

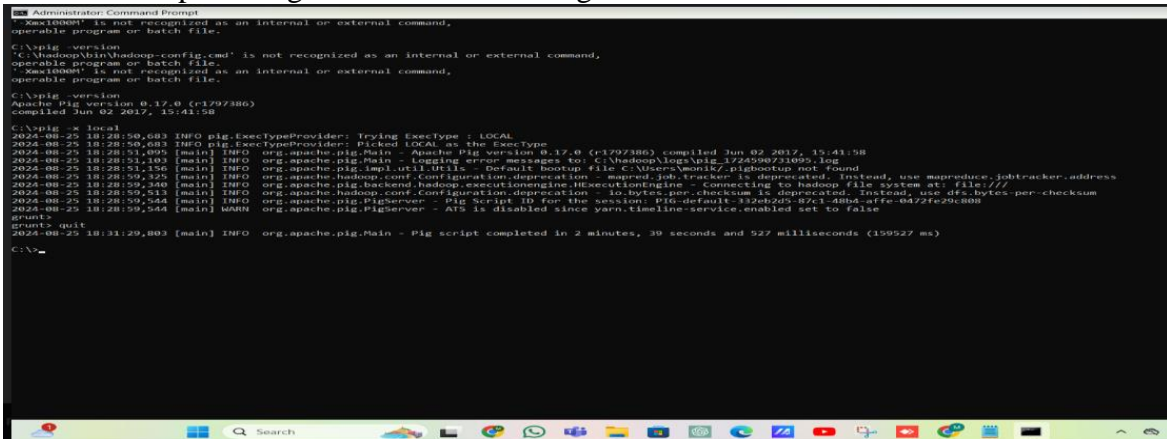
Ex:5 Create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode

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.



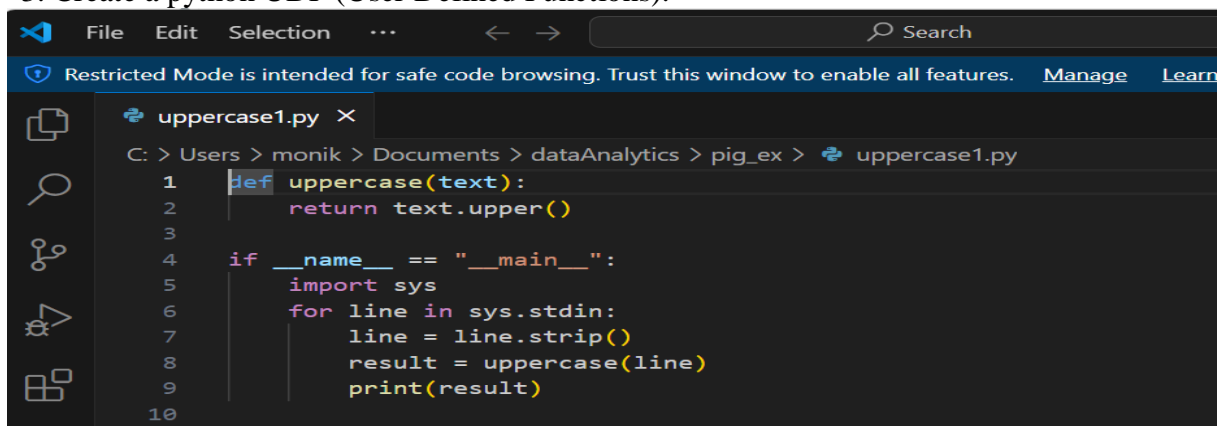
```
Administrator Command Prompt
C:\> pig -version
'Apache Pig' is not recognized as an internal or external command,
operable program or batch file.

C:\> pig -version
'Apache Pig' is not recognized as an internal or external command,
operable program or batch file.

C:\> pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58

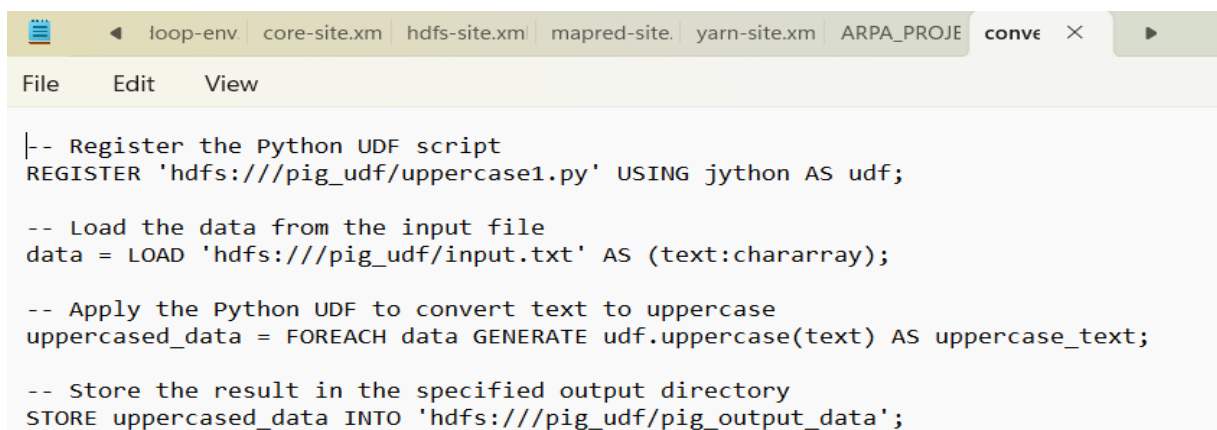
C:\> pig -x local
2024-08-25 18:28:50.683 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-25 18:28:50.683 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-08-25 18:28:51.095 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-08-25 18:28:51.103 [main] INFO org.apache.pig.Main - Logging error messages to: C:\Hadoop\log\pig-172259873895.log
2024-08-25 18:28:51.156 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file C:\Users\monik/.pigbootstrap not found
2024-08-25 18:28:59.345 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-25 18:28:59.340 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-08-25 18:28:59.513 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes.per.checksum
2024-08-25 18:28:59.544 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-332eb2d5-87c1-48b4-affe-0472fe29c808
2024-08-25 18:28:59.544 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> quit
2024-08-25 18:31:29.803 [main] INFO org.apache.pig.Main - Pig script completed in 2 minutes, 39 seconds and 527 milliseconds (159527 ms)
C:\>
```

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



```
File Edit Selection ... Search
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn
uppercase1.py X
C: > Users > monik > Documents > dataAnalytics > pig_ex > uppercase1.py
1 def uppercase(text):
2     return text.upper()
3
4 if __name__ == "__main__":
5     import sys
6     for line in sys.stdin:
7         line = line.strip()
8         result = uppercase(line)
9         print(result)
10
```

4. Jython should be installed as Pig will use it to interpret the Python UDFs.
5. Create a Pig script that registers and uses the Python UDF.



```
loop-env core-site.xml hdfs-site.xml mapred-site yarn-site.xml ARPA_PROJE conve X
File Edit View
|-- 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';
```

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

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

OUTPUT:

```
Administrator: Command Prompt
2024-08-27 20:23:32,504 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-08-27 20:23:32,516 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-08-27 20:23:32,521 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-08-27 20:23:32,521 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the path specified.
-- file path: tmp/temp479226616/tmp-1994028954/part-m-00000
2024-08-27 20:23:32,585 [main] WARN org.apache.hadoop.io.nativeio.NativeIO - NativeIO.getStat error (3): The system cannot find the path specified.
-- file path: tmp/temp479226616/tmp-1994028954/_SUCCESS
2024-08-27 20:23:32,678 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-08-27 20:23:32,680 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(HI)
(HELLO)
(MONIKA)
(GOOD)
(MORNING)
(NIGHT)
2024-08-27 20:23:32,777 [main] INFO org.apache.pig.Main - Pig script completed in 8 seconds and 468 milliseconds (8468 ms)
C:\Users\monik\Documents\dataAnalytics\pig_ex>

Administrator: Command Prompt
C:\Users\monik\Documents\dataAnalytics\pig_ex>pig -x local convert_to_uppercase.pig
2024-08-27 20:23:24,614 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-27 20:23:24,614 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-08-27 20:23:25,044 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-08-27 20:23:25,056 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1724770405044.1.log
2024-08-27 20:23:25,072 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2024-08-27 20:23:25,277 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\monik/.pigbootup not found
2024-08-27 20:23:25,339 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-27 20:23:25,354 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-08-27 20:23:25,370 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-convert_to_uppercase.pig-ae9d574c-43d7-42b0-a0b8-e7660cd66a4d
2024-08-27 20:23:25,370 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-08-27 20:23:25,449 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=C:\Users\monik\AppData\Local\Temp\pig_jython_4779704809615964497
2024-08-27 20:23:29,900 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: myfuncs.to_uppercase
2024-08-27 20:23:30,084 [main] INFO org.apache.pig.scripting.jython.JythonFunction - Schema 'word:chararray' defined for func to_uppercase
2024-08-27 20:23:30,241 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-08-27 20:23:30,288 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeF
```

