CS4830 – Big Data Laboratory

Assignment 5-Lab 8

Ishan Chokshi – BE19B018

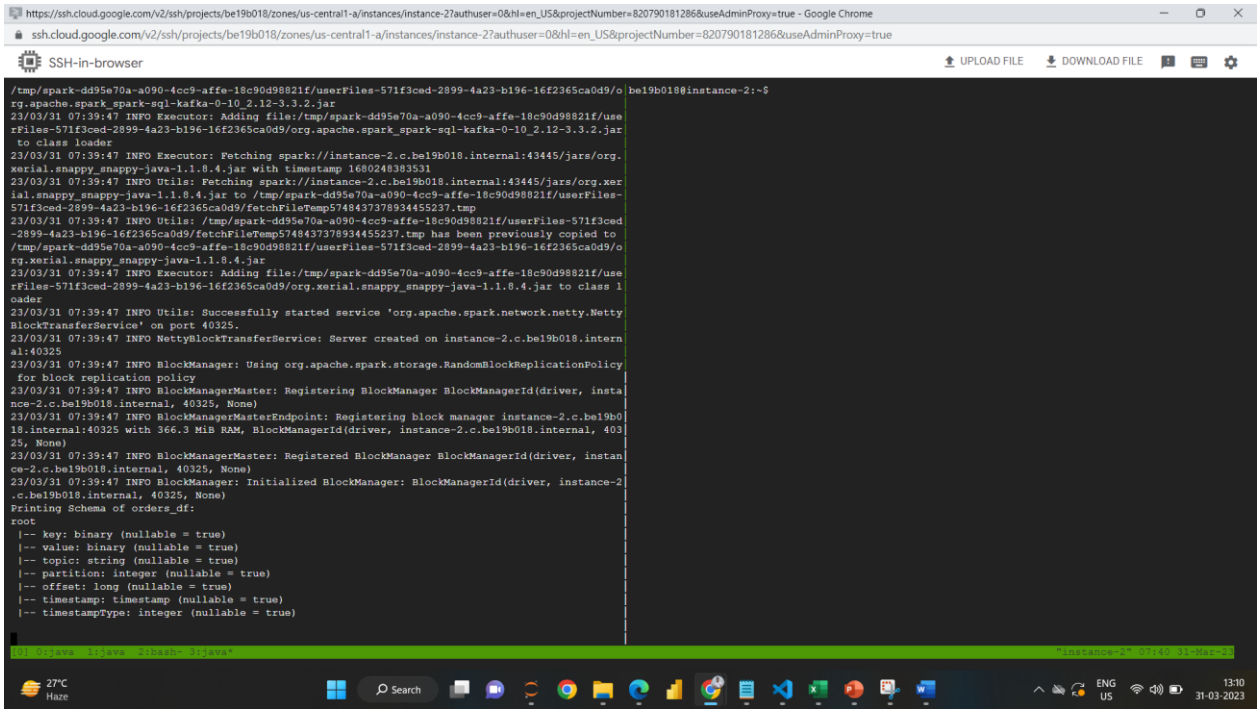
January-May 2023

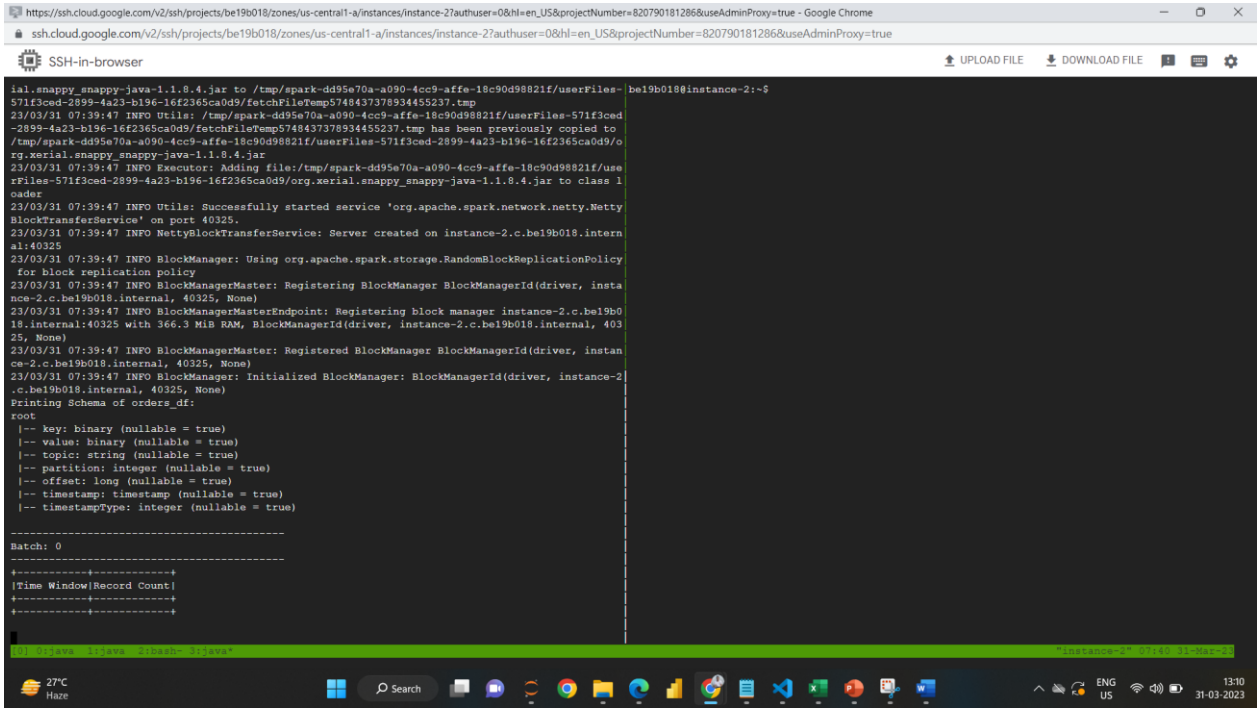
Indian Institute of Technology Madras

Ishan Chokshi
BE19B018

1. **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.**

    a. First we follow the steps given in the handout until the creation of the topic. I have set the topic name as 'test-topic'
    b. Next, I generated synthetic data having the same schema as the one used in the demo, using a python file. 100 records were created for this data and the data was stored in a csv file.
    c. This csv file was uploaded to a bucket names 'a5_lab8'.
    d. Now, I created two python files in the VM: producer_final.py and streaming_file.py
    e. Streaming_file.py uses spark streaming to read from Kafka every 5 seconds. The time window is 10 seconds and sliding interval is taken as 5 seconds. For easy output interpretation, I have selected only the timestamp and number of records streamed in every window to be displayed.
    f. To ensure only a maximum of 100 records are written, I have added break statement to terminate the csv content iterator.
    g. The producer_final.py file is reading the csv file uploaded to the bucket and iterating through its contents. Every 10 seconds, it is sending the data to the streaming_file.py file. Sleep time is set to 10 seconds.
    h. Screenshots of the output have been attached below in a sequential order from running the streaming_file.py to the end of stream data processing:

