

Spark SQL and streaming Spark

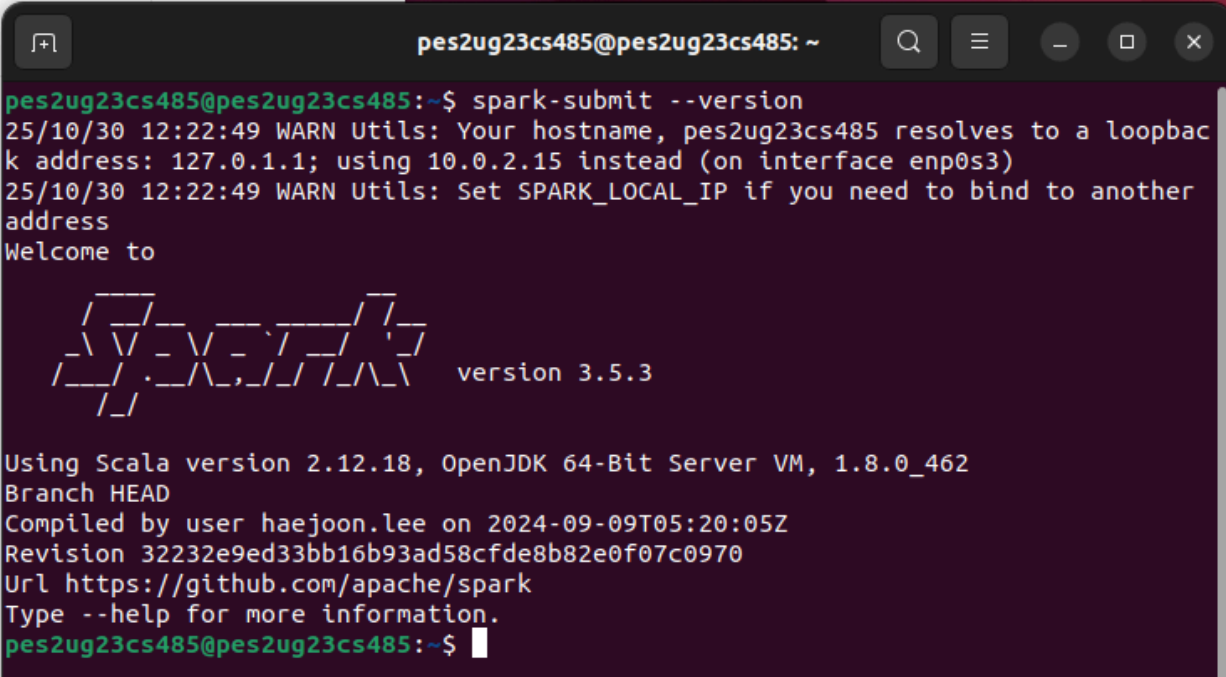
Handson

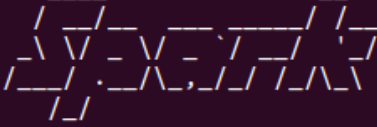
Name: Rithvik Rajesh Matta

SRN: PES2UG23CS485

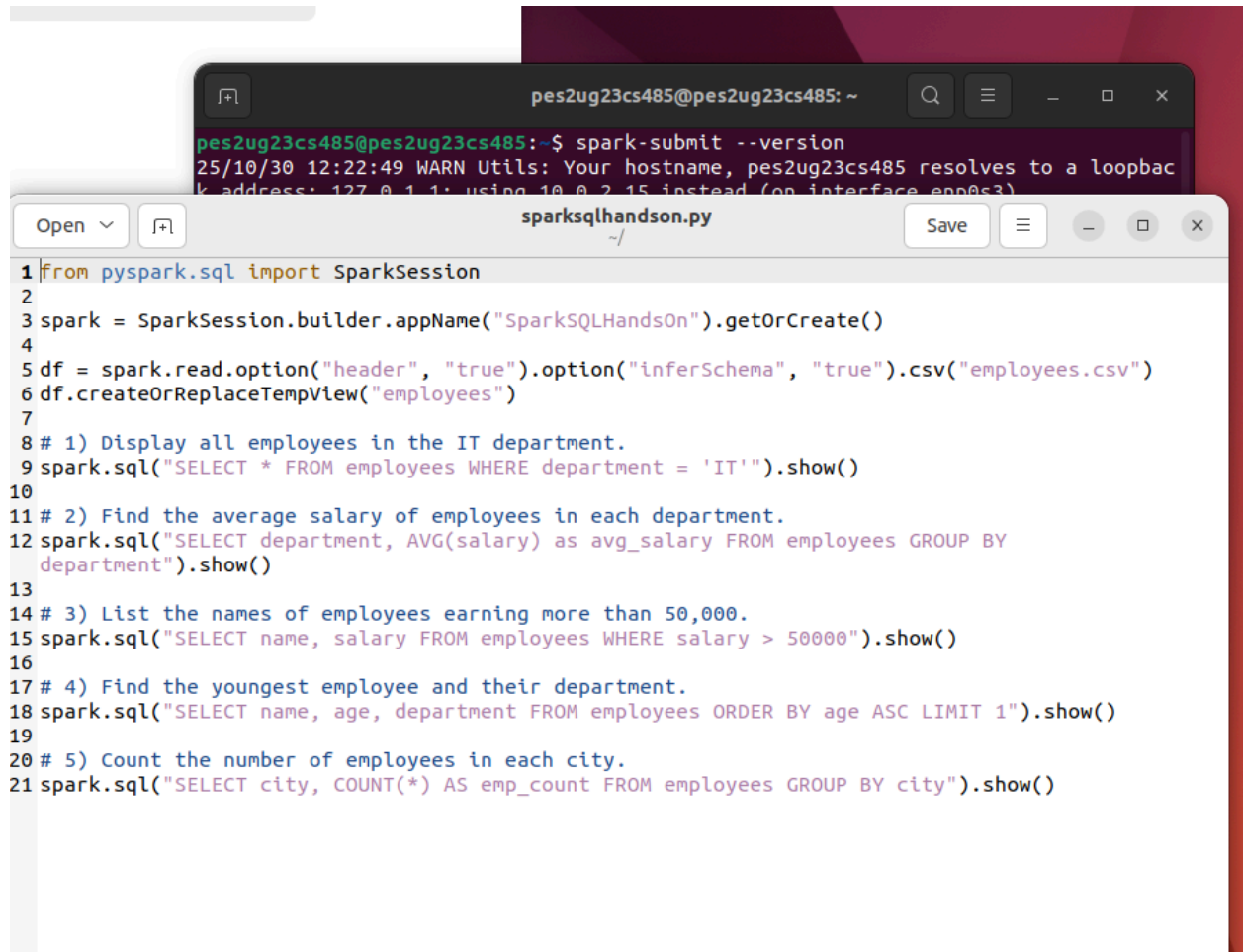
Class: 5H

Screenshot 1

A terminal window with a dark purple background. The title bar shows the user 'pes2ug23cs485' and the host 'pes2ug23cs485'. The terminal displays the output of the 'spark-submit --version' command. It includes warning messages about hostname resolution and Spark local IP settings. A 'Welcome to' message is followed by the Spark logo (a stylized 'S' made of triangles) and the text 'version 3.5.3'. Below this, it lists the Scala version (2.12.18), OpenJDK version (64-Bit Server VM, 1.8.0_462), the branch (HEAD), the compiler user (haejeon.lee), the revision hash, and the GitHub URL for Apache Spark. The prompt returns to the user's shell.

```
pes2ug23cs485@pes2ug23cs485: ~  
pes2ug23cs485@pes2ug23cs485:~$ spark-submit --version  
25/10/30 12:22:49 WARN Utils: Your hostname, pes2ug23cs485 resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)  
25/10/30 12:22:49 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
Welcome to  
 version 3.5.3  
  
Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 1.8.0_462  
Branch HEAD  
Compiled by user haejeon.lee on 2024-09-09T05:20:05Z  
Revision 32232e9ed33bb16b93ad58cfde8b82e0f07c0970  
Url https://github.com/apache/spark  
Type --help for more information.  
pes2ug23cs485@pes2ug23cs485:~$
```

