EXP NO: 4   **CREATE UDF IN PIG**

$start-all.sh

$ jps

$wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz

$ tar xvzf pig-0.16.0.tar.gz

$nano ~/.bashrc

$mv pig-0.16.0 pig

$pig

$cd DA-Lab
$mkdir exp4
$cd exp4
$nano sample.txt

$nano demo_pig.pig

$hdfs dfs -mkdir /exp4

$hdfs dfs -copyFromLocal ~/DA-Lab/exp4/sample.txt /exp4
$pig demo_pig.pig

$nano uppercase_udf.py

$hdfs dfs -copyFromLocal ~/DA-Lab/exp4/uppercase_udf.py /exp4

$nano udf_example.pig

$pig -f udf_example.pig

```
                                               hadoop@fedora:~/da/exp4                    Q  ≡  ×

        at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.base/java.lang.reflect.Method.invoke(Method.java:566)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:328)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
2024-09-11 14:33:30,316 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <file demo_pig.pig, line 5, column
 0>  mismatched input 'DUMP' expecting SEMI_COLON
Details at logfile: /home/hadoop/da/exp4/pig_1726045407699.log
2024-09-11 14:33:30,380 [main] INFO  org.apache.pig.Main - Pig script completed in 2 seconds and 918 milliseconds (291
8 ms)
[hadoop@fedora exp4]$ nano demo_pig.pig
[hadoop@fedora exp4]$ pig demo_pig.pig
2024-09-11 14:34:17,296 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-11 14:34:17,299 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-11 14:34:17,299 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-11 14:34:17,420 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016,
23:10:49
2024-09-11 14:34:17,420 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/da/exp4/pig_1726045
457396.log
2024-09-11 14:34:18,225 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not
found
2024-09-11 14:34:18,369 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecat
ed. Instead, use mapreduce.jobtracker.address
2024-09-11 14:34:18,369 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
 Instead, use fs.defaultFS
2024-09-11 14:34:18,370 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to ha
doop file system at: hdfs://localhost:9000
2024-09-11 14:34:19,515 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
 Instead, use fs.defaultFS
2024-09-11 14:34:19,575 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-114288
6c-f5d8-4f00-9076-ce9b4317a5b1
2024-09-11 14:34:19,576 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled se
t to false
2024-09-11 14:34:20,618 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
 Instead, use fs.defaultFS
2024-09-11 14:34:21,209 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKN
OWN
2024-09-11 14:34:21,271 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
 Instead, use fs.defaultFS
2024-09-11 14:34:21,294 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... wil
l not generate code.
2024-09-11 14:34:21,378 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[A
ddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, Me
rgeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, S
plitFilter, StreamTypeCastInserter]}
2024-09-11 14:34:21,553 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of s
ize 1048576000 to monitor. collectionUsageThreshold = 734003200, usageThreshold = 734003200
2024-09-11 14:34:21,671 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File co
ncatenation threshold: 100 optimistic? false
2024-09-11 14:34:21,727 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size before optimization: 1
2024-09-11 14:34:21,728 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size after optimization: 1
2024-09-11 14:34:21,793 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
 Instead, use fs.defaultFS
2024-09-11 14:34:22,033 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-met
```

$hdfs dfs -cat /exp4/output/*

```
                    GC time elapsed (ms)=0
                    Total committed heap usage (bytes)=204472320
        File Input Format Counters
                    Bytes Read=0
        File Output Format Counters
                    Bytes Written=0
2024-09-11 14:48:31,941 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local125569218_0001_m_000000_0
2024-09-11 14:48:31,942 [Thread-24] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2024-09-11 14:48:32,245 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2024-09-11 14:48:32,245 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_local125569218_0001]
2024-09-11 14:48:35,761 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-11 14:48:35,775 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-11 14:48:35,777 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-09-11 14:48:35,777 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2024-09-11 14:48:35,781 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-11 14:48:35,850 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-09-11 14:48:35,858 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion     UserId StartedAt       FinishedAt      Features
3.3.6  0.16.0 hadoop 2024-09-11 14:48:24     2024-09-11 14:48:35     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTimeAvgReduceTime      MedianReducetime        Alias   Feature Outputs
job_local125569218_0001 1       0       n/a     n/a     n/a     n/a     0       0       0       0       data,uppercased_data MAP_ONLY        hdfs:///pig/pig_output_data,

Input(s):
Successfully read 4 records (42400493 bytes) from: "hdfs:///pig/sample.txt"

Output(s):
Successfully stored 4 records (42400294 bytes) in: "hdfs:///pig/pig_output_data"

Counters:
Total records written : 4
Total bytes written : 42400294
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local125569218_0001


2024-09-11 14:48:35,869 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-11 14:48:35,877 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-11 14:48:35,884 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-11 14:48:35,942 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-11 14:48:36,008 [main] INFO  org.apache.pig.Main - Pig script completed in 23 seconds and 530 milliseconds (23530 ms)
[hadoop@fedora exp4]$ hdfs dfs -cat /pig/pig_output_data/*
1,JOHN
2,JANE
3,JOE
4,EMMA
[hadoop@fedora exp4]$
```