# Week – 4 PySpark ETL Assignment

Piyush Kumar Yadav

## To perform:

**PySpark ETL Pipeline**

**Goal:**

Align with the Python work to process large sales data using PySpark and generate enriched insights.

**Tasks:**

Read raw CSVs from HDFS or local folder.

Transform and clean the data (handle missing values, duplicates, etc.).

Enrich data by calculating KPIs such as:

- Monthly Revenue
- Profit Margin (%)
- Region-wise Sales
- Average Order Value
- (Come up with 3–4 more KPIs as needed)

Write aggregated results to Parquet or a managed table.

(Optional) Integrate with Kafka for streaming order ingestion.

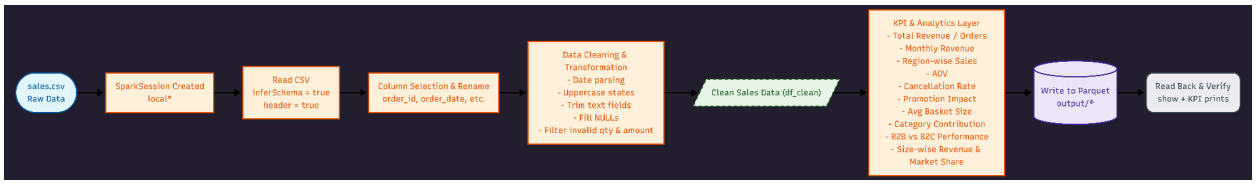
**Tech:** PySpark, HDFS, Kafka (optional)

**Deliverables:** PySpark script or notebook + pipeline diagram + screenshots (Spark job output, DAG view, Parquet output sample)

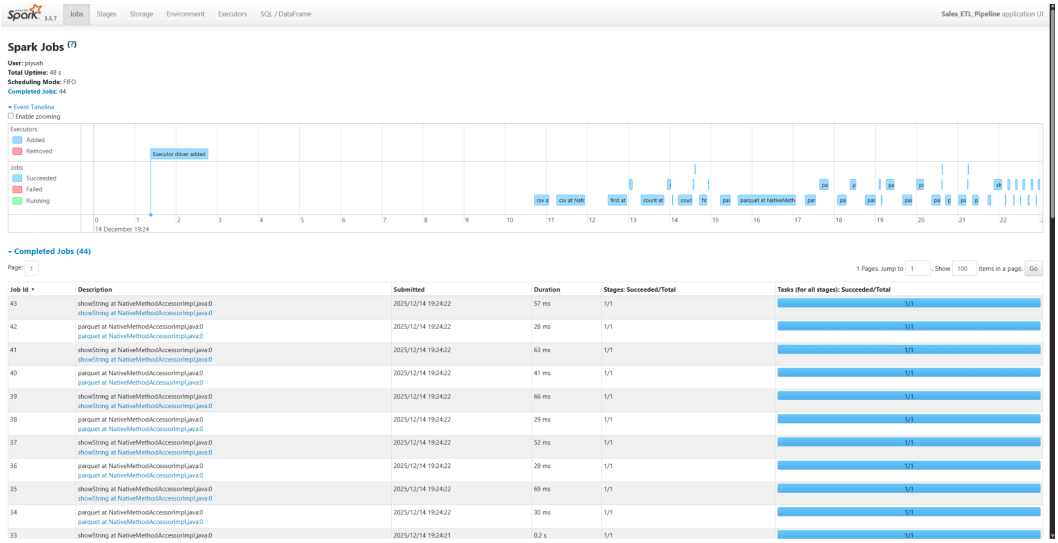
Link to .py: [Piyush_PySpark_ETL_assignment_week4.py]

Or :
https://wgcp-my.sharepoint.com/:u:/g/personal/piyush_kumar_blend360_com/IQCyiGq8REuBQpt2LD1GvF3SAWKxwy4N-yAgdjhzdc4TCJI?e=lB0Msh

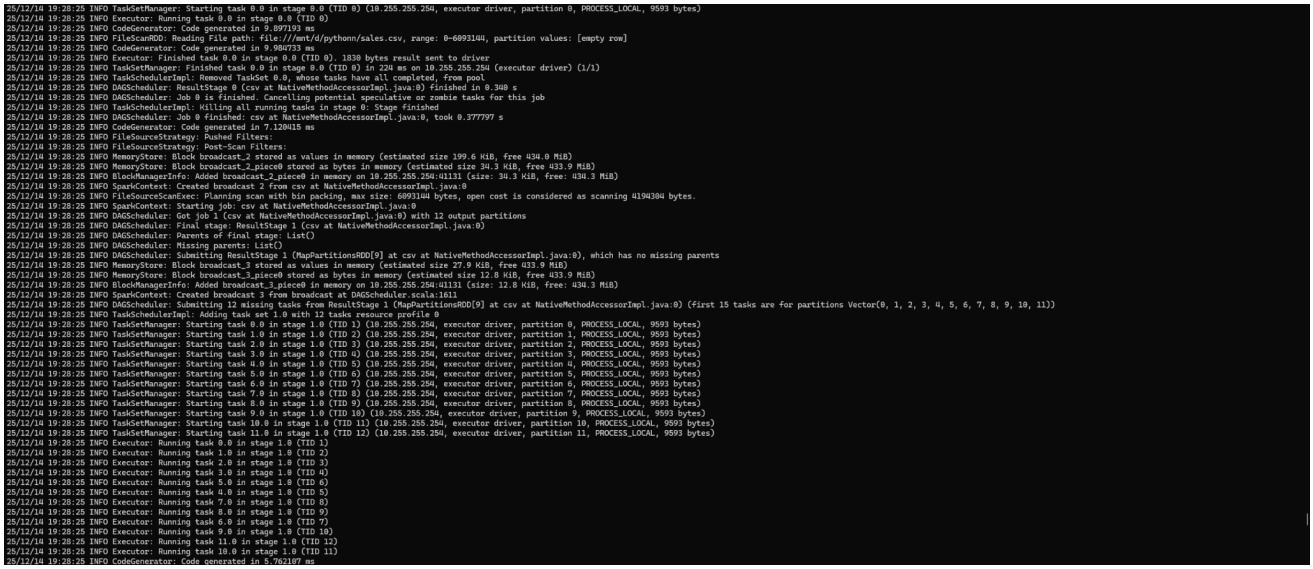
Pipeline:

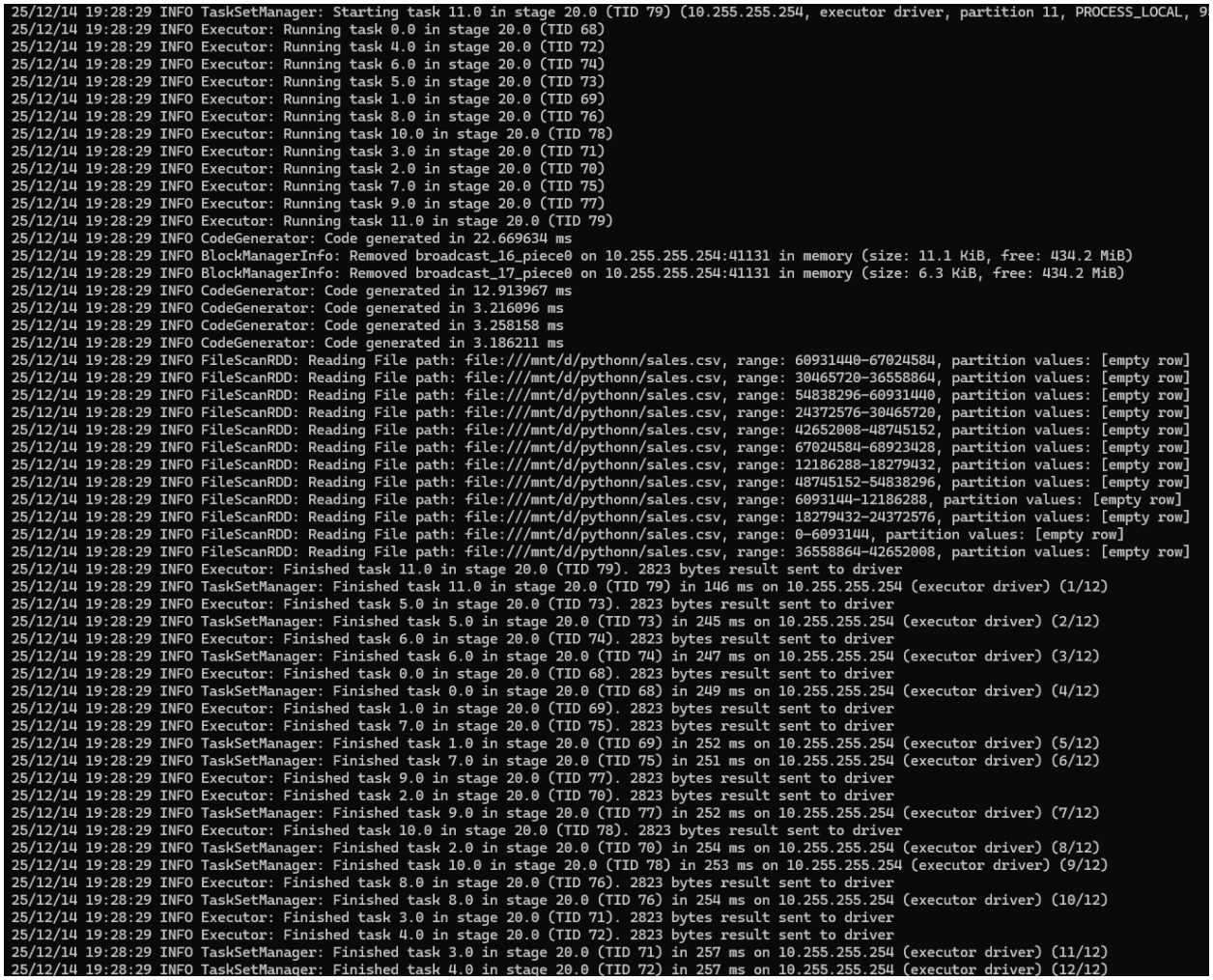


Utilizing PySpark with csv from the local folder, and using a PySpark script to perform operations and aggregating results into parquet output file(s).

