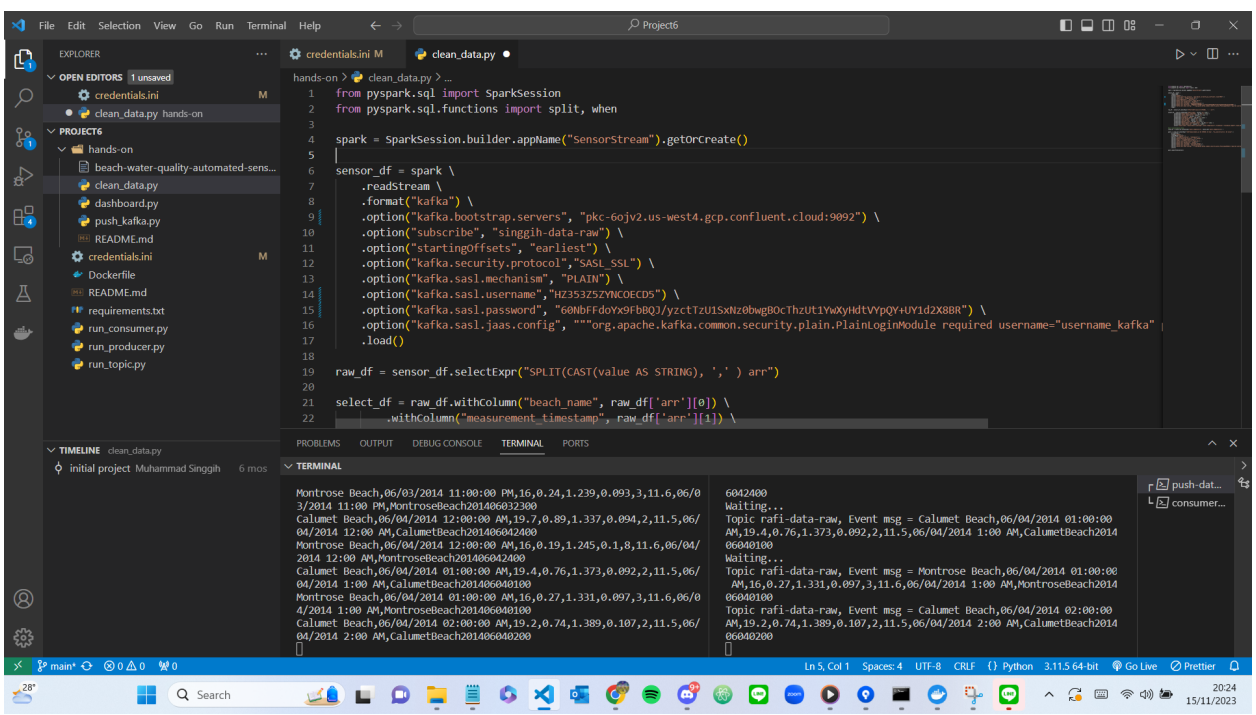
```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import split, when

spark = SparkSession.builder.appName("SensorStream").getOrCreate()

sensor_df = spark \
    .readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.gcp.confluent.cloud:9092") \
    .option("subscribe", "rafi-data-raw") \
    .option("startingOffsets", "earliest") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR") \
    .option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka" password="password_kafl
    .load()

raw_df = sensor_df.selectExpr("SPLIT(CAST(value AS STRING), ',' ) arr")

select_df = raw_df.withColumn("beach_name", raw_df['arr'][0]) \
        .withColumn("measurement_timestamp", raw_df['arr'][1]) \
        .withColumn("water_temperature", raw_df['arr'][2] ) \
        .withColumn("turbidity", raw_df['arr'][3]) \
        .withColumn("transducer_depth", raw_df['arr'][4]) \
        .withColumn("wave_height", raw_df['arr'][5]) \
        .withColumn("wave_period", raw_df['arr'][6]) \
        .withColumn("battery_life", raw_df['arr'][7]) \
        .withColumn("measurement_timestamp_label", raw_df['arr'][8]) \
        .withColumn("measurement_id", raw_df['arr'][9]) \
        .select("beach_name","measurement_timestamp","water_temperature","turbidity","transducer_depth","wave_height","wave_period","battery_life","measur

#set transformation here
clean_df = select_df.withColumn("water_temperature", select_df['water_temperature'] )
```



```python
    .load()

raw_df = sensor_df.selectExpr("SPLIT(CAST(value AS STRING), ',' ) arr")

select_df = raw_df.withColumn("beach_name", raw_df['arr'][0]) \
        .withColumn("measurement_timestamp", raw_df['arr'][1]) \
        .withColumn("water_temperature", raw_df['arr'][2] ) \
        .withColumn("turbidity", raw_df['arr'][3]) \
        .withColumn("transducer_depth", raw_df['arr'][4]) \
        .withColumn("wave_height", raw_df['arr'][5]) \
        .withColumn("wave_period", raw_df['arr'][6]) \
        .withColumn("battery_life", raw_df['arr'][7]) \
        .withColumn("measurement_timestamp_label", raw_df['arr'][8]) \
        .withColumn("measurement_id", raw_df['arr'][9]) \
        .select("beach_name","measurement_timestamp","water_temperature","turbidity","transducer_depth","wave_height","wave_period","battery_life","measur

#set transformation here
clean_df = select_df.withColumn("water_temperature", select_df['water_temperature'] )

query = clean_df.selectExpr("CAST(measurement_id AS STRING) AS key", "to_json(struct(*)) AS value") \
    .writeStream \
    .format("kafka") \
    .outputMode("append") \
    .option("checkpointLocation", "checkpoint") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.gcp.confluent.cloud:9092") \
    .option("topic", "rafi-data-clean") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR") \
    .option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka" password="password_kafi
    .start()

query.awaitTermination()
```

