Load data from Kafka to Hadoop

Steps to run the python file to load data from Kafka

1. Use WinScp to transfer the file from local computer to the cluster.

2. Spark submit the python file

>> export SPARK_KAFKA_VERSION=0.10
>> spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py

4. Use another python file to clean the loaded Kafka data to a more structured format as csv file format.
>> spark_local_flatten.py

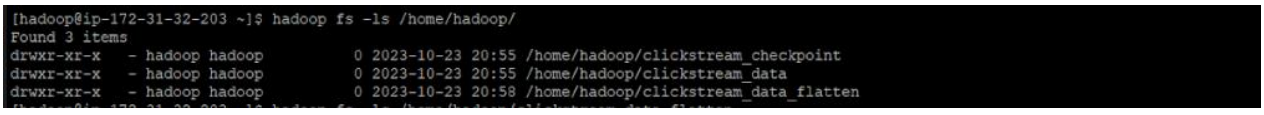
5. Spark submit the python file
>> spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_local_flatten.py

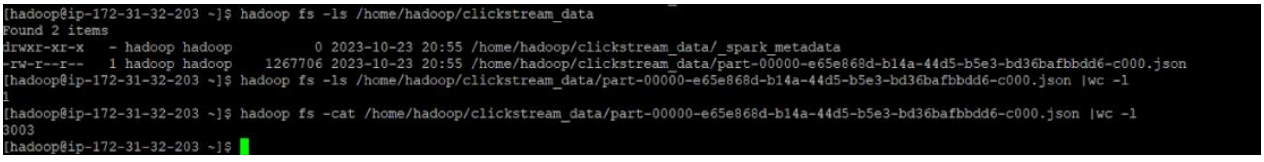
Steps to load the data into Hadoop

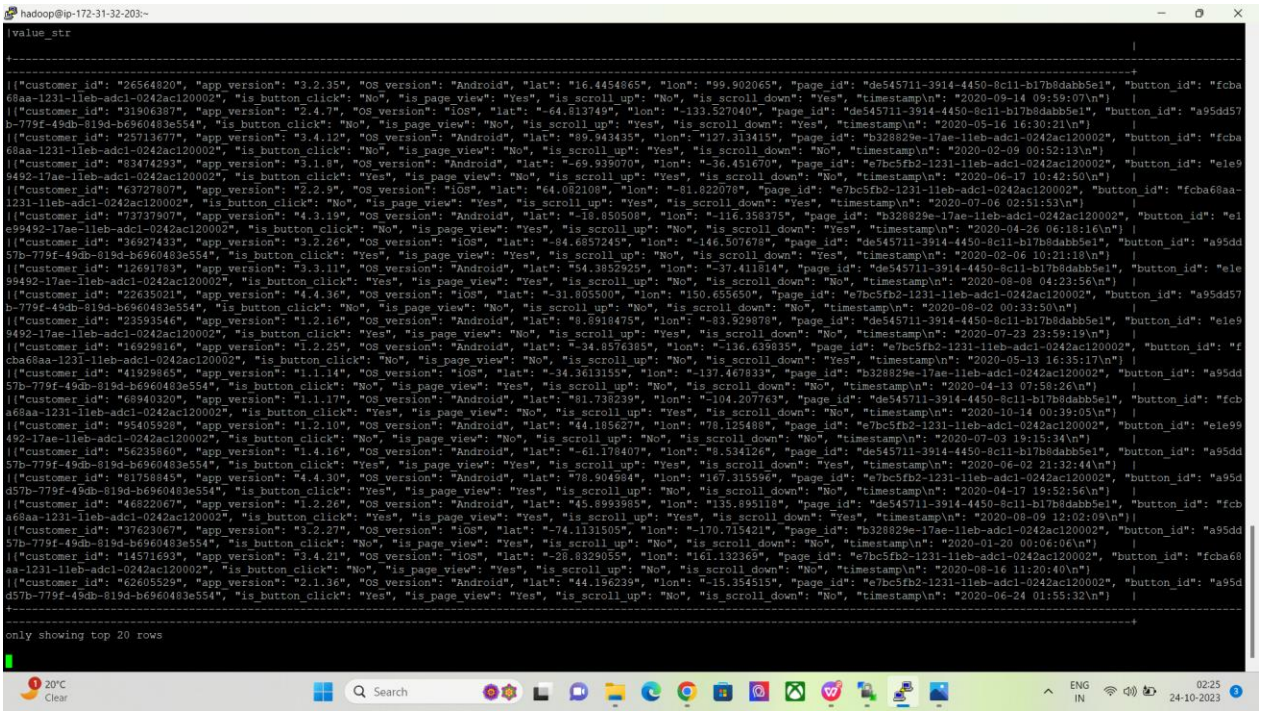
The data is already loaded into Hdfs using the Python file.

Screenshot of the data

Data loaded in .json format from kafka stream.







Clean Data file.