

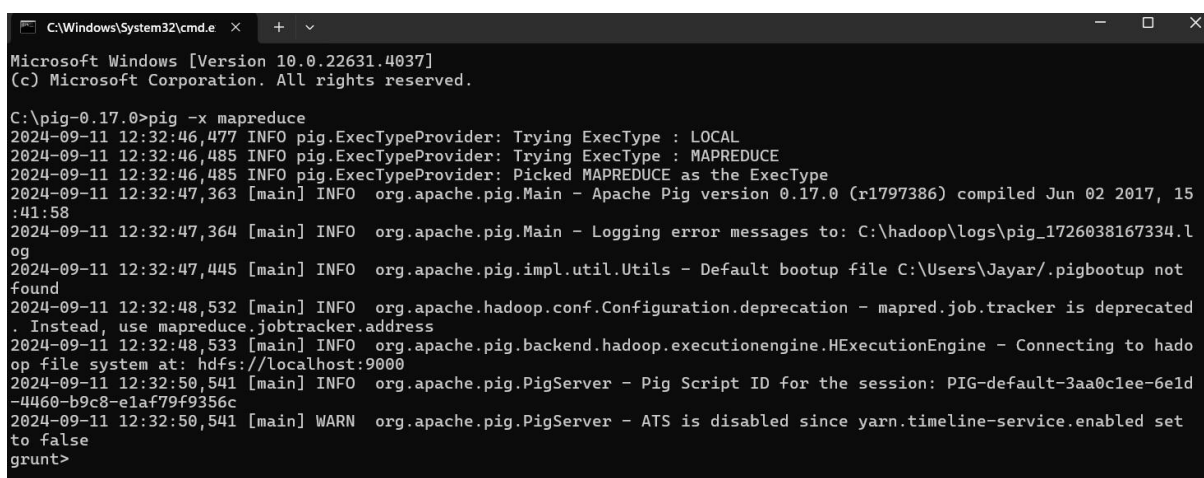
Create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode

AIM:

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

PROCEDURE:

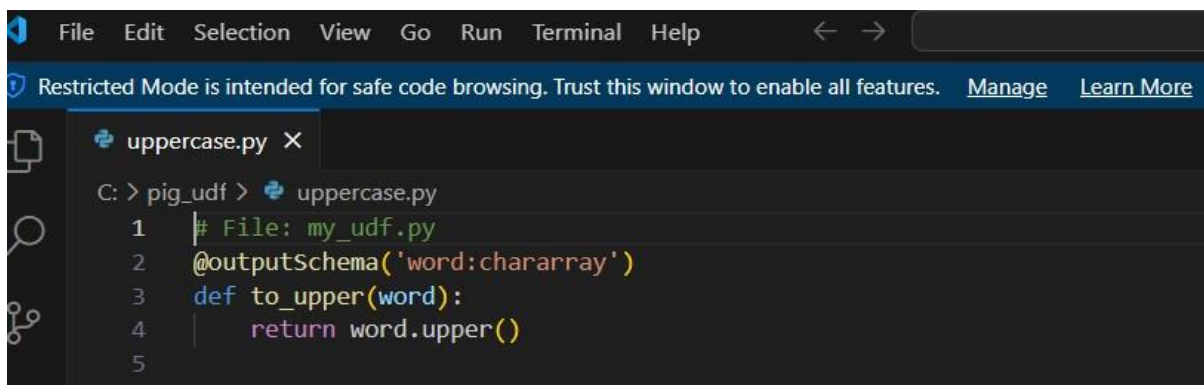
1. Ensure that Apache Pig is installed and configured.



```
C:\Windows\System32\cmd.e X + v
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\pig-0.17.0>pig -x mapreduce
2024-09-11 12:32:46,477 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-11 12:32:46,485 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-11 12:32:46,485 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-11 12:32:47,363 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-11 12:32:47,364 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1726038167334.log
2024-09-11 12:32:47,445 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\Jayar/.pigbootup not found
2024-09-11 12:32:48,532 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-11 12:32:48,533 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-11 12:32:50,541 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-3aa0c1ee-6e1d-4460-b9c8-e1af79f9356c
2024-09-11 12:32:50,541 [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).



```
File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

uppercase.py X
C: > pig_udf > uppercase.py
1 # File: my_udf.py
2 @outputSchema('word:chararray')
3 def to_upper(word):
4     return word.upper()
5
```

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.

```

uppercase.py  pigscript.pig X
C: > pig_udf > pigscript.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/data.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:

```

C:\Windows\System32\cmd.exe X + v
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\pig-0.17.0>pig -x mapreduce
2024-09-11 12:32:46,477 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-11 12:32:46,485 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-11 12:32:46,485 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-11 12:32:47,363 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-11 12:32:47,364 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1726038167334.log
2024-09-11 12:32:47,445 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\Jayar/.pigbootup not found
2024-09-11 12:32:48,532 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-09-11 12:32:48,533 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-11 12:32:50,541 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-3aa0c1ee-6e1d-4460-b9c8-e1af79f9356c
2024-09-11 12:32:50,541 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>

```

```

C:\> Administrator: Command Prompt

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

C:\Windows\System32>cd..

C:\Windows>cd..

C:\>hadoop fs -cat /pig/output/part-m-00000
HELLO
HI
RATHI
WELCOME

```

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities +

Browse Directory

/pig Go!

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	Rathi	supergroup	27 B	Aug 27 12:20	1	128 MB	data.txt	
<input type="checkbox"/>	drwxr-xr-x	Rathi	supergroup	0 B	Aug 27 13:47	0	0 B	output	

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2023.

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities +

Browse Directory

/pig/output Go!

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	Rathi	supergroup	27 B	Aug 27 12:20	1	128 MB	data.txt	
<input type="checkbox"/>	-rw-r--r--	Rathi	supergroup	0 B	Aug 27 13:47	0	0 B	output	

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2023.

File information - part-m-00000

Download Head the file (first 32K) Tail the file (last 32K)

Block Information -- Block 0

Block ID: 1073741917
Block Pool ID: BP-562178099-192.168.56.1-1723879711863
Generation Stamp: 1093
Size: 25
Availability:
• LAPTOP-G64VVBVO

File contents

```
HELLO
HI
RATHI
WELCOME
```

RESULT:

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