Saya lupa untuk melakukan screenshot terminal pada step sebelumnya, hanya ada screenshot untuk langkah-langkah setelah ini.

Screenshot Terminal:



```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import split, when

spark = SparkSession.builder.appName("SensorStream").getOrCreate()

sensor_df = spark \
    .readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.gcp.confluent.cloud:9092") \
    .option("subscribe", "singgih-data-raw") \
    .option("startingOffsets", "earliest") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR") \
    .option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka" p
    .load()

raw_df = sensor_df.selectExpr("SPLIT(CAST(value AS STRING), ',' ) arr")

select_df = raw_df.withColumn("beach_name", raw_df['arr'][0]) \
        .withColumn("measurement_timestamp", raw_df['arr'][1]) \
```

```
Montrose Beach,06/03/2014 11:00:00 PM,16,0.24,1.239,0.093,3,11.6,06/03/2014 11:00 PM,MontroseBeach201406032300
Calumet Beach,06/04/2014 12:00:00 AM,19.7,0.89,1.337,0.094,2,11.5,06/04/2014 12:00 AM,CalumetBeach201406042400
Montrose Beach,06/04/2014 12:00:00 AM,16,0.19,1.245,0.1,8,11.6,06/04/2014 12:00 AM,MontroseBeach201406042400
Calumet Beach,06/04/2014 01:00:00 AM,19.4,0.76,1.373,0.092,2,11.5,06/04/2014 1:00 AM,CalumetBeach201406040100
Montrose Beach,06/04/2014 01:00:00 AM,16,0.27,1.331,0.097,3,11.6,06/04/2014 1:00 AM,MontroseBeach201406040100
Calumet Beach,06/04/2014 02:00:00 AM,19.2,0.74,1.389,0.107,2,11.5,06/04/2014 2:00 AM,CalumetBeach201406040200
```

```
6042400
Waiting...
Topic rafi-data-raw, Event msg = Calumet Beach,06/04/2014 01:00:00 AM,19.4,0.76,1.373,0.092,2,11.5,06/04/2014 1:00 AM,CalumetBeach201406040100
Waiting...
Topic rafi-data-raw, Event msg = Montrose Beach,06/04/2014 01:00:00 AM,16,0.27,1.331,0.097,3,11.6,06/04/2014 1:00 AM,MontroseBeach201406040100
Topic rafi-data-raw, Event msg = Calumet Beach,06/04/2014 02:00:00 AM,19.2,0.74,1.389,0.107,2,11.5,06/04/2014 2:00 AM,CalumetBeach201406040200
```

```python
        .select("beach_name","measurement_timestamp","water_temperature","turbidity","transducer_depth","wave_height","wave_period",

#set transformation here
clean_df = select_df.withColumn("water_temperature", select_df['water_temperature'] )

query = clean_df.selectExpr("CAST(measurement_id AS STRING) AS key", "to_json(struct(*)) AS value") \
    .writeStream \
    .format("kafka") \
    .outputMode("append") \
    .option("checkpointLocation", "checkpoint") \
    .option("kafka.bootstrap.servers", "url_kafka") \
    .option("topic", "rafi-data-clean") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","username_kafka") \
    .option("kafka.sasl.password", "password_kafka") \
    .option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka"
    .start()

query.awaitTermination()
```

```
PS C:\Users\Rafi Muhammad\Documents\Bootcamp_DataEngineer\Project6> wsl
rafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcrafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_
rafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$
```

```python
    .format("kafka") \
    .outputMode("append") \
    .option("checkpointLocation", "checkpoint") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.gcp.confluent.cloud:9092") \
    .option("topic", "rafi-data-clean") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR") \
    .option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka"
    .start()

query.awaitTermination()
```

```
rafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$ docker exec -it my_container python /project/
run_consumer.py /project/credentials.ini singgih-data-raw
Waiting...
Waiting...
Waiting...
^Crafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$ docker exec -it my_container python /projec
t/run_consumer.py /project/credentials.ini rafi-data-raw
Waiting...
Waiting...
Waiting...
Waiting...
^Crafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$ docker exec -it my_container spark-submit -
-packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.3 /project/hands-on/clean_data.py
```

```
.format("kafka") \
.outputMode("append") \
.option("checkpointLocation", "checkpoint") \
.option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.gcp.confluent.cloud:9092") \
.option("topic", "rafi-data-clean") \
.option("kafka.security.protocol","SASL_SSL") \
.option("kafka.sasl.mechanism", "PLAIN") \
.option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
.option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR") \
.option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka"
```

```
t/run_consumer.py /project/credentials.ini rafi-data-raw
Waiting...
Waiting...
Waiting...
Waiting...
^Crafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$ docker exec -it my_container spark-submit -
-packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.3 /project/hands-on/clean_data.py
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/usr/local/lib/python3.8/site-packages/pyspark/jars/spark-unsaf
e_2.12-3.0.3.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
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/python3.8/site-packages/pyspark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivyset
tings.xml
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-33403308-b1a9-4ef2-adc0-fc1b33236124;1.0
        confs: [default]
```

```
^Crafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$ docker exec -it my_container spark-submit -
-packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.3 /project/hands-on/clean_data.py
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/usr/local/lib/python3.8/site-packages/pyspark/jars/spark-unsaf
e_2.12-3.0.3.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
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/python3.8/site-packages/pyspark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivyset
tings.xml
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-33403308-b1a9-4ef2-adc0-fc1b33236124;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.3 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;1.1.8.2 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
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.12/3.0.3/spark-sql-kafka-0-10_2.12-3.0.3.jar ...
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-token-provider-kafka-0-10_2.12/3.0.3/spark-token-provider-kafka-0-10_2.12-
3.0.3.jar ...
        [SUCCESSFUL ] org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.3!spark-token-provider-kafka-0-10_2.12.jar (1233ms)
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/2.4.1/kafka-clients-2.4.1.jar ...
        [SUCCESSFUL ] org.apache.kafka#kafka-clients;2.4.1!kafka-clients.jar (8312ms)
downloading https://repo1.maven.org/maven2/org/apache/commons/commons-pool2/2.6.2/commons-pool2-2.6.2.jar ...
        [SUCCESSFUL ] org.apache.commons#commons-pool2;2.6.2!commons-pool2.jar (486ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.0.0.jar ...
        [SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (179ms)
downloading https://repo1.maven.org/maven2/com/github/luben/zstd-jni/1.4.4-3/zstd-jni-1.4.4-3.jar ...
        [SUCCESSFUL ] com.github.luben#zstd-jni;1.4.4-3!zstd-jni.jar (10692ms)
downloading https://repo1.maven.org/maven2/org/lz4/lz4-java/1.7.1/lz4-java-1.7.1.jar ...
