

EXP NO: 4

CREATE UDF IN PIG

\$start-all.sh

```

(hadoop@kali)~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop 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 [kali]
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-08-29 04:59:16,429 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting resourcemanager
Starting nodemanagers

```

\$ jps

```

(hadoop@kali)~$ jps
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
14436 NodeManager
16772 Jps
13830 SecondaryNameNode
14311 ResourceManager
13597 DataNode
13471 NameNode

```

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

\$ tar xvfz pig-0.16.0.tar.gz

```

kali-linux-2023.4-vmware-amd64 - VMware Workstation 17 Player (Non-commercial use only)
Player
(hadoop@kali)~$ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz
--2024-08-29 10:55:35-- https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org) ... 144.156.2.132
Connecting to dlcdn.apache.org (dlcdn.apache.org)|2004:1e42::644|443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17729333 (169M) [application/x-gzip]
Saving to: 'pig-0.16.0.tar.gz'

pig-0.16.0.tar.gz      100%[=====] 169.07M  26.5MB/s   in 6.5s

2024-08-29 10:55:41 (26.0 MB/s) - 'pig-0.16.0.tar.gz' saved [17729333/17729333]

(hadoop@kali)~$ 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/math/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/state/
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/avro/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/storage/hbase/
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/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/

```

\$nano ~/.bashrc

```
#PIG settings
export PIG_HOME=/home/hadoop/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_HOME/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
#export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
#PIG setting ends
```

\$mv pig-0.16.0 pig \$pig

```
kali-linux-2023.4-vmware-amd64 - VMware Workstation 17 Player (Non-commercial use only)
Bayer
File Actions Edit View Help
pig 0.16.0/tutorial/src/org/apache/pig/tutorial/tutorialUtil.java
pig 0.16.0/bin/pig
pig 0.16.0/bin/pig.cmd
pig 0.16.0/bin/pig.py
(hadoop@kali) ~
$ ls
(hadoop@kali) ~
$ mv pig-0.16.0 pig
(hadoop@kali) ~
$ ls
(hadoop@kali) ~
$ nano ~/.bashrc
(hadoop@kali) ~
$ source ~/.bashrc
(hadoop@kali) ~
$ java -jar pig.jar
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
55024 ResourceManager
54624 SecondaryNameNode
52128 NameNode
54393 DataNode
64903 Jps
54267 NameNode
(hadoop@kali) ~
$ pig
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-08-29 18:58:45.222 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-29 18:58:45.276 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-29 18:58:45.277 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-29 18:58:45.477 INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746538) compiled Jun 01 2016, 21:10:49
2024-08-29 18:58:45.477 INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig-17249452588897.0002
2024-08-29 18:58:45.562 INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/pigbootstrap not found
2024-08-29 18:58:45.662 INFO org.apache.pig.hadoop.conf.Configuration.deprecation - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-08-29 18:58:46.722 INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-29 18:58:46.722 INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-29 18:58:46.724 INFO org.apache.pig.backend.hadoop.executionengine.MExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-08-29 18:58:46.975 INFO org.apache.pig.PigServer - Pig script ID for the session: PIG-default-5208285-2350-7ba3ba-cdbb2f4bfc2
2024-08-29 18:58:48.975 INFO org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
Shim's quit
2024-08-29 18:59:00.426 INFO org.apache.pig.Main - Pig script completed in 15 seconds and 392 milliseconds (15392 ms)
(hadoop@kali) ~
```

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

```
hadoop@kali: ~/DA-Lab/exp4
File Actions Edit View Help
GNU nano 7.2 sample.txt
1,John
2,Jane
3,Joe
4,Emma
```

\$nano demo\_pig.pig

```
hadoop@kali: ~/DA-Lab/exp4
File Actions Edit View Help
GNU nano 7.2 demo_pig.pig
-- Load the data from HDFS
data = LOAD '/exp4/sample.txt' USING PigStorage(',') AS (id:int, name: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





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

