Load data from Kafka to Hadoop

<Steps to run the python file to load data from Kafka>

1. We start with first importing the modules

*C:\Users\rancy\Downloads\CabRides_Kafka_to_Hadoop.py - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

new 2 | CabRides_Kafka_to_Hadoop.py

```python


# Script is to get all the cab rides related click stream data residing in given Kafka server to Hadoop cluster


# we start with first importing the modules
import os,sys
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
```

2. Set required environment variables needed to get the data

```python
# Next step is to set required environment variables needed to get the data
os.environ["PYSPARK_PYTHON"] = "/opt/cloudera/parcels/Anaconda/bin/python"
os.environ["JAVA_HOME"] = "/usr/java/jdk1.8.0_161/jre"
os.environ["SPARK_HOME"] = "/opt/cloudera/parcels/SPARK2-2.3.0.cloudera2-1.cdh5.13.3.p0.316101/lib/spark2/"
os.environ["PYLIB"] = os.environ["SPARK_HOME"] + "/python/lib"
sys.path.insert(0, os.environ["PYLIB"] + "/py4j-0.10.6-src.zip")
sys.path.insert(0, os.environ["PYLIB"] + "/pyspark.zip")
```

3. Next initialize Spark session

```python
# We will now initialize Spark session
spark = SparkSession \
        .builder \
        .appName("Kafka-to-local") \
        .getOrCreate()
```

4. Read Data from kafka server from given Kafka server details

```python
# From the given server connection details connect to kafka and get the stream in a dataframe
  # Read Input from kafka
streamdf = spark.readStream \
        .format("kafka") \
        .option("kafka.bootstrap.servers", "18.211.252.152:9092") \
        .option("startingOffsets", "earliest") \
        .option("subscribe", "de-capstone3") \
        .load()
```

5. Keep relevant field ‘value’ rename it to ‘value_str’ and drop other irrelevant fields

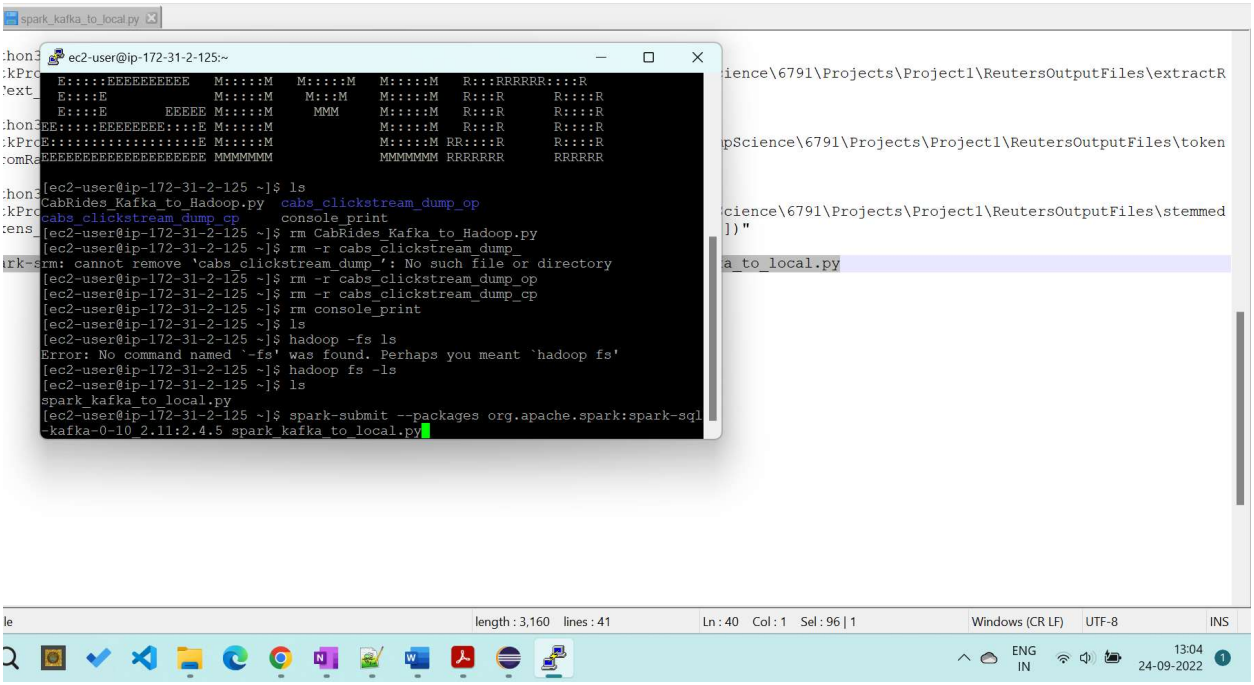
```python
# get only relevant fields and drop others
streamdf= streamdf \
       .withColumn('value_str',streamdf['value'].cast('string').alias('key_str')).drop('value') \
       .drop('key','topic','partition','offset','timestamp','timestampType')
```

<Steps to load the data into Hadoop>

1. Write the click stream to folder ‘cabs_clickstream_dump_op’ in Hadoop

```
#Writing the click stream to a folder in Hadoop
streamdf.writeStream \
  .format("json") \
  .outputMode("append") \
  .option("path", "cabs_clickstream_dump_op") \
  .option("checkpointLocation", "cabs_clickstream_dump_cp") \
  .start() \
  .awaitTermination()
```

2. Submit Spark Job



<Screenshot of the data>