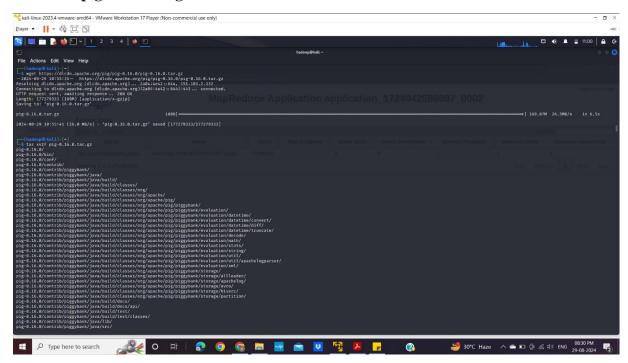**EXP NO: 4** **CREATE UDF IN PIG**

$start-all.sh

```
karthickragav@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as karthickragav in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [fedora]
Starting resourcemanager
Starting nodemanagers
```

$ jps

```
karthickragav@fedora:~$ jps
4916 Jps
4487 NodeManager
3609 NameNode
4009 SecondaryNameNode
4347 ResourceManager
3775 DataNode
```

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

$ tar xvzf pig-0.16.0.tar.gz
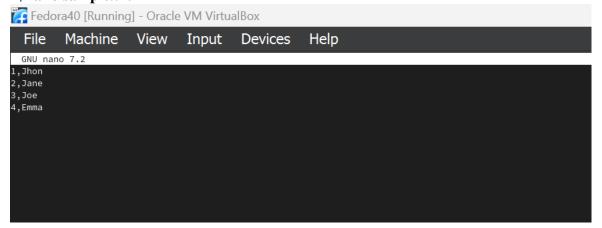

$nano ~/.bashrc

```
#PIG settings
export PIG_HOME=~/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
```

$mv pig-0.16.0 pig

$pig

```
karthickragav@fedora:~/dalab/exp3$ pig
2024-09-24 09:58:34,607 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-24 09:58:34,610 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-24 09:58:34,610 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-24 09:58:35,003 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-24 09:58:35,003 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/karthickragav/dalab/exp3/pig_1727152114994.log
2024-09-24 09:58:35,102 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/karthickragav/.pigbootup not found
2024-09-24 09:58:35,821 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-24 09:58:35,822 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 09:58:35,822 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-24 09:58:36,755 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 09:58:37,163 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-07881d21-6cd2-41e7-86b1-cf39a23d9a64
2024-09-24 09:58:37,164 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> quit
2024-09-24 09:58:53,976 [main] INFO  org.apache.pig.Main - Pig script completed in 19 seconds and 312 milliseconds (19312 ms)
karthickragav@fedora:~/dalab/exp3$
```

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

```
Fedora40 [Running] - Oracle VM VirtualBox
File   Machine   View   Input   Devices   Help
  GNU nano 7.2
1,Jhon
2,Jane
3,Joe
4,Emma
```

$nano demo_pig.pig

```
Fedora40 [Running] - Oracle VM VirtualBox
File   Machine   View   Input   Devices   Help
  GNU nano 7.2
-- Load the data from HDFS
data = LOAD '/piginput/sample.txt' USING PigStorage(',') AS (id:int,text:chararray);
-- Dump the data to check if it was loaded correctly
DUMP data;
```

$hdfs dfs -mkdir /exp4

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

$pig demo_pig.pig

```
karthickragav@fedora:~/dalab/exp4$ pig demo_pig.pig
2024-09-24 10:02:03,353 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-24 10:02:03,356 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-24 10:02:03,356 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-24 10:02:03,546 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-24 10:02:03,547 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/karthickragav/dalab/exp4/pig_1727152323530.log
2024-09-24 10:02:04,157 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/karthickragav/.pigbootup not found
2024-09-24 10:02:04,297 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-24 10:02:04,297 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 10:02:04,297 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-24 10:02:05,167 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 10:02:05,621 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-c299a4b7-1b2d-432b-b695-11069555f134
2024-09-24 10:02:05,622 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-24 10:02:06,249 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 10:02:06,944 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-24 10:02:07,117 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 10:02:07,160 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-24 10:02:07,432 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, Merg
eFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-09-24 10:02:07,817 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-09-24 10:02:08,132 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-24 10:02:08,208 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-24 10:02:08,209 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-24 10:02:08,375 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 10:02:08,790 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2024-09-24 10:02:09,189 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2024-09-24 10:02:09,189 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2024-09-24 10:02:09,319 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2024-09-24 10:02:09,440 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2024-09-24 10:02:09,440 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-24 10:02:09,440 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2024-09-24 10:02:09,494 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2024-09-24 10:02:09,573 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2024-09-24 10:02:11,096 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/karthickragav/pig/pig-0.16.0-core-h2.jar to DistributedCache through /tmp/temp-1205491644/tmp-1680
833006/pig-0.16.0-core-h2.jar
2024-09-24 10:02:11,322 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/karthickragav/pig/lib/automaton-1.11-8.jar to DistributedCache through /tmp/temp-1205491644/tmp575
679758/automaton-1.11-8.jar
2024-09-24 10:02:12,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/karthickragav/pig/lib/antlr-runtime-3.4.jar to DistributedCache through /tmp/temp-1205491644/tmp39
5398926/antlr-runtime-3.4.jar
2024-09-24 10:02:12,777 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/karthickragav/pig/lib/joda-time-2.9.3.jar to DistributedCache through /tmp/temp-1205491644/tmp-155
```

$nano uppercase_udf.py

Fedora40 [Running] - Oracle VM VirtualBox

File   Machine   View   Input   Devices   Help

```
GNU nano 7.2
def uppercase(text):
    return text.upper()


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

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

```
karthickragav@fedora:~/dalab/exp4$ hdfs dfs -ls /piginput
Found 1 items
-rw-r--r--   1 karthickragav supergroup         27 2024-09-04 12:28 /piginput/sample.txt
karthickragav@fedora:~/dalab/exp4$
```

$nano udf_example.pig

Fedora40 [Running] - Oracle VM VirtualBox

File   Machine   View   Input   Devices   Help

```
  GNU nano 7.2                                                         udf_example.pig
REGISTER 'hdfs:///udfs/uppercase_udf.py' USING jython AS udf;

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

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

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

$pig -f udf_example.pig

Success!

Job Stats (time in seconds):
JobId   Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime   Alias Feature Outputs
job_local270706031_0001 1   0   n/a   n/a   n/a   n/a   0   0   0   0   data MAP_ONLY   hdfs://localhost:9000/tmp/temp-1205491644/tmp1413721026,

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

Output(s):
Successfully stored 4 records (5378257 bytes) in: "hdfs://localhost:9000/tmp/temp-1205491644/tmp1413721026"

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_local270706031_0001

2024-09-24 10:02:23,664 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-24 10:02:23,671 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-24 10:02:23,690 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-24 10:02:23,712 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-24 10:02:23,729 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-24 10:02:23,741 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-24 10:02:23,894 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-09-24 10:02:23,899 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Jhon)
(2,Jane)
(3,Joe)
(4,Emma)
2024-09-24 10:02:24,314 [main] INFO org.apache.pig.Main - Pig script completed in 20 seconds and 988 milliseconds (20988 ms)

$hdfs dfs -cat /exp4/output/*

Successfully read 4 records (42400495 bytes) from: "hdfs:///piginput/sample.txt"

Output(s):
Successfully stored 4 records (42400297 bytes) in: "hdfs:///home/hadoop/pig_output_data"

Counters:
Total records written : 4
Total bytes written : 42400297
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1373331387_0001

2024-09-04 12:42:06,312 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-04 12:42:06,318 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-04 12:42:06,336 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-04 12:42:06,448 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-04 12:42:06,724 [main] INFO org.apache.pig.Main - Pig script completed in 43 seconds and 342 milliseconds (43342 ms)
karthickragav@fedora:~/dalab/exp4$ hdfs dfs -ls /
Found 7 items
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:41 /home
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:28 /piginput
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:42 /tmp
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:37 /udfs
drwxr-xr-x   - karthickragav supergroup   0 2024-09-02 11:02 /weatherdata
drwxr-xr-x   - karthickragav supergroup   0 2024-09-01 21:13 /word_count_in_py
drwxr-xr-x   - karthickragav supergroup   0 2024-09-01 20:59 /word_count_in_python
karthickragav@fedora:~/dalab/exp4$ hdfs dfs -ls /tmp
karthickragav@fedora:~/dalab/exp4$ hdfs dfs -ls /home
Found 1 items
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:41 /home/hadoop
karthickragav@fedora:~/dalab/exp4$ ls
demo_pig.pig   pig_1725290479440.log pig_1725290702550.log pig_1725432772737.log udf_example.pig
pig_1725289163833.log pig_1725290532876.log pig_1725432499143.log sample.txt   uppercase_udf.py
karthickragav@fedora:~/dalab/exp4$ cd
karthickragav@fedora:~$ hdfs dfs -ls /home
Found 1 items
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:41 /home/hadoop
karthickragav@fedora:~$ hdfs dfs -ls /home/hadoop
Found 1 items
drwxr-xr-x   - karthickragav supergroup   0 2024-09-04 12:42 /home/hadoop/pig_output_data
karthickragav@fedora:~$ hdfs dfs -ls /home/hadoop/pig_output_data
Found 2 items
-rw-r--r--   1 karthickragav supergroup   0 2024-09-04 12:42 /home/hadoop/pig_output_data/_SUCCESS
-rw-r--r--   1 karthickragav supergroup   27 2024-09-04 12:42 /home/hadoop/pig_output_data/part-m-00000
karthickragav@fedora:~$ hdfs dfs -ls /home/hadoop/pig_output_data/part-m-00000
-rw-r--r--   1 karthickragav supergroup   27 2024-09-04 12:42 /home/hadoop/pig_output_data/part-m-00000
karthickragav@fedora:~$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m-00000
1,JHON
2,JANE
3,JOE
4,EMMA
karthickragav@fedora:~$