        [SUCCESSFUL ] org.lz4#lz4-java;1.7.1!lz4-java.jar (1756ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.8.2/snappy-java-1.1.8.2.jar ...
        [SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.8.2!snappy-java.jar(bundle) (5043ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.30/slf4j-api-1.7.30.jar ...
        [SUCCESSFUL ] org.slf4j#slf4j-api;1.7.30!slf4j-api.jar (197ms)
:: resolution report :: resolve 274734ms :: artifacts dl 805262ms
```

```
        found org.apache.commons#commons-pool2;2.6.2 in central
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.12/3.0.3/spark-sql-kafka-0-10_2.12-3.0.3.jar ...
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-token-provider-kafka-0-10_2.12/3.0.3/spark-token-provider-kafka-0-10_2.12-
3.0.3.jar ...
        [SUCCESSFUL ] org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.3!spark-token-provider-kafka-0-10_2.12.jar (1233ms)
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/2.4.1/kafka-clients-2.4.1.jar ...
        [SUCCESSFUL ] org.apache.kafka#kafka-clients;2.4.1!kafka-clients.jar (8312ms)
downloading https://repo1.maven.org/maven2/org/apache/commons/commons-pool2/2.6.2/commons-pool2-2.6.2.jar ...
        [SUCCESSFUL ] org.apache.commons#commons-pool2;2.6.2!commons-pool2.jar (486ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.0.0.jar ...
        [SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (179ms)
downloading https://repo1.maven.org/maven2/com/github/luben/zstd-jni/1.4.4-3/zstd-jni-1.4.4-3.jar ...
        [SUCCESSFUL ] com.github.luben#zstd-jni;1.4.4-3!zstd-jni.jar (10692ms)
downloading https://repo1.maven.org/maven2/org/lz4/lz4-java/1.7.1/lz4-java-1.7.1.jar ...
        [SUCCESSFUL ] org.lz4#lz4-java;1.7.1!lz4-java.jar (1756ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.8.2/snappy-java-1.1.8.2.jar ...
        [SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.8.2!snappy-java.jar(bundle) (5043ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.30/slf4j-api-1.7.30.jar ...
        [SUCCESSFUL ] org.slf4j#slf4j-api;1.7.30!slf4j-api.jar (197ms)
:: resolution report :: resolve 274734ms :: artifacts dl 805262ms
        :: 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-sql-kafka-0-10_2.12;3.0.3 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.3 from central in [default]
        org.lz4#lz4-java;1.7.1 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;1.1.8.2 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   9   |   9   |   9   |   0   ||   9   |   8   |
        ---------------------------------------------------------------------

:: problems summary ::
:::: WARNINGS
        [FAILED     ] org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3!spark-sql-kafka-0-10_2.12.jar: Downloaded file size doesn't m
atch expected Content Length for https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.12/3.0.3/spark-sql-kafka-0-10_2.12-3
.0.3.jar. Please retry. (776764ms)
```

```
        [FAILED     ] org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3!spark-sql-kafka-0-10_2.12.jar: Downloaded file size doesn't m
atch expected Content Length for https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.12/3.0.3/spark-sql-kafka-0-10_2.12-3
.0.3.jar. Please retry. (776764ms)

        [FAILED     ] org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3!spark-sql-kafka-0-10_2.12.jar: Downloaded file size doesn't m
atch expected Content Length for https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.12/3.0.3/spark-sql-kafka-0-10_2.12-3
.0.3.jar. Please retry. (776764ms)

        ==== central: tried

          https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.12/3.0.3/spark-sql-kafka-0-10_2.12-3.0.3.jar

