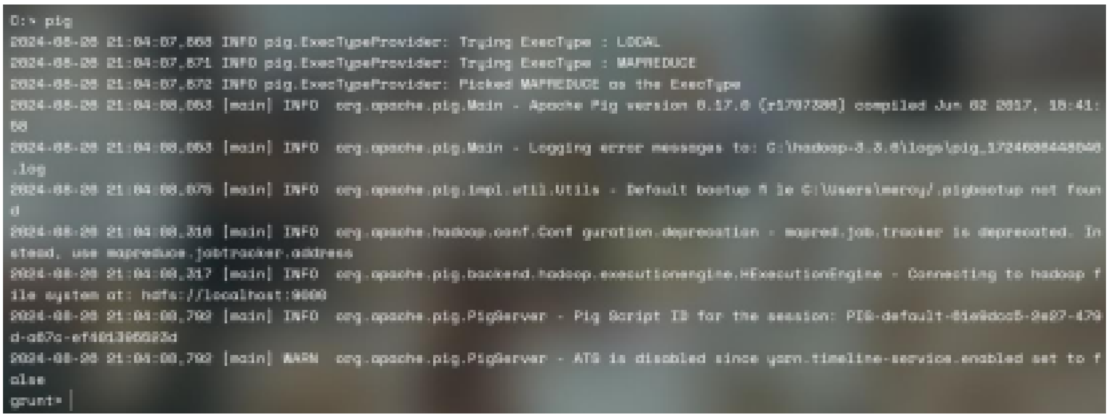
Create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode

AIM:

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

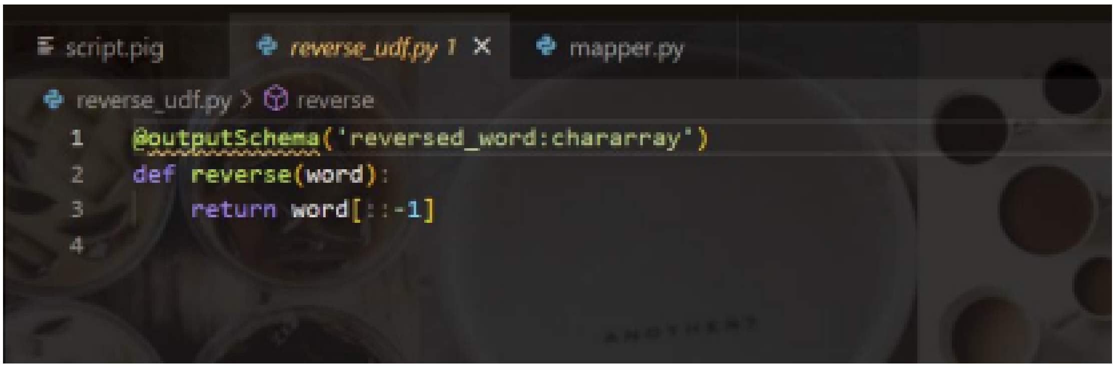
PROCEDURE:

1. Ensure that Apache Pig is installed and configured.



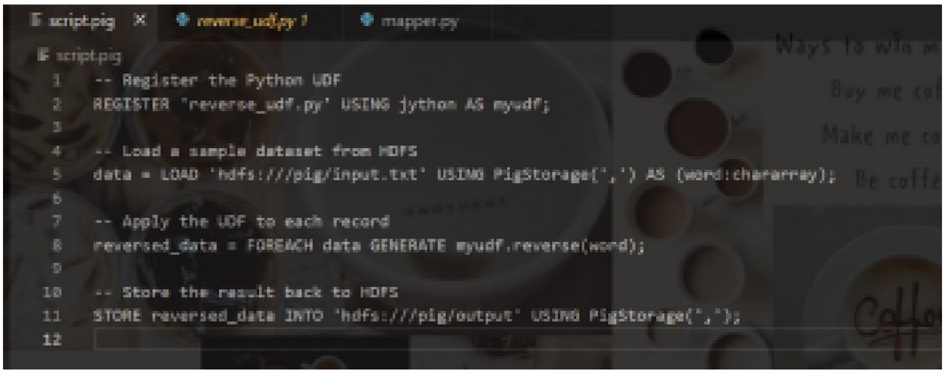
```
D:> pig
2024-08-28 21:04:07,868 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-28 21:04:07,871 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-28 21:04:07,872 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-28 21:04:08,063 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:
58
2024-08-28 21:04:08,063 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.0\logs\pig_1724858448048
.log
2024-08-28 21:04:08,075 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\meroy/.pigbootup not foun
d
2024-08-28 21:04:08,316 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. In
stead, use mapreduce.jobtracker.address
2024-08-28 21:04:08,317 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
ile system at: hdfs://localhost:9000
2024-08-28 21:04:08,792 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-81e8dcc5-9e27-479
d-a87c-ef4013966924
2024-08-28 21:04:08,792 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to f
alse
grunt>
```

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



```
@outputSchema('reversed_word:chararray')
def reverse(word):
    return word[::-1]
```

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.

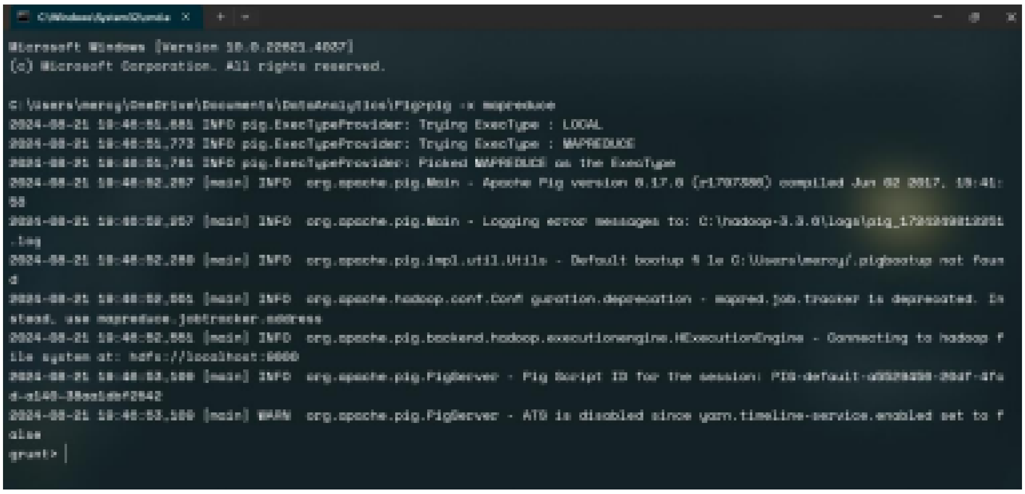


```
-- Register the Python UDF
REGISTER 'reverse_udf.py' USING jython AS myudf;

-- 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(',');
```

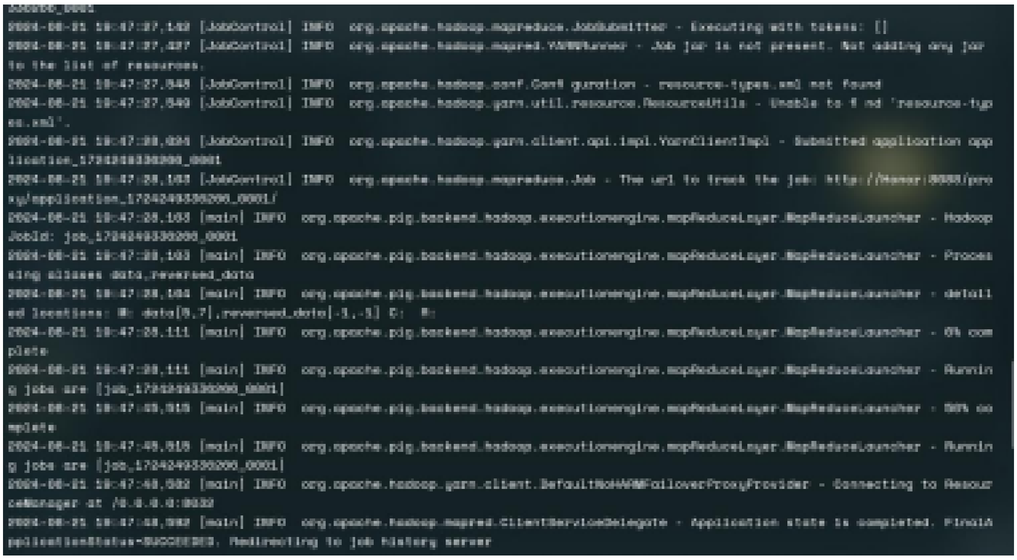
5. Execute the Pig Script in MapReduce Mode using the command:

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

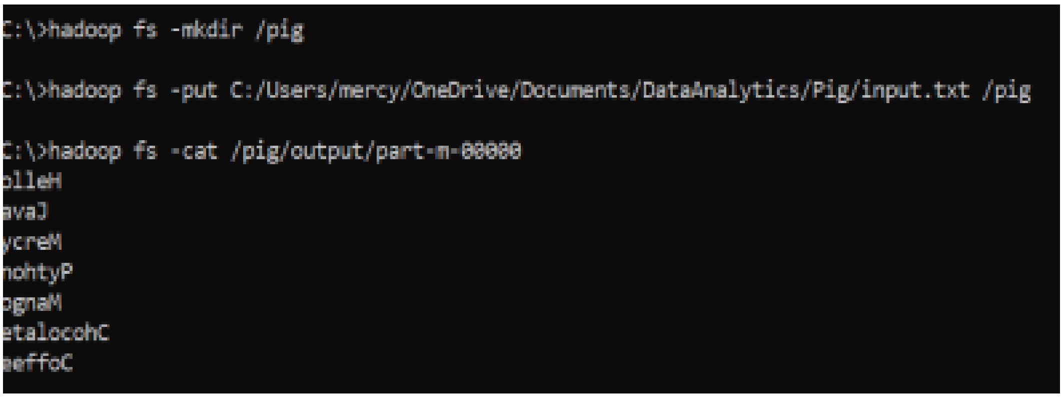
OUTPUT:



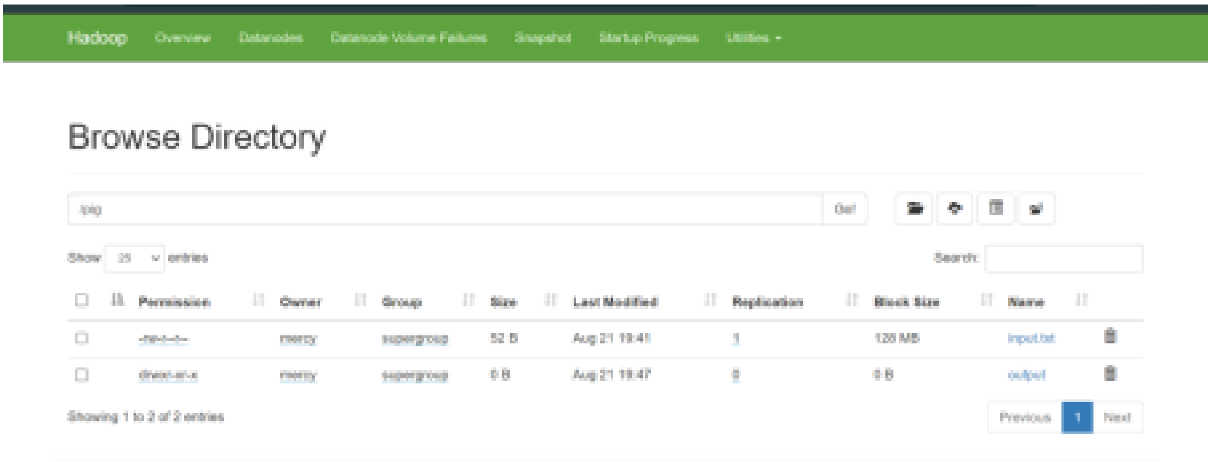
```
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mercy\OneDrive\Documents\DataAnalytics\Pig>pig -x mapreduce
2024-08-21 19:46:51,681 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-21 19:46:51,773 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-21 19:46:51,781 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-21 19:46:52,287 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:
58
2024-08-21 19:46:52,287 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.0\logs\pig_1724249812281
.log
2024-08-21 19:46:52,288 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\mercy/.pigbootup not foun
d
2024-08-21 19:46:52,551 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. In
stead, use mapreduce.jobtracker.address
2024-08-21 19:46:52,551 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
ile system at: hdfs://localhost:9000
2024-08-21 19:46:53,109 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-a9928498-20df-4fa
d-a140-38aa1dbf2642
2024-08-21 19:46:53,109 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to f
alse
grunt>
```



```
336206_0001
2024-08-21 19:47:27,142 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2024-08-21 19:47:27,427 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar
to the list of resources.
2024-08-21 19:47:27,548 [JobControl] INFO  org.apache.hadoop.conf.Configuration - resource-types.xml not found
2024-08-21 19:47:27,549 [JobControl] INFO  org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-typ
es.xml'.
2024-08-21 19:47:28,026 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application app
lication_1724249336206_0001
2024-08-21 19:47:28,103 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://Honor:8088/pro
xy/application_1724249336206_0001/
2024-08-21 19:47:28,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Hadoop
JobId: job_1724249336206_0001
2024-08-21 19:47:28,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Proces
sing aliases data,reversed_data
2024-08-21 19:47:28,104 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detail
ed locations: M: data[5,7],reversed_data[-1,-1] C:  R:
2024-08-21 19:47:28,111 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% com
plete
2024-08-21 19:47:28,111 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Runnin
g jobs are [job_1724249336206_0001]
2024-08-21 19:47:45,515 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% co
mplete
2024-08-21 19:47:45,515 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Runnin
g jobs are [job_1724249336206_0001]
2024-08-21 19:47:48,582 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to Resour
ceManager at /0.0.0.0:8032
2024-08-21 19:47:48,592 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCEEDED. Redirecting to job history server
```



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

C:\>hadoop fs -put C:/Users/mercy/OneDrive/Documents/DataAnalytics/Pig/input.txt /pig

C:\>hadoop fs -cat /pig/output/part-m-00000
olleH
avaJ
ycreM
nohtyP
ognaM
etalocohC
eeffoC
```



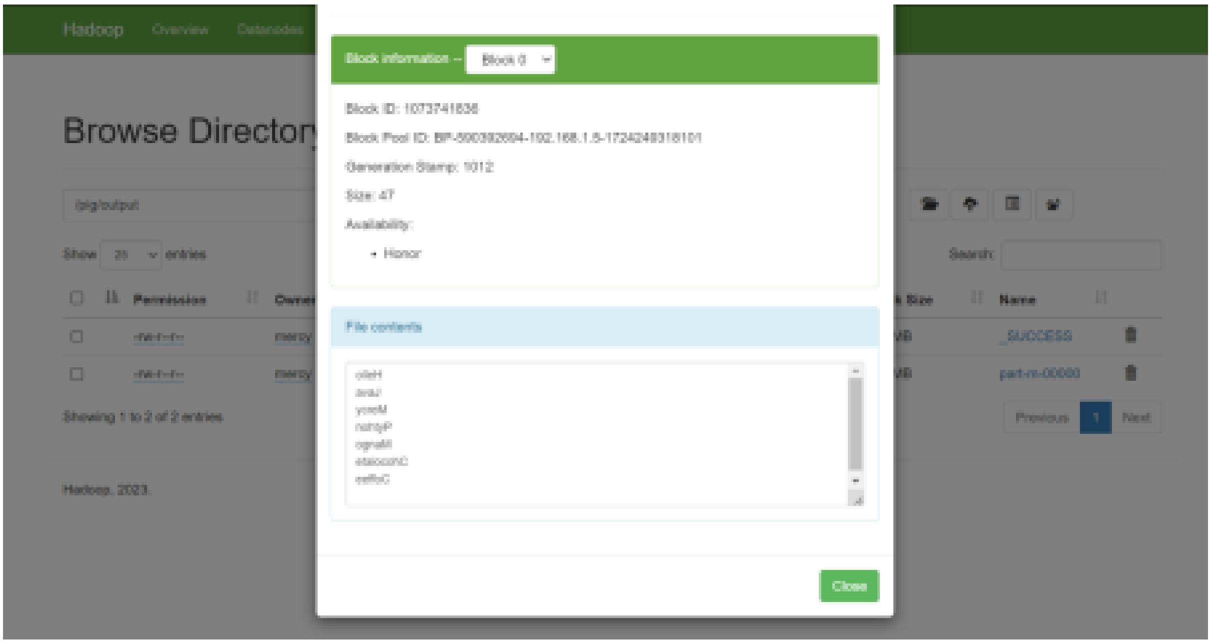
Browse Directory

/pig Go!

Show 25 entries Search:

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	-rw-r--r--	mercy	supergroup	52 B	Aug 21 19:41	1	128 MB	input.txt
☐	drwxr-xr-x	mercy	supergroup	0 B	Aug 21 19:47	0	0 B	output

Showing 1 to 2 of 2 entries



Browse Directory

/pig/output

Show 25 entries Search:

☐	Permission	Owner	Block Size	Name
☐	-rw-r--r--	mercy	MB	_SUCCESS
☐	-rw-r--r--	mercy	MB	part-m-00000

Showing 1 to 2 of 2 entries

Hadoop, 2023.

Block information -- Block 0

Block ID: 1073741836

Block Pool ID: BP-590392694-192.168.1.5-1724248318101

Generation Stamp: 1012

Size: 47

Availability:

- Honor

File contents

```
olleH
avaJ
ycreM
nohtyP
ognaM
etalocohC
eeffoC
```

RESULT:

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