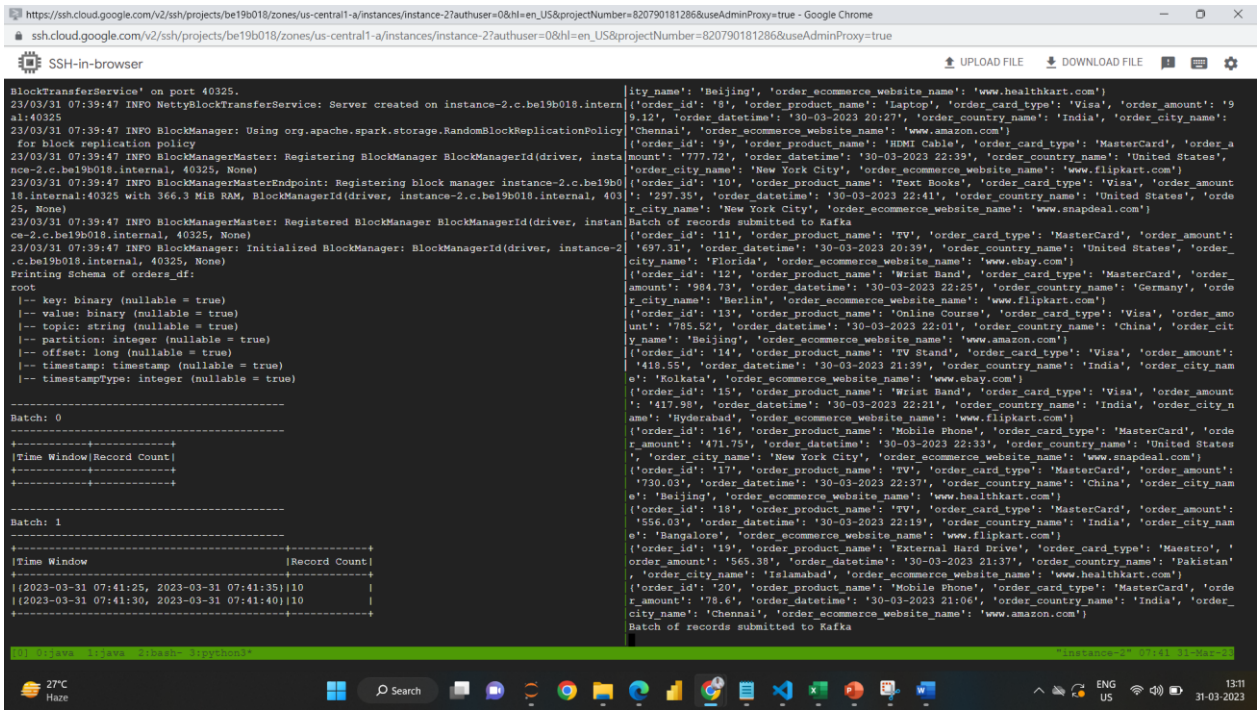

https://ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
be19b018@instance-2:~$
/tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.apache.spark_spark-sql-kafka-0-10_2.12-3.3.2.jar
23/03/31 07:39:47 INFO Executor: Adding file:/tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.apache.spark_spark-sql-kafka-0-10_2.12-3.3.2.jar to class loader
23/03/31 07:39:47 INFO Executor: Fetching spark://instance-2.c.be19b018.internal:43445/jars/org.xerial.snappy_snappy-java-1.1.8.4.jar with timestamp 1680248383531
23/03/31 07:39:47 INFO Utils: Fetching spark://instance-2.c.be19b018.internal:43445/jars/org.xerial.snappy_snappy-java-1.1.8.4.jar to /tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/fetchFileTemp5748437378934455237.tmp
23/03/31 07:39:47 INFO Utils: /tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/fetchFileTemp5748437378934455237.tmp has been previously copied to /tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.xerial.snappy_snappy-java-1.1.8.4.jar
23/03/31 07:39:47 INFO Executor: Adding file:/tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.xerial.snappy_snappy-java-1.1.8.4.jar to class loader
23/03/31 07:39:47 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40325.
23/03/31 07:39:47 INFO NettyBlockTransferService: Server created on instance-2.c.be19b018.internal:40325
23/03/31 07:39:47 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
23/03/31 07:39:47 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
23/03/31 07:39:47 INFO BlockManagerMasterEndpoint: Registering block manager instance-2.c.be19b018.internal:40325 with 366.3 MiB RAM, BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
23/03/31 07:39:47 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
23/03/31 07:39:47 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
Printing Schema of orders_df:
root
|-- key: binary (nullable = true)
|-- value: binary (nullable = true)
|-- topic: string (nullable = true)
|-- partition: integer (nullable = true)
|-- offset: long (nullable = true)
|-- timestamp: timestamp (nullable = true)
|-- timestampType: integer (nullable = true)
[0] 0:java 1:java 2:bash- 3:java*
"instance-2" 07:40 31-Mar-23
27°C
Haze
Search
ENG
US
13:10
31-03-2023