File information - part-m-00000

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

Block information -- Block 0

Block ID: 1073741887
Block Pool ID: BP-176020507-192.168.56.1-1724161408547
Generation Stamp: 1063
Size: 35
Availability:
• 192.168.56.1

File contents

```
HI  
HELLO  
MONIKA  
GOOD  
MORNING  
NIGHT
```

Administrator: Command Prompt

```
C:\>pig -x mapreduce C:\Users\monik\Documents\data\analytics\pig_ex\convert_to_uppercase.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 bootstrap file C:\Users\monik\pig\bootstrap not found  
2024-09-01 14:07:26,986 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2024-09-01 14:07:26,987 [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_uppercase.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 14:07:33,789 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - No schema defined for function 'uppercase' in C:\Users\monik\AppData\Local\Temp\pig438707595742128871tmp\uppercase.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.mapreduce.LiteMapReduceEngine - File concatenation threshold: 100 optimistic? false  
2024-09-01 14:07:34,164 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - MR plan size before optimization: 1  
2024-09-01 14:07:34,164 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - MR plan size after optimization: 1  
2024-09-01 14:07:34,280 [main] INFO org.apache.hadoop.yarn.client.DefaultHARMFailoverProxyProvider - 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.MRScripScriptState - 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.mapreduce.LiteMapReduceEngine - 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.fileoutformat.compress  
2024-09-01 14:07:34,583 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - 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 14:07:35,386 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - Added jar file:C:\pig-0.17.0\pig-0.17.0-core-h2.jar to DistributedCache through /tmp/temp-1057786754/tmp-428922270/pig-0.17.0-core-h2.jar  
2024-09-01 14:07:35,554 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - Added jar file:C:\pig-0.17.0\lib\jython-standalone-2.7.0.jar to DistributedCache through /tmp/temp-1057786754/tmp-1010967281/jython-standalone-2.7.0.jar  
2024-09-01 14:07:35,586 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - Added jar file:C:\pig-0.17.0\lib\automaton-1.11-8.jar to DistributedCache through /tmp/temp-1057786754/tmp-1026972090/automaton-1.11-8.jar  
2024-09-01 14:07:35,628 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - Added jar file:C:\pig-0.17.0\lib\antlr-runtime-3.4.jar to DistributedCache through /tmp/temp-1057786754/tmp-1644729665/antlr-runtime-3.4.jar  
2024-09-01 14:07:35,671 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - Added jar file:C:\pig-0.17.0\lib\joda-time-2.9.3.jar to DistributedCache through /tmp/temp-1057786754/tmp-14604636/joda-time-2.9.3.jar  
2024-09-01 14:07:35,744 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - Added jar file:C:\Users\monik\AppData\Local\Temp\PigScriptUDF-1e074a601b1d4156f3b7f51deb9405a8.jar to DistributedCache through /tmp/temp-1057786754/tmp-460489526/PigScriptUDF-1e074a601b1d4156f3b7f51deb9405a8.jar  
2024-09-01 14:07:35,761 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - 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 cache  
2024-09-01 14:07:35,775 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig:schematuple.class] with classes to deserialize []  
2024-09-01 14:07:35,843 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.LiteMapReduceEngine - 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