                ::::::::::::::::::::::::::::::::::::::::::::

                ::              FAILED DOWNLOADS            ::

                :: ^ see resolution messages for details  ^ ::

                ::::::::::::::::::::::::::::::::::::::::::::

                :: org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3!spark-sql-kafka-0-10_2.12.jar

                ::::::::::::::::::::::::::::::::::::::::::::


:: USE VERBOSE OR DEBUG MESSAGE LEVEL FOR MORE DETAILS
Exception in thread "main" java.lang.RuntimeException: [download failed: org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3!spark-sql-kafka-0-
10_2.12.jar]
        at org.apache.spark.deploy.SparkSubmitUtils$.resolveMavenCoordinates(SparkSubmit.scala:1389)
        at org.apache.spark.deploy.DependencyUtils$.resolveMavenDependencies(DependencyUtils.scala:54)
        at org.apache.spark.deploy.SparkSubmit.prepareSubmitEnvironment(SparkSubmit.scala:308)
        at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:871)
        at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:180)
        at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:203)
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:90)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:1007)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:1016)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
rafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$
```

```python
    .format("kafka") \
    .outputMode("append") \
    .option("checkpointLocation", "checkpoint") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.gcp.confluent.cloud:9092") \
    .option("topic", "rafi-data-clean") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzU1SxNz0bwgBOcThzUt1YwXyHdtVYpQY+UY1d2X8BR") \
    .option("kafka.sasl.jaas.config", """org.apache.kafka.common.security.plain.PlainLoginModule required username="username_kafka" |
    .start()

query.awaitTermination()
```

```
rafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Project6$ docker exec -it my_container python /project/
run_consumer.py /project/credentials.ini rafi-data-clean
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
```

```python
    .format("kafka") \
    .outputMode("append") \
    .option("checkpointLocation", "checkpoint") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.g
    .option("topic", "rafi-data-clean") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzL
    .option("kafka.sasl.jaas.config", """org.apache.kafka.cc
    .start()

query.awaitTermination()
```

```
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
Waiting...
```

```
2014 1:00 AM,CalumetBeach201407040100
Montrose Beach,07/03/2014 07:00:00 PM,19.1,1.04,1.783,0.19,4,10.5,07/
03/2014 7:00 PM,MontroseBeach201407031900
Osterman Beach,07/03/2014 07:00:00 PM,18.9,1.06,1.455,0.166,4,11,07/0
3/2014 7:00 PM,OstermanBeach201407031900
Ohio Street Beach,07/03/2014 07:00:00 PM,17.7,1.14,1.857,0.169,4,10.6
,07/03/2014 7:00 PM,OhioStreetBeach201407031900
Calumet Beach,07/03/2014 08:00:00 PM,20.8,2.06,1.491,0.111,4,10.3,07/
03/2014 8:00 PM,CalumetBeach201407032000
63rd Street Beach,07/03/2014 08:00:00 PM,18.9,1.95,1.795,0.196,4,10.8
,07/03/2014 8:00 PM,63rdStreetBeach201407032000
```

```
Topic rafi-data-raw, Event msg = Osterman Beach,07/03/2014 07:00:00
 PM,18.9,1.06,1.455,0.166,4,11,07/03/2014 7:00 PM,OstermanBeach2014
07031900
Waiting...
Topic rafi-data-raw, Event msg = Ohio Street Beach,07/03/2014 07:00
:00 PM,17.7,1.14,1.857,0.169,4,10.6,07/03/2014 7:00 PM,OhioStreetBe
ach201407031900
Waiting...
Topic rafi-data-raw, Event msg = Calumet Beach,07/03/2014 08:00:00
PM,20.8,2.06,1.491,0.111,4,10.3,07/03/2014 8:00 PM,CalumetBeach2014
07032000
```