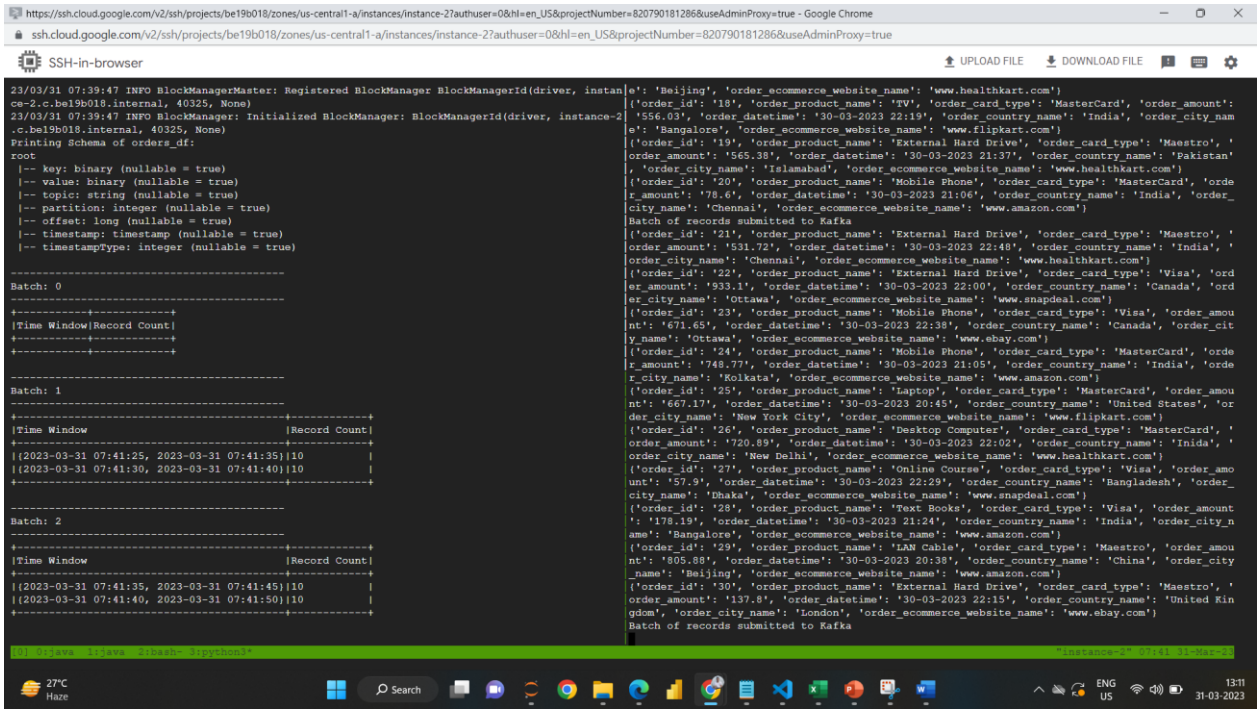

https://ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
be19b018@instance-2:~$
ial.snappy_snappy-java-1.1.8.4.jar to /tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/fetchFileTemp5748437378934455237.tmp
23/03/31 07:39:47 INFO Utils: /tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/fetchFileTemp5748437378934455237.tmp has been previously copied to /tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.xerial.snappy_snappy-java-1.1.8.4.jar
23/03/31 07:39:47 INFO Executor: Adding file:/tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.xerial.snappy_snappy-java-1.1.8.4.jar to class loader
23/03/31 07:39:47 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40325.
23/03/31 07:39:47 INFO NettyBlockTransferService: Server created on instance-2.c.be19b018.internal:40325
23/03/31 07:39:47 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
23/03/31 07:39:47 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
23/03/31 07:39:47 INFO BlockManagerMasterEndpoint: Registering block manager instance-2.c.be19b018.internal:40325 with 366.3 MiB RAM, BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
23/03/31 07:39:47 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
23/03/31 07:39:47 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)
Printing Schema of orders_df:
root
|-- key: binary (nullable = true)
|-- value: binary (nullable = true)
|-- topic: string (nullable = true)
|-- partition: integer (nullable = true)
|-- offset: long (nullable = true)
|-- timestamp: timestamp (nullable = true)
|-- timestampType: integer (nullable = true)