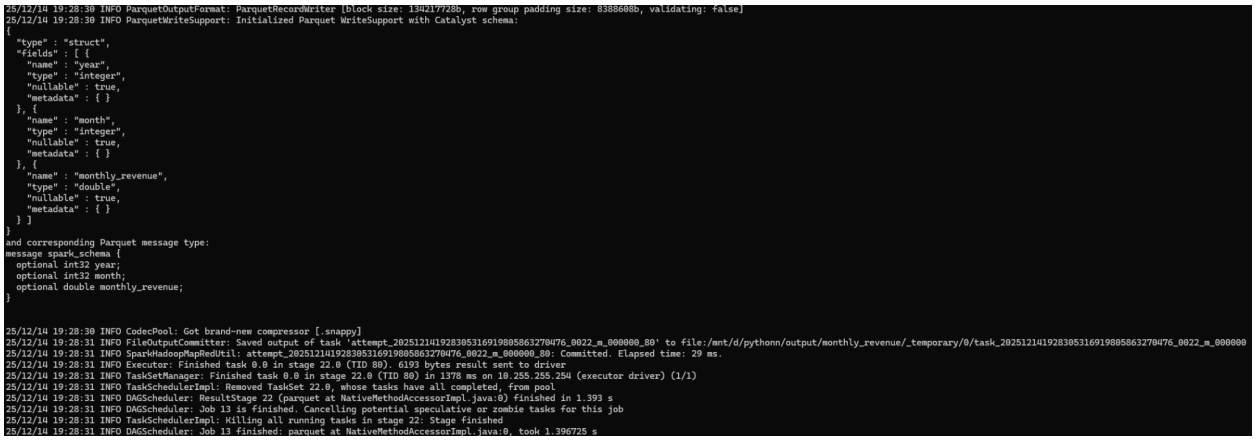


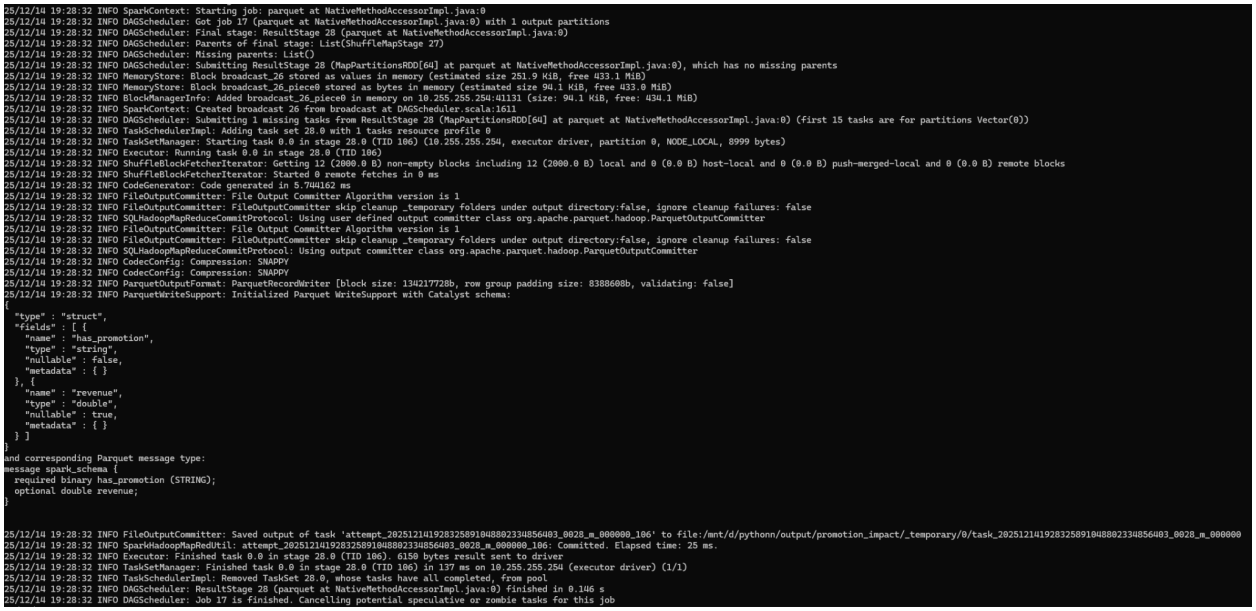
Spark job execution(write,.show()) and KPI results:



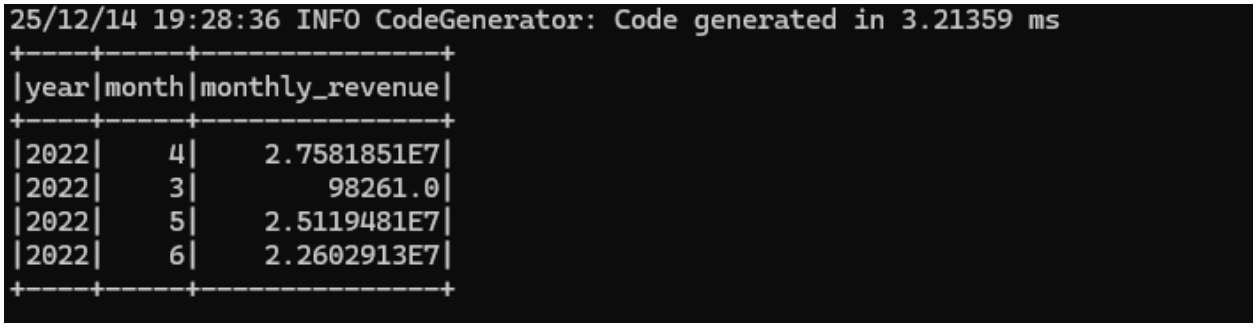



-parquet write (writing aggregate results)




-show() (reading parquet files)



```
25/12/14 19:28:36 INFO CodeGenerator: Code generated in 3.208849 ms
+----------------+-------------+
|           state|state_revenue|
+----------------+-------------+
| DADRA AND NAGAR|      39276.0|
|          SIKKIM|     134847.0|
|       MEGHALAYA|     111234.0|
|              NL|        961.0|
|     WEST BENGAL|    3357170.0|
|       NEW DELHI|      45609.0|
|             GOA|     622220.0|
|    CHHATTISGARH|     541811.0|
|       RAJASTHAN|    1680195.0|
|             NULL|      17531.0|
|         TRIPURA|      86799.0|
|           DELHI|    4185475.0|
|HIMACHAL PRADESH|     470548.0|
|         GUJARAT|    2618903.0|
|           BIHAR|    1351529.0|
|      CHANDIGARH|     203354.0|
|       KARNATAKA|    1.01531E7|
| JAMMU & KASHMIR|     432712.0|
|   UTTAR PRADESH|    6494393.0|
|          MANIPUR|     210922.0|
+----------------+-------------+
only showing top 20 rows
```

```
25/12/14 19:28:36 INFO DAGScheduler: Job 37 finished: showString at NativeMethodAccessorImpl.java:0, took 0.054281 s
+-------------+-----------+
| has_promotion|    revenue|
+-------------+-----------+
|Promotion Used|5.3588435E7|
|  No Promotion|2.1814071E7|
```