credentials.ini M | clean_data.py M

hands-on > clean_data.py > query

```python
    .writeStream \
    .format("kafka") \
    .outputMode("append") \
    .option("checkpointLocation", "checkpoint") \
    .option("kafka.bootstrap.servers", "pkc-6ojv2.us-west4.g
    .option("topic", "rafi-data-clean") \
    .option("kafka.security.protocol","SASL_SSL") \
    .option("kafka.sasl.mechanism", "PLAIN") \
    .option("kafka.sasl.username","HZ353Z5ZYNCOECD5") \
    .option("kafka.sasl.password", "60NbFFdoYx9FbBQJ/yzctTzl
    .option("kafka.sasl.jaas.config", """org.apache.kafka.co
    .start()

query.awaitTermination()
```

wsl

```
Waiting...
Waiting...
Waiting...
```

File Edit Selection View Go Run Terminal Help

Project6

credentials.ini M | push_kafka.py | clean_data.py M

hands-on > push_kafka.py > ...

```python
        print("ERROR: Message failed delivery: {}".format(err))
        else:
            print("Topic {}, Event msg = {}".format(msg.topic(), msg.value().decode('utf-8')))

    data = open('/project/hands-on/beach-water-quality-automated-sensors-1.csv','r').read().split("\n")
    for msg in data:
        producer.produce(args.topic, msg, callback=delivery_callback)
        print(msg)
        time.sleep(1)

    # Block until the messages are sent.
    producer.poll(10000)
    producer.flush()
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

TERMINAL

```
entials.ini rafi-data-raw
Beach Name,Measurement Timestamp,Water Temperature,Turbidity,Transducer Depth,Wave Heigh
t,Wave Period,Battery Life,Measurement Timestamp Label,Measurement ID
Montrose Beach,08/30/2013 08:00:00 AM,20.3,1.18,0.891,0.08,3,9.4,8/30/2013 8:00 AM,Montr
oseBeach201308300800
Ohio Street Beach,05/26/2016 01:00:00 PM,14.4,1.23,,0.111,4,12.4,05/26/2016 1:00 PM,Ohio
StreetBeach201605261300
Calumet Beach,09/03/2013 04:00:00 PM,23.2,3.63,1.201,0.174,6,9.4,9/3/2013 4:00 PM,Calume
tBeach201309031600
Calumet Beach,05/28/2014 12:00:00 PM,16.2,1.26,1.514,0.147,4,11.7,5/28/2014 12:00 PM,Cal
umetBeach201405281200
```

```
Topic rafi-data-raw, Event msg = Beach Name,Measurement Timestamp,Water Temperature,Tur
bidity,Transducer Depth,Wave Height,Wave Period,Battery Life,Measurement Timestamp Labe
l,Measurement ID
Topic rafi-data-raw, Event msg = Montrose Beach,08/30/2013 08:00:00 AM,20.3,1.18,0.891,
0.08,3,9.4,8/30/2013 8:00 AM,MontroseBeach201308300800
Topic rafi-data-raw, Event msg = Ohio Street Beach,05/26/2016 01:00:00 PM,14.4,1.23,,0.
111,4,12.4,05/26/2016 1:00 PM,OhioStreetBeach201605261300
Topic rafi-data-raw, Event msg = Calumet Beach,05/28/2014 12:00:00 PM,16.2,1.26,1.514,0
.147,4,11.7,5/28/2014 12:00 PM,CalumetBeach201405281200
Topic rafi-data-raw, Event msg = Calumet Beach,09/03/2013 04:00:00 PM,23.2,3.63,1.201,0
.174,6,9.4,9/3/2013 4:00 PM,CalumetBeach201309031600
```

wsl
wsl

main* 0 0 0 Ln 1, Col 1 Spaces: 4 UTF-8 CRLF {} Python 3.11.5 64-bit Go Live Prettier

31°C Cerah Search 16:44 16/11/2023

File Edit Selection View Go Run Terminal Help

Project6

credentials.ini M | push_kafka.py | clean_data.py M

hands-on > push_kafka.py > ...

