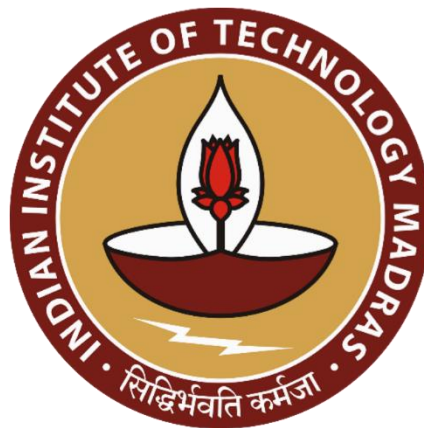


CS4830 – Big Data Laboratory

Assignment 5-Lab 8

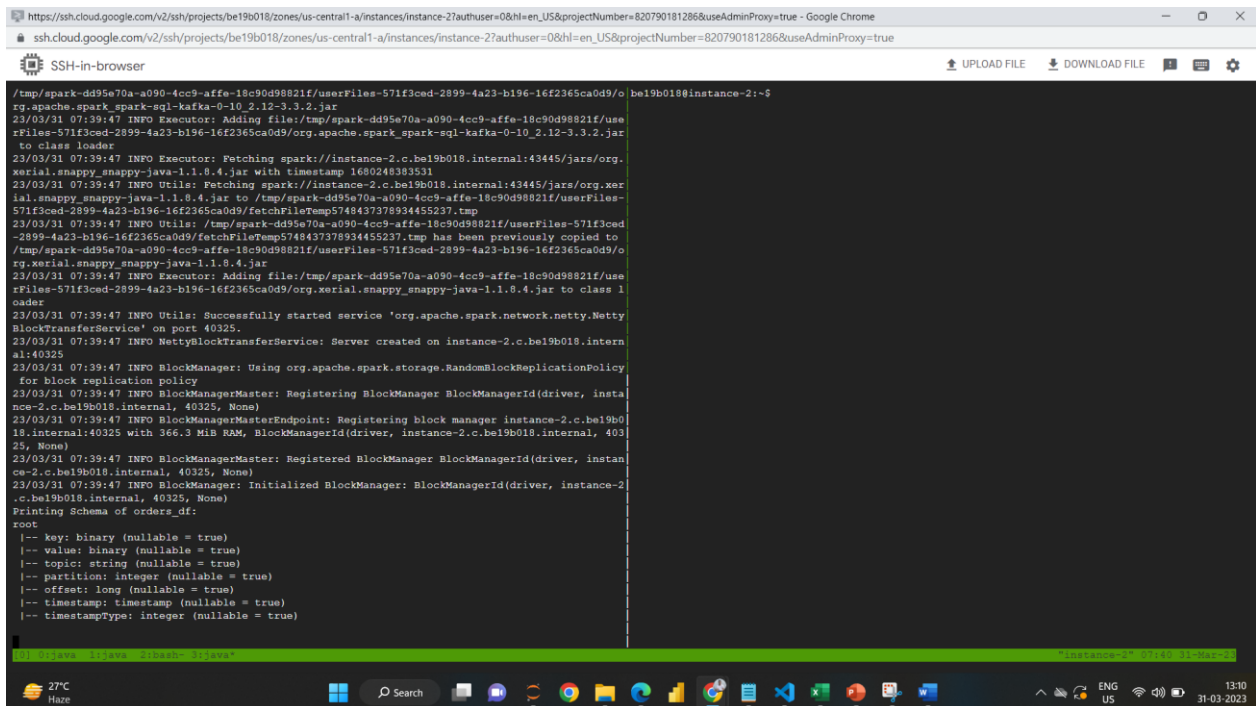


Ishan Chokshi – BE19B018

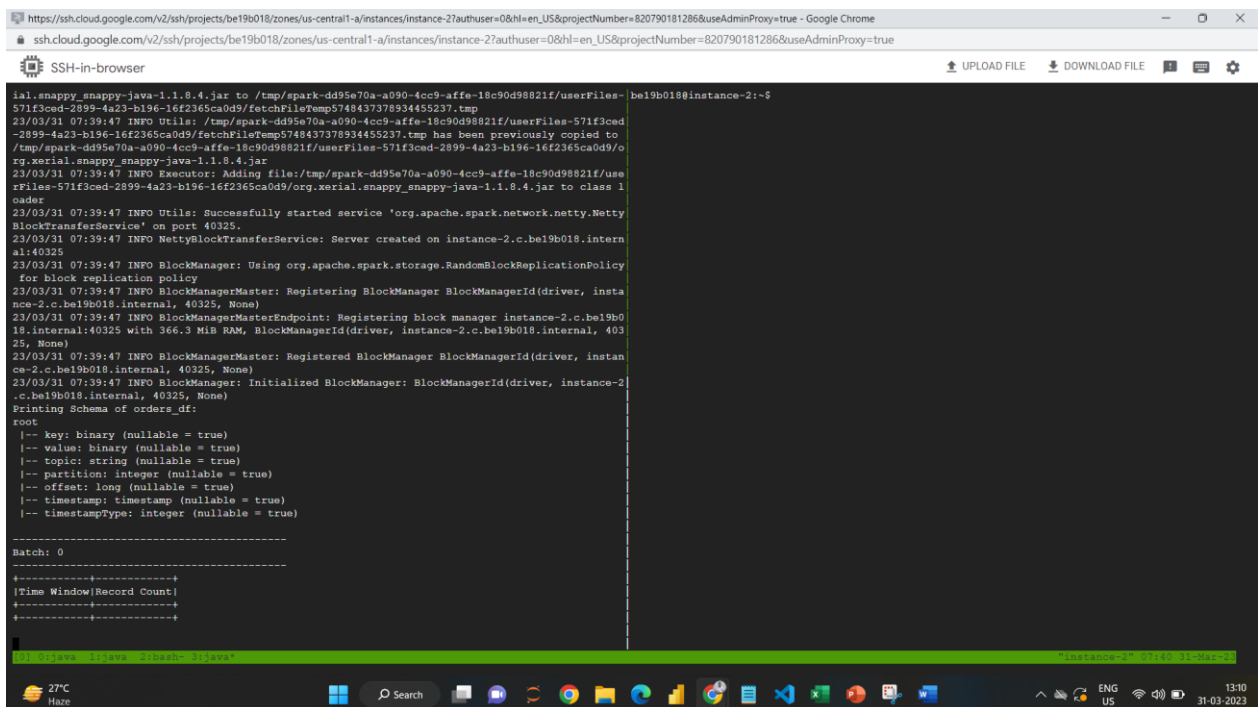
January-May 2023

Indian Institute of Technology Madras

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:



The screenshot shows a terminal window titled "SSH-in-browser" with a URL bar indicating a connection to Google Cloud. The terminal output displays logs for a Spark application. Key messages include: "INFO Executor: Adding file:/tmp/spark-dd95e70a-a090-4cc9-affe-18c90d98821f/userFiles-571f3ced-2899-4a23-b196-16f2365ca0d9/org.apache.spark.sql-kafka-0-10.2.12-3.3.2.jar to class loader", "INFO Utils: Fetching spark://instance-2.c.be19b018.internal:43445/jars/org.xerial.snappy.snappy-java-1.1.8.4.jar with timestamp 1680248383531", "INFO NettyBlockTransferService: Server created on instance-2.c.be19b018.internal:40325", and "INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, instance-2.c.be19b018.internal, 40325, None)". The logs also show the initialization of the BlockManager and the printing of a schema for orders_df.