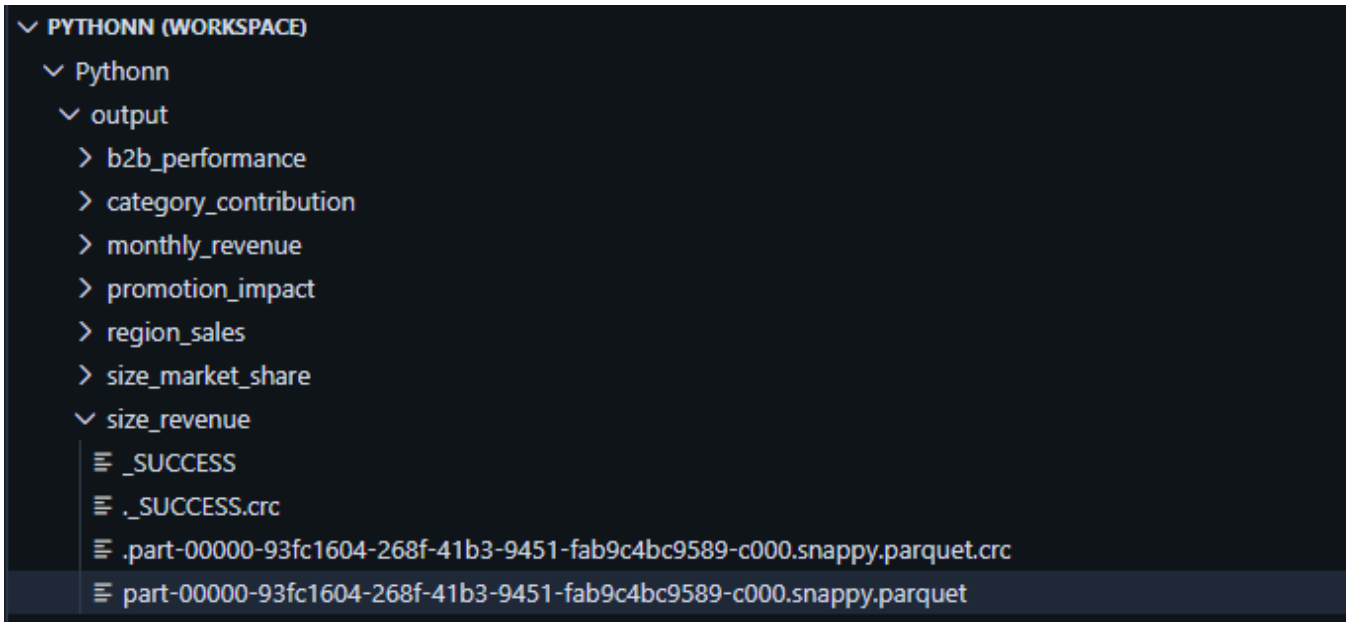
```
25/12/14 19:28:36 INFO DAGScheduler: Job 39 finished: showString at NativeMethodAccessorImpl.java:0, took 0.070654 s
+------------+----------------+------------------+
|     category|category_revenue|category_percentage|
+------------+----------------+------------------+
|         Set|     3.7662424E7|             49.95|
|       kurta|     2.0452141E7|             27.12|
|Western Dress|     1.0629096E7|              14.1|
|         Top|       5203733.0|               6.9|
| Ethnic Dress|        760711.0|              1.01|
|      Blouse|        434751.0|              0.58|
|      Bottom|        140226.0|              0.19|
|       Saree|        118509.0|              0.16|
|     Dupatta|           915.0|               0.0|
+------------+----------------+------------------+
```

```
25/12/14 19:28:37 INFO DAGScheduler: Job 41 finished: showString at NativeMethodAccessorImpl.java:0, took 0.077308 s
+------+-----------+----------------+
|is_b2b|    revenue|market_share_pct|
+------+-----------+----------------+
|  true|   579930.0|            0.77|
| false|7.4822576E7|           99.23|
```

```
25/12/14 19:28:37 INFO Executor: Running task 0.0 in stage 67.0 (TID 176)
25/12/14 19:28:37 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/output/size_market_share/part-00000-9b942c4b-ea73-4568-8540-4f0d59d10c37-c000.snappy
25/12/14 19:28:37 INFO Executor: Finished task 0.0 in stage 67.0 (TID 176). 2007 bytes result sent to driver
25/12/14 19:28:37 INFO TaskSetManager: Finished task 0.0 in stage 67.0 (TID 176) in 49 ms on 10.255.255.254 (executor driver) (1/1)
25/12/14 19:28:37 INFO TaskSchedulerImpl: Removed TaskSet 67.0, whose tasks have all completed, from pool
25/12/14 19:28:37 INFO DAGScheduler: ResultStage 67 (showString at NativeMethodAccessorImpl.java:0) finished in 0.052 s
25/12/14 19:28:37 INFO DAGScheduler: Job 43 is finished. Cancelling potential speculative or zombie tasks for this job
25/12/14 19:28:37 INFO TaskSchedulerImpl: Killing all running tasks in stage 67: Stage finished
25/12/14 19:28:37 INFO DAGScheduler: Job 43 finished: showString at NativeMethodAccessorImpl.java:0, took 0.053264 s
+----+------------+----------------+
|Size|size_revenue|market_share_pct|
+----+------------+----------------+
|   M| 1.3302545E7|           17.64|
|   L| 1.2684744E7|           16.82|
|  XL| 1.1917813E7|           15.81|
| XXL| 1.0258118E7|            13.6|
|   S| 1.0186895E7|           13.51|
| 3XL|   8795455.0|           11.66|
|  XS|   6759023.0|            8.96|
| 6XL|    562718.0|            0.75|
| 5XL|    414809.0|            0.55|
| 4XL|    325318.0|            0.43|
|Free|    195068.0|            0.26|
+----+------------+----------------+

AOV: 711.67
Cancellation Rate: 16.22%
Average Basket Size: 1.08
```

Output directory structure:

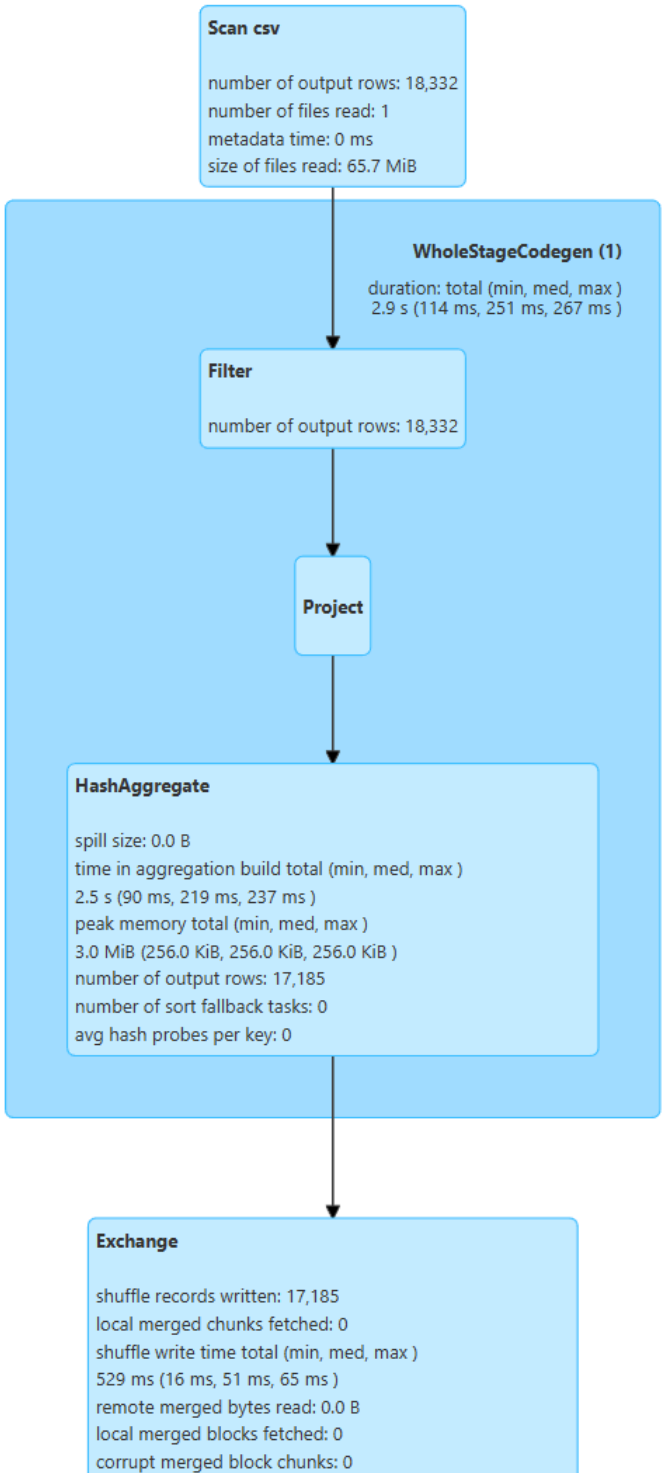


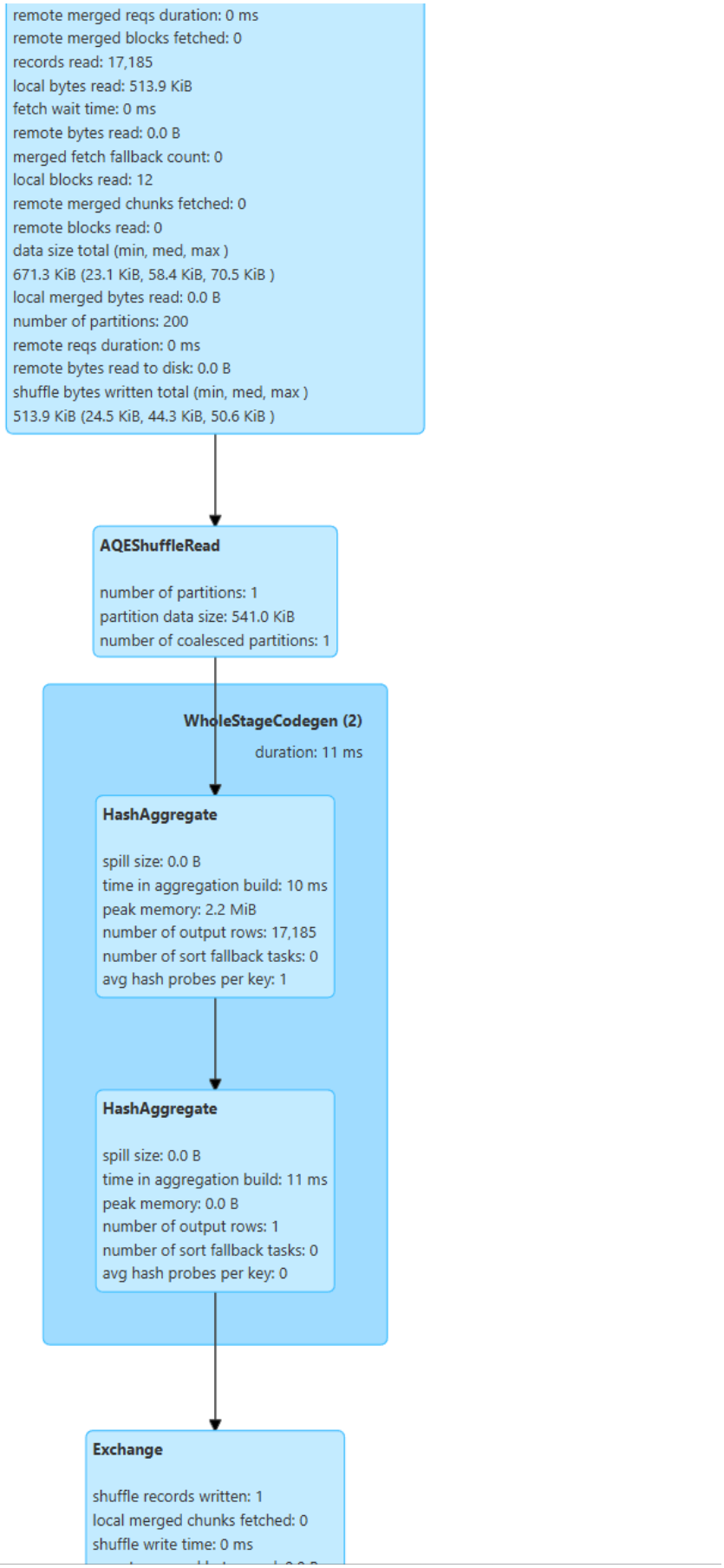

Sample query exec (DAG) view: (next page)

