

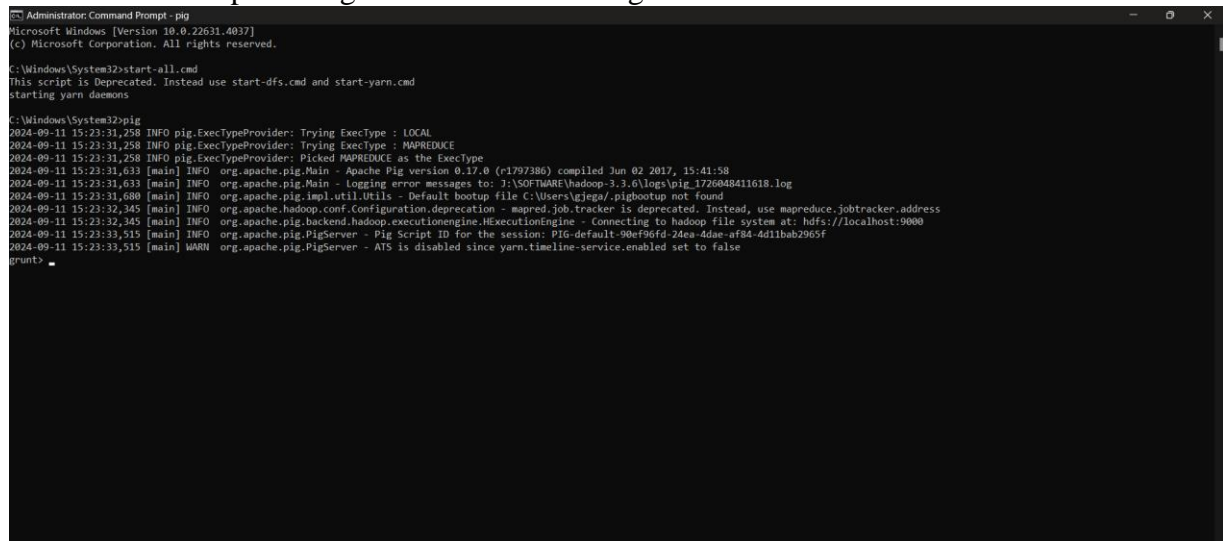
Create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode

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.

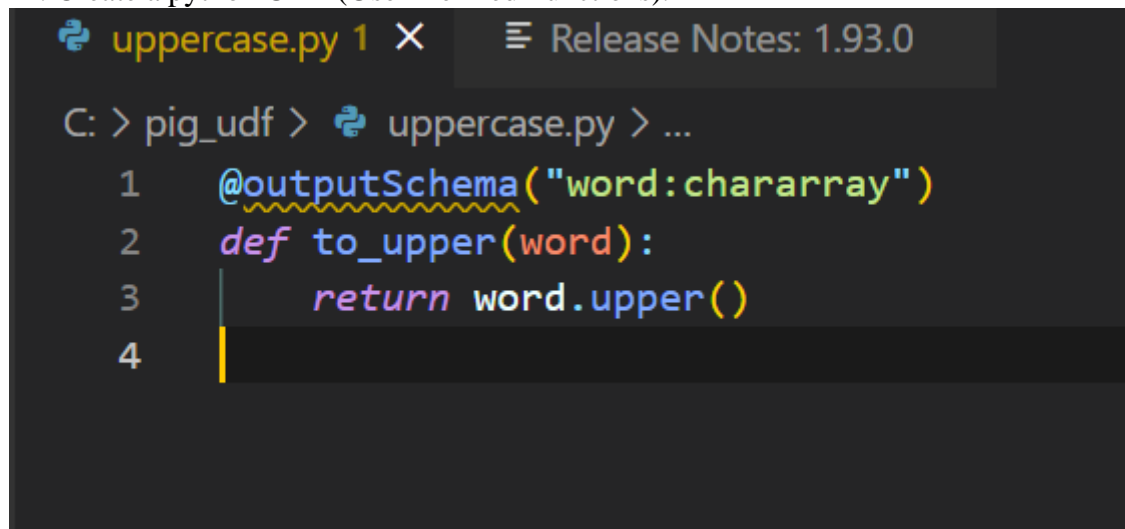


```
Administrator Command Prompt - pig
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>start-all.cmd
This script is deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Windows\System32>pig
2024-09-11 15:23:31,258 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-11 15:23:31,258 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-11 15:23:31,258 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-11 15:23:31,633 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-11 15:23:31,633 [main] INFO org.apache.pig.Main - logging error messages to: J:\SOFTWARE\hadoop-3.3.6\logs\pig_1726048411618.log
2024-09-11 15:23:31,680 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file C:\Users\tejga/.pigbootstrap not found
2024-09-11 15:23:32,345 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-11 15:23:32,345 [main] INFO org.apache.pig.backend.hadoop.executionengine.MExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-11 15:23:33,515 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-90ef96fd-24ea-4dae-af84-4d11bab2965f
2024-09-11 15:23:33,515 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
prunt>
```

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



```
uppercase.py 1 X Release Notes: 1.93.0
C: > pig_udf > uppercase.py > ...
1 @outputSchema("word:chararray")
2 def to_upper(word):
3     return word.upper()
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.

```

C: > pig_udf > script.pig
1  |-- Register the Python UDF script
2  REGISTER 'uppercase.py' USING jython AS myudf;
3
4  -- Load the input file from HDFS
5  data = LOAD 'hdfs:///pig/word.txt' USING PigStorage(',') AS (line: chararray);
6
7  -- Apply the UDF to convert each line to uppercase
8  uppercased_data = FOREACH data GENERATE myudf.to_upper(line);
9
10 -- Store the result in HDFS
11 STORE uppercased_data INTO 'hdfs:///pig/output' USING PigStorage(',');

```

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

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

OUTPUT:

```

Administrator: Command Prompt - pig
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>start-all.cmd
This script is deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Windows\System32>pig
2024-09-11 15:23:31,258 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-11 15:23:31,258 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-11 15:23:31,258 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-11 15:23:31,633 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-11 15:23:31,633 [main] INFO org.apache.pig.Main - Logging error messages to: J:\SOFTWARE\hadoop-3.3.6\logs\pig_1726048411618.log
2024-09-11 15:23:31,680 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file C:\Users\gjege/.pigbootstrap not found
2024-09-11 15:23:32,345 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-11 15:23:32,345 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-11 15:23:32,515 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-90ef96fd-24e8-4d8e-af84-4d11bab2965f
2024-09-11 15:23:33,515 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>

```

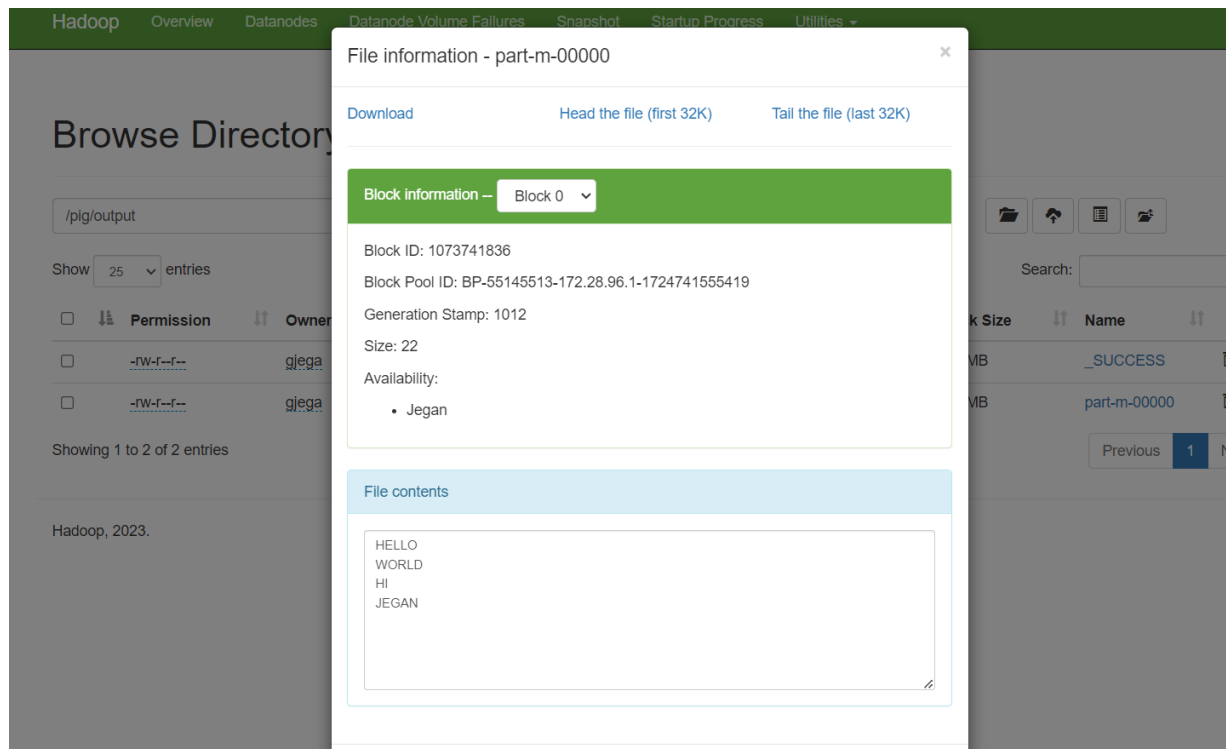
```

Administrator: Apache Hadoop Distribution
at sun.nio.ch.Net.kbind(Inet4Method)
at sun.nio.ch.Net.kbind(Inet4Method)
at sun.nio.ch.Net.kbind(Inet4Method)
at sun.nio.ch.ServerSocketAdaptor.kbind(ServerSocketAdaptor.kbind:74)
at org.apache.hadoop.ipc.Server.kbind(Server.kbind:1050)
2024-09-27 12:24:41,994 INFO nodemanager.NodeManager: SHUTDOWN_MSG:
SHUTDOWN_MSG: Shutting down NodeManager at 2024/09/27 12:24:41
*****
C:\Windows\System32>ps
PID Name
48560 java
47152 DataNode
63736 NodeManager
25676 ResourceManager

C:\Windows\System32>hadoop fs -ls /
-rw-r--r-- 1 hadoop hdfs 102400 2024-09-27 12:24:41 /pig
-rw-r--r-- 1 hadoop hdfs 102400 2024-09-27 12:24:41 /pig/output

C:\Windows\System32>hadoop fs -cat /pig/output/part-m-000000
HELLO
PIGEM

```

**RESULT:**

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