

## 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 21:04:07.668 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-28 21:04:07.671 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-28 21:04:07.672 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-28 21:04:08.663 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1707388) compiled Jun 02 2017, 18:41:
58
2024-08-28 21:04:08.663 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.0\logs\pig_1724688448668
.log
2024-08-28 21:04:08.676 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file C:\Users\mearoy/.pigbootstrap not found
2024-08-28 21:04:08.310 [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.MExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9800
2024-08-28 21:04:08.792 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PID-default-81e8d0c5-2e27-479
d-a87c-ef481366523d
2024-08-28 21:04:08.792 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
  
```

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

```

script.pig  reverse_udf.py 1 X  mapper.py
reverse_udf.py > reverse
1  @outputSchema('reversed_word:chararray')
2  def reverse(word):
3      return word[::-1]
4
  
```

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.

```

script.pig  X  reverse_udf.py 1  mapper.py
script.pig
1  -- Register the Python UDF
2  REGISTER 'reverse_udf.py' USING jython AS myudf;
3
4  -- Load a sample dataset from HDFS
5  data = LOAD 'hdfs:///pig/input.txt' USING PigStorage(',') AS (word:chararray);
6
7  -- Apply the UDF to each record
8  reversed_data = FOREACH data GENERATE myudf.reverse(word);
9
10 -- Store the result back to HDFS
11 STORE reversed_data INTO 'hdfs:///pig/output' USING PigStorage(',');
12
  
```

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

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

**OUTPUT:**

```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 10.0.22621.4027]
(c) Microsoft Corporation. All rights reserved.