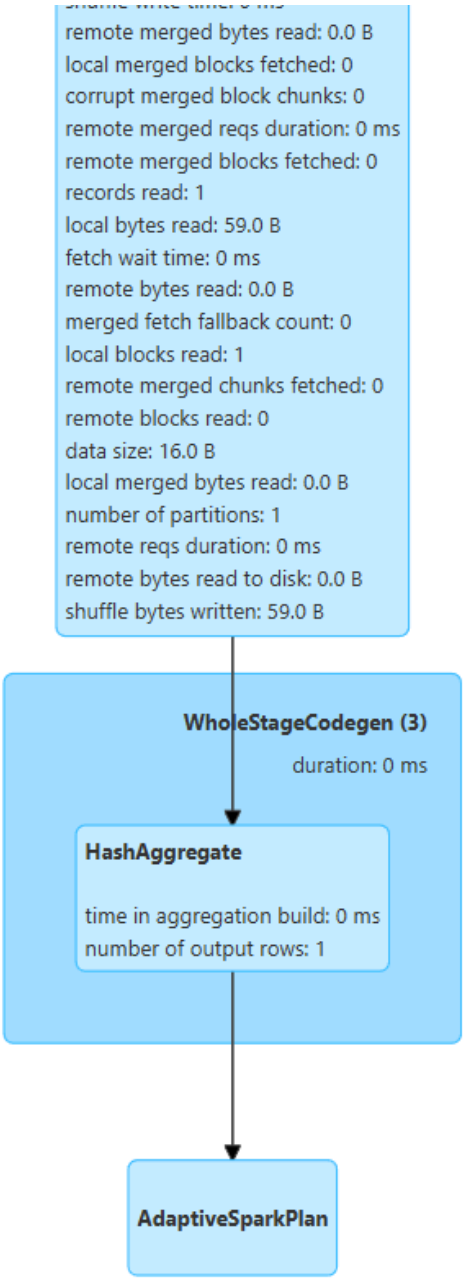
# Details for Query 3

**Submitted Time:** 2025/12/14 19:28:28
**Duration:** 0.5 s
**Succeeded Jobs:** 7 8 9

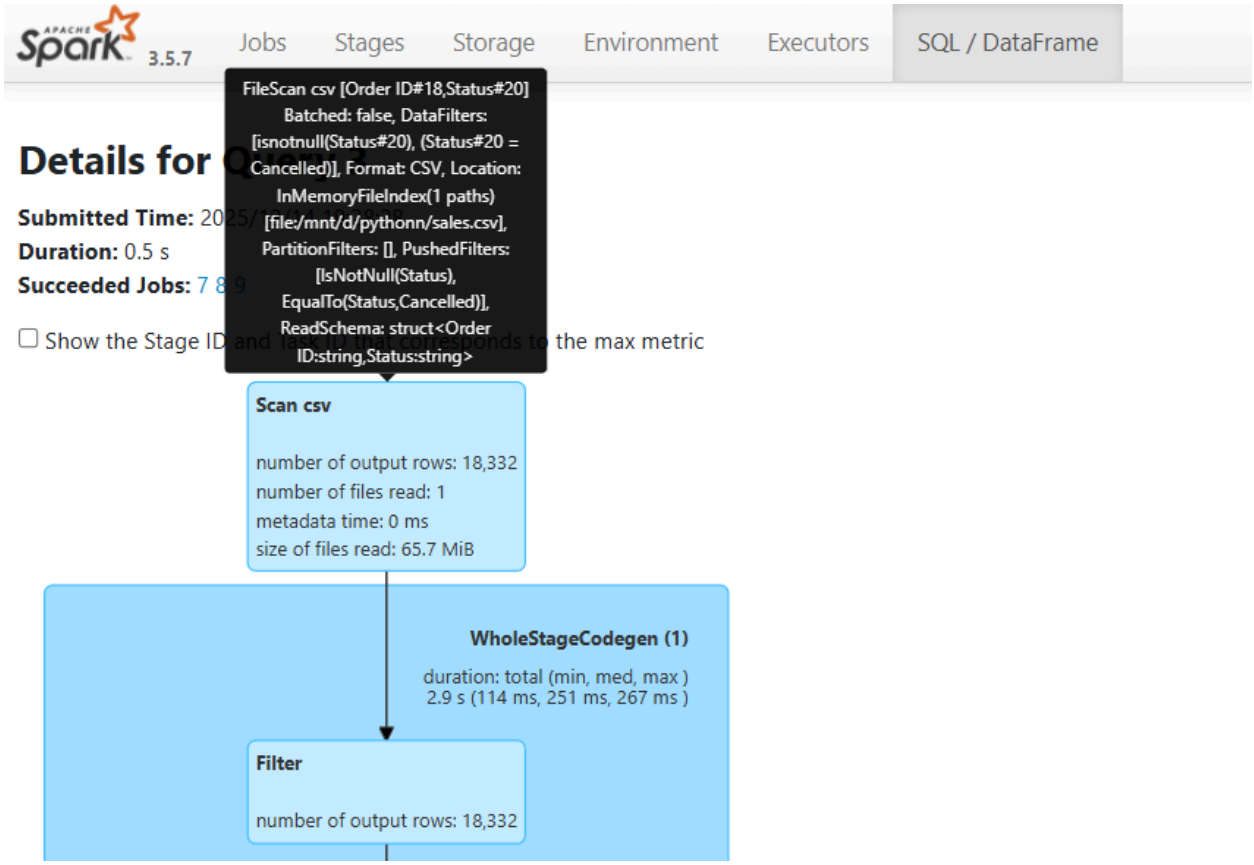
☐ Show the Stage ID and Task ID that corresponds to the max metric



