

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)

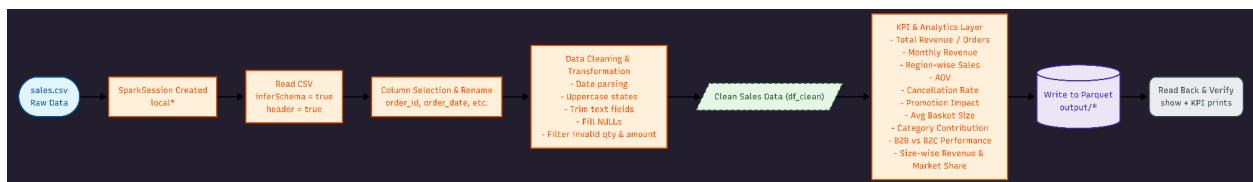
Deliverables: PySpark script or notebook + pipeline diagram + screenshots (Spark job output, DAG view, Parquet output sample)

Link to .py: [Piyush_PySpark_ETL_assignment_week4.py](#)

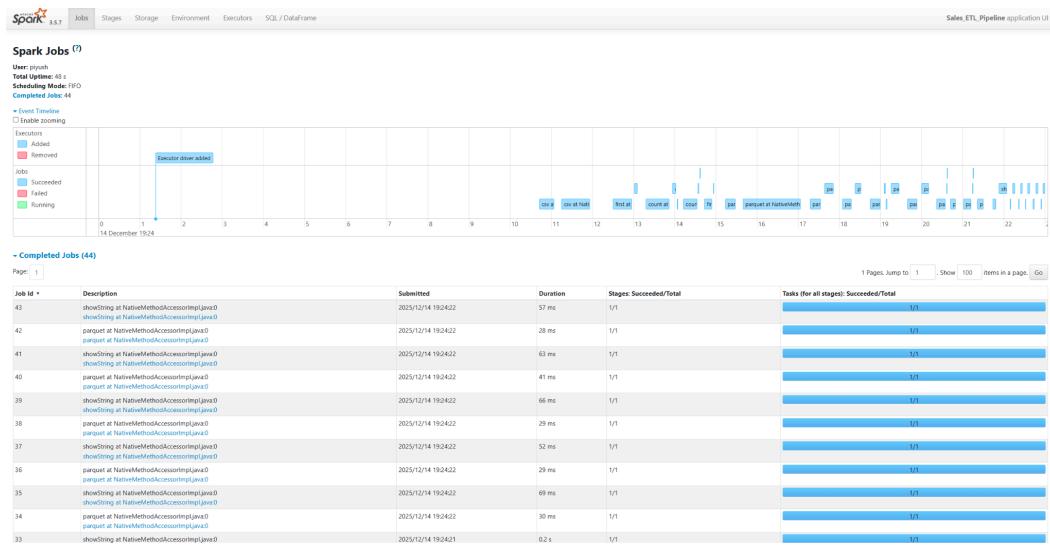
Or :

https://wgcp-my.sharepoint.com/:u/g/personal/piyush_kumar_blend360_com/IQCyiGq8REuBQpt2LD1GvF3SAWKxwy4N-yAgdjhzdc4TCJI?e=IB0Msh

