

Exp. No : 4

User Defined Function (UDF) in PIG

1. Create sample.txt

```

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

2. Create demo_pig.pig file

```

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 17:39:19,442 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-14 17:39:19,443 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig.172615759436.log
2024-09-14 17:39:19,734 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/pig.bootstrap not found
2024-09-14 17:39:19,792 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:19,793 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:19,793 [main] INFO org.apache.pig.backend.hadoop.executionengine.MExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-14 17:39:20,323 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:20,344 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo-pig-pig-e3733cfe-bcb1-4e40-9dc3-de73928ff058
2024-09-14 17:39:20,345 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-14 17:39:20,351 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,079 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig Features used in the script: UNKNOWN
2024-09-14 17:39:21,107 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,122 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schemaTuple] was not set... will not generate code.
2024-09-14 17:39:21,179 [main] INFO org.apache.pig.newplan.logical.optimizer.LocalPlanOptimizer - [RULES_ENABLED[AddressEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]]
2024-09-14 17:39:21,283 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (ps Old Gen) of size 699440192 to monitor, collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-09-14 17:39:21,361 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.HMCompiler - File concatenation threshold: 100 optimistic false
2024-09-14 17:39:21,386 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-14 17:39:21,386 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-14 17:39:21,413 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 17:39:21,554 [main] INFO org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2024-09-14 17:39:21,690 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2024-09-14 17:39:21,690 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2024-09-14 17:39:21,717 [main] INFO org.apache.pig.tools.pigstats.mapreduce.HMScriptState - Pig script settings are added to the job
2024-09-14 17:39:21,731 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2024-09-14 17:39:21,731 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-14 17:39:21,731 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2024-09-14 17:39:21,737 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - This job cannot be converted run in-process
2024-09-14 17:39:21,755 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2024-09-14 17:39:21,984 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/home/hadoop/pig/pig-0.16.0-core-h2.jar to DistributedCache through /tmp/temp1678667203/tmp1489688726/joda-tne-2.9.3.jar
2024-09-14 17:39:22,827 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/home/hadoop/pig/lib/autotaton-1.11-8.jar to DistributedCache through /tmp/temp1678667203/tmp-926487665/autotaton-1.11-8.jar
2024-09-14 17:39:22,828 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/home/hadoop/pig/lib/antlr-runtime-3.4.jar to DistributedCache through /tmp/temp1678667203/tmp-1222872531/antlr-runtime-3.4.jar
2024-09-14 17:39:22,887 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Added jar file:/home/hadoop/pig/lib/joda-tne-2.9.3.jar to DistributedCache through /tmp/temp1678667203/tmp1489688726/joda-tne-2.9.3.jar
2024-09-14 17:39:23,000 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.JobControlCompiler - Setting up single store job
2024-09-14 17:39:23,013 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schemaTuple] is false, will not generate code.
2024-09-14 17:39:23,013 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-09-14 17:39:23,013 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schemaTuple.classes] with classes to deserialize []
2024-09-14 17:39:23,067 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce_layer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2024-09-14 17:39:23,085 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-14 17:39:23,143 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2024-09-14 17:39:23,226 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or JobsetJar(String).
2024-09-14 17:39:23,491 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2024-09-14 17:39:23,533 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input files to process : 1
2024-09-14 17:39:23,533 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2024-09-14 17:39:23,635 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2024-09-14 17:39:23,697 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
  
```

3. Execute demo_pig.pig

```

Bytes Read=0
File Output Format Counters
  Bytes Written=0
2024-09-14 17:39:25,363 [LocalJobRunner Map Task Executor #9] 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 17:39:29,944 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
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  hdfs://localhost:9000/tmp/tempt1670667203/tmp-932804698,

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

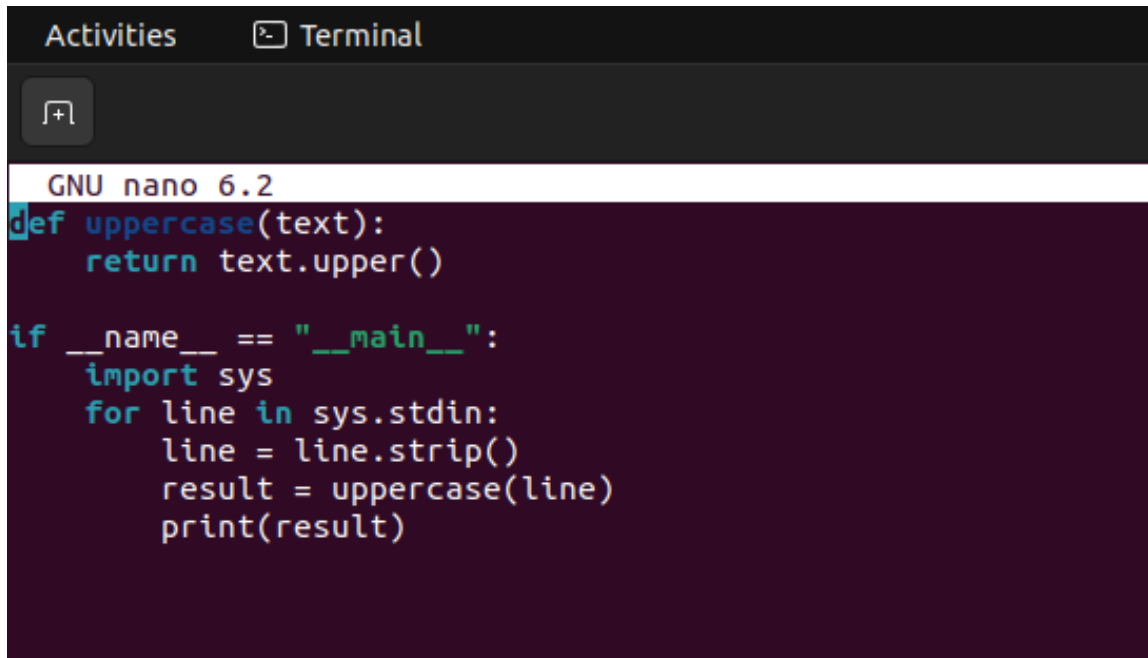
Output(s):
Successfully stored 4 records (5378257 bytes) in: "hdfs://localhost:9000/tmp/tempt1670667203/tmp-932804698"

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.MapReduceUtil - Total input paths to process : 1
(1,John)
(2,June)
(3,Joe)
(4,Emma)

```



```

Activities  Terminal

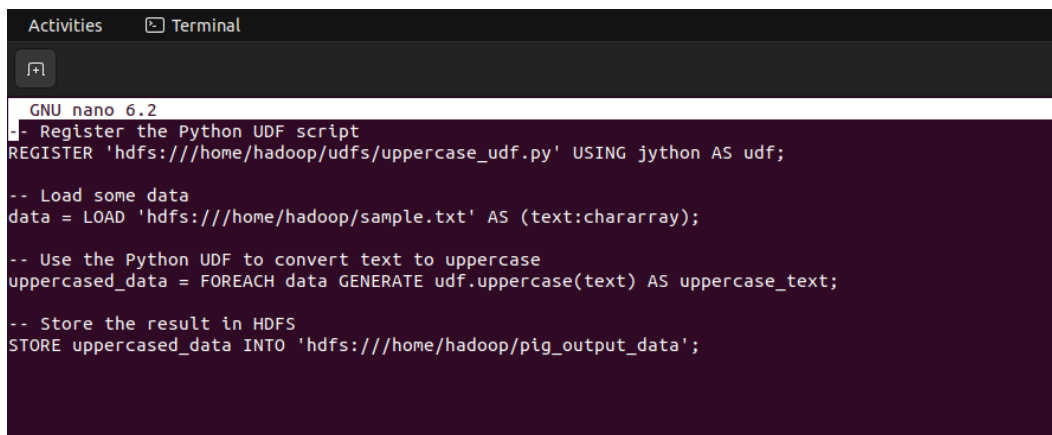
GNU nano 6.2
def uppercase(text):
    return text.upper()

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

```

4. Upload uppercase_udf.py file to HDFS Storage.

5. Create udf_example.pig



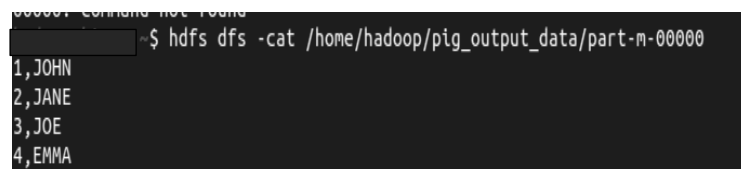
```
Activities Terminal
GNU nano 6.2
-- Register the Python UDF script
REGISTER 'hdfs:///home/hadoop/udfs/uppercase_udf.py' USING jython AS udf;

-- 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';
```

Output :



```
000001: Command not found
$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m-000000
1,JOHN
2,JANE
3,JOE
4,EMMA
```