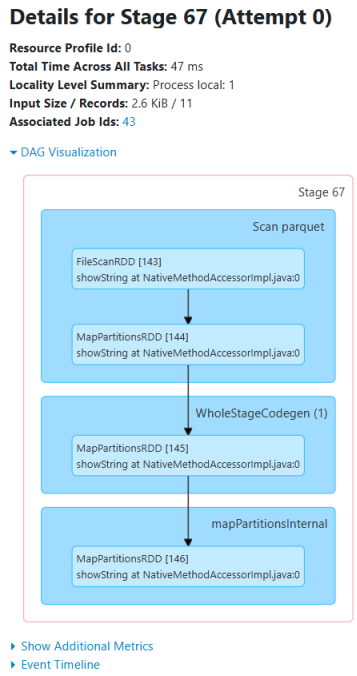

remote merged reqs duration: 0 ms
remote merged blocks fetched: 0
records read: 17,185
local bytes read: 513.9 KiB
fetch wait time: 0 ms
remote bytes read: 0.0 B
merged fetch fallback count: 0
local blocks read: 12
remote merged chunks fetched: 0
remote blocks read: 0
data size total (min, med, max )
671.3 KiB (23.1 KiB, 58.4 KiB, 70.5 KiB )
local merged bytes read: 0.0 B
number of partitions: 200
remote reqs duration: 0 ms
remote bytes read to disk: 0.0 B
shuffle bytes written total (min, med, max )
513.9 KiB (24.5 KiB, 44.3 KiB, 50.6 KiB )
AQEShuffleRead
number of partitions: 1
partition data size: 541.0 KiB
number of coalesced partitions: 1
WholeStageCodegen (2)
duration: 11 ms
HashAggregate
spill size: 0.0 B
time in aggregation build: 10 ms
peak memory: 2.2 MiB
number of output rows: 17,185
number of sort fallback tasks: 0
avg hash probes per key: 1
HashAggregate
spill size: 0.0 B
time in aggregation build: 11 ms
peak memory: 0.0 B
number of output rows: 1
number of sort fallback tasks: 0
avg hash probes per key: 0
Exchange
shuffle records written: 1
local merged chunks fetched: 0
shuffle write time: 0 ms

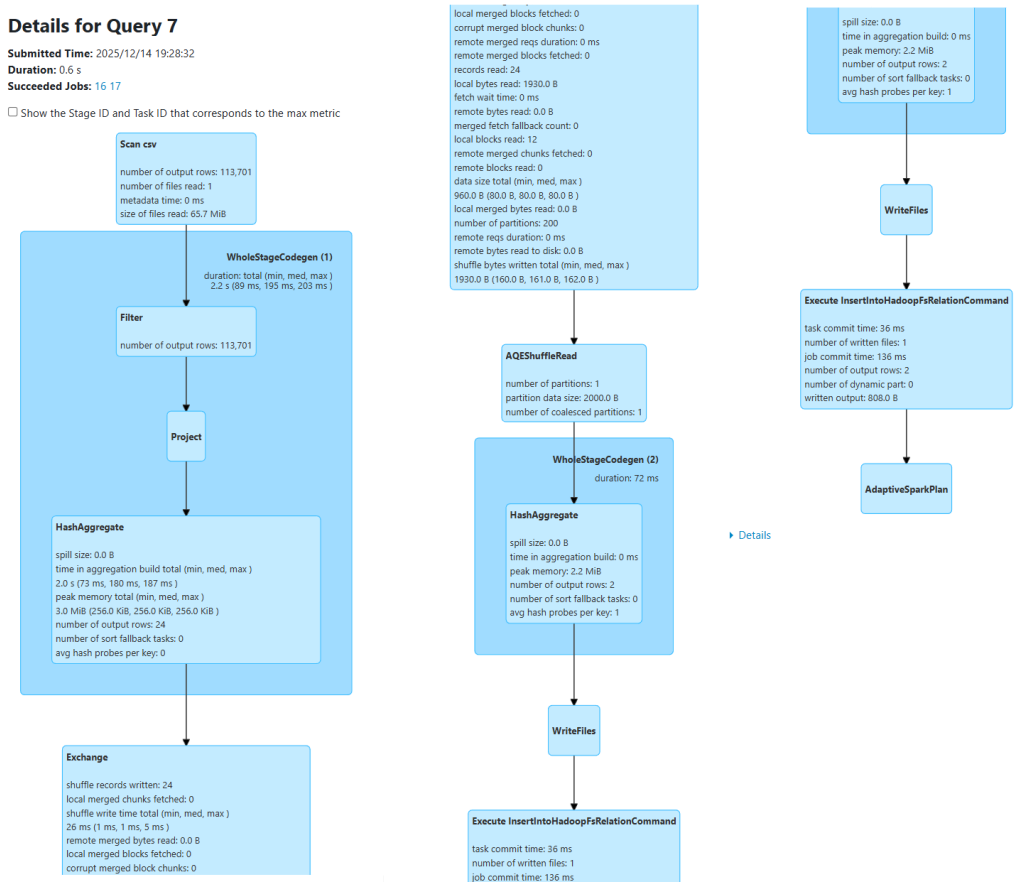

remote merged bytes read: 0.0 B
local merged blocks fetched: 0
corrupt merged block chunks: 0
remote merged reqs duration: 0 ms
remote merged blocks fetched: 0
records read: 1
local bytes read: 59.0 B
fetch wait time: 0 ms
remote bytes read: 0.0 B
merged fetch fallback count: 0
local blocks read: 1
remote merged chunks fetched: 0
remote blocks read: 0
data size: 16.0 B
local merged bytes read: 0.0 B
number of partitions: 1
remote reqs duration: 0 ms
remote bytes read to disk: 0.0 B
shuffle bytes written: 59.0 B
WholeStageCodegen (3)
duration: 0 ms
HashAggregate
time in aggregation build: 0 ms
number of output rows: 1
AdaptiveSparkPlan

Checking for cancelled orders in the depiction.

