

Ex: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. Ensure that Apache Pig is installed and configured.

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

C:\Windows\System32>pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58

C:\Windows\System32>java -version
java version "1.8.0_421"
Java(TM) SE Runtime Environment (build 1.8.0_421-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.421-b09, mixed mode)

C:\Windows\System32>hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1bc78238728da9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /C:/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar

C:\Windows\System32>hadoop fs -mkdir /piguser
mkdir: Call from LAPTOP-H3TCD9BP/192.168.56.1 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused: no further information; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused

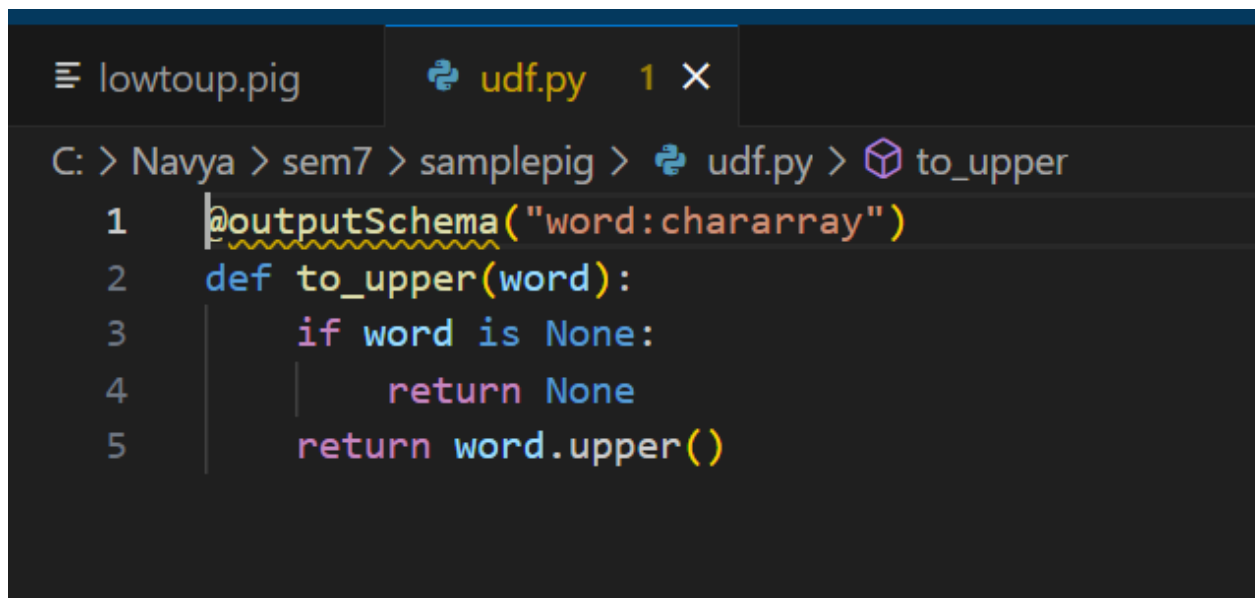
C:\Windows\System32>hadoop fs -mkdir /userpigipig
mkdir: Call from LAPTOP-H3TCD9BP/192.168.56.1 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused: no further information; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused

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>hadoop fs -mkdir /piguser
mkdir: Cannot create directory /piguser. Name node is in safe mode.

C:\Windows\System32>jps
17280 NameNode
22468 ResourceManager
36596 Jps
18556 Jps
8644 DataNode
```

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



```
lowtoup.pig  udf.py  1 X
C: > Navya > sem7 > samplepig > udf.py > to_upper
1  @outputSchema("word:chararray")
2  def to_upper(word):
3      if word is None:
4          return None
5      return word.upper()
```

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.

5. the Pig Script in MapReduce Mode using the command: `pig -x`

`mapreduce script.pig`

OUTPUT:

```
C:\Windows\System32>jps
17280 NameNode
22468 NodeManager
36596 ResourceManager
10556 Jps
9644 DataNode

C:\Windows\System32>hadoop fs -mkdir /userpigg

C:\Windows\System32>hadoop fs -put C:\Navya\sem7\samplepig\input1.txt /userpigg

C:\Windows\System32>hadoop fs -put C:\Navya\sem7\samplepig\lowtoup.pig /userpigg

C:\Windows\System32>pig -x mapreduce lowtoup.pig
2024-08-27 20:18:46,131 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-27 20:18:46,133 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-27 20:18:46,133 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-27 20:18:46,314 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-08-27 20:18:46,314 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1724770126309.log
2024-08-27 20:18:46,570 [main] ERROR org.apache.pig.Main - ERROR 2997: Encountered IOException. File lowtoup.pig does not exist
Details at logfile: C:\hadoop\logs\pig_1724770126309.log
2024-08-27 20:18:46,590 [main] INFO org.apache.pig.Main - Pig script completed in 494 milliseconds (494 ms)
```

```
C:\Windows\System32>pig -x mapreduce hdfs://localhost:9000/userpigg/lowtoup.pig
2024-08-27 20:20:09,851 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-27 20:20:09,852 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
```

```
C:\Windows\System32>hadoop fs -rm /userpigg/lowtoup.pig
Deleted /userpigg/lowtoup.pig

C:\Windows\System32>hadoop fs -put C:\Navya\sem7\samplepig\lowtoup.pig /userpigg

C:\Windows\System32>pig -x mapreduce hdfs://localhost:9000/userpigg/lowtoup.pig
2024-08-27 20:28:18,051 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-27 20:28:18,052 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-27 20:28:18,052 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-27 20:28:18,236 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-08-27 20:28:18,237 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1724770698231.log
2024-08-27 20:28:19,116 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\navya/.pigbootup not found
```

```
sleepTime:1000 MILLISECONDS)
2024-08-27 20:37:51,024 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:37:54,874 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:37:57,925 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:38:00,966 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:38:04,019 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:38:07,081 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:38:10,130 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:38:13,178 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1000 MILLISECONDS)
2024-08-27 20:38:15,338 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2024-08-27 20:38:15,339 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-08-27 20:38:15,364 [main] INFO org.apache.pig.Main - Pig script completed in 9 minutes, 57 seconds and 346 milliseconds (597346 ms)
```

```
C:\Windows\System32>
```

File information - part-m-00000

Download

Head the file (first 32K)

Tail the file (last 32K)

Block information – Block 0

Block ID: 1073741877

Block Pool ID: BP-2005220528-192.168.56.1-1723478856842

Generation Stamp: 1053

Size: 18

Availability:

- LAPTOP-H3TCD9BP

File contents

HI HELLO GOOD DAY

Close

RESULT:

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