This screenshot is another instance of the terminal window, showing similar logs to the first one. It includes the same file addition and Spark utility messages. However, the "INFO NettyBlockTransferService" message specifies "Server created on instance-2.c.be19b018.internal:40325". The "INFO BlockManagerMaster" message also specifies "instance-2.c.be19b018.internal, 40325, None". The logs conclude with the same schema printing for orders_df.

```
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

23/03/31 07:39:47 INFO MettlyBlockTransferService: 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|
+-----+-----+
|2023-03-31 07:41:25, 2023-03-31 07:41:35|110|
+-----+-----+

Batch: 1
-----+-----+
|Time Window|Record Count|
+-----+-----+
|2023-03-31 07:41:25, 2023-03-31 07:41:35|110|
|2023-03-31 07:41:30, 2023-03-31 07:41:40|110|
+-----+-----+

[1] C:\Users\Ishan> cd C:\ProgramData\Python38\Scripts
C:\ProgramData\Python38\Scripts>

27°C
Haze
```

```
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

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|
+-----+-----+
|2023-03-31 07:41:25, 2023-03-31 07:41:35|110|
|2023-03-31 07:41:30, 2023-03-31 07:41:40|110|
+-----+-----+

Batch: 1
-----+-----+
|Time Window|Record Count|
+-----+-----+
|2023-03-31 07:41:35, 2023-03-31 07:41:45|110|
|2023-03-31 07:41:40, 2023-03-31 07:41:50|110|
+-----+-----+

Batch: 2
-----+-----+
|Time Window|Record Count|
+-----+-----+
|2023-03-31 07:41:35, 2023-03-31 07:41:45|110|
|2023-03-31 07:41:40, 2023-03-31 07:41:50|110|
+-----+-----+

[1] C:\Users\Ishan> cd C:\ProgramData\Python38\Scripts
C:\ProgramData\Python38\Scripts>

27°C
Haze
```

```
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

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)|10|

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|

[0] 27°C
Haze

ty name': 'Hyderabad', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '58', 'order_product_name': 'Online Course', 'order_card_type': 'Maestro', 'order_
amount': '569.05', 'order_datetime': '30-03-2023 21:27', 'order_country_name': 'Singapore', 'or
der_city_name': 'Singapore', 'order_ecommerce_website_name': 'www.healthkart.com')
['order_id': '59', 'order_product_name': 'TV Stand', 'order_card_type': 'Visa', 'order_amount':
'268.71', 'order_datetime': '30-03-2023 22:11', 'order_country_name': 'France', 'order_city_na
me': 'Paris', 'order_ecommerce_website_name': 'www.snapdeal.com')
['order_id': '60', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Maestro', 'order_a
mount': '358.91', 'order_datetime': '30-03-2023 21:28', 'order_country_name': 'Canada', 'order_
city_name': 'Ottawa', 'order_ecommerce_website_name': 'www.flipkart.com')
Batch of records submitted to Kafka
['order_id': '61', 'order_product_name': 'Wrist Watch', 'order_card_type': 'Maestro', 'order_am
ount': '64.99', 'order_datetime': '30-03-2023 20:15', 'order_country_name': 'India', 'order_cit
y_name': 'Mumbai', 'order_ecommerce_website_name': 'www.healthkart.com')
['order_id': '62', 'order_product_name': 'Pen Drive', 'order_card_type': 'Visa', 'order_amount'
: '659.18', 'order_datetime': '30-03-2023 21:48', 'order_country_name': 'China', 'order_city_na
me': 'Beijing', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '63', 'order_product_name': 'TV Stand', 'order_card_type': 'Maestro', 'order_amoun
t': '707.29', 'order_datetime': '30-03-2023 20:10', 'order_country_name': 'India', 'order_city_
name': 'Chennai', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '64', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Maestro', 'order_a
mount': '316.05', 'order_datetime': '30-03-2023 22:48', 'order_country_name': 'Pakistan', 'orde
r_city_name': 'Islamabad', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '65', 'order_product_name': 'HDMI Cable', 'order_card_type': 'Visa', 'order_amount'
: '191.56', 'order_datetime': '30-03-2023 22:49', 'order_country_name': 'United Kingdom', 'ord
er_city_name': 'London', 'order_ecommerce_website_name': 'www.snapdeal.com')
['order_id': '66', 'order_product_name': 'TV Stand', 'order_card_type': 'Maestro', 'order_amoun
t': '37.94', 'order_datetime': '30-03-2023 20:36', 'order_country_name': 'Thailand', 'order_cit