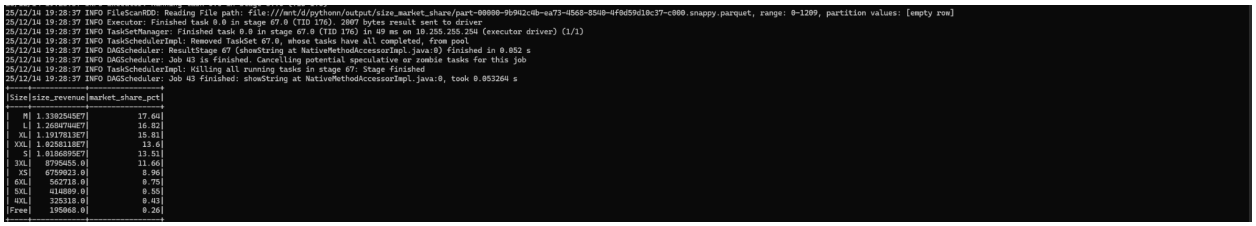


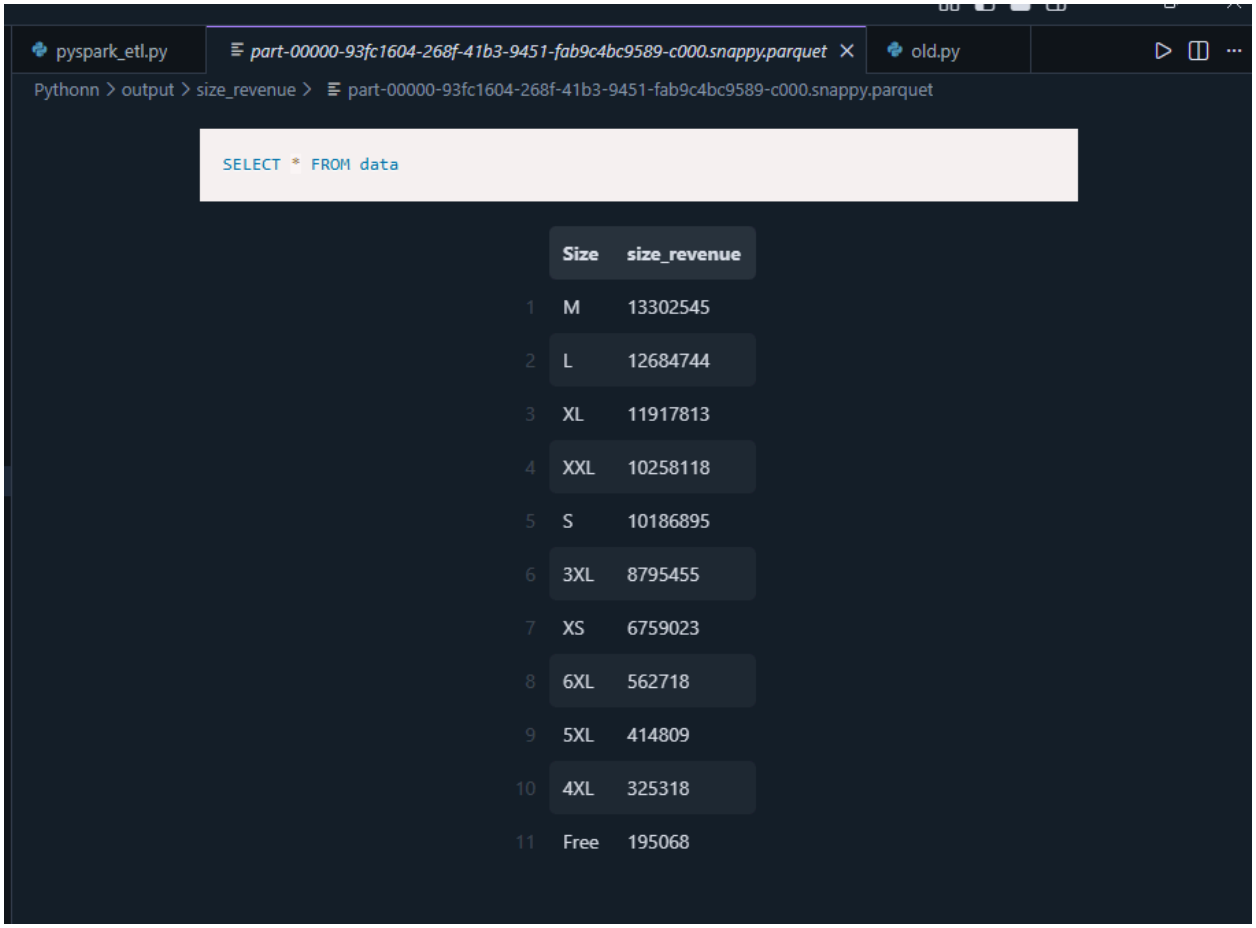

Job 43(sample read parquet and show):






## Parquet output sample (already included in spark job exec)




pyspark_etl.py
part-00000-93fc1604-268f-41b3-9451-fab9c4bc9589-c000.snappy.parquet
old.py
Pythonn > output > size_revenue > part-00000-93fc1604-268f-41b3-9451-fab9c4bc9589-c000.snappy.parquet
SELECT * FROM data
Size size_revenue
1 M 13302545
2 L 12684744
3 XL 11917813
4 XXL 10258118
5 S 10186895
6 3XL 8795455
7 XS 6759023
8 6XL 562718
9 5XL 414809
10 4XL 325318
11 Free 195068