```python
        print("ERROR: Message failed delivery: {}".format(err))
        else:
            print("Topic {}, Event msg = {}".format(msg.topic(), msg.value().decode('utf-8')))

    data = open('/project/hands-on/beach-water-quality-automated-sensors-1.csv','r').read().split("\n")
    for msg in data:
        producer.produce(args.topic, msg, callback=delivery_callback)
        print(msg)
        time.sleep(1)

    # Block until the messages are sent.
    producer.poll(10000)
    producer.flush()
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

TERMINAL

```
Montrose Beach,05/28/2014 09:00:00 PM,14.5,7.11,1.374,0.361,5,11.9,5/28/2014 9:00 PM,Mon
troseBeach201405282100
^CTraceback (most recent call last):
  File "/project/hands-on/push_kafka.py", line 36, in <module>
    time.sleep(1)
KeyboardInterrupt
rafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/Pro
ject6$ docker exec -it my_container python /project/hands-on/push_kafka.py /project/cred
entials.ini rafi-data-clean
Beach Name,Measurement Timestamp,Water Temperature,Turbidity,Transducer Depth,Wave Heigh
t,Wave Period,Battery Life,Measurement Timestamp Label,Measurement ID
```

```
Waiting...
Topic rafi-data-raw, Event msg = Montrose Beach,05/28/2014 09:00:00 PM,14.5,7.11,1.374,
0.361,5,11.9,5/28/2014 9:00 PM,MontroseBeach201405282100
Waiting...
^Crafimuhammad@Rafimuhammad:/mnt/c/Users/Rafi Muhammad/Documents/Bootcamp_DataEngineer/
Project6$ docker exec -it my_container python /project/run_consumer.py /project/credent
ials.ini rafi-data-clean
Waiting...
Waiting...
Waiting...
Waiting...
```

wsl
wsl

main* 0 0 0 Ln 1, Col 1 Spaces: 4 UTF-8 CRLF {} Python 3.11.5 64-bit Go Live Prettier