Screenshot 2



The image shows a terminal window and a code editor. The terminal window, titled 'pes2ug23cs485@pes2ug23cs485: ~', displays the output of the command 'spark-submit --version', which is '25/10/30 12:22:49 WARN Utils: Your hostname, pes2ug23cs485 resolves to a loopback address: 127.0.0.1; using 10.0.2.15 instead (on interface enp0s3)'. The code editor, titled 'sparksqllhandson.py', contains the following Python code:

```
1 from pyspark.sql import SparkSession
2
3 spark = SparkSession.builder.appName("SparkSQLHandsOn").getOrCreate()
4
5 df = spark.read.option("header", "true").option("inferSchema", "true").csv("employees.csv")
6 df.createOrReplaceTempView("employees")
7
8 # 1) Display all employees in the IT department.
9 spark.sql("SELECT * FROM employees WHERE department = 'IT']").show()
10
11 # 2) Find the average salary of employees in each department.
12 spark.sql("SELECT department, AVG(salary) as avg_salary FROM employees GROUP BY
13 department").show()
14
15 # 3) List the names of employees earning more than 50,000.
16 spark.sql("SELECT name, salary FROM employees WHERE salary > 50000").show()
17
18 # 4) Find the youngest employee and their department.
19 spark.sql("SELECT name, age, department FROM employees ORDER BY age ASC LIMIT 1").show()
20
21 # 5) Count the number of employees in each city.
22 spark.sql("SELECT city, COUNT(*) AS emp_count FROM employees GROUP BY city").show()
```

Screenshot 3

```

pool
25/10/30 13:41:11 INFO DAGScheduler: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0) finished in 0.199 s
25/10/30 13:41:11 INFO DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
25/10/30 13:41:11 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
25/10/30 13:41:11 INFO DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 0.222435 s
25/10/30 13:41:12 INFO CodeGenerator: Code generated in 19.457448 ms
+-----+
|emp_id|  name|age|department|salary|  city|
+-----+
|  102|  Bob| 35|         IT| 60000| Chennai|
|  104| David| 40|         IT| 80000| Bangalore|
|  107| Grace| 38|         IT| 70000| Mumbai|
|  110| Judy| 26|         IT| 52000| Chennai|
|  114| Nina| 41|         IT| 85000| Delhi|
|  118| Raj| 42|         IT| 90000| Bangalore|
|  121| Usha| 28|         IT| 61000| Hyderabad|
|  124|Xavier| 31|         IT| 56000| Chennai|
|  127| Amit| 34|         IT| 65000| Delhi|
|  131|Edward| 40|         IT| 88000| Hyderabad|
|  136| Jatin| 39|         IT| 77000| Mumbai|
|  139| Manoj| 42|         IT| 93000| Chennai|
|  144| Rohit| 40|         IT| 87000| Hyderabad|
|  147| Uday| 39|         IT| 81000| Pune|
|  150| Yusuf| 43|         IT| 95000| Delhi|
|  194| Rani| 30|         IT| 66000| Hyderabad|
+-----+
25/10/30 13:41:12 INFO FileSourceStrategy: Pushed Filters:
25/10/30 13:41:12 INFO FileSourceStrategy: Post-Scan Filters:
25/10/30 13:41:12 INFO CodeGenerator: Code generated in 64.895502 ms
25/10/30 13:41:12 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 350.2 KiB, free 364.8 MiB)
25/10/30 13:41:12 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 34.2 KiB, free 364.8 MiB)
25/10/30 13:41:12 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on 10.0.2.15:40477 (size: 34.2 KiB, free: 366.2 MiB)
25/10/30 13:41:12 INFO SparkContext: Created broadcast 6 from showString at NativeMethodAccessorImpl

```

Screenshot 4

```

pes2ug23cs485@pes2ug23cs485: ~
25/10/30 13:41:13 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 4) in 161 ms on 10.0.2.15
(executor driver) (1/1)
25/10/30 13:41:13 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from
pool
25/10/30 13:41:13 INFO DAGScheduler: ResultStage 5 (showString at NativeMethodAccessorImpl.java:0) f
inished in 0.223 s
25/10/30 13:41:13 INFO DAGScheduler: Job 4 is finished. Cancelling potential speculative or zombie t
asks for this job
25/10/30 13:41:13 INFO TaskSchedulerImpl: Killing all running tasks in stage 5: Stage finished
25/10/30 13:41:13 INFO DAGScheduler: Job 4 finished: showString at NativeMethodAccessorImpl.java:0,
took 0.263563 s
25/10/30 13:41:13 INFO CodeGenerator: Code generated in 16.745358 ms
+-----+
|department|      avg_salary|
+-----+
|    Sales|60416.666666666664|
|      HR| 44363.63636363636|
|   Finance| 68666.66666666667|
|     Admin| 45090.90909090909|
|Marketing|      56300.0|
|       IT|      75375.0|
|   Support| 47272.72727272727|
|      R&D| 74727.27272727272|
+-----+
25/10/30 13:41:13 INFO FileSourceStrategy: Pushed Filters: IsNotNull(salary),GreaterThan(salary,5000
0)
25/10/30 13:41:13 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(salary#21),(salary#21 > 5000
0)
25/10/30 13:41:13 INFO CodeGenerator: Code generated in 16.826947 ms
25/10/30 13:41:13 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 350
.2 KiB, free 364.3 MiB)
25/10/30 13:41:13 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated si
ze 34.2 KiB, free 364.3 MiB)
25/10/30 13:41:13 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on 10.0.2.15:40477 (size
: 34.2 KiB, free: 366.1 MiB)
25/10/30 13:41:13 INFO SparkContext: Created broadcast 9 from showString at NativeMethodAccessorImpl
.java:0
25/10/30 13:41:13 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes,
open cost is considered as scanning 4194304 bytes.
(estimated size 34.2 KiB, free 364.2 MiB)

```

Screenshot 5

```
ge: 0-3310, partition values: [empty row]
25/10/30 13:41:13 INFO CodeGenerator: Code generated in 19.753051 ms
25/10/30 13:41:13 INFO Executor: Finished task 0.0 in stage 6.0 (TID 5). 1883 bytes result sent to driver
25/10/30 13:41:13 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 5) in 70 ms on 10.0.2.15 (executor driver) (1/1)
25/10/30 13:41:13 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
25/10/30 13:41:13 INFO DAGScheduler: ResultStage 6 (showString at NativeMethodAccessorImpl.java:0) finished in 0.129 s
25/10/30 13:41:13 INFO DAGScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
25/10/30 13:41:13 INFO TaskSchedulerImpl: Killing all running tasks in stage 6: Stage finished
25/10/30 13:41:13 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 0.176788 s
+-----+-----+
| name|salary|
+-----+-----+
| Bob| 60000|
| David| 80000|
| Frank| 55000|
| Grace| 70000|
| Ian| 90000|
| Judy| 52000|
| Linda| 51000|
| Mark| 65000|
| Nina| 85000|
| Paul| 72000|
| Queen| 55000|
| Raj| 90000|
| Sara| 60000|
| Tom| 58000|
| Usha| 61000|
| Vinod| 95000|
| Xavier| 56000|
| Zara| 72000|
| Amit| 65000|
| Beena| 60000|
+-----+-----+
only showing top 20 rows
```

Screenshot 6

```
pes2ug23cs485@pes2ug23cs485: ~
25/10/30 13:41:14 INFO TaskSchedulerImpl: Adding task set 7.0 with 1 tasks resource profile 0
25/10/30 13:41:14 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 6) (10.0.2.15, executor driver, partition 0, PROCESS_LOCAL, 9602 bytes)
25/10/30 13:41:14 INFO Executor: Running task 0.0 in stage 7.0 (TID 6)
25/10/30 13:41:14 INFO CodeGenerator: Code generated in 8.78188 ms
25/10/30 13:41:14 INFO FileScanRDD: Reading File path: file:///home/pes2ug23cs485/employees.csv, range: 0-3310, partition values: [empty row]
25/10/30 13:41:14 INFO CodeGenerator: Code generated in 14.329719 ms
25/10/30 13:41:14 INFO Executor: Finished task 0.0 in stage 7.0 (TID 6). 1495 bytes result sent to driver
25/10/30 13:41:14 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 6) in 123 ms on 10.0.2.15 (executor driver) (1/1)
25/10/30 13:41:14 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
25/10/30 13:41:14 INFO DAGScheduler: ResultStage 7 (showString at NativeMethodAccessorImpl.java:0) finished in 0.156 s
25/10/30 13:41:14 INFO DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job
25/10/30 13:41:14 INFO TaskSchedulerImpl: Killing all running tasks in stage 7: Stage finished
25/10/30 13:41:14 INFO DAGScheduler: Job 6 finished: showString at NativeMethodAccessorImpl.java:0, took 0.198422 s
25/10/30 13:41:14 INFO CodeGenerator: Code generated in 11.708225 ms
25/10/30 13:41:14 INFO CodeGenerator: Code generated in 11.851345 ms
+-----+
|name|age|department|
+-----+
|Judy| 26|          IT|
+-----+

25/10/30 13:41:14 INFO FileSourceStrategy: Pushed Filters:
25/10/30 13:41:14 INFO FileSourceStrategy: Post-Scan Filters:
25/10/30 13:41:14 INFO CodeGenerator: Code generated in 35.920951 ms
25/10/30 13:41:14 INFO MemoryStore: Block broadcast_13 stored as values in memory (estimated size 350.2 KiB, free 363.5 MiB)
25/10/30 13:41:14 INFO MemoryStore: Block broadcast_13_piece0 stored as bytes in memory (estimated size 34.2 KiB, free 363.5 MiB)
25/10/30 13:41:14 INFO BlockManagerInfo: Added broadcast_13_piece0 in memory on 10.0.2.15:40477 (size: 34.2 KiB, free: 366.0 MiB)
25/10/30 13:41:14 INFO SparkContext: Created broadcast 13 from showString at NativeMethodAccessorImpl.java:0
```

Screenshot 7

```
pes2ug23cs485@pes2ug23cs485: ~  
25/10/30 13:41:14 INFO DAGScheduler: ResultStage 10 (showString at NativeMethodAccessorImpl.java:0)  
finished in 0.183 s  
25/10/30 13:41:14 INFO DAGScheduler: Job 8 is finished. Cancelling potential speculative or zombie t  
asks for this job  
25/10/30 13:41:14 INFO TaskSchedulerImpl: Killing all running tasks in stage 10: Stage finished  
25/10/30 13:41:14 INFO DAGScheduler: Job 8 finished: showString at NativeMethodAccessorImpl.java:0,  
took 0.197313 s  
+-----+  
|      city|emp_count|  
+-----+  
|Bangalore|      20|  
|Chennai|    16|  
|Mumbai|    18|  
|Pune|      11|  
|Delhi|     20|  
|Hyderabad|  15|  
+-----+  
25/10/30 13:41:15 INFO BlockManagerInfo: Removed broadcast_10_piece0 on 10.0.2.15:40477 in memory (s  
ize: 7.8 KiB, free: 366.0 MiB)  
25/10/30 13:41:15 INFO BlockManagerInfo: Removed broadcast_7_piece0 on 10.0.2.15:40477 in memory (si  
ze: 20.1 KiB, free: 366.1 MiB)  
25/10/30 13:41:15 INFO BlockManagerInfo: Removed broadcast_8_piece0 on 10.0.2.15:40477 in memory (si  
ze: 21.4 KiB, free: 366.1 MiB)  
25/10/30 13:41:15 INFO SparkContext: Invoking stop() from shutdown hook  
25/10/30 13:41:15 INFO SparkContext: SparkContext is stopping with exitCode 0.  
25/10/30 13:41:15 INFO BlockManagerInfo: Removed broadcast_12_piece0 on 10.0.2.15:40477 in memory (s  
ize: 7.3 KiB, free: 366.1 MiB)  
25/10/30 13:41:15 INFO BlockManagerInfo: Removed broadcast_9_piece0 on 10.0.2.15:40477 in memory (si  
ze: 34.2 KiB, free: 366.1 MiB)  
25/10/30 13:41:15 INFO BlockManagerInfo: Removed broadcast_4_piece0 on 10.0.2.15:40477 in memory (si  
ze: 34.2 KiB, free: 366.1 MiB)  
25/10/30 13:41:15 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040  
25/10/30 13:41:15 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!  
25/10/30 13:41:15 INFO MemoryStore: MemoryStore cleared  
25/10/30 13:41:15 INFO BlockManager: BlockManager stopped  
25/10/30 13:41:15 INFO BlockManagerMaster: BlockManagerMaster stopped  
25/10/30 13:41:15 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordina  
tor stopped!  
25/10/30 13:41:15 INFO SparkContext: Successfully stopped SparkContext  
as shown above)
```

Screenshot 8


```
pes2ug23cs485@pes2ug23cs485: ~  
Mumbai| 18|  
Pune| 11|  
  
Open sparkstream.py Save  
1 from pyspark.sql import SparkSession  
2 from pyspark.sql.functions import split, window, col  
3  
4 # Start Spark  
5 spark = SparkSession.builder.appName("StreamingHandsOn").getOrCreate()  
6  
7 # Read stream from socket  
8 lines = spark.readStream.format("socket").option("host", "localhost").option("port",  
9 9999).load()  
10  
11 # Split CSV fields  
12 columns = split(lines.value, ",")  
13 df = lines.select(  
14     columns.getItem(0).alias("timestamp"),  
15     columns.getItem(1).alias("userId"),  
16     columns.getItem(2).alias("action"),  
17     columns.getItem(3).alias("page")  
18 )  
19 # Query 1: Count total clicks per page  
20 page_counts = df.groupBy("page").count()  
21  
22 # Query 2: Count clicks per page in 10-second windows  
23 time_window_counts = df.groupBy(  
24     window(col("timestamp"), "10 seconds"),  
25     col("page")  
26 ).count()  
27  
28 # Output to console  
29 '''  
30 query = page_counts.writeStream.outputMode("complete").format("console").start()  
31 query.awaitTermination()  
32 '''  
33  
34 query = time_window_counts.writeStream.outputMode("complete").format("console").start()  
35 query.awaitTermination()  
  
Python 2 Tab Width: 8 Ln 1, Col 1 INS
```