```
(hadoop@kali)-[~/hadoop/bin]
$ ./hdfs dfs -ls /exp4
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-09-21 00:26:01,736 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-08-30 05:07 /exp4/output
-rw-r--r--   1 hadoop supergroup        27 2024-08-30 04:43 /exp4/sample.txt
-rw-r--r--   1 hadoop supergroup       172 2024-08-30 05:02 /exp4/uppercase_udf.py
```

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

```
File Actions Edit View Help
GNU nano 7.2 udf_example.pig
-- Register the Python UDF script
REGISTER 'hdfs:///exp4/uppercase_udf.py' USING jython AS udf;
-- Load some data
data = LOAD 'hdfs:///exp4/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:///exp4/output';
```

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

```
kali:linux-2023.4-vmware-amd64 - VMware Workstation 17 Player (Non-commercial use only)
Player
hadoop@kali: ~/DA-Lab/exp4
File Actions Edit View Help
$ nano udf_example.pig
$ pig -f udf_example.pig
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-08-30 05:08:18,591 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-30 05:08:18,601 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-30 05:08:18,601 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-30 05:08:18,636 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/DA-Lab/exp4/pig17240077/err.log
2024-08-30 05:08:18,685 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-08-30 05:08:20,127 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/pigbootstrap not found
2024-08-30 05:08:20,382 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-30 05:08:20,383 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:20,382 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-08-30 05:08:22,438 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:22,513 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-1dc0cc3c-3e11-47e3-a9e8-561b5b6b38a1
2024-08-30 05:08:22,513 [main] WARN org.apache.pig.PigServer - AFS is disabled since yarn.timeline-service.enabled set to false
2024-08-30 05:08:22,725 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:24,296 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir/tmp/pig_jython_2781842653894475217
2024-08-30 05:08:26,818 [main] WARN org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.reminders is empty. This is not expected unless on testing.
2024-08-30 05:08:26,844 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: udf_uppercase
2024-08-30 05:08:27,784 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:37,867 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:38,407 [main] INFO org.apache.pig.scripting.jython.JythonFunction - No schema defined for function 'uppercase' in /tmp/pig211342629392933085tmp/uppercase_udf.py
2024-08-30 05:08:38,818 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:38,743 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2024-08-30 05:08:38,838 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-08-30 05:08:39,858 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:39,982 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schema tuple] was not set ... will not generate code.
2024-08-30 05:08:39,994 [main] INFO org.apache.pig.plan.logical.optimizer.LogicalRelationalOptimizer - [RULES_ENABLED: [AddProject, ColumnRenamePrune, ConstantCalculator, GroupConstParallelLetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]]]
2024-08-30 05:08:39,983 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699138848 to monitor: collectionUsageThreshold = 489396640, usageThreshold = 489396640
2024-08-30 05:08:39,984 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MRCCompiler - File concatenation threshold: 100 optimizer: false
2024-08-30 05:08:39,988 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-08-30 05:08:39,989 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-08-30 05:08:39,994 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

```

kali-linux-2023.4-vmware-amd64 - VMware Workstation 17 Player (Non-commercial use only)
Player
File Actions Edit View Help
NOS)
2024-09-20 05:18:12,729 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:13,731 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:14,733 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:15,735 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:16,738 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:17,740 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:17,859 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-09-20 05:18:18,862 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:19,864 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:20,866 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:21,868 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:22,871 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:23,874 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:24,876 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:25,879 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:26,881 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:27,883 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECO
NOS)
2024-09-20 05:18:27,987 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2024-09-20 05:18:27,988 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-20 05:18:28,188 [main] INFO org.apache.pig.Main - Pig script completed in 4 minutes, 9 seconds and 747 milliseconds (249747 ms)

hadoop@kali:~/BA-Lab/exp4$ cd ../..
hadoop@kali:~/hadoop/bin$ ./hdfs dfs -cat /exp4/output/part-m-00000
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-09-20 05:12:25,900 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1,JOHN
2,JANE
3,JOE
4,EMMA

hadoop@kali:~/hadoop/bin$

```

**\$hdfs dfs -cat /exp4/output/\***

```

(hadoop@kali)~[~/hadoop/bin]
$ ./hdfs dfs -cat /exp4/output/*
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-09-21 00:33:32,731 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
1,JOHN
2,JANE
3,JOE
4,EMMA

```