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:

```

25/12/14 19:28:25 INFO TaskSetManager: Starting Task 0.0 in stage 0.0 (TID 0) (10.255.255.254, executor driver, partition 0, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
25/12/14 19:28:25 INFO CodeGenerator: Code generated in 9.897193 ms
25/12/14 19:28:25 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 0-6093104, partition values: [empty row]
25/12/14 19:28:25 INFO CodeGenerator: Code generated in 9.980733 ms
25/12/14 19:28:25 INFO Executor: Finished Task 0.0 in stage 0.0 (TID 0). 1830 bytes result sent to driver
25/12/14 19:28:25 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 146 ms on 10.255.255.254 (executor driver) (1/1)
25/12/14 19:28:25 INFO TaskScheduler: Removed TaskSet 0.0, whose tasks have all completed. From Partition 0
25/12/14 19:28:25 INFO DAGScheduler: ResultStage 0 (csv at NativeMethodAccessoImpl.java:8) finished in 0.3948 s
25/12/14 19:28:25 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/12/14 19:28:25 INFO TaskScheduler: Removed TaskSet 0.0, whose tasks have all completed. From Partition 0
25/12/14 19:28:25 INFO DAGScheduler: Job 0 finished: csv at NativeMethodAccessoImpl.java:8, took 0.377797 s
25/12/14 19:28:25 INFO CodeGenerator: Code generated in 7.128415 ms
25/12/14 19:28:25 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 0-6093104, partition values: [empty row]
25/12/14 19:28:25 INFO FileScannerStrategy: Post-Scan Filters:
25/12/14 19:28:25 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size: 199.6 KB, free: 434.0 MB)
25/12/14 19:28:25 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size: 199.6 KB, free: 433.9 MB)
25/12/14 19:28:25 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 10.255.255.254:41131 (size: 199.6 KB, free: 433.9 MB)
25/12/14 19:28:25 INFO SparkContext: Created broadcast 2 from csv at NativeMethodAccessoImpl.java:8
25/12/14 19:28:25 INFO FileScanRDD: Planning scan with bin packing, max size: 6697114 bytes, open cost is considered as scanning 14194384 bytes.
25/12/14 19:28:25 INFO DAGScheduler: Got job 1 (csv at NativeMethodAccessoImpl.java:8) with 12 output partitions
25/12/14 19:28:25 INFO DAGScheduler: Final stage: ResultStage 1 (csv at NativeMethodAccessoImpl.java:8)
25/12/14 19:28:25 INFO DAGScheduler: Missing parents: List()
25/12/14 19:28:25 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[9] at csv at NativeMethodAccessoImpl.java:8), which has no missing parents
25/12/14 19:28:25 INFO DAGScheduler: Removed broadcast_2 from memory (size: 199.6 KB, free: 433.9 MB)
25/12/14 19:28:25 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size: 12.9 KB, free: 433.9 MB)
25/12/14 19:28:25 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 10.255.255.254:41131 (size: 12.9 KB, free: 433.9 MB)
25/12/14 19:28:25 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1611
25/12/14 19:28:25 INFO DAGScheduler: Got job 2 (csv at NativeMethodAccessoImpl.java:8) with 12 output partitions
25/12/14 19:28:25 INFO DAGScheduler: Adding task 1.0 in stage 0 (TID 1) at csv at NativeMethodAccessoImpl.java:8
25/12/14 19:28:25 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1) (10.255.255.254, executor driver, partition 0, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
25/12/14 19:28:25 INFO CodeGenerator: Code generated in 5.762105 ms
25/12/14 19:28:25 INFO TaskSetManager: Starting task 2.0 in stage 1.0 (TID 2) (10.255.255.254, executor driver, partition 2, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 2.0 in stage 1.0 (TID 2)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 3.0 in stage 1.0 (TID 3) (10.255.255.254, executor driver, partition 3, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 3.0 in stage 1.0 (TID 3)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 4.0 in stage 1.0 (TID 4) (10.255.255.254, executor driver, partition 4, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 4.0 in stage 1.0 (TID 4)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 5.0 in stage 1.0 (TID 5) (10.255.255.254, executor driver, partition 5, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 5.0 in stage 1.0 (TID 5)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 6.0 in stage 1.0 (TID 6) (10.255.255.254, executor driver, partition 6, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 6.0 in stage 1.0 (TID 6)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 7.0 in stage 1.0 (TID 7) (10.255.255.254, executor driver, partition 7, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 7.0 in stage 1.0 (TID 7)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 8.0 in stage 1.0 (TID 8) (10.255.255.254, executor driver, partition 8, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 8.0 in stage 1.0 (TID 8)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 9.0 in stage 1.0 (TID 9) (10.255.255.254, executor driver, partition 9, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 9.0 in stage 1.0 (TID 9)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 10.0 in stage 1.0 (TID 10) (10.255.255.254, executor driver, partition 10, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 10.0 in stage 1.0 (TID 10)
25/12/14 19:28:25 INFO TaskSetManager: Starting task 11.0 in stage 1.0 (TID 11) (10.255.255.254, executor driver, partition 11, PROCESS_LOCAL, 9993 bytes)
25/12/14 19:28:25 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
25/12/14 19:28:25 INFO Executor: Running task 1.0 in stage 1.0 (TID 2)
25/12/14 19:28:25 INFO Executor: Running task 2.0 in stage 1.0 (TID 3)
25/12/14 19:28:25 INFO Executor: Running task 3.0 in stage 1.0 (TID 4)
25/12/14 19:28:25 INFO Executor: Running task 5.0 in stage 1.0 (TID 6)
25/12/14 19:28:25 INFO Executor: Running task 6.0 in stage 1.0 (TID 7)
25/12/14 19:28:25 INFO Executor: Running task 7.0 in stage 1.0 (TID 8)
25/12/14 19:28:25 INFO Executor: Running task 8.0 in stage 1.0 (TID 9)
25/12/14 19:28:25 INFO Executor: Running task 9.0 in stage 1.0 (TID 10)
25/12/14 19:28:25 INFO Executor: Running task 10.0 in stage 1.0 (TID 11)
25/12/14 19:28:25 INFO Executor: Running task 11.0 in stage 1.0 (TID 12)
25/12/14 19:28:25 INFO Executor: Running task 0.0 in stage 1.0 (TID 11)
25/12/14 19:28:25 INFO CodeGenerator: Code generated in 5.762105 ms

25/12/14 19:28:29 INFO TaskSetManager: Starting task 11.0 in stage 20.0 (TID 79) (10.255.255.254, executor driver, partition 11, PROCESS_LOCAL, 9)
25/12/14 19:28:29 INFO Executor: Running task 0.0 in stage 20.0 (TID 68)
25/12/14 19:28:29 INFO Executor: Running task 4.0 in stage 20.0 (TID 72)
25/12/14 19:28:29 INFO Executor: Running task 6.0 in stage 20.0 (TID 74)
25/12/14 19:28:29 INFO Executor: Running task 5.0 in stage 20.0 (TID 73)
25/12/14 19:28:29 INFO Executor: Running task 1.0 in stage 20.0 (TID 69)
25/12/14 19:28:29 INFO Executor: Running task 8.0 in stage 20.0 (TID 76)
25/12/14 19:28:29 INFO Executor: Running task 10.0 in stage 20.0 (TID 78)
25/12/14 19:28:29 INFO Executor: Running task 3.0 in stage 20.0 (TID 71)
25/12/14 19:28:29 INFO Executor: Running task 2.0 in stage 20.0 (TID 79)
25/12/14 19:28:29 INFO Executor: Running task 7.0 in stage 20.0 (TID 75)
25/12/14 19:28:29 INFO Executor: Running task 9.0 in stage 20.0 (TID 77)
25/12/14 19:28:29 INFO Executor: Running task 11.0 in stage 20.0 (TID 79)
25/12/14 19:28:29 INFO CodeGenerator: Code generated in 22.669634 ms
25/12/14 19:28:29 INFO BlockManagerInfo: Removed broadcast_16_piece0 on 10.255.255.254:41131 in memory (size: 11.1 KiB, free: 434.2 MiB)
25/12/14 19:28:29 INFO BlockManagerInfo: Removed broadcast_17_piece0 on 10.255.255.254:41131 in memory (size: 6.3 KiB, free: 434.2 MiB)
25/12/14 19:28:29 INFO CodeGenerator: Code generated in 12.913967 ms
25/12/14 19:28:29 INFO CodeGenerator: Code generated in 3.216096 ms
25/12/14 19:28:29 INFO CodeGenerator: Code generated in 3.258158 ms
25/12/14 19:28:29 INFO CodeGenerator: Code generated in 3.186211 ms
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 60931440-67024584, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 30465720-36558864, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 54838296-60931440, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 24372576-30465720, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 42652008-48745152, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 67024584-68923428, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 12186288-182794932, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 48745152-54838296, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 6093144-12186288, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 18279432-24372576, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 0-6093144, partition values: [empty row]
25/12/14 19:28:29 INFO FileScanRDD: Reading File path: file:///mnt/d/pythonn/sales.csv, range: 36558864-42652008, partition values: [empty row]
25/12/14 19:28:29 INFO Executor: Finished task 11.0 in stage 20.0 (TID 79). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 11.0 in stage 20.0 (TID 79) in 146 ms on 10.255.255.254 (executor driver) (1/12)
25/12/14 19:28:29 INFO Executor: Finished task 5.0 in stage 20.0 (TID 73). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 5.0 in stage 20.0 (TID 73) in 245 ms on 10.255.255.254 (executor driver) (2/12)
25/12/14 19:28:29 INFO Executor: Finished task 6.0 in stage 20.0 (TID 74). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 6.0 in stage 20.0 (TID 74) in 247 ms on 10.255.255.254 (executor driver) (3/12)
25/12/14 19:28:29 INFO Executor: Finished task 0.0 in stage 20.0 (TID 68). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 0.0 in stage 20.0 (TID 68) in 249 ms on 10.255.255.254 (executor driver) (4/12)
25/12/14 19:28:29 INFO Executor: Finished task 1.0 in stage 20.0 (TID 69). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 7.0 in stage 20.0 (TID 75). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 1.0 in stage 20.0 (TID 69) in 252 ms on 10.255.255.254 (executor driver) (5/12)
25/12/14 19:28:29 INFO TaskSetManager: Finished task 7.0 in stage 20.0 (TID 75) in 251 ms on 10.255.255.254 (executor driver) (6/12)
25/12/14 19:28:29 INFO Executor: Finished task 9.0 in stage 20.0 (TID 77). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO Executor: Finished task 2.0 in stage 20.0 (TID 78). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 9.0 in stage 20.0 (TID 77) in 252 ms on 10.255.255.254 (executor driver) (7/12)
25/12/14 19:28:29 INFO Executor: Finished task 10.0 in stage 20.0 (TID 78). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 2.0 in stage 20.0 (TID 70) in 254 ms on 10.255.255.254 (executor driver) (8/12)
25/12/14 19:28:29 INFO TaskSetManager: Finished task 10.0 in stage 20.0 (TID 78) in 253 ms on 10.255.255.254 (executor driver) (9/12)
25/12/14 19:28:29 INFO Executor: Finished task 8.0 in stage 20.0 (TID 76). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 8.0 in stage 20.0 (TID 76) in 254 ms on 10.255.255.254 (executor driver) (10/12)
25/12/14 19:28:29 INFO Executor: Finished task 3.0 in stage 20.0 (TID 71). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO Executor: Finished task 4.0 in stage 20.0 (TID 72). 2823 bytes result sent to driver
25/12/14 19:28:29 INFO TaskSetManager: Finished task 3.0 in stage 20.0 (TID 71) in 257 ms on 10.255.255.254 (executor driver) (11/12)
25/12/14 19:28:29 INFO TaskSetManager: Finished task 4.0 in stage 20.0 (TID 72) in 257 ms on 10.255.255.254 (executor driver) (12/12)