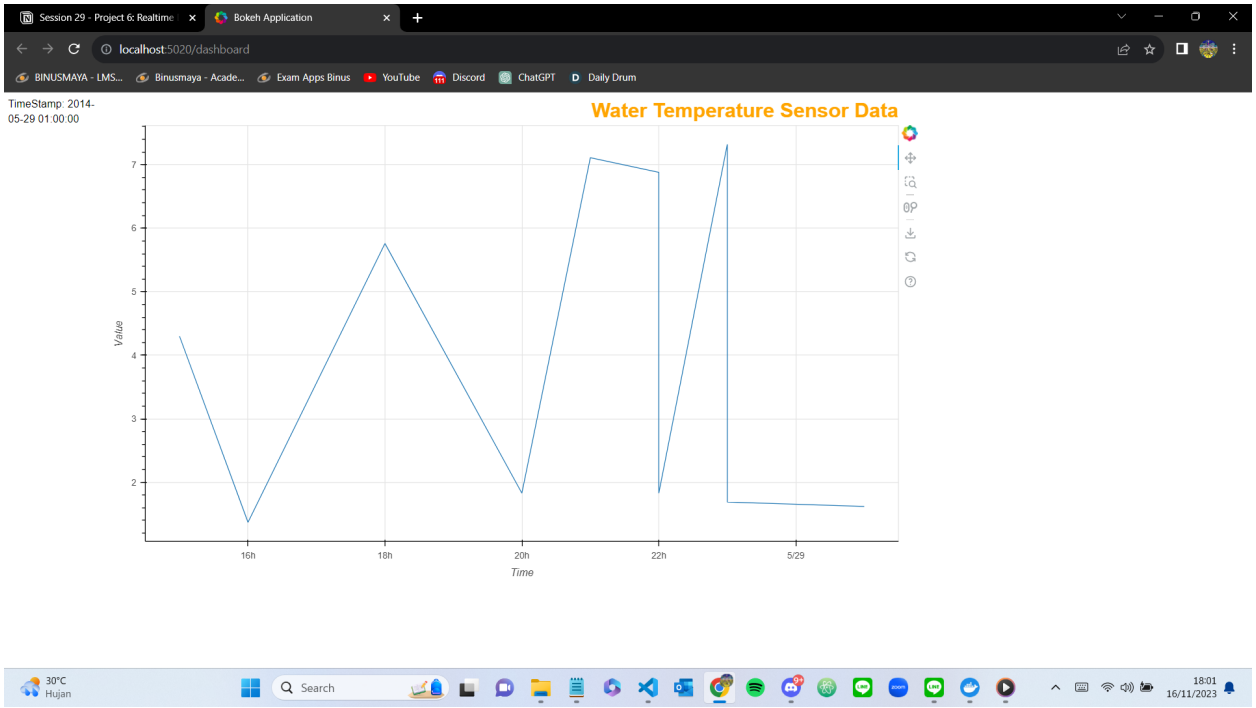
31°C Cerah Search 16:45 16/11/2023

Untuk run script for consume data from topic `clean-data`

```
python3 -m bokeh serve --port 5020 --show hands-on/dashboard.py
```

dashboard.py file:

```python
import random
import pandas as pd
from bokeh.driving import count
from bokeh.models import ColumnDataSource
from bokeh.plotting import curdoc, figure, show
from bokeh.models import DatetimeTickFormatter
from bokeh.models.widgets import Div
from bokeh.layouts import column,row
import time
from datetime import datetime
from confluent_kafka import Consumer


UPDATE_INTERVAL = 1
ROLLOVER = 10 # Number of displayed data points

source = ColumnDataSource({"x": [], "y": []})
consumer = Consumer({'bootstrap.servers': 'pkc-6ojv2.us-west4.gcp.confluent.cloud:9092', 'security.protocol': 'SASL_SSL', 'sasl.mechanisms': 'PLAIN', 'sasl.username': 'HZ
# Set up a callback to handle the '--reset' flag.

# Subscribe to topic
consumer.subscribe(['rafi-data-raw'])
div = Div(
    text='',
    width=120,
    height=35
)



@count()
def update(x):
    msg_value=None
    while True:
        msg = consumer.poll(1.0)
        if msg is None:
            print("Waiting...")
```

```python
        if msg is None:
            print("Waiting...")
            return
        elif msg.error():
            print("ERROR: %s".format(msg.error()))
            return
        else:
            print("Topic {}, Event msg = {}".format(msg.topic(), msg.value().decode('utf-8')))
            msg_value=msg
            break

    values=msg_value.value().decode("utf-8").split(',')
    x=pd.to_datetime(values[1])

    print(x)

    div.text = "TimeStamp: "+str(x)


    y = values[3]
    print(y)

    source.stream({"x": [x], "y": [y]},ROLLOVER)

p = figure(title="Water Temperature Sensor Data",x_axis_type = "datetime",width=1000)
p.line("x", "y", source=source)

p.xaxis.formatter=DatetimeTickFormatter(hourmin = ['%H:%M'])
p.xaxis.axis_label = 'Time'
p.yaxis.axis_label = 'Value'
p.title.align = "right"
p.title.text_color = "orange"
p.title.text_font_size = "25px"

doc = curdoc()
#doc.add_root(p)

doc.add_root(
    row(children=[div,p])
)
doc.add_periodic_callback(update, UPDATE_INTERVAL)
```

Hasil Graph Water Temperature Sensor Data: