Naresh kumar V
210701323

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

Aim:

To create UDF in Apache Pig and execute it in MapReduce/HDFS mode.
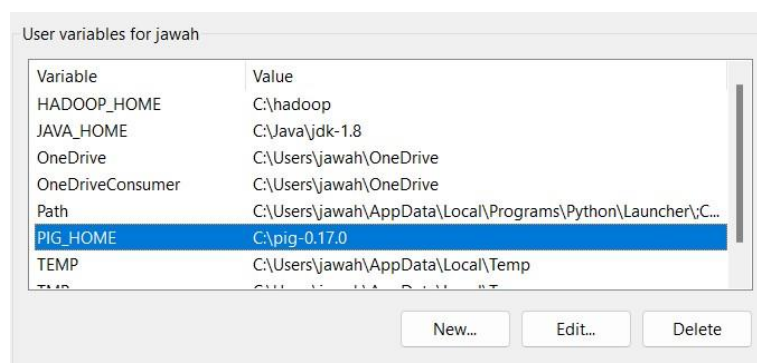
Procedure:

1.Pig Installation:

Step 1: Install pig from https://dlcdn.apache.org/pig/pig-0.17.0/

Click on pig-0.17.0.tar.gz and extract the downloaded files to c drive

Step 2: Set up environment variables for PIG_HOME and set path in user variables to bin folder of pig

Step 3: To check if pig is installed, open command prompt as administrator and run the command pig

PIG INSTALLATION IS DONE

2.Create UDF

Step 1