```

-parquet write (writing aggregate results)

```

25/12/14 19:28:30 INFO ParquetOutputFormat: ParquetRecordWriter [block size: 134217728B, row group padding size: 8388608B, validating: false]
25/12/14 19:28:30 INFO ParquetWriteSupport: Initialized Parquet WriteSupport with Catalyst schema:
{
  "type" : "struct",
  "fields" : [ {
    "name" : "year",
    "type" : "integer",
    "nullable" : true,
    "metadata" : {}
  }, {
    "name" : "month",
    "type" : "integer",
    "nullable" : true,
    "metadata" : {}
  }, {
    "name" : "monthly_revenue",
    "type" : "double",
    "nullable" : true,
    "metadata" : {}
  } ]
}
and corresponding Parquet message type:
message spark_schema {
  optional int32 year;
  optional int32 month;
  optional double monthly_revenue;
}

25/12/14 19:28:32 INFO CodecPool: Got brand-new compressor [.snappy]
25/12/14 19:28:33 INFO FileOutputCommitter: Saved output of task @attempt_202512141928305316919805863270476_0022_m_000000 to file:/mnt/d/pythonn/output/monthly_revenue/_temporary/0/task_202512141928305316919805863270476_0022_m_000000
25/12/14 19:28:33 INFO SparkHadoopFileRDD: Committed. Elapsed time: 29 ms.
25/12/14 19:28:33 INFO TaskSetManager: Finished task 0.0 in stage 22.0 (TID 88) in 1.978 ms on 10.255.255.254 (executor driver) (1/1)
25/12/14 19:28:31 INFO TaskSetManager: Finished task 0.0 in stage 22.0 (TID 88) in 1.978 ms on 10.255.255.254 (executor driver) (1/1)
25/12/14 19:28:31 INFO DAGSchedulerImpl: Removed TaskSet 22.0, whose tasks have all completed, from pool
25/12/14 19:28:31 INFO DAGScheduler: ResultStage 22 (parquet at NativeMethodAccesso... java:0) finished in 1.393 s
25/12/14 19:28:31 INFO DAGScheduler: Cancelling potential speculative tasks for this job
25/12/14 19:28:31 INFO DAGSchedulerImpl: Killing all running tasks in stage 22. Stage finished
25/12/14 19:28:31 INFO DAGScheduler: Job 13 finished: parquet at NativeMethodAccesso... java:0, took 1.396725 s

25/12/14 19:28:32 INFO SparkContext: Starting job: parquet at NativeMethodAccesso... java:0
25/12/14 19:28:32 INFO DAGScheduler: Got job 13 of parquet at NativeMethodAccesso... java:0 with 1 output partitions
25/12/14 19:28:32 INFO DAGScheduler: Submitting ResultStage 28 (parquet at NativeMethodAccesso... java:0)
25/12/14 19:28:32 INFO DAGScheduler: Parents of final stage: List(DriverStage 27)
25/12/14 19:28:32 INFO DAGScheduler: Missing parents: List()
25/12/14 19:28:32 INFO DAGScheduler: Submitting ResultStage 28 (MapPartitionsRDD[64] at parquet at NativeMethodAccesso... java:0), which has no missing parents
25/12/14 19:28:32 INFO DAGScheduler: ResultStage 28 was triggered by partitions (estimated size: 221.9 KB, free: 432.1 MB)
25/12/14 19:28:32 INFO MemoryStore: Block broadcast_26_piece0 stored as bytes in memory (estimated size 94.1 KB, free: 434.1 MB)
25/12/14 19:28:32 INFO BlockManagerInfo: Added broadcast_26_piece0 in memory on 10.255.255.254:41131 (size: 94.1 KB, free: 434.1 MB)
25/12/14 19:28:32 INFO SparkContext: Created broadcast 26 from broadcast at DAGScheduler.scala:161
25/12/14 19:28:32 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 28 (MapPartitionsRDD[64] at parquet at NativeMethodAccesso... java:0) (first 15 tasks are for partitions Vector(0))
25/12/14 19:28:32 INFO TaskSchedulerImpl: Adding missing task for partitions profile
25/12/14 19:28:32 INFO TaskSetManager: Starting task 0.0 in stage 28.0 (TID 106) (10.255.255.254, executor driver, partition 0, NODE_LOCAL, 8999 bytes)
25/12/14 19:28:32 INFO Executor: Running task 0.0 in stage 28.0 (TID 106)
25/12/14 19:28:32 INFO ShuffleBlockFetcherIterator: Getting 12 (2090.0 B) non-empty blocks including 12 (2090.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
25/12/14 19:28:32 INFO ContainerCoordinator: Queued 1 remote fetches in 5.74142 s
25/12/14 19:28:32 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
25/12/14 19:28:32 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
25/12/14 19:28:32 INFO FileOutputCommitter: Using default output committer class org.apache.parquet.hadoop.ParquetOutputCommitter
25/12/14 19:28:32 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
25/12/14 19:28:32 INFO TaskSetManagerImpl: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
25/12/14 19:28:32 INFO SQLHadoopReduceCommitProtocol: Using output committer class org.apache.parquet.hadoop.ParquetOutputCommitter
25/12/14 19:28:32 INFO CodeConfig: Compression: SNAPPY
25/12/14 19:28:32 INFO CodeConfig: Compression: SNAPPY
25/12/14 19:28:32 INFO ParquetOutputFormat: ParquetRecordWriter [block size: 134217728B, row group padding size: 8388608B, validating: false]
25/12/14 19:28:32 INFO ParquetWriteSupport: Initialized Parquet WriteSupport with Catalyst schema:
{
  "type" : "struct",
  "fields" : [ {
    "name" : "has_promotion",
    "type" : "string",
    "nullable" : false,
    "metadata" : {}
  }, {
    "name" : "revenue",
    "type" : "double",
    "nullable" : true,
    "metadata" : {}
  } ]
}
and corresponding Parquet message type:
message spark_schema {
  required binary has_promotion (STRING);
  optional double revenue;
}

25/12/14 19:28:32 INFO FileOutputCommitter: Saved output of task @attempt_2025121419283258910488802334856403_0028_m_000000 to file:/mnt/d/pythonn/output/promotion_impact/_temporary/0/task_2025121419283258910488802334856403_0028_m_000000
25/12/14 19:28:32 INFO SparkHadoopMapPartUtil: attempt_2025121419283258910488802334856403_0028_m_000000: Committed. Elapsed time: 25 ms.
25/12/14 19:28:32 INFO TaskSetManager: Finished task 0.0 in stage 28.0 (TID 106) in 1.271 ms on 10.255.255.254 (executor driver) (1/1)
25/12/14 19:28:32 INFO TaskSetManager: Finished task 0.0 in stage 28.0 (TID 106) in 1.271 ms on 10.255.255.254 (executor driver) (1/1)
25/12/14 19:28:32 INFO DAGScheduler: ResultStage 28 (parquet at NativeMethodAccesso... java:0) finished in 0.446 s
25/12/14 19:28:32 INFO DAGScheduler: Job 17 is finished. Cancelling potential speculative or zombie tasks for this job