Screenshot 9


```
asks for this job
25/10/30 14:41:36 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
25/10/30 14:41:36 INFO DAGScheduler: Job 1 finished: start at NativeMethodAccessorImpl.java:0, took
51.522434 s
25/10/30 14:41:36 INFO WriteToDataSourceV2Exec: Data source write support MicroBatchWrite[epoch: 1,
writer: ConsoleWriter[numRows=20, truncate=true]] is committing.
-----
Batch: 1
-----
25/10/30 14:41:36 INFO CodeGenerator: Code generated in 10.737162 ms
25/10/30 14:41:37 INFO BlockManagerInfo: Removed broadcast_7_piece0 on 10.0.2.15:45243 in memory (si
ze: 32.4 KiB, free: 366.2 MiB)
25/10/30 14:41:39 INFO CodeGenerator: Code generated in 17.397676 ms
+-----+-----+
|      page|count|
+-----+-----+
| products|   27|
| homepage|   54|
|      cart|   19|
+-----+-----+

25/10/30 14:41:39 INFO WriteToDataSourceV2Exec: Data source write support MicroBatchWrite[epoch: 1,
writer: ConsoleWriter[numRows=20, truncate=true]] committed.
25/10/30 14:41:39 INFO CheckpointFileManager: Writing atomically to file:/tmp/temporary-7199d270-a75
3-4661-94f0-31cfe3160802/commits/1 using temp file file:/tmp/temporary-7199d270-a753-4661-94f0-31cfe
3160802/commits/.1.9c1382d7-f2df-4d07-870d-0b6f16681bea.tmp
25/10/30 14:41:41 INFO CheckpointFileManager: Renamed temp file file:/tmp/temporary-7199d270-a753-46
61-94f0-31cfe3160802/commits/.1.9c1382d7-f2df-4d07-870d-0b6f16681bea.tmp to file:/tmp/temporary-7199
d270-a753-4661-94f0-31cfe3160802/commits/1
25/10/30 14:41:41 INFO MicroBatchExecution: Streaming query made progress: {
  "id" : "23cf957d-4d19-4ae3-b5ea-8f899fc218dc",
  "runId" : "f1e7580c-af85-446b-aeb3-de6b68ce29c7",
  "name" : null,
  "timestamp" : "2025-10-30T09:10:44.794Z",
  "batchId" : 1,
  "numInputRows" : 100,
  "inputRowsPerSecond" : 1.3413996163597095,
  "processedRowsPerSecond" : 1.7734584212673135,
  "durationMs" : {
    "addBatch" : 54982,
```

Screenshot 10

pes2ug23cs485@pes2ug23cs485: ~

pes2ug23cs485@pes2ug23cs485: ~

25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	StateStore:	Retrieved reference to StateStoreCoordinator:	org.apache.spark.sql.execution.streaming.state.StateStoreCoordinatorRef
25/10/30	14:18:25	INFO	CodeGenerator:	Code generated in	92.21746 ms

```

+-----+-----+
| window | page | count |
+-----+-----+
| {2025-10-01 09:01... | products | 3 |
| {2025-10-01 09:00... | cart | 1 |
| {2025-10-01 09:00... | cart | 3 |
| {2025-10-01 09:00... | products | 3 |
| {2025-10-01 09:01... | homepage | 6 |
| {2025-10-01 09:00... | cart | 2 |
| {2025-10-01 09:01... | cart | 2 |
| {2025-10-01 09:00... | homepage | 4 |
| {2025-10-01 09:00... | homepage | 5 |
| {2025-10-01 09:00... | cart | 2 |
| {2025-10-01 09:00... | products | 2 |
| {2025-10-01 09:01... | homepage | 5 |
| {2025-10-01 09:01... | cart | 1 |
| {2025-10-01 09:01... | homepage | 1 |
| {2025-10-01 09:01... | homepage | 5 |
| {2025-10-01 09:01... | homepage | 6 |
| {2025-10-01 09:01... | cart | 2 |
| {2025-10-01 09:01... | products | 3 |
| {2025-10-01 09:01... | cart | 2 |
| {2025-10-01 09:00... | products | 3 |
+-----+-----+
only showing top 20 rows

```