Ex:5 Create UDF (User Defined Functions) in Apache Pig and execute it inMapReduce/HDFS mode

AIM:

To create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode.

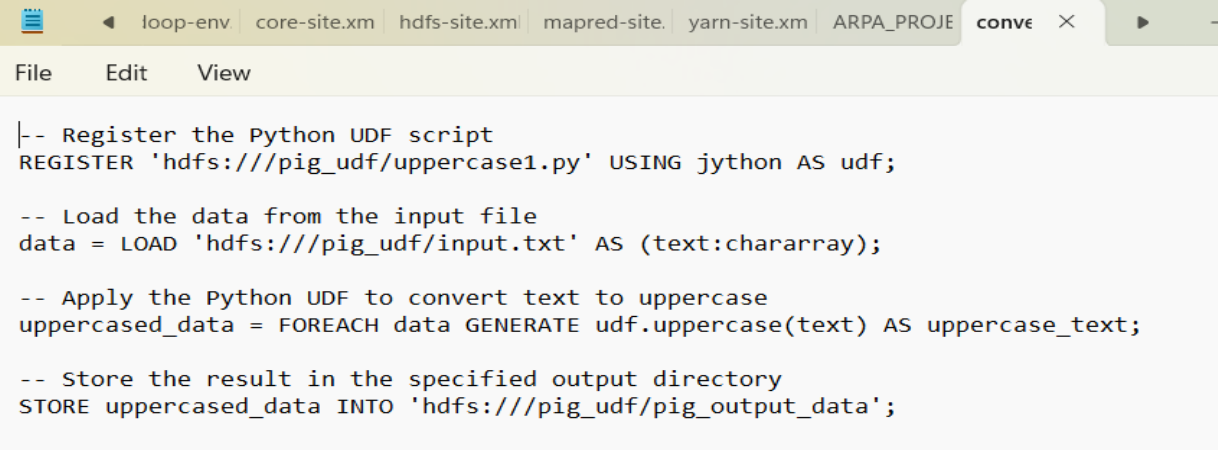
PROCEDURE:

1. Ensure that Apache Pig is installed and configured.

2. Create a python UDF (User Defined Functions).

3. Jython should be installed as Pig will use it to interpret the Python UDFs.

4. Create a Pig script that registers and uses the Python UDF.

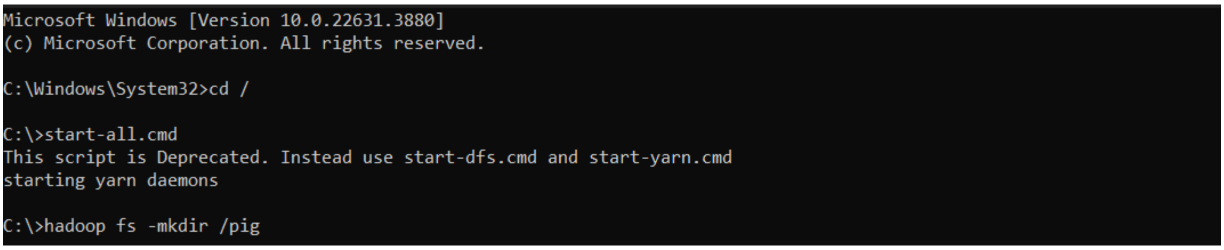


```
-- Register the Python UDF script
REGISTER 'hdfs:///pig_udf/uppercase1.py' USING jython AS udf;

-- Load the data from the input file
data = LOAD 'hdfs:///pig_udf/input.txt' AS (text:chararray);

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

-- Store the result in the specified output directory
STORE uppercased_data INTO 'hdfs:///pig_udf/pig_output_data';
```