```

-show() (reading parquet files)

```

25/12/14 19:28:36 INFO CodeGenerator: Code generated in 3.21359 ms
+---+---+-----+
|year|month|monthly_revenue|
+---+---+-----+
|2022|    4|      2.7581851E7|
|2022|    3|        98261.0|
|2022|    5|      2.5119481E7|
|2022|    6|      2.2602913E7|
+---+---+-----+

```

```
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|
+-----+-----+-----+
```

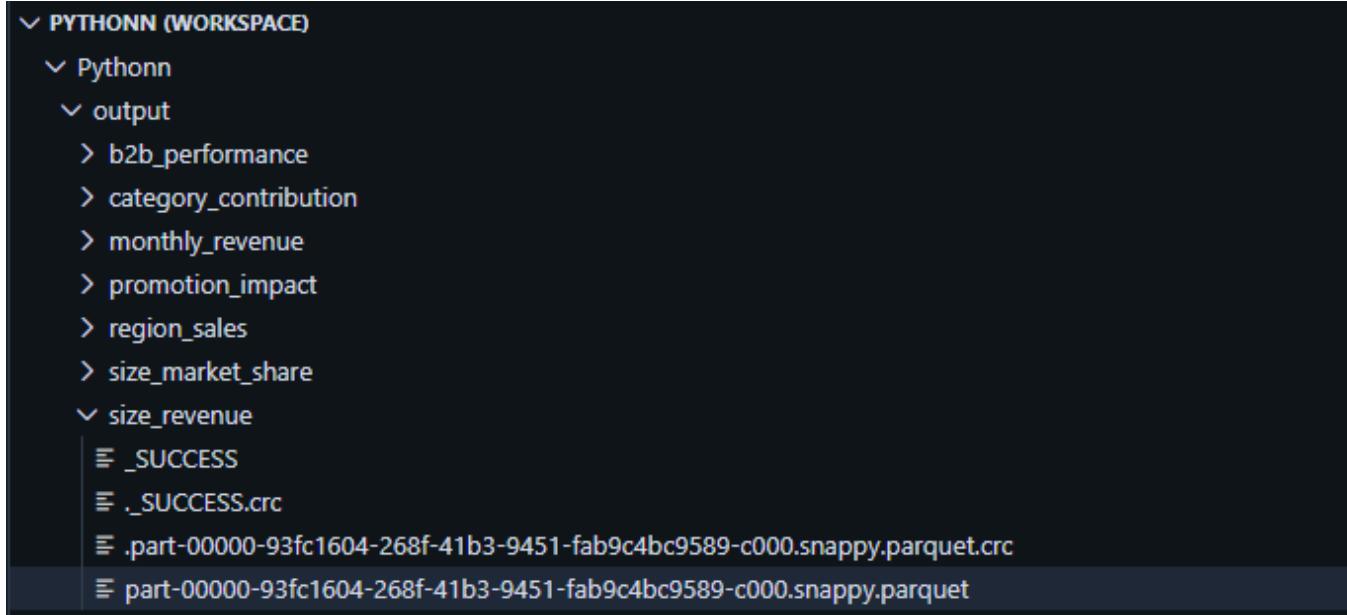
```
25/12/14 19:28:37 INFO DAGScheduler: Job 41 finished: showString at NativeMethodAccessorImpl.java:0, took 0.077308 s
+-----+-----+
|is_b2b| revenue|market_share_pct|
+-----+-----+
| true| 579930.0| 0.77|
| false|7.4822576E7| 99.23|
```

```

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|
| SXL| 414889.0|      0.55|
| 4XL| 325318.0|      0.43|
| FFree| 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)

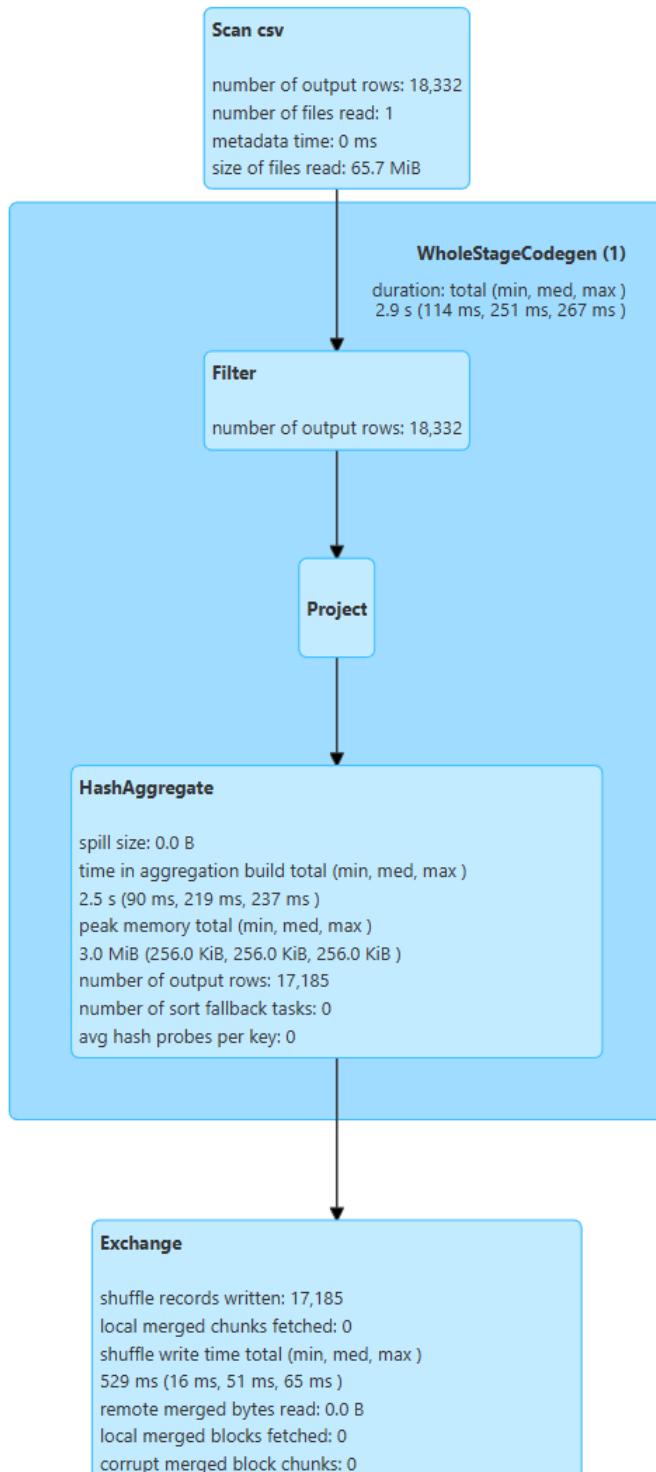
Details for Query 3

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
```