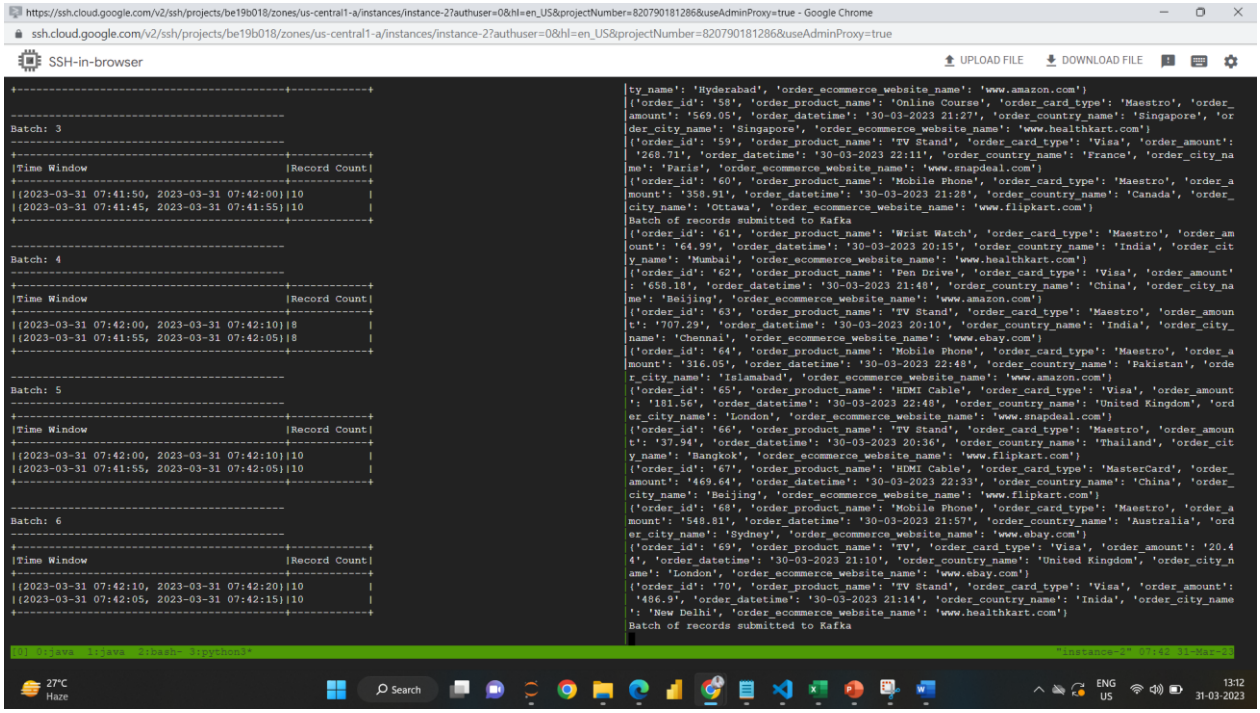
Batch: 0

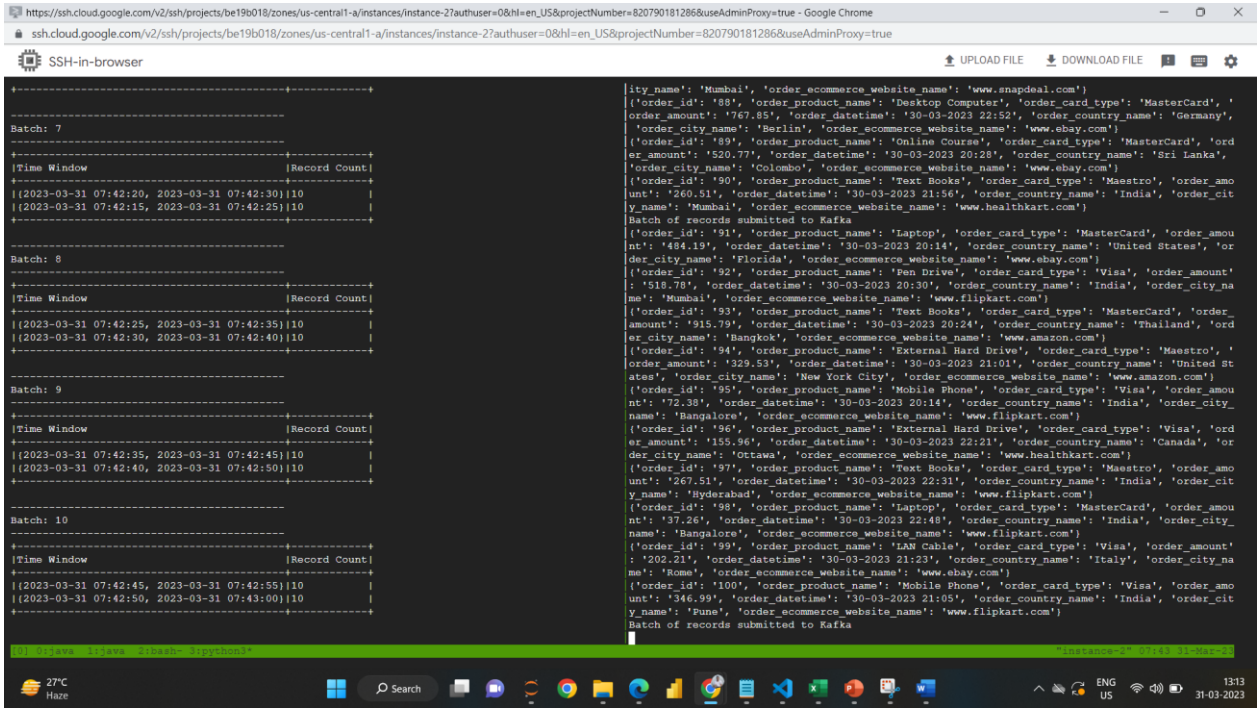
+-----------+------------+
|Time Window|Record Count|
+-----------+------------+
+-----------+------------+
[0] 0:java 1:java 2:bash- 3:java*
"instance-2" 07:40 31-Mar-23
27°C
Haze
Search
ENG
US
13:10
31-03-2023

