

## EXP NO: 4

## CREATE UDF IN PIG

## \$start-all.sh

```
kamal@kamal-VirtualBox: ~
kamal@kamal-VirtualBox:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as kamal in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 4790. Stop it first and ensure /tmp/hadoop-kamal-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 4916. Stop it first and ensure /tmp/hadoop-kamal-datanode.pid file is empty before retry.
Starting secondary namenodes [kamal-VirtualBox]
kamal-VirtualBox: secondarynamenode is running as process 5112. Stop it first and ensure /tmp/hadoop-kamal-secondarynamenode.pid file is empty before retry.
Starting resource manager
resource manager is running as process 5333. Stop it first and ensure /tmp/hadoop-kamal-resource manager.pid file is empty before retry.
Starting node managers
localhost: nodemanager is running as process 5455. Stop it first and ensure /tmp/hadoop-kamal-nodemanager.pid file is empty before retry.
```

## \$jps

```
kamal@kamal-VirtualBox:~$ jps
4916 DataNode
5333 ResourceManager
4790 NameNode
6919 Jps
5112 SecondaryNameNode
5455 NodeManager
```

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

## \$tar xvfz pig-0.16.0.tar.gz

```
kamal@kamal-VirtualBox: ~
kamal@kamal-VirtualBox:~$ tar xvfz pig-0.16.0.tar.gz
pig-0.16.0/
pig-0.16.0/bin/
pig-0.16.0/conf/
pig-0.16.0/contrib/
pig-0.16.0/contrib/piggybank/
pig-0.16.0/contrib/piggybank/java/
pig-0.16.0/contrib/piggybank/java/build/
pig-0.16.0/contrib/piggybank/java/build/classes/
pig-0.16.0/contrib/piggybank/java/build/classes/org/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/convert/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/diff/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/truncate/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/decode/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/math/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/stats/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/string/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/util/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/util/apachelogparser/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/xml/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/allloader/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/apachelog/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/avro/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/hiverc/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/partition/
pig-0.16.0/contrib/piggybank/java/build/docs/
pig-0.16.0/contrib/piggybank/java/build/docs/api/
pig-0.16.0/contrib/piggybank/java/build/test/
pig-0.16.0/contrib/piggybank/java/build/test/classes/
pig-0.16.0/contrib/piggybank/java/lib/
pig-0.16.0/contrib/piggybank/java/src/
pig-0.16.0/contrib/piggybank/java/src/main/
pig-0.16.0/contrib/piggybank/java/src/main/java/
pig-0.16.0/contrib/piggybank/java/src/main/java/org/
```

## \$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

```
kamal@kamal-VirtualBox:~$ pig
2024-10-12 13:01:39,155 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 13:01:39,163 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 13:01:39,164 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-12 13:01:39,283 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 13:01:39,283 [main] INFO org.apache.pig.Main - Logging error messages to: /home/kamal/pig_1728718299270.log
2024-10-12 13:01:39,352 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/kamal/.pigbootup not found
2024-10-12 13:01:39,836 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
se mapreduce.jobtracker.address
2024-10-12 13:01:39,836 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:01:39,836 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://localhost:9000
2024-10-12 13:01:40,659 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:01:40,683 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-a878a9a9-2f4d-4742-850e-6
f70e04790cb
2024-10-12 13:01:40,683 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

## \$cd DA-Lab

## \$mkdir exp4

## \$cd exp4

## \$nano sample.txt



```
kamal@kamal-VirtualBox: ~/DA_LAB/exp4
GNU nano 6.2 sample.txt
1,Jane
2,John
3,Emma
4,Peter
```

## \$nano demo\_pig.pig



