Ex no: 4

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

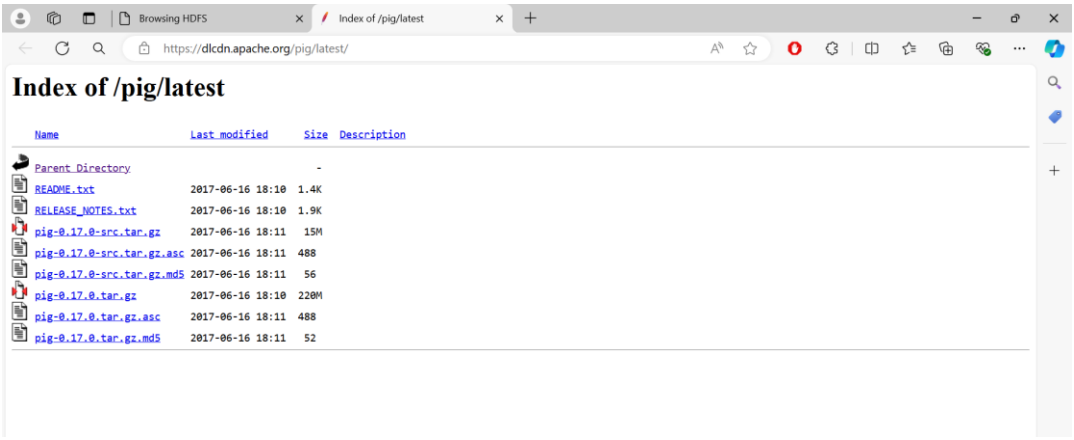
Aim:

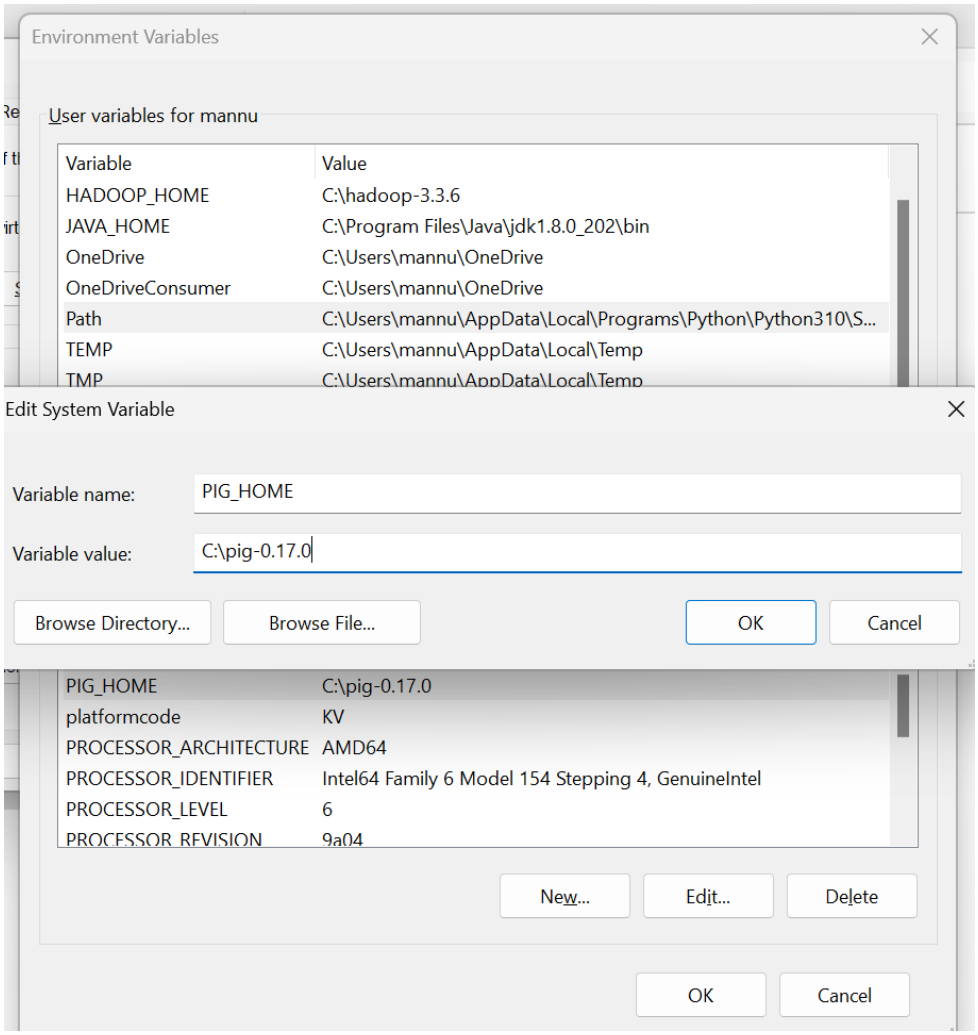
To create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode.

