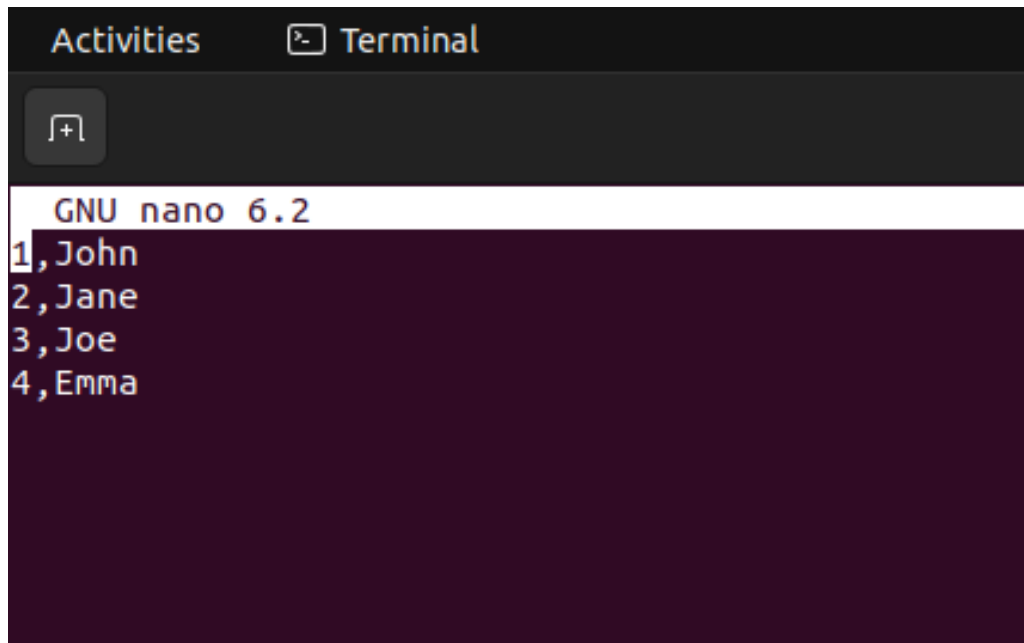


Exp. No : 4

User Defined Function (UDF) in PIG

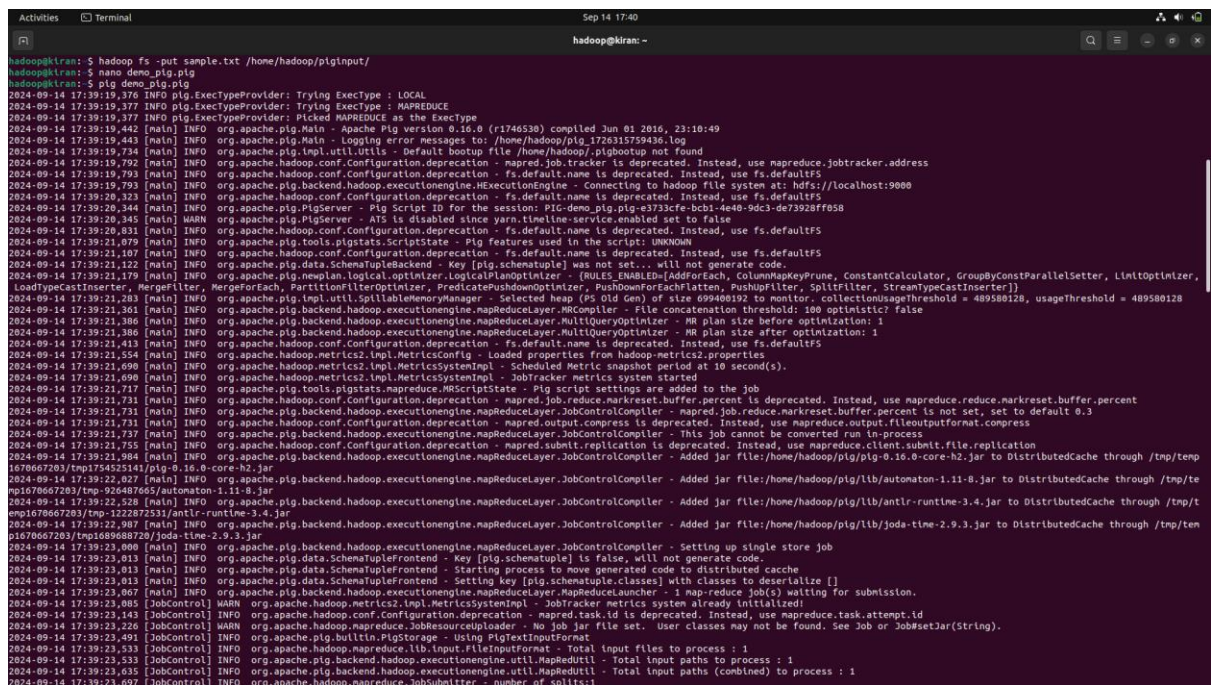
1. Create sample.txt



```

Activities Terminal
GNU nano 6.2
1, John
2, Jane
3, Joe
4, Emma
  
