Continuous Internal Assessment Test - II

Name : Shyam Sundar S
Roll No : 24MX347

Step 1 :
Run the command: `hdfs dfs -ls/data/input`
And create the `sales.csv` and `weather.csv` files.

Step 2:

Load the local csv files into Ubuntu hdfs storage

Screenshot of the output :

```
devcontainers@SHYAM:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
localhost: datanode is running as process 12611.  Stop it first and ensure /tmp/hadoop-devcontainers-datanode.pid file is empty before retry.
Starting secondary namenodes [SHYAM]
SHYAM: secondarynamenode is running as process 12784.  Stop it first and ensure /tmp/hadoop-devcontainers-secondarynamenode.pid file is empty before retry.
devcontainers@SHYAM:~$ start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 13087.  Stop it first and ensure /tmp/hadoop-devcontainers-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 13224.  Stop it first and ensure /tmp/hadoop-devcontainers-nodemanager.pid file is empty before retry.
devcontainers@SHYAM:~$ hdfs dfs -mkdir -p /data/input
devcontainers@SHYAM:~$ hdfs dfs -put sales.csv /data/input/
devcontainers@SHYAM:~$ hdfs dfs -put weather.csv /data/input/
devcontainers@SHYAM:~$ hdfs dfs -ls /data/input/
Found 2 items
-rw-r--r--   1 devcontainers supergroup        197 2025-10-13 04:17 /data/input/sales.csv
-rw-r--r--   1 devcontainers supergroup        178 2025-10-13 04:17 /data/input/weather.csv
devcontainers@SHYAM:~$ nano analysis.pig
devcontainers@SHYAM:~$ pig -x mapreduce analysis.pig
2025-10-13 04:18:30,695 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-10-13 04:18:30,697 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-10-13 04:18:30,697 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-10-13 04:18:30,758 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-10-13 04:18:30,758 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/devcontainers/pig_1760329110753.log
2025-10-13 04:18:30,999 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/devcontainers/.pigbootup not found
2025-10-13 04:18:31,060 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-10-13 04:18:31,060 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2025-10-13 04:18:31,530 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-analysis.pig-acfeda3e-6c18-4c2c-95b1-acbf9e771d95
2025-10-13 04:18:31,530 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
```

Pig Analysis Generation using pig Latin script Screenshot:

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.6   0.17.0  devcontainers   2025-10-13 04:33:27     2025-10-13 04:33:52     HASH_JOIN,GROUP_BY,ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local1030950535_0003        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     sorted_sales    SAMPLER
job_local1268590773_0002        2       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     sales_weather_join,sunny_day_sales,weather       HASH_JOIN,MULTI_QUERY   hdfs
:///data/output/sunny_day_sales,
job_local1355744409_0001        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     high_value_sales,sales,sales_by_product,total_sales     MULTI_QUERY,COMBINER    hdfs
:///data/output/high_value_sales,hdfs:///data/output/total_sales,
job_local1571309557_0004        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     sorted_sales    ORDER_BY,COMBINER
job_local832496057_0005 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     avg_sales_by_condition,grouped_by_condition      GROUP_BY,COMBINER       hdfs:///data
/output/avg_sales_by_condition,
job_local856747376_0006 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     sorted_sales            hdfs:///data/output/top_3_sales,

Input(s):
Successfully read 7 records (11511958 bytes) from: "hdfs:///data/input/sales.csv"
Successfully read 7 records from: "hdfs:///data/input/weather.csv"

Output(s):
Successfully stored 2 records (53 bytes) in: "hdfs:///data/output/high_value_sales"
Successfully stored 3 records (35 bytes) in: "hdfs:///data/output/total_sales"
Successfully stored 3 records (69071962 bytes) in: "hdfs:///data/output/top_3_sales"
Successfully stored 3 records (162 bytes) in: "hdfs:///data/output/sunny_day_sales"
Successfully stored 4 records (57560267 bytes) in: "hdfs:///data/output/avg_sales_by_condition"

Counters:
Total records written : 15
Total bytes written : 126632479
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1355744409_0001 ->      job_local1030950535_0003,job_local1268590773_0002,
job_local1030950535_0003        ->      job_local1571309557_0004,
job_local1571309557_0004        ->      job_local856747376_0006,
job_local856747376_0006
job_local1268590773_0002        ->      job_local832496057_0005,
job_local832496057_0005


2025-10-13 04:33:52,213 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-10-13 04:33:52,214 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-10-13 04:33:52,216 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-10-13 04:33:52,220 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-10-13 04:33:52,222 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
```

Step 3:
Queries of the Script :

What is the total sales amount for each product?

Which sales transactions had a value greater than $200?

Which sales occurred on days when the weather was sunny?

What is the average sales amount for each weather condition?

What are the top 3 highest sales transactions by amount?

Five Queries Output Screenshot:

Run the command: `hdfs dfs -cat /data/output/total_sales/part-r-00000`