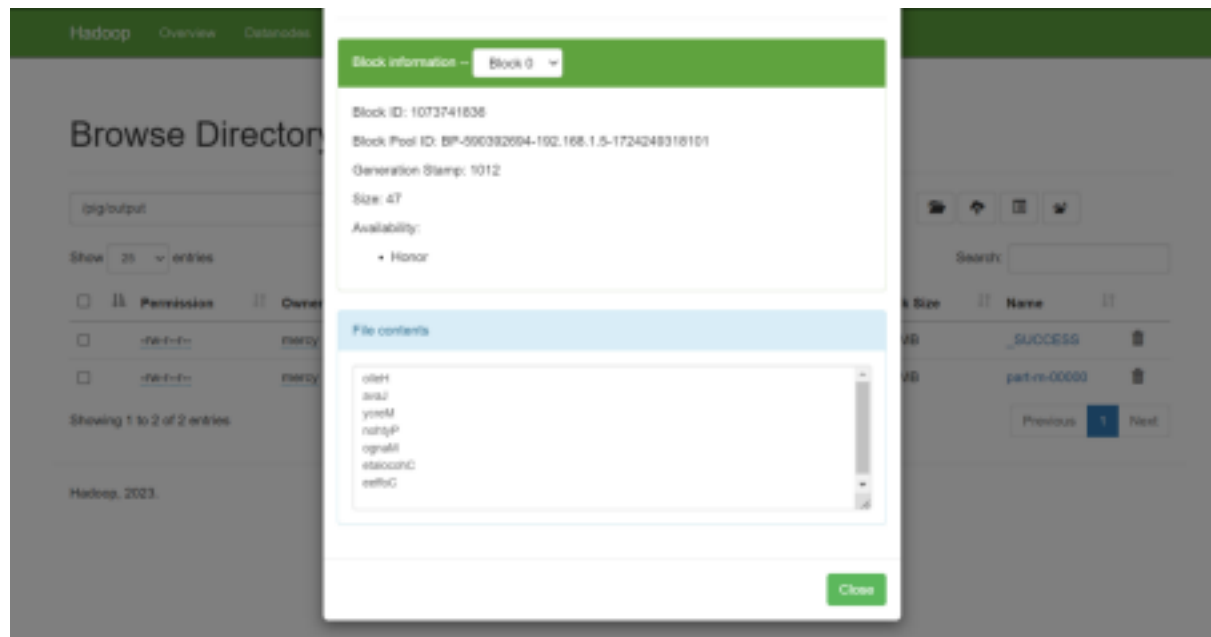
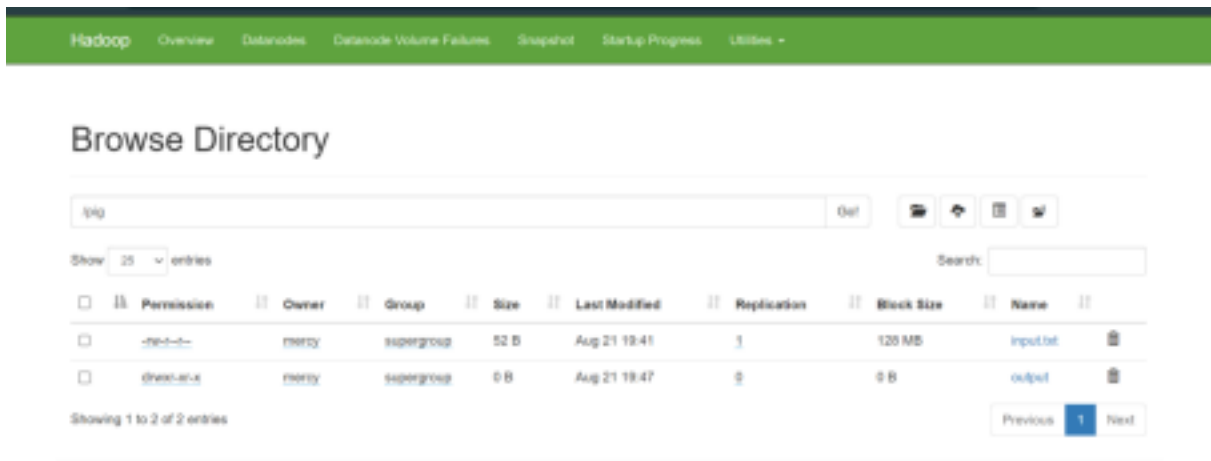
C:\Users\mercy\OneDrive\Documents\DataAnalytics\Pig>pig -x mapreduce
2024-08-21 19:46:55.885 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-21 19:46:55.773 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-21 19:46:55.785 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-21 19:46:52.287 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1707886) compiled Jun 02 2017, 19:41:
58
2024-08-21 19:46:52.957 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.0\logs\pig_1794249330000_0001
.log
2024-08-21 19:46:52.889 [main] INFO org.apache.pig.impl.util.Util - Default bootstrap $ is C:\Users\mercy\pigbootstrap not found
2024-08-21 19:46:52.885 [main] INFO org.apache.hadoop.conf.Configuration - deprecated: mapred.job.tracker is deprecated. In
stead, use mapreduce.jobtracker.address
2024-08-21 19:46:52.885 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
ile system at: hdfs://localhost:9000
2024-08-21 19:46:53.589 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-00000000-0000-0000-0000-000000000000
2024-08-21 19:46:53.589 [main] WARN org.apache.pig.PigServer - ATB is disabled since yarn.timeline-service.enabled set to f
alse
grant> |

2024-08-21 19:47:27.183 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2024-08-21 19:47:27.337 [JobControl] INFO org.apache.hadoop.mapred.YarnRunner - Job jar is not present. Not adding any jar
to the list of resources.
2024-08-21 19:47:27.343 [JobControl] INFO org.apache.hadoop.conf.Configuration - resource-types.xml not found
2024-08-21 19:47:27.349 [JobControl] INFO org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-typ
es.xml'.
2024-08-21 19:47:28.094 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application app
lication_1794249330000_0001
2024-08-21 19:47:28.183 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://mercy:8088/pro
xy/application_1794249330000_0001/
2024-08-21 19:47:28.183 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MapReduceLauncher - Hadoop
JobId: job_1794249330000_0001
2024-08-21 19:47:28.183 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MapReduceLauncher - Proces
sing aliases data, reversed_data
2024-08-21 19:47:28.183 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MapReduceLauncher - detail
ed locations: W: data[0,7],reversed_data[1,1] C: W:
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_1794249330000_0001]
2024-08-21 19:47:28.515 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MapReduceLauncher - 50% co
mplete
2024-08-21 19:47:28.515 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreducelayer.MapReduceLauncher - Runnin
g jobs are [job_1794249330000_0001]
2024-08-21 19:47:28.582 [main] INFO org.apache.hadoop.yarn.client.DefaultHadoopFailoverProxyProvider - Connecting to Resour
ceManager at /0.0.0.0:8032
2024-08-21 19:47:28.982 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCESS, 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
```



## RESULT:

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