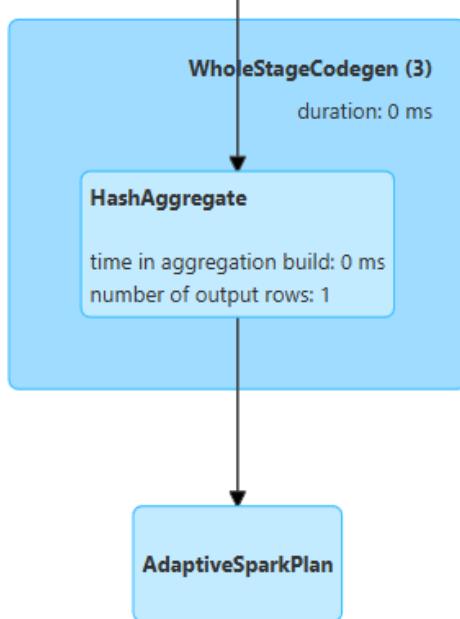
HashAggregate

```
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
```

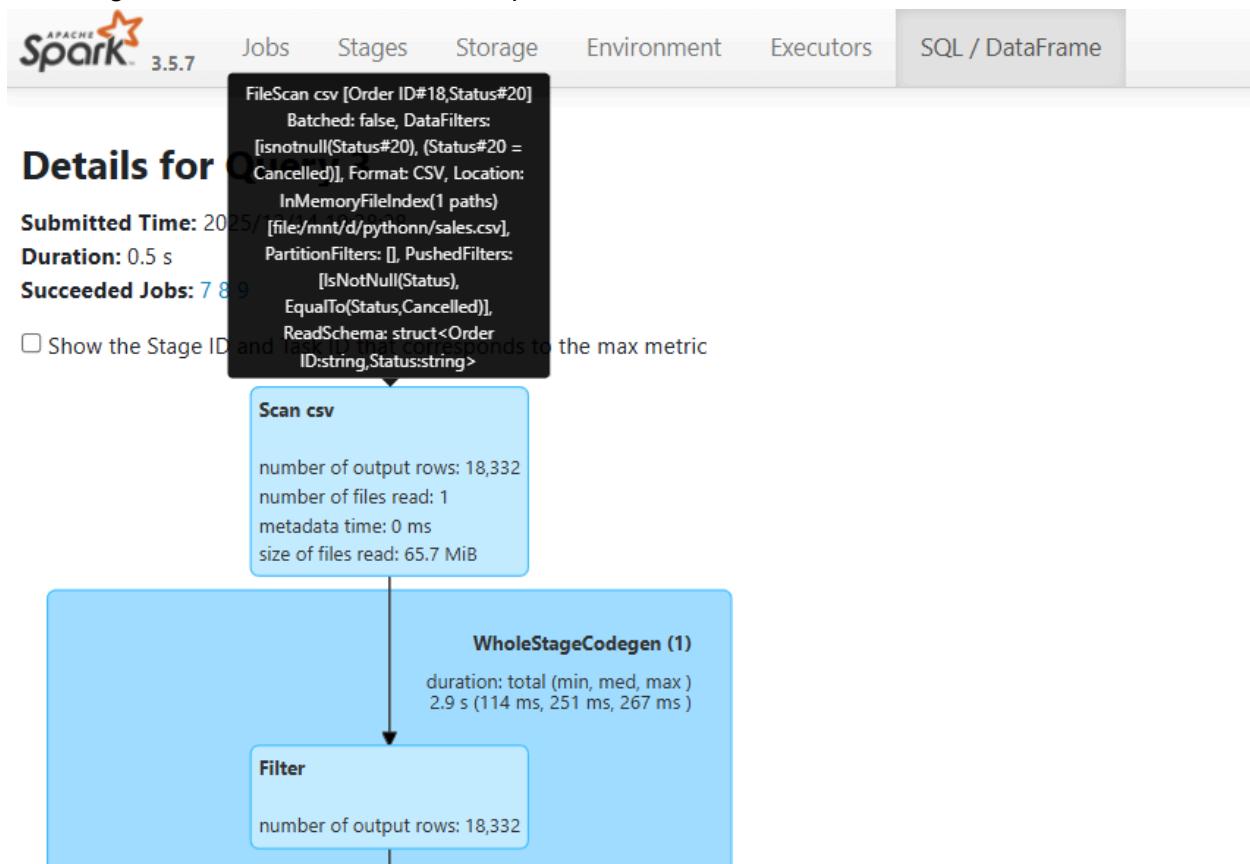
Exchange

```
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
```