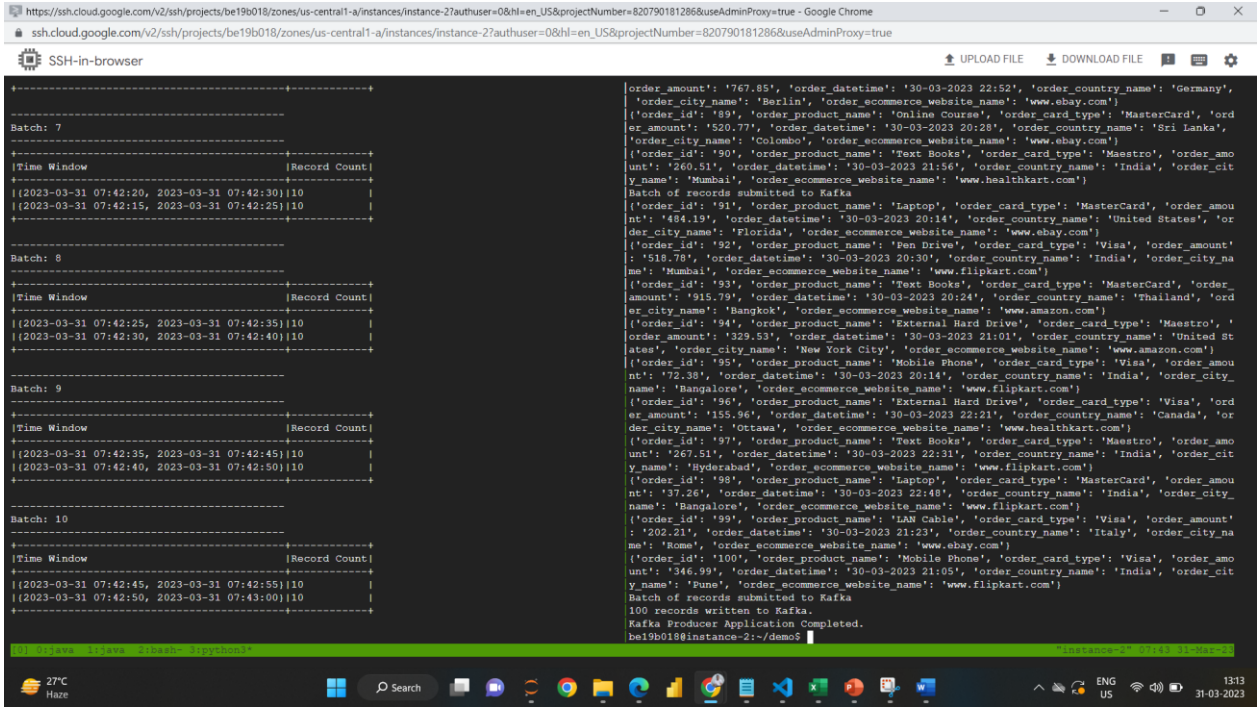
Ishan Chokshi
BE19B018






https://ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Batch: 3
Time Window	Record Count
(2023-03-31 07:41:50, 2023-03-31 07:42:00)	10
(2023-03-31 07:41:45, 2023-03-31 07:41:55)	10
Batch: 4
Time Window	Record Count
(2023-03-31 07:42:00, 2023-03-31 07:42:10)	8
(2023-03-31 07:41:55, 2023-03-31 07:42:05)	8
Batch: 5
Time Window	Record Count
(2023-03-31 07:42:00, 2023-03-31 07:42:10)	10
(2023-03-31 07:41:55, 2023-03-31 07:42:05)	10
Batch: 6
Time Window	Record Count
(2023-03-31 07:42:10, 2023-03-31 07:42:20)	10
(2023-03-31 07:42:05, 2023-03-31 07:42:15)	10
Batch of records submitted to Kafka
Batch of records submitted to Kafka
Search
27°C
Haze
ENG
US
13:12
31-03-2023


https://ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Batch: 7
Time Window	Record Count
(2023-03-31 07:42:20, 2023-03-31 07:42:30)	10
(2023-03-31 07:42:15, 2023-03-31 07:42:25)	10
Batch: 8
Time Window	Record Count
(2023-03-31 07:42:25, 2023-03-31 07:42:35)	10
(2023-03-31 07:42:30, 2023-03-31 07:42:40)	10
Batch: 9
Time Window	Record Count
(2023-03-31 07:42:35, 2023-03-31 07:42:45)	10
(2023-03-31 07:42:40, 2023-03-31 07:42:50)	10
Batch: 10
Time Window	Record Count
(2023-03-31 07:42:45, 2023-03-31 07:42:55)	10
(2023-03-31 07:42:50, 2023-03-31 07:43:00)	10
Batch of records submitted to Kafka
Search
27°C
Haze
ENG
US
13:13
31-03-2023


https://ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/be19b018/zones/us-central1-a/instances/instance-2?authuser=0&hl=en_US&projectNumber=820790181286&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Batch: 7
Time Window	Record Count
(2023-03-31 07:42:20, 2023-03-31 07:42:30)	10
(2023-03-31 07:42:15, 2023-03-31 07:42:25)	10
Batch: 8
Time Window	Record Count
(2023-03-31 07:42:25, 2023-03-31 07:42:35)	10
(2023-03-31 07:42:30, 2023-03-31 07:42:40)	10
Batch: 9
Time Window	Record Count
(2023-03-31 07:42:35, 2023-03-31 07:42:45)	10
(2023-03-31 07:42:40, 2023-03-31 07:42:50)	10
Batch: 10
Time Window	Record Count
(2023-03-31 07:42:45, 2023-03-31 07:42:55)	10
(2023-03-31 07:42:50, 2023-03-31 07:43:00)	10
Batch of records submitted to Kafka
100 records written to Kafka.
Kafka Producer Application Completed.
be19b018@instance-2:~/demo$
27°C
Haze
Search
ENG
US
13:13
31-03-2023