y_name': 'Bangkok', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '67', 'order_product_name': 'HDMI Cable', 'order_card_type': 'MasterCard', 'order_
amount': '469.64', 'order_datetime': '30-03-2023 22:33', 'order_country_name': 'China', 'order_
city_name': 'Beijing', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '68', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Maestro', 'order_a
mount': '548.81', 'order_datetime': '30-03-2023 21:57', 'order_country_name': 'Australia', 'ord
er_city_name': 'Sydney', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '69', 'order_product_name': 'TV', 'order_card_type': 'Visa', 'order_amount': '20.4
4', 'order_datetime': '30-03-2023 21:10', 'order_country_name': 'United Kingdom', 'order_cit
y_name': 'London', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '70', 'order_product_name': 'TV Stand', 'order_card_type': 'Visa', 'order_amount':
'486.9', 'order_datetime': '30-03-2023 21:14', 'order_country_name': 'India', 'order_city_name'
: 'New Delhi', 'order_ecommerce_website_name': 'www.healthkart.com')
Batch of records submitted to Kafka
```

```
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

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|

[0] 27°C
Haze

ity name': 'Mumbai', 'order_ecommerce_website_name': 'www.snapdeal.com')
['order_id': '88', 'order_product_name': 'Desktop Computer', 'order_card_type': 'MasterCard', '
order_amount': '767.85', 'order_datetime': '30-03-2023 22:52', 'order_country_name': 'Germany', '
order_city_name': 'Berlin', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '89', 'order_product_name': 'Online Course', 'order_card_type': 'MasterCard', 'ord
er_amount': '520.77', 'order_datetime': '30-03-2023 20:28', 'order_country_name': 'Sri Lanka',
'order_city_name': 'Colombo', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '90', 'order_product_name': 'Text Books', 'order_card_type': 'Maestro', 'order_amo
unt': '260.51', 'order_datetime': '30-03-2023 21:56', 'order_country_name': 'India', 'order_cit
y_name': 'Mumbai', 'order_ecommerce_website_name': 'www.healthkart.com')
Batch of records submitted to Kafka
['order_id': '91', 'order_product_name': 'Laptop', 'order_card_type': 'MasterCard', 'order_amou
nt': '484.19', 'order_datetime': '30-03-2023 20:14', 'order_country_name': 'United States', 'or
der_city_name': 'Florida', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '92', 'order_product_name': 'Pen Drive', 'order_card_type': 'Visa', 'order_amount'
: '518.78', 'order_datetime': '30-03-2023 20:30', 'order_country_name': 'India', 'order_city_na
me': 'Mumbai', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '93', 'order_product_name': 'Text Books', 'order_card_type': 'MasterCard', 'order_
amount': '915.79', 'order_datetime': '30-03-2023 20:24', 'order_country_name': 'Thailand', 'ord
er_city_name': 'Bangkok', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '94', 'order_product_name': 'External Hard Drive', 'order_card_type': 'Maestro', '
order_amount': '329.53', 'order_datetime': '30-03-2023 21:01', 'order_country_name': 'United St
ates', 'order_city_name': 'New York City', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '95', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Visa', 'order_amo
unt': '72.38', 'order_datetime': '30-03-2023 20:14', 'order_country_name': 'India', 'order_cit
y_name': 'Bangalore', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '96', 'order_product_name': 'External Hard Drive', 'order_card_type': 'Visa', 'ord
er_amount': '153.96', 'order_datetime': '30-03-2023 22:21', 'order_country_name': 'Canada', 'or
der_city_name': 'Ottawa', 'order_ecommerce_website_name': 'www.healthkart.com')
['order_id': '97', 'order_product_name': 'Text Books', 'order_card_type': 'Maestro', 'order_amo
unt': '267.51', 'order_datetime': '30-03-2023 22:31', 'order_country_name': 'India', 'order_cit
y_name': 'Hyderabad', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '98', 'order_product_name': 'Laptop', 'order_card_type': 'MasterCard', 'order_amo
unt': '37.26', 'order_datetime': '30-03-2023 22:48', 'order_country_name': 'India', 'order_cit