Checking for cancelled orders in the depiction.

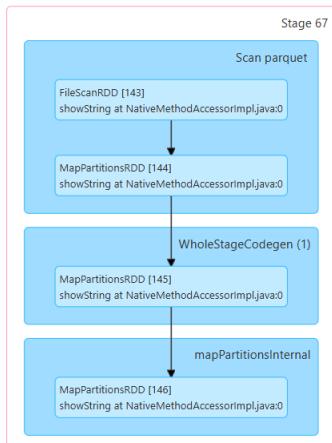


Job 43(sample read parquet and show):

Details for Stage 67 (Attempt 0)

Resource Profile Id: 0
Total Time Across All Tasks: 47 ms
Locality Level Summary: Process local: 1
Input Size / Records: 2.6 KiB / 11
Associated Job Ids: 43

▼ DAG Visualization

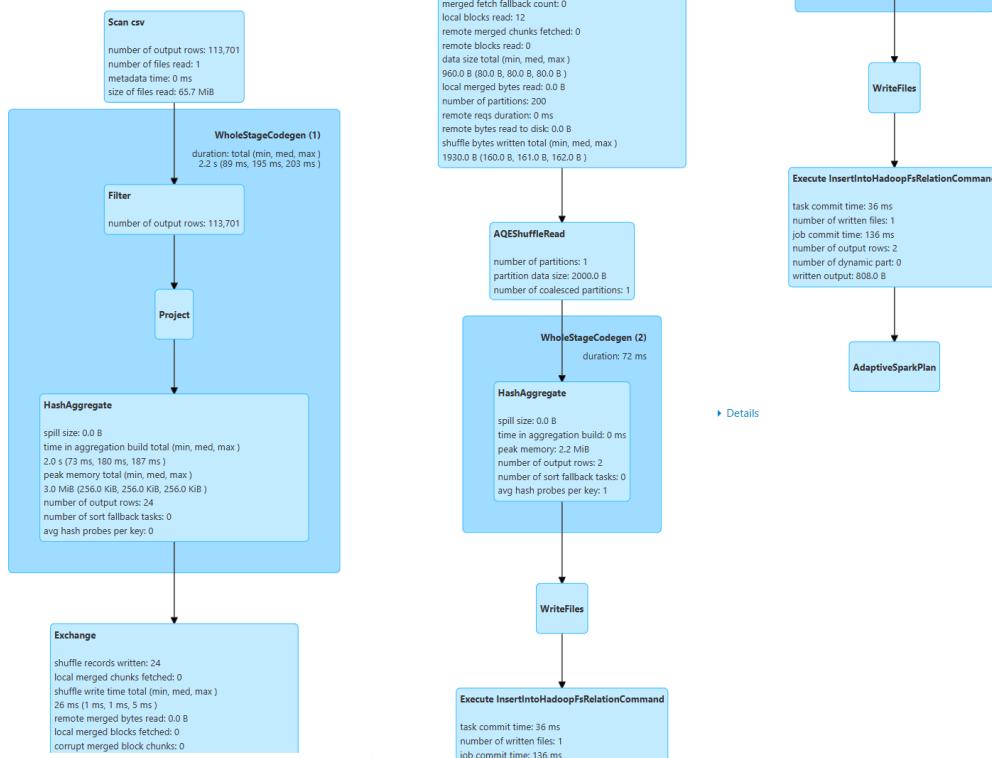


► Show Additional Metrics
► Event Timeline

Details for Query 7

Submitted Time: 2025/12/14 19:28:32
Duration: 0.6 s
Succeeded Jobs: 16/17

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



Parquet output sample (already included in spark job exec)

size		size_revenue	market_share_pct
M	1.389545E7	17.64	
L	1.197813E7	10.53	
XL	1.197813E7	18.81	
XXL	1.258511E7	13.8	
S	1.197813E7	17.53	
3XL	6795455.0	11.66	
K5	6759923.0	8.96	
6XL	414889.0	0.75	
S5XL	414889.0	8.55	
4XL	325318.0	8.43	
P	19000.0	0.28	

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