```

2. Create demo_pig.pig file



```

Activities Terminal
Sep 14 17:40
hadoop@kiran:~$
hadoop@kiran:~$ hadoop fs -put sample.txt /home/hadoop/piginput/
hadoop@kiran:~$ nano demo_pig.pig
2024-09-14 17:39:19,376 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-14 17:39:19,377 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-14 17:39:19,377 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-14 17:39:19,392 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Logging error Messages to: /home/hadoop/pig/1726315759436.log
2024-09-14 17:39:19,393 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:19,393 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-14 17:39:19,393 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:19,393 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Pig Script ID for the session: pig-demo_pig.pig-e3733cfe-bcbl-4e4b-9d3-6e7328ff058
2024-09-14 17:39:19,393 [main] WARN org.apache.hadoop.mapreduce.JobTracker - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-14 17:39:19,393 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Key [pig.schematuple] was not set... will not generate code.
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - (RULES_ENABLED: [AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushNupFilter, SplitFilter, StreamTypeCastInserter])
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Selected heap (PS Old Gen) of size 699408192 to monitor, collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - File concatenation threshold: 100 optimistic false
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - MR plan size before optimization: 1
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - MR plan size after optimization: 1
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Loaded properties from hadoop-metrics2.properties
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Scheduled Metric snapshot period at 10 second(s).
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - JobTracker metrics system started
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Pig script settings are added to the job
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Added jar file:/home/hadoop/pig/lib/autotom-1.11-8.jar to DistributedCache through /tmp/tenp1676667203/tp-926487665/autotom-1.11-8.jar
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Added jar file:/home/hadoop/pig/lib/antlr-runtime-3.4.jar to DistributedCache through /tmp/tenp1676667203/tp-926487665/antlr-runtime-3.4.jar
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Added jar file:/home/hadoop/pig/lib/joda-time-2.9.3.jar to DistributedCache through /tmp/tenp1676667203/tp-926487665/joda-time-2.9.3.jar
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Setting up single store job
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Key [pig.schematuple] is false, will not generate code.
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Starting process to move generated code to distributed cache
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Setting key [pig.schematuple.classes] with classes to deserialize []
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - 1 map-reduce Job(s) waiting for submission.
2024-09-14 17:39:21,107 [main] WARN org.apache.hadoop.mapreduce.JobTracker - JobTracker metrics system already initialized!
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - fs.defaultFS is deprecated. Instead, use fs.default
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Using PigTextInputFormat
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Total input files to process : 1
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Total input paths to process : 1
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - Total input paths (combined) to process : 1
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.mapreduce.JobTracker - number of splits: 1
  
