

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

AIM:

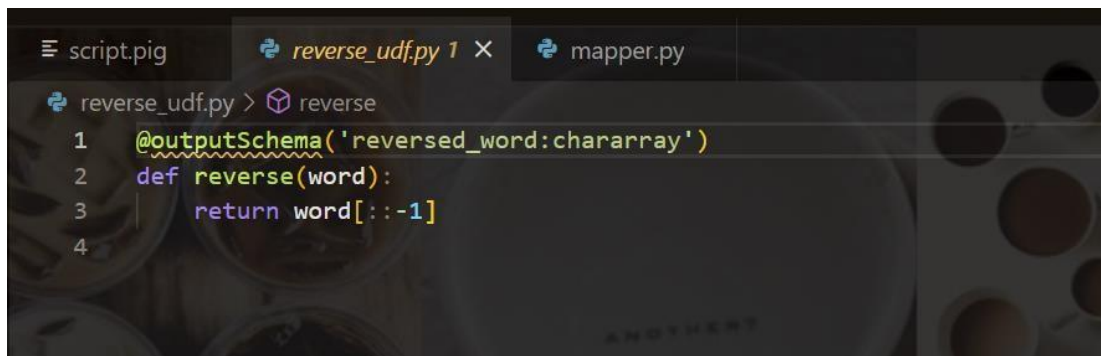
To create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode.

PROCEDURE:

1. Ensure that Apache Pig is installed and configured.

```
C:\> pig
2024-08-26 21:04:07,868 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-26 21:04:07,871 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-26 21:04:07,872 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-26 21:04:08,053 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 16:41:58
2024-08-26 21:04:08,053 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.0\logs\pig_1724686448046.log
2024-08-26 21:04:08,075 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\mercy/.pigbootup not found
2024-08-26 21:04:08,316 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-26 21:04:08,317 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-08-26 21:04:08,792 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-81e9d0c5-2e27-479d-a67c-ef401395523d
2024-08-26 21:04:08,792 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

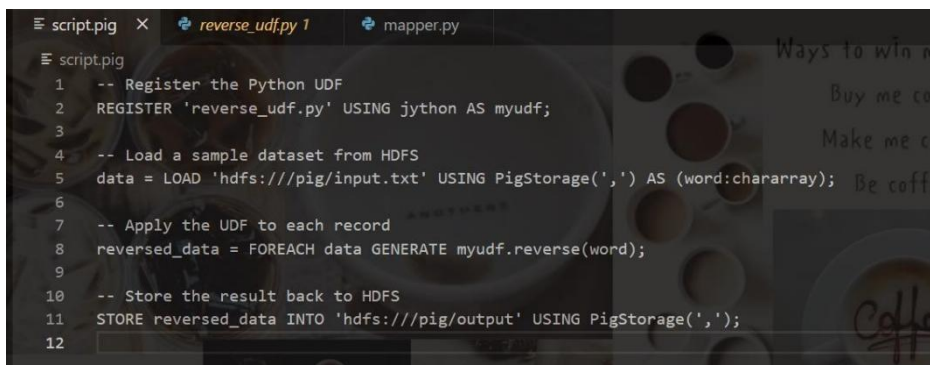
2. Create a python UDF (User Defined Functions).



```
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
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 10.0.22621.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mercy\OneDrive\Documents\DataAnalytics\Pig>pig -x mapreduce
2024-08-21 19:46:51,081 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 19:46:52,257 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1707386) compiled Jun 02 2017, 15:41:58
2024-08-21 19:46:52,257 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.0\logs\pig_1724249812251.log
2024-08-21 19:46:52,289 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\mercy\.pigbootup not found
2024-08-21 19:46:52,551 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-21 19:46:52,551 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file 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-a5529456-20df-4fcd-a140-38aa1dbf2542
2024-08-21 19:46:53,109 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> |

```

```

336266_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-types.xml'.
2024-08-21 19:47:28,024 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1724249336266_0001
2024-08-21 19:47:28,103 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://Honor:8088/proxy/application_1724249336266_0001/
2024-08-21 19:47:28,103 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Hadoop JobId: job_1724249336266_0001
2024-08-21 19:47:28,103 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases data.reversed_data
2024-08-21 19:47:28,104 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed 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% complete
2024-08-21 19:47:28,111 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1724249336266_0001]
2024-08-21 19:47:45,515 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2024-08-21 19:47:45,515 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1724249336266_0001]
2024-08-21 19:47:48,582 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2024-08-21 19:47:48,592 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. Final applicationStatus=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

```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/

Show entries Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name | <input type="checkbox"/> |
|--------------------------|------------|-------|------------|------|---------------|-------------|------------|----------------|--------------------------|
| <input type="checkbox"/> | drwxr-xr-x | mohan | supergroup | 0 B | Aug 12 19:38 | 0 | 0 B | input | <input type="checkbox"/> |
| <input type="checkbox"/> | drwxr-xr-x | mohan | supergroup | 0 B | Aug 12 19:42 | 0 | 0 B | out | <input type="checkbox"/> |
| <input type="checkbox"/> | drwxr-xr-x | mohan | supergroup | 0 B | Sep 09 13:55 | 0 | 0 B | tmp | <input type="checkbox"/> |
| <input type="checkbox"/> | drwxr-xr-x | mohan | supergroup | 0 B | Sep 09 13:57 | 0 | 0 B | user | <input type="checkbox"/> |
| <input type="checkbox"/> | drwxr-xr-x | mohan | supergroup | 0 B | Aug 23 09:00 | 0 | 0 B | weather | <input type="checkbox"/> |
| <input type="checkbox"/> | drwxr-xr-x | mohan | supergroup | 0 B | Aug 23 09:05 | 0 | 0 B | weather_output | <input type="checkbox"/> |

Showing 1 to 6 of 6 entries

Hadoop, 2023.

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/pig/output

Show entries Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name | <input type="checkbox"/> |
|--------------------------|------------|-------|------------|------|---------------|-------------|------------|--------------|--------------------------|
| <input type="checkbox"/> | -rw-r--r-- | mohan | supergroup | 46 B | Aug 23 09:00 | 1 | 131072 B | _SUCCESS | <input type="checkbox"/> |
| <input type="checkbox"/> | -rw-r--r-- | mohan | supergroup | 46 B | Aug 23 09:05 | 1 | 131072 B | part-m-00000 | <input type="checkbox"/> |

Showing 1 to 2 of 2 entries

Hadoop, 2023.

File information - part-m-00000

[Download](#) [Head the file \(first 32K\)](#) [Tail the file \(last 32K\)](#)

Block information --

Block ID: 1073741886
 Block Pool ID: BP-208583806-192.168.56.1-1723471191118
 Generation Stamp: 1062
 Size: 46
 Availability:
 • LAPTOP-7AIDGIAK

File contents

```
HELLO WORLD
APACHE PIG
USER DEFINED FUNCTIONS
```

RESULT:

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