y_name': 'Bangalore', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '99', 'order_product_name': 'LAN Cable', 'order_card_type': 'Visa', 'order_amount'
: '202.21', 'order_datetime': '30-03-2023 21:23', 'order_country_name': 'Italy', 'order_city_na
me': 'Rome', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '100', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Visa', 'order_amo
unt': '346.99', 'order_datetime': '30-03-2023 21:05', 'order_country_name': 'India', 'order_cit
y_name': 'Pune', 'order_ecommerce_website_name': 'www.flipkart.com')
Batch of records submitted to Kafka
```

```
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

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 |

order_amount': '767.85', 'order_datetime': '30-03-2023 22:52', 'order_country_name': 'Germany',
'order_city_name': 'Berlin', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '89', 'order_product_name': 'Online Course', 'order_card_type': 'MasterCard', 'ord
er_amount': '520.77', 'order_datetime': '30-03-2023 20:28', 'order_country_name': 'Sri Lanka',
'order_city_name': 'Colombo', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '90', 'order_product_name': 'Text Books', 'order_card_type': 'Maestro', 'order_amo
unt': '260.51', 'order_datetime': '30-03-2023 21:56', 'order_country_name': 'India', 'order_cit
y_name': 'Mumbai', 'order_ecommerce_website_name': 'www.healthkart.com')
Batch of records submitted to Kafka
['order_id': '91', 'order_product_name': 'Laptop', 'order_card_type': 'MasterCard', 'order_amo
unt': '484.19', 'order_datetime': '30-03-2023 20:14', 'order_country_name': 'United States', 'or
der_city_name': 'Florida', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '92', 'order_product_name': 'Pen Drive', 'order_card_type': 'Visa', 'order_amount'
: '518.78', 'order_datetime': '30-03-2023 20:30', 'order_country_name': 'India', 'order_city_na
me': 'Mumbai', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '93', 'order_product_name': 'Text Books', 'order_card_type': 'MasterCard', 'order
_amount': '915.79', 'order_datetime': '30-03-2023 20:24', 'order_country_name': 'Thailand', 'ord
er_city_name': 'Bangkok', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '94', 'order_product_name': 'External Hard Drive', 'order_card_type': 'Maestro', '
order_amount': '329.53', 'order_datetime': '30-03-2023 21:01', 'order_country_name': 'United St
ates', 'order_city_name': 'New York City', 'order_ecommerce_website_name': 'www.amazon.com')
['order_id': '95', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Visa', 'order_amo
unt': '72.39', 'order_datetime': '30-03-2023 20:14', 'order_country_name': 'India', 'order_cit
y_name': 'Bangalore', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '96', 'order_product_name': 'External Hard Drive', 'order_card_type': 'Visa', 'ord
er_amount': '155.96', 'order_datetime': '30-03-2023 22:21', 'order_country_name': 'Canada', 'or
der_city_name': 'Ottawa', 'order_ecommerce_website_name': 'www.healthkart.com')
['order_id': '97', 'order_product_name': 'Text Books', 'order_card_type': 'Maestro', 'order_amo
unt': '267.51', 'order_datetime': '30-03-2023 22:31', 'order_country_name': 'India', 'order_cit
y_name': 'Hyderabad', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '98', 'order_product_name': 'Laptop', 'order_card_type': 'MasterCard', 'order_amo
unt': '37.26', 'order_datetime': '30-03-2023 22:48', 'order_country_name': 'India', 'order_cit
y_name': 'Bangalore', 'order_ecommerce_website_name': 'www.flipkart.com')
['order_id': '99', 'order_product_name': 'LAN Cable', 'order_card_type': 'Visa', 'order_amo
unt': '202.21', 'order_datetime': '30-03-2023 21:23', 'order_country_name': 'Italy', 'order_cit
y_name': 'Rome', 'order_ecommerce_website_name': 'www.ebay.com')
['order_id': '100', 'order_product_name': 'Mobile Phone', 'order_card_type': 'Visa', 'order_amo
unt': '346.99', 'order_datetime': '30-03-2023 21:05', 'order_country_name': 'India', 'order_cit
y_name': 'Pune', 'order_ecommerce_website_name': 'www.flipkart.com')
Batch of records submitted to Kafka
100 records written to Kafka.
Kafka Producer Application Completed.
be19b018@instance-2:~/demo$
```