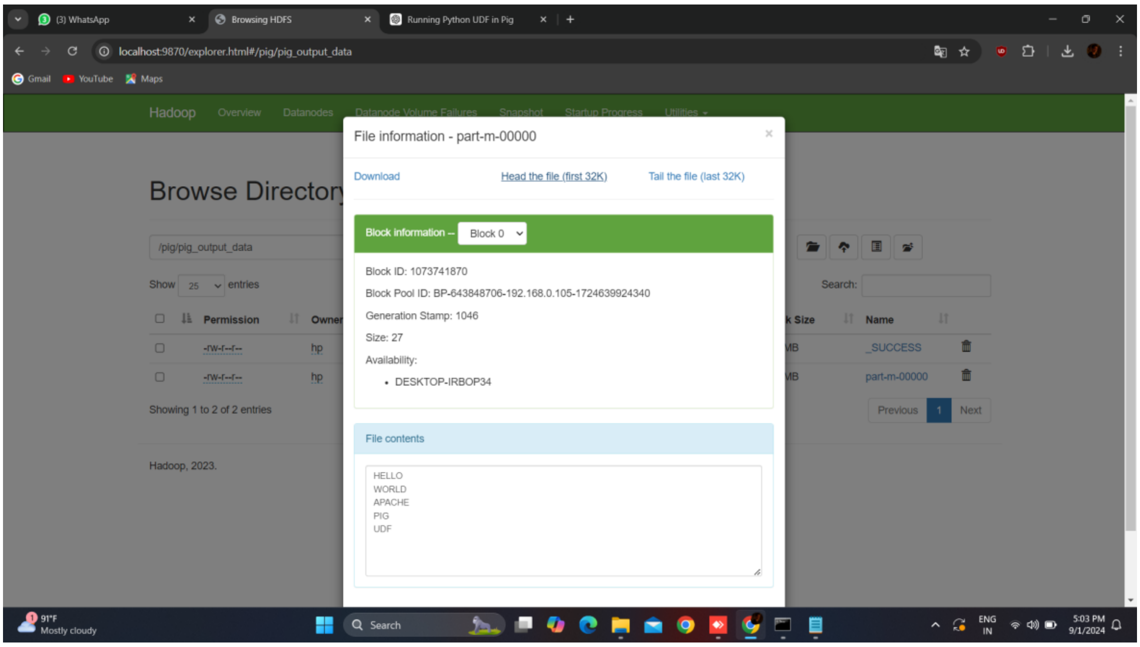
5. Execute the Pig Script in MapReduce Mode using the command:

```
pig -x mapreduce script.pig
```

OUTPUT:



```
Microsoft Windows [Version 10.0.22631.3880]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd /

C:\>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\>hadoop fs -mkdir /pig
```



File information - part-m-00000

Download | Head the file (first 32K) | Tail the file (last 32K)

