

```
#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: ~
$ mv pig-0.16.0 pig
hadoop@kali: ~
$ nano -f ./bashrc
hadoop@kali: ~
$ source ~/.bashrc
hadoop@kali: ~
$ ps
  PID   PPID  STAT SCHED PRI   NA    user   command
  55021  55021  S    S    20    hadoop  java -Djava.library.path=/usr/lib64 -Dwt.useSystemAAFontSettings=on -Dswing.aatext=true
  55021  55021  S    S    20    hadoop  SecondaryNameNode
  55138  55021  S    S    20    hadoop  NodeManager
  54993  55021  S    S    20    hadoop  DataNode
  64953  55021  S    S    20    hadoop  JPS
  54267  55021  S    S    20    hadoop  NameNode
hadoop@kali: ~
$ pig
Picked up _JAVA_OPTIONS: -Dwt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-08-29 18:58:45.222 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-29 18:58:45.226 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-29 18:58:45.227 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-29 18:58:45.477 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 22:10:49
2024-08-29 18:58:45.477 [main] INFO org.apache.pig.Main - logging error messages to: /home/hadoop/pig-172494258897.0002.log
2024-08-29 18:58:45.503 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/pig/bootstrap not found
2024-08-29 18:58:46.667 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-08-29 18:58:46.722 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-29 18:58:46.722 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-29 18:58:46.724 [main] INFO org.apache.pig.backend.hadoop.executionengine.MExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-08-29 18:58:48.972 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-29 18:58:48.972 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-5c78e2d3-235b-475a-a3b8-cb08a274bfc2
2024-08-29 18:58:48.972 [main] WARN org.apache.pig.PigServer - AIs is disabled since yarn.timeline-service.enabled set to false
grunt> quit
2024-08-29 18:59:08.426 [main] 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

[illegible]

\$nano uppercase_udf.py

```

File Actions Edit View 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)

```

uppercase_udf.py

\$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

```
hadoop@kali: ~/DA-Lab/exp4
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)
hadoop@kali: ~/DA-Lab/exp4
File Actions Edit View Help
(hadoop@kali)-[~/DA-Lab/exp4]
$ nano uppercase_udf.py
(hadoop@kali)-[~/DA-Lab/exp4]
$ cd ../..
(hadoop@kali)-[~/hadoop/bin]
$ ./hdfs dfs -copyFromLocal ~/DA-Lab/exp4/uppercase_udf.py /exp4
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-08-30 05:02:54,541 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
(hadoop@kali)-[~/hadoop/bin]
$ cd ../..
(hadoop@kali)-[~/DA-Lab/exp4]
$ nano udf_example.pig
(hadoop@kali)-[~/DA-Lab/exp4]
$ 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 05:08:18,606 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746536) compiled Jun 01 2016, 23:10:49
2024-08-30 05:08:18,606 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/DA-Lab/exp4/pig-1725000778810.log
2024-08-30 05:08:19,685 [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 05:08:20,383 [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-1725000778810-4763-90b-503338633a1
2024-08-30 05:08:22,513 [main] WARN org.apache.pig.PigServer - ATS 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 05:08:24,298 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python cache dir /tmp/pig_jython.3781842658994475167
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 05:08:37,764 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:37,807 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:38,583 [main] INFO org.apache.pig.scripting.jython.JythonFunction - No schema defined for function uppercase in /tmp/pig3133262932933085tmp/uppercase_udf.py
2024-08-30 05:08:38,583 [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 05:08:38,836 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig Features used in the script: UNKNOWN
2024-08-30 05:08:38,958 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-30 05:08:39,382 [main] INFO org.apache.pig.data.SchemaTupleBackend - key (pig.schemaTuple) was not set... will not generate code.
2024-08-30 05:08:39,332 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES_ENABLED:AdaptorEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, M
ergeSorter, MergeSorterEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownWorkForEachPlaten, PushDownFilter, SplitFilter, StreamTypeCastInserter]]
2024-08-30 05:08:39,443 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 692318840 to monitor, collectionUsageThreshold = 489396640, usageThreshold = 489396640
2024-08-30 05:08:39,684 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-08-30 05:08:39,856 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-08-30 05:08:39,859 [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-vmware-amd64 - VMware Workstation 17 Player (Non-commercial use only)
Player
hadoop@kali:~/hadoop/bin
2024-08-30 05:10:12,729 [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
NDS)
2024-08-30 05:10:13,731 [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
NDS)
2024-08-30 05:10:14,733 [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
NDS)
2024-08-30 05:10:15,735 [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
NDS)
2024-08-30 05:10:16,738 [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
NDS)
2024-08-30 05:10:17,740 [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
NDS)
2024-08-30 05:10:17,859 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-08-30 05:10:18,862 [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
NDS)
2024-08-30 05:10:19,864 [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
NDS)
2024-08-30 05:10:20,866 [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
NDS)
2024-08-30 05:10:21,868 [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
NDS)
2024-08-30 05:10:22,871 [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
NDS)
2024-08-30 05:10:23,874 [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
NDS)
2024-08-30 05:10:24,876 [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
NDS)
2024-08-30 05:10:25,879 [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
NDS)
2024-08-30 05:10:26,881 [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
NDS)
2024-08-30 05:10:27,883 [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
NDS)
2024-08-30 05:10:27,987 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2024-08-30 05:10:27,988 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success!
2024-08-30 05:10:28,108 [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/bin

(hadoop@kali)~/hadoop/bin
$ ./hdfs dfs -cat /exp4/output/part-r-00000
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
2024-08-30 05:12:25,908 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

```