```
kamal@kamal-VirtualBox: ~/DA_LAB/exp4
GNU nano 6.2 demo_pig.pig
-- 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

```

kamal@kamal-VirtualBox: ~/DA_LAB/exp4
kamal@kamal-VirtualBox:~/DA_LAB/exp4$ nano demo_pig.pig
kamal@kamal-VirtualBox:~/DA_LAB/exp4$ pig demo_pig.pig
2024-10-12 13:03:17,987 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 13:03:17,988 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 13:03:17,988 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-12 13:03:18,050 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 13:03:18,050 [main] INFO org.apache.pig.Main - Logging error messages to: /home/kamal/DA_LAB/exp4/pig_1728718398043.log
2024-10-12 13:03:18,394 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/kamal/.pigbootstrap not found
2024-10-12 13:03:18,465 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
se mapreduce.jobtracker.address
2024-10-12 13:03:18,465 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:18,465 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://localhost:9000
2024-10-12 13:03:19,051 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:19,090 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-700fdbd0-3e5f-4d8e-9
2e0-925abc9b6155
2024-10-12 13:03:19,091 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-10-12 13:03:19,898 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:20,495 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-10-12 13:03:20,556 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:20,613 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
code.
2024-10-12 13:03:20,703 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Col
umnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, Par
titionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-10-12 13:03:20,857 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192
to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-10-12 13:03:20,972 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation th
reshold: 100 optimistic? false
2024-10-12 13:03:21,038 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
before optimization: 1
2024-10-12 13:03:21,039 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
after optimization: 1
2024-10-12 13:03:21,080 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:21,176 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManage

```

## \$nano uppercase\_udf.py

```

GNU nano 6.2                                     uppercase_udf.py
import sys
def uppercase(text):
    return text.upper()
if __name__ == "__main__":
    for line in sys.stdin:
        line = line.strip()
        result = uppercase(line)
        print(result)

```



**\$hdfs dfs -copyFromLocal ~/DA-Lab/exp4/uppercase\_udf.py /exp4**

**\$nano udf\_example.pig**

```
kamal@kamal-VirtualBox: ~/DA_LAB/exp4
GNU nano 6.2 udf_example.pig
-- Register the Python UDF script
REGISTER 'hdfs:///piginput/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:///piginput/pig_output_data';
```

**\$pig -f udf\_example.pig**

```
kamal@kamal-VirtualBox: ~/DA_LAB/exp4
kamal@kamal-VirtualBox:~/DA_LAB/exp4$ nano demo_pig.pig
kamal@kamal-VirtualBox:~/DA_LAB/exp4$ pig demo_pig.pig
2024-10-12 13:03:17,987 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 13:03:17,988 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 13:03:17,988 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-12 13:03:18,050 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 13:03:18,050 [main] INFO org.apache.pig.Main - Logging error messages to: /home/kamal/DA_LAB/exp4/pig_1728718398043.log
2024-10-12 13:03:18,394 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/kamal/.pigbootup not found
2024-10-12 13:03:18,465 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
se mapreduce.jobtracker.address
2024-10-12 13:03:18,465 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:18,465 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://localhost:9000
2024-10-12 13:03:19,051 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:19,090 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-700fbbd0-3e5f-4d8e-9
2e0-925abc9b6155
2024-10-12 13:03:19,091 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-10-12 13:03:19,898 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:20,495 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-10-12 13:03:20,556 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:20,613 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
code.
2024-10-12 13:03:20,703 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Col
umnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, Par
titionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-10-12 13:03:20,857 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192
to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-10-12 13:03:20,972 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation th
reshold: 100 optimistic? false
2024-10-12 13:03:21,038 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
before optimization: 1
2024-10-12 13:03:21,039 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
after optimization: 1
2024-10-12 13:03:21,080 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-10-12 13:03:21,176 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMAFailoverProxyProvider - Connecting to ResourceManage
```

## Shdfs dfs -cat /exp4/output/\*

```

2024-09-04 12:42:06,212 [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 12:20 /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_1725298479449.log pig_1725298702559.log pig_172543272737.log udf_example.pig
pig_1725289103833.log pig_1725298532876.log pig_1725432499143.log sample.txt uppercase_udf.py
karthickragav@fedora:~/dalab/exp4$ cd
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$ 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:~/dalab/exp4$ 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:~/dalab/exp4$ 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:~/dalab/exp4$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m-00000
1,JOHN
2,JANE
3,JOE
4,EMMA
karthickragav@fedora:~/dalab/exp4$

```