```

3. Execute demo_pig.pig

```

Activities  Terminal
Sep 14 17:40
hadoop@kiran: ~
File Output Format Counters
  Bytes Read=8
  Bytes Written=0
2024-09-14 17:39:25,363 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local264353967_0001_n_000000_0
2024-09-14 17:39:25,364 [Thread-19] INFO org.apache.hadoop.mapred.LocalJobRunner - Map task executor complete.
2024-09-14 17:39:25,562 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - 50% complete
2024-09-14 17:39:25,562 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - Running jobs are [job_local264353967_0001]
2024-09-14 17:39:29,930 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:29,943 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:29,944 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-09-14 17:39:29,946 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:29,995 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - 100% complete
2024-09-14 17:39:30,000 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.3.6  0.16.0  hadoop  2024-09-14 17:39:21  2024-09-14 17:39:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  Feature Outputs
job_local264353967_0001  1  0  n/a  n/a  n/a  n/a  0  0  data  MAP_ONLY  hdf://localhost:9000/tmp/temp1670667283/tmp-932804698

Input(s):
Successfully read 4 records (5378234 bytes) from: "/home/hadoop/pigInput/sample.txt"

Output(s):
Successfully stored 4 records (5378257 bytes) in: "hdfs://localhost:9000/tmp/temp1670667283/tmp-932804698"

Counters:
Total records written : 4
Total bytes written : 5378257
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local264353967_0001

2024-09-14 17:39:30,005 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:30,008 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:30,010 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:30,020 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - Success!
2024-09-14 17:39:30,024 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:30,025 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-14 17:39:30,038 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-09-14 17:39:30,038 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-09-14 17:39:30,123 [main] INFO org.apache.pig.Main - Pig script completed in 10 seconds and 796 milliseconds (10796 ms)
hadoop@kiran: ~

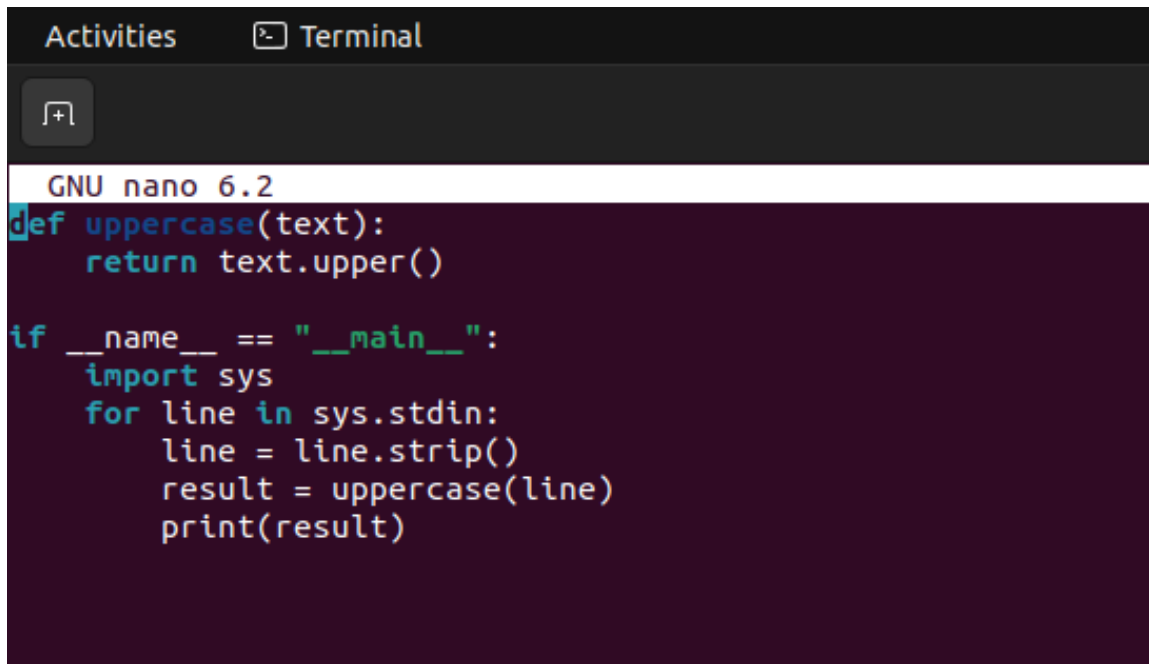
```

4. Create uppercase_udf.py

```

Activities  Terminal
(4,Emma)
2024-09-14 17:39:30,123 [main] INFO org.apache.pig.Main - Pig script completed
hadoop@kiran:~$ nano uppercase_udf.py
hadoop@kiran:~$ nano uppercase_udf.py
hadoop@kiran:~$ hadoop fs -mkdir /home/hadoop/udfs
hadoop@kiran:~$ hdfs dfs -put uppercase_udf.py /home/hadoop/udfs/
hadoop@kiran:~$ nano udf_example.pig

```



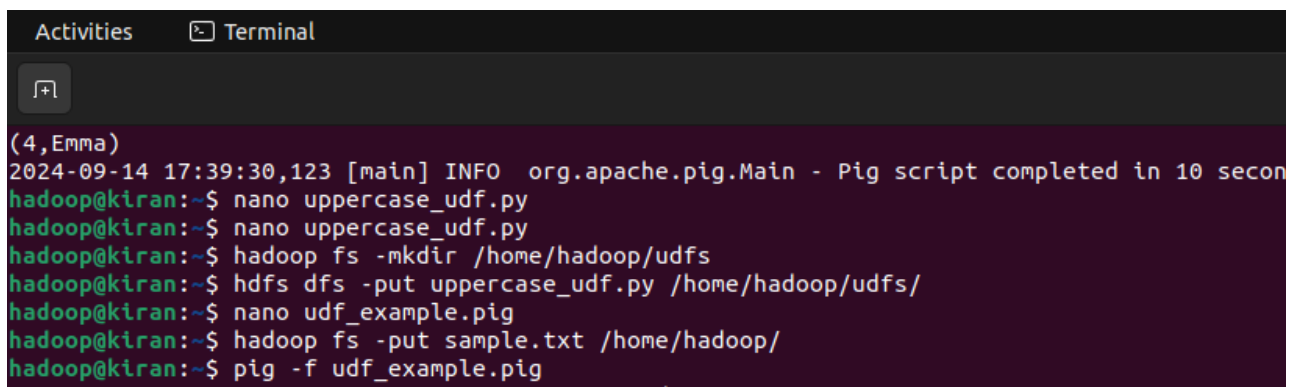
```

GNU nano 6.2
def uppercase(text):
    return text.upper()

if __name__ == "__main__":
    import sys
    for line in sys.stdin:
        line = line.strip()
        result = uppercase(line)
        print(result)

```

5. Upload uppercase_udf.py file to HDFS Storage.

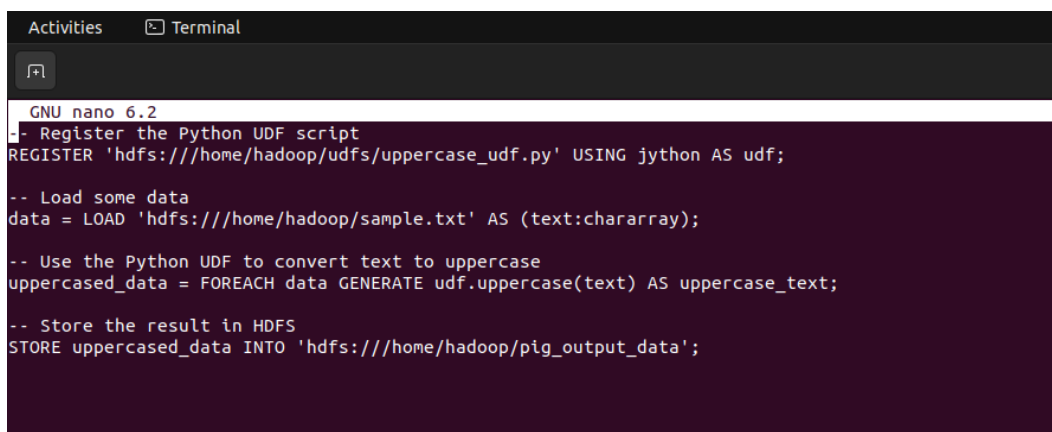


```

(4, Emma)
2024-09-14 17:39:30,123 [main] INFO  org.apache.pig.Main - Pig script completed in 10 seconds
hadoop@kiran:~$ nano uppercase_udf.py
hadoop@kiran:~$ nano uppercase_udf.py
hadoop@kiran:~$ hadoop fs -mkdir /home/hadoop/udfs
hadoop@kiran:~$ hdfs dfs -put uppercase_udf.py /home/hadoop/udfs/
hadoop@kiran:~$ nano udf_example.pig
hadoop@kiran:~$ hadoop fs -put sample.txt /home/hadoop/
hadoop@kiran:~$ pig -f udf_example.pig

```

6. Create udf_example.pig



```

GNU nano 6.2
-- Register the Python UDF script
REGISTER 'hdfs:///home/hadoop/udfs/uppercase_udf.py' USING jython AS udf;

-- Load some data
data = LOAD 'hdfs:///home/hadoop/sample.txt' AS (text:chararray);

-- Use the Python UDF to convert text to uppercase
uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;

-- Store the result in HDFS
STORE uppercased_data INTO 'hdfs:///home/hadoop/pig_output_data';

```

7. Execute udf_example.pig

```
hadoop@kiran:~$ nano udf_example.pig
hadoop@kiran:~$ hdfs dfs -ls /home/hadoop/pig_output_data
Found 2 items
-rw-r--r-- 1 hadoop supergroup      0 2024-09-14 17:46 /home/hadoop/pig_output_data/_SUCCESS
-rw-r--r-- 1 hadoop supergroup    27 2024-09-14 17:46 /home/hadoop/pig_output_data/part-m-00000
hadoop@kiran:~$
```

Output :

```
000001: Command not found
hadoop@kiran:~$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m-00000
1,JOHN
2,JANE
3,JOE
4,EMMA
hadoop@kiran:~$
```