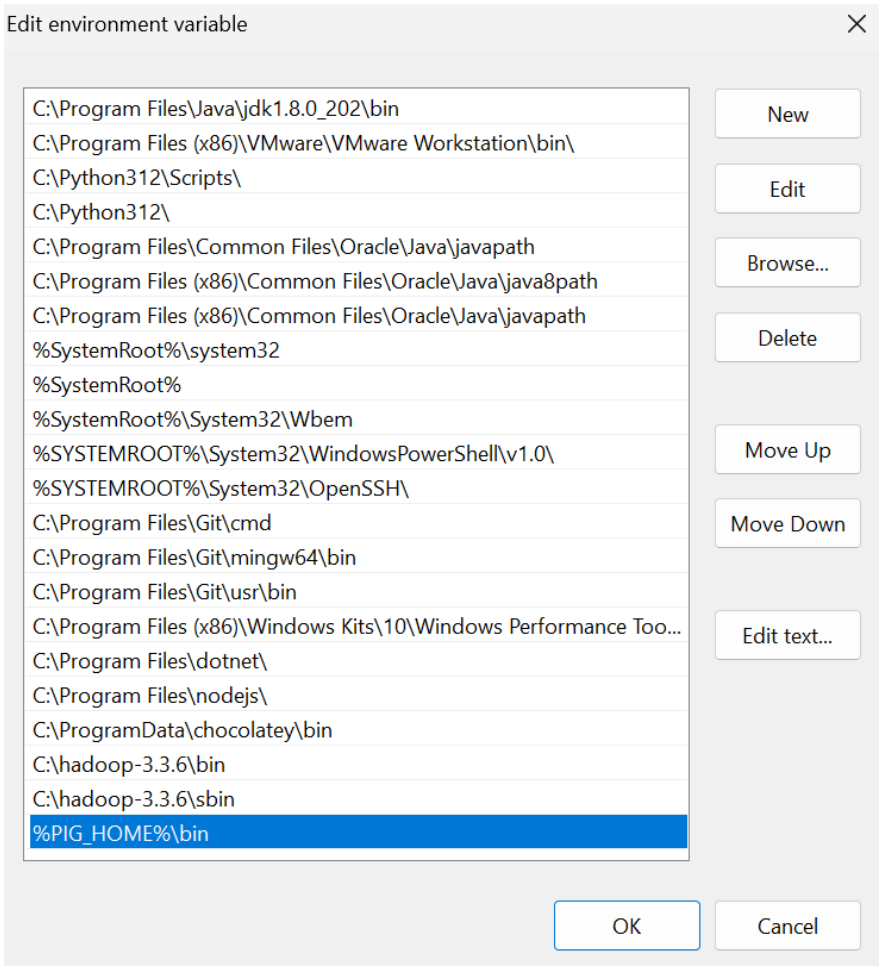
Procedure:

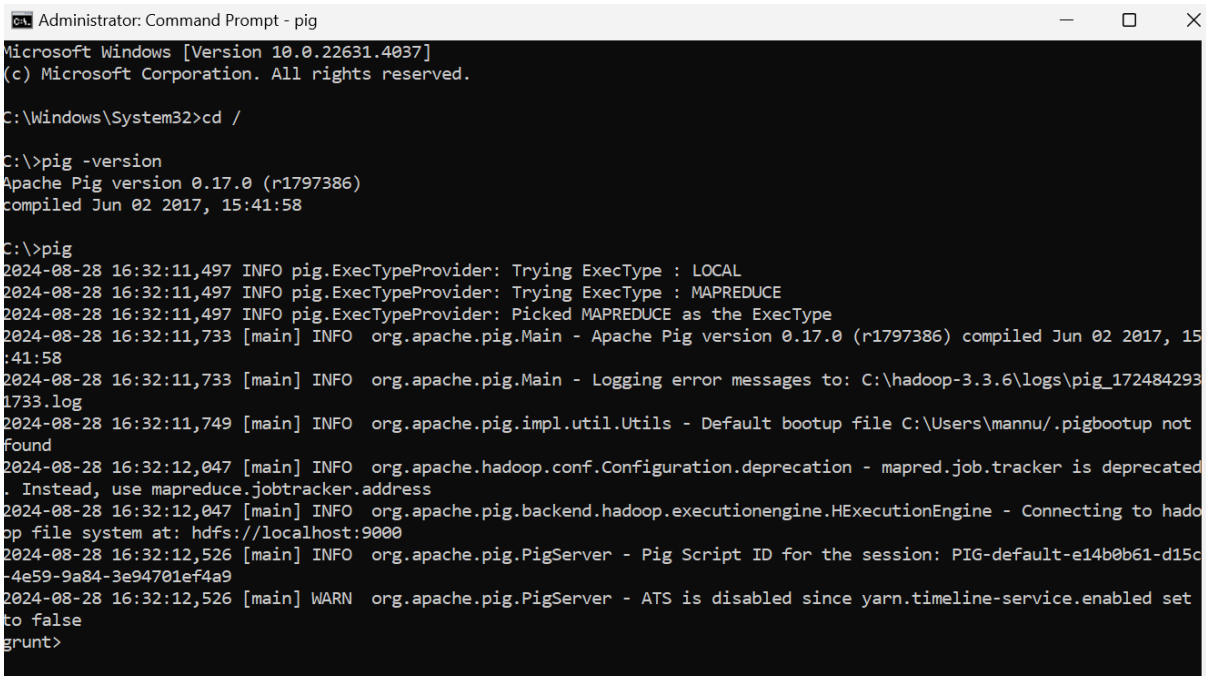
1. Install and configure Apache Pig.
2. Create a python User Defined Function (UDF).
3. Ensure Jython is installed as Pig uses it to interpret the python UDFs.
4. Create a Pig Script that registers and uses the python UDF.
5. Start the hadoop services.
6. Create a new directory for pig and put the input text file to the directory.
7. Start Apache pig in command prompt.
8. Run the script in MapReduce mode using the command “exec script.py”
9. Now open the output folder in the pig directory and take a screenshot of the output.