Block information -- Block 0

Block ID: 1073741870

Block Pool ID: BP-643848706-192.168.0.105-1724639924340

Generation Stamp: 1046

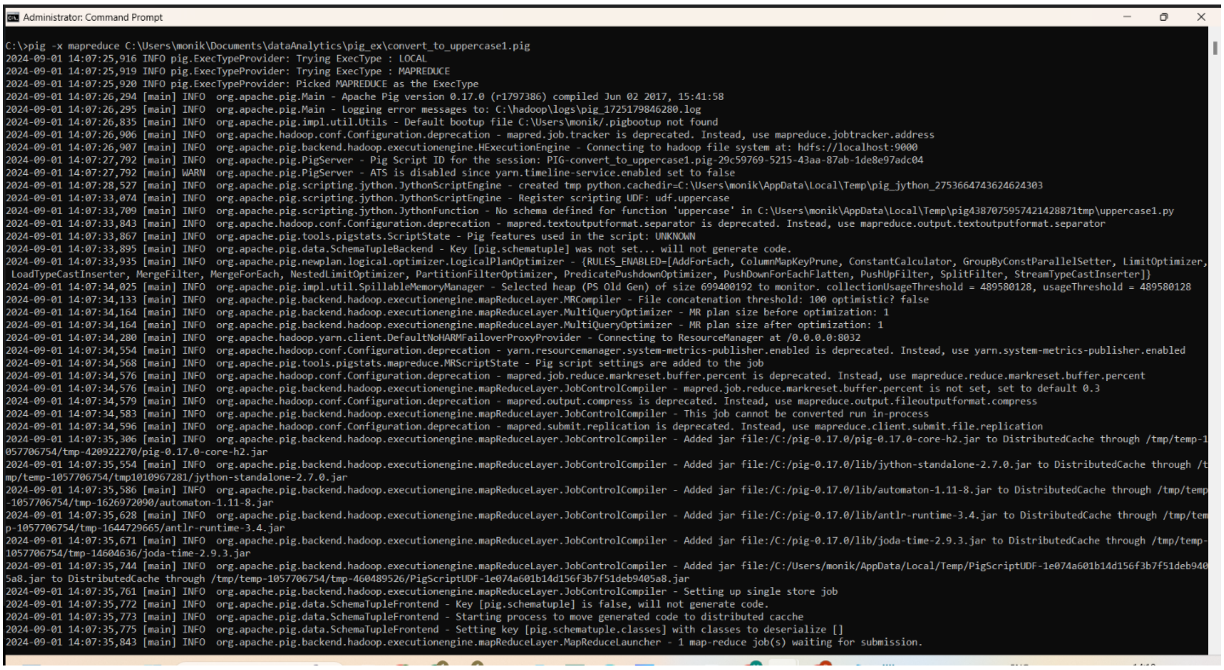
Size: 27

Availability:

- DESKTOP-IRBOP34

File contents

```
HELLO
WORLD
APACHE
PIG
UDF
```



Administrator: Command Prompt

```
C:\>pig -x mapreduce C:\Users\monik\Documents\dataAnalytics\pig_ex\convert_to_uppercase1.pig
2024-09-01 14:07:25,916 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-01 14:07:25,919 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-01 14:07:25,920 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-01 14:07:26,294 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-01 14:07:26,295 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1725179846280.log
2024-09-01 14:07:26,835 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\monik/.pigbootup not found
2024-09-01 14:07:26,906 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-01 14:07:26,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-01 14:07:27,792 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-convert_to_uppercase1.pig-29c59769-5215-43aa-87ab-1de8e97adc04
2024-09-01 14:07:27,792 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-01 14:07:28,527 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=C:\Users\monik\AppData\Local\Temp\pig_jython_2753664743624624303
2024-09-01 14:07:33,074 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: udf.uppercase
2024-09-01 14:07:33,709 [main] INFO  org.apache.pig.scripting.jython.JythonFunction - No schema defined for function 'uppercase' in C:\Users\monik\AppData\Local\Temp\pig4387075957421428871tmp\uppercase1.py
2024-09-01 14:07:33,843 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2024-09-01 14:07:33,867 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-01 14:07:33,895 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-01 14:07:33,935 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer,
 LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-09-01 14:07:34,025 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2024-09-01 14:07:34,133 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-01 14:07:34,164 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-01 14:07:34,164 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-01 14:07:34,280 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2024-09-01 14:07:34,554 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-09-01 14:07:34,568 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2024-09-01 14:07:34,576 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2024-09-01 14:07:34,576 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-01 14:07:34,579 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2024-09-01 14:07:34,583 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2024-09-01 14:07:34,596 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2024-09-01 14:07:35,306 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/C:/pig-0.17.0/pig-0.17.0-core-h2.jar to DistributedCache through /tmp/temp-1
057706754/tmp-420922270/pig-0.17.0-core-h2.jar
2024-09-01 14:07:35,554 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/C:/pig-0.17.0/lib/jython-standalone-2.7.0.jar to DistributedCache through /t
mp/temp-1057706754/tmp1010967281/jython-standalone-2.7.0.jar
2024-09-01 14:07:35,586 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/C:/pig-0.17.0/lib/automaton-1.11-8.jar to DistributedCache through /tmp/temp
-1057706754/tmp-1626972090/automaton-1.11-8.jar
2024-09-01 14:07:35,628 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/C:/pig-0.17.0/lib/antlr-runtime-3.4.jar to DistributedCache through /tmp/tem
p-1057706754/tmp-1644729665/antlr-runtime-3.4.jar
2024-09-01 14:07:35,671 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/C:/pig-0.17.0/lib/joda-time-2.9.3.jar to DistributedCache through /tmp/temp-
1057706754/tmp-14604636/joda-time-2.9.3.jar
2024-09-01 14:07:35,744 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/C:/Users/monik/AppData/Local/Temp/PigScriptUDF-1e074a601b14d156f3b7f51deb940
5a8.jar to DistributedCache through /tmp/temp-1057706754/tmp-460489526/PigScriptUDF-1e074a601b14d156f3b7f51deb9405a8.jar
2024-09-01 14:07:35,761 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-01 14:07:35,772 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-01 14:07:35,773 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2024-09-01 14:07:35,775 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2024-09-01 14:07:35,843 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
```

RESULT:

Thus, to create a UDF in Apache Pig and execute in MapReduce mdoe has been executed successfully .