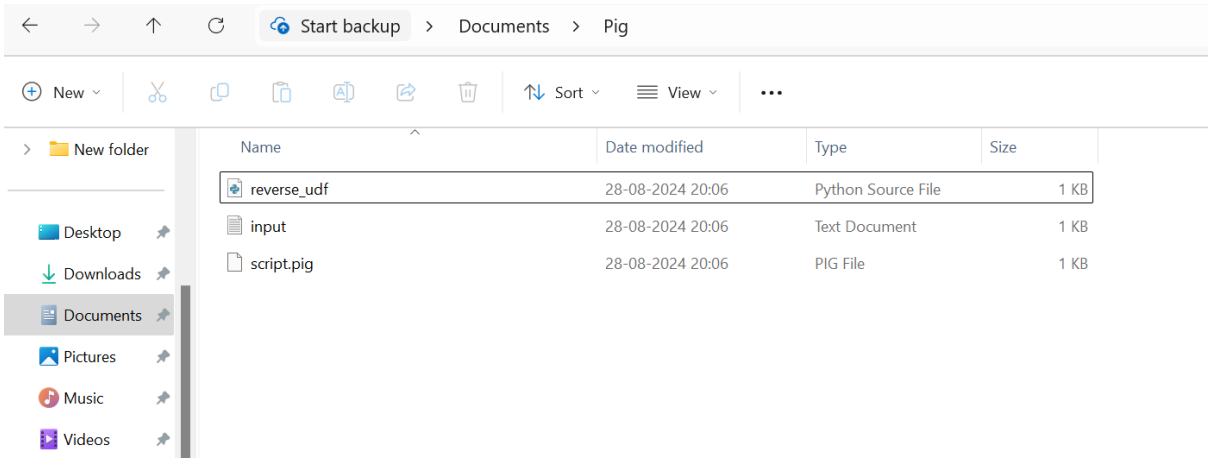
Output Screenshots:

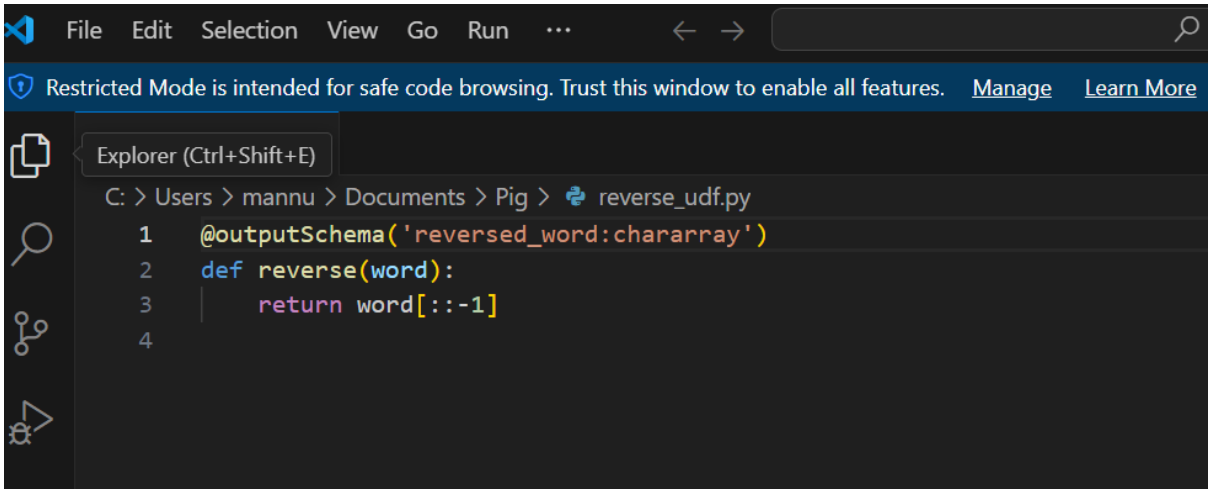


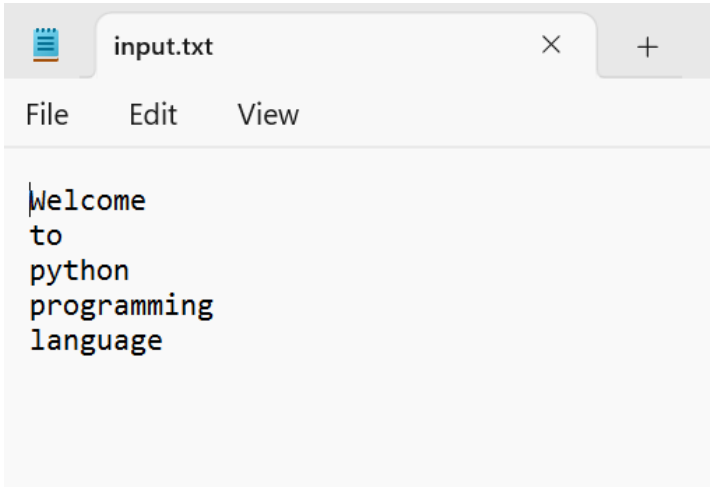

Environment Variables
User variables for mannu
Variable
Value
HADOOP_HOME
C:\hadoop-3.3.6
JAVA_HOME
C:\Program Files\Java\jdk1.8.0_202\bin
OneDrive
C:\Users\mannu\OneDrive
OneDriveConsumer
C:\Users\mannu\OneDrive
Path
C:\Users\mannu\AppData\Local\Programs\Python\Python310\S...
TEMP
C:\Users\mannu\AppData\Local\Temp
TMP
C:\Users\mannu\AppData\Local\Temp
Edit System Variable
Variable name:
PIG_HOME
Variable value:
C:\pig-0.17.0
Browse Directory...
Browse File...
OK
Cancel
PIG_HOME
C:\pig-0.17.0
platformcode
KV
PROCESSOR_ARCHITECTURE
AMD64
PROCESSOR_IDENTIFIER
Intel64 Family 6 Model 154 Stepping 4, GenuineIntel
PROCESSOR_LEVEL
6
PROCESSOR_REVISION
9a04
New...
Edit...
Delete
OK
Cancel


Edit environment variable
C:\Program Files\Java\jdk1.8.0_202\bin
C:\Program Files (x86)\VMware\VMware Workstation\bin\
C:\Python312\Scripts\
C:\Python312\
C:\Program Files\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\Common Files\Oracle\Java\java8path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files\Git\cmd
C:\Program Files\Git\mingw64\bin
C:\Program Files\Git\usr\bin
C:\Program Files (x86)\Windows Kits\10\Windows Performance Too...
C:\Program Files\dotnet\
C:\Program Files\nodejs\
C:\ProgramData\chocolatey\bin
C:\hadoop-3.3.6\bin
C:\hadoop-3.3.6\sbin
%PIG_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel


Administrator: Command Prompt - pig
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.
C:\Windows\System32>cd /
C:\>pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
C:\>pig
2024-08-28 16:32:11,497 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-28 16:32:11,497 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-28 16:32:11,497 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-28 16:32:11,733 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2024-08-28 16:32:11,733 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.6\logs\pig_172484293
1733.log
2024-08-28 16:32:11,749 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\mannu/.pigbootup not
found
2024-08-28 16:32:12,047 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-08-28 16:32:12,047 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://localhost:9000
2024-08-28 16:32:12,526 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-e14b0b61-d15c
-4e59-9a84-3e94701ef4a9
2024-08-28 16:32:12,526 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set
to false
grunt>


Start backup
Documents
Pig
New
Sort
View
New folder
Desktop
Downloads
Documents
Pictures
Music
Videos
Name
Date modified
Type
Size
reverse_udf
28-08-2024 20:06
Python Source File
1 KB
input
28-08-2024 20:06
Text Document
1 KB
script.pig
28-08-2024 20:06
PIG File
1 KB


File Edit Selection View Go Run
Restricted Mode is intended for safe code browsing. Trust this window to enable all features.
Manage
Learn More
Explorer (Ctrl+Shift+E)
C: > Users > mannu > Documents > Pig > reverse_udf.py
@outputSchema('reversed_word:chararray')
def reverse(word):
return word[::-1]


input.txt
File Edit View
Welcome
to
python
programming
language

script.pig

File Edit View

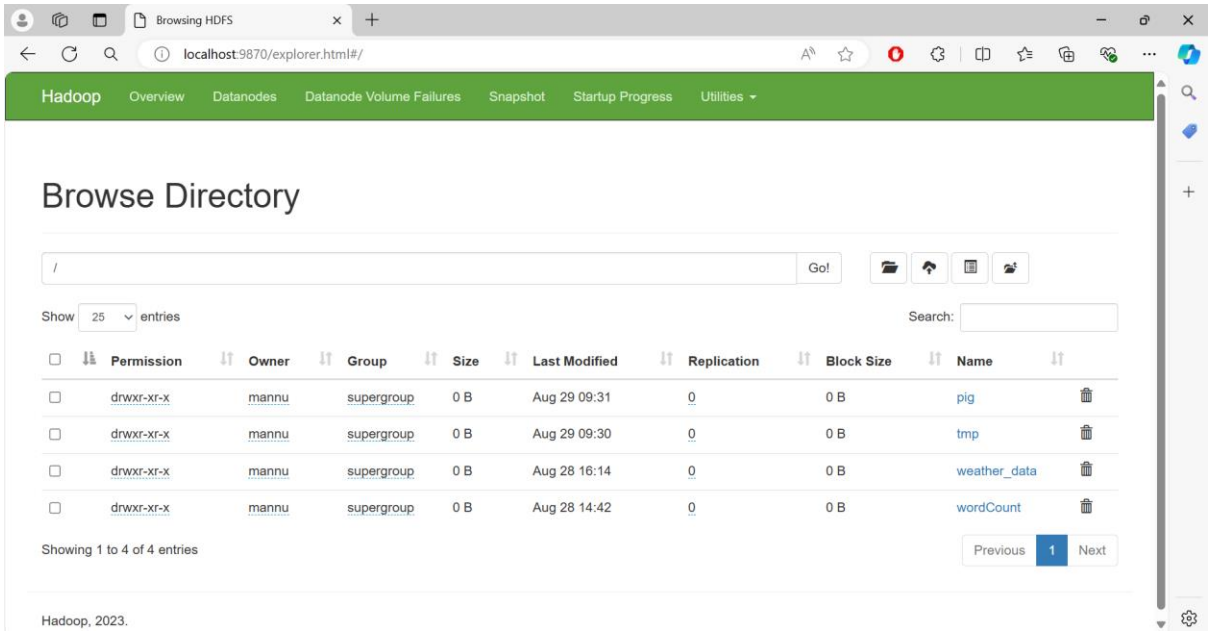
```
-- Register the Python UDF
REGISTER 'reverse_udf.py' USING jython AS myudf;

-- Load a sample dataset from HDFS
data = LOAD 'hdfs:///pig/input.txt' USING PigStorage(',') AS (word:chararray);

-- Apply the UDF to each record
reversed_data = FOREACH data GENERATE myudf.reverse(word);

-- Store the result back to HDFS
STORE reversed_data INTO 'hdfs:///pig/output' USING PigStorage(',');
```

C:\Windows\System32\cmd.e

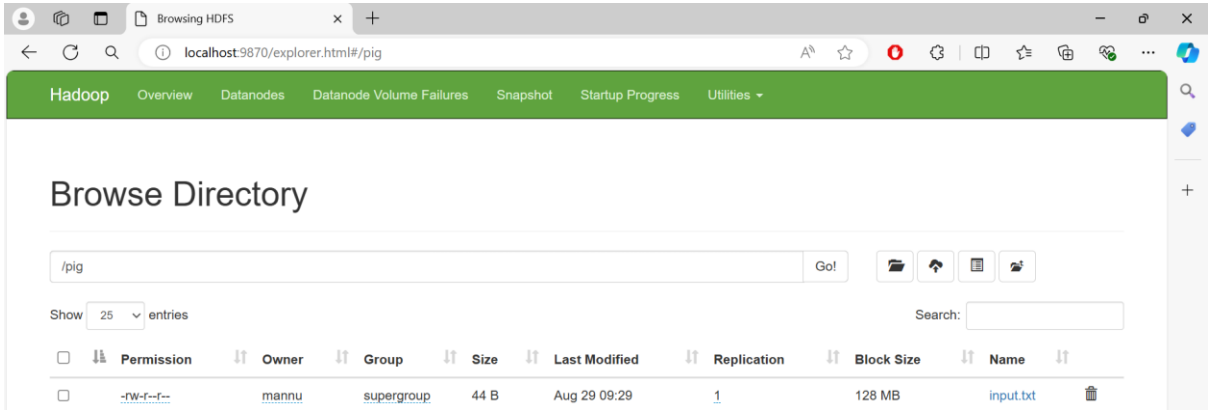
```
Microsoft Windows [Version 10.0.22631.4112]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mannu\Documents\Pig>pig -x mapreduce
2024-08-29 09:24:53,641 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-29 09:24:53,641 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-29 09:24:53,641 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-29 09:24:53,893 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2024-08-29 09:24:53,893 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.6\logs\pig_172490369
3893.log
2024-08-29 09:24:53,924 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\mannu/.pigbootup not
found
2024-08-29 09:24:54,209 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-08-29 09:24:54,209 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://localhost:9000
2024-08-29 09:24:54,633 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-bff0ebd8-9e54
-4071-ac51-1fa0b5a5d537
2024-08-29 09:24:54,633 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set
to false
grunt>
```

Administrator: Command Prompt

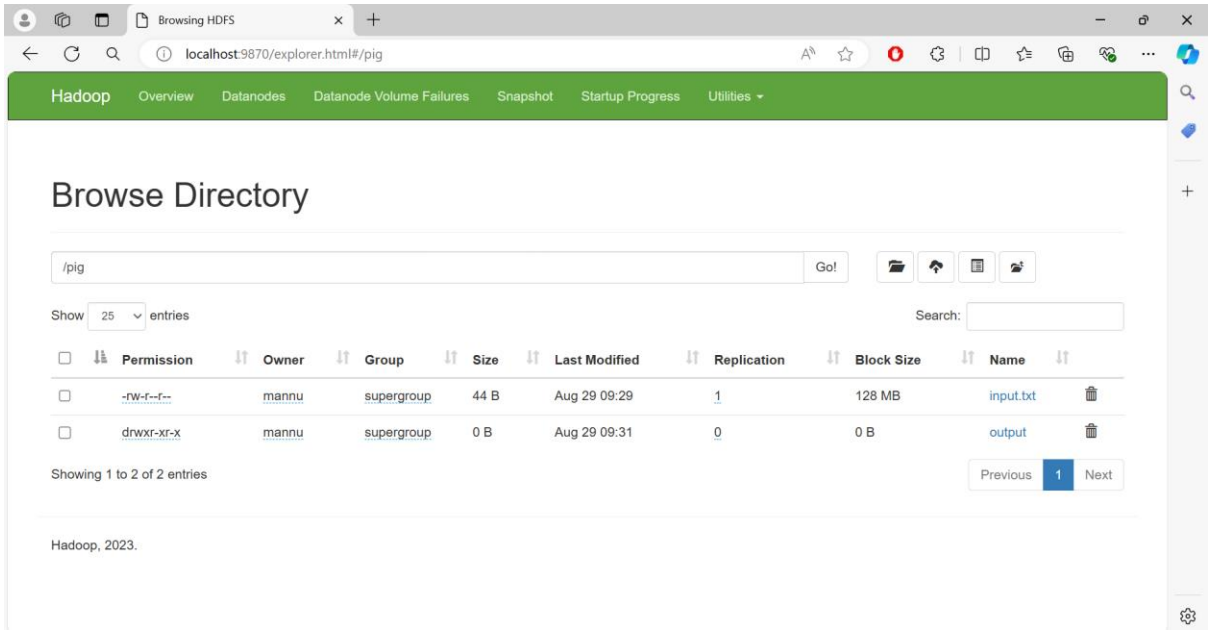
```
Microsoft Windows [Version 10.0.22631.4112]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd /

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

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

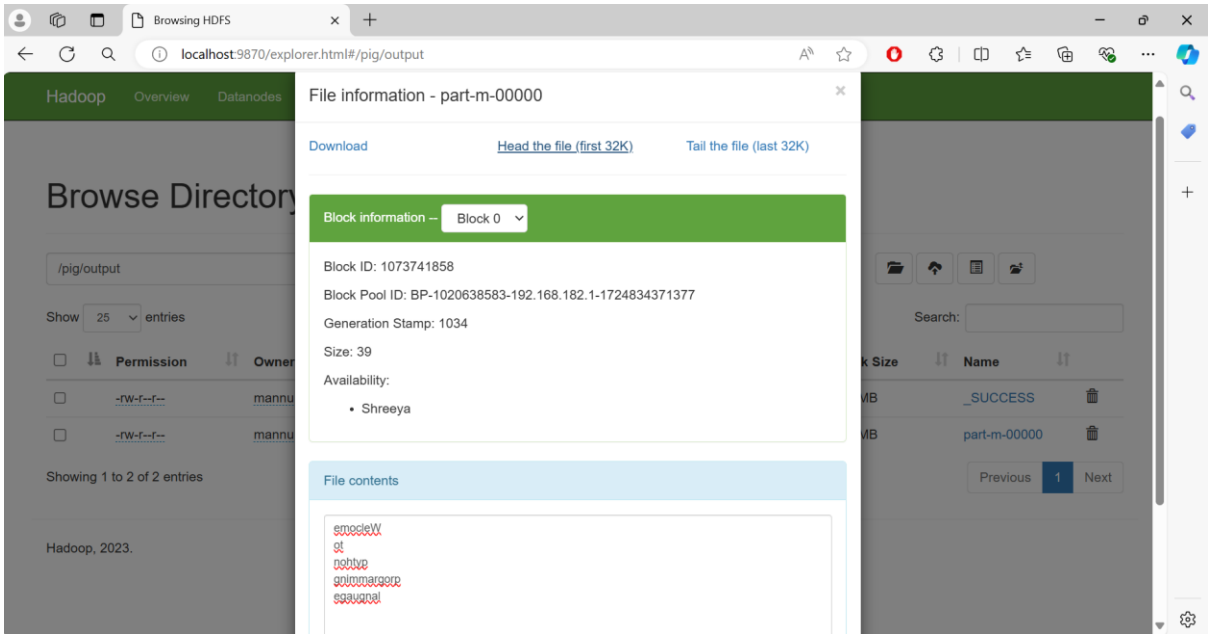



```
C:\>hadoop fs -put C:\\Users\\mannu\\Documents\\Pig\\input.txt /pig

C:\>
```




```
grunt> exec script.pig
2024-08-29 09:30:50,531 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=C:
\Users\mannu\AppData\Local\Temp\pig_jython_3682435500788065235
2024-08-29 09:30:54,013 [main] WARN  org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.remainders is emp
ty. This is not expected unless on testing.
2024-08-29 09:30:54,608 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: myudf.
reverse
2024-08-29 09:30:55,134 [main] INFO  org.apache.pig.scripting.jython.JythonFunction - Schema 'reversed_word:chararray' d
efined for func reverse
2024-08-29 09:30:55,503 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separato
r is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2024-08-29 09:30:55,530 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOW
N
2024-08-29 09:30:55,566 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will
not generate code.
2024-08-29 09:30:55,615 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[Add
ForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeF
ilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten,
 PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-08-29 09:30:55,698 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of siz
e 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
```






```
C:\>hdfs dfs -cat /pig/output/part-m-00000
emocleW
ot
nohtyp
gnimmargorp
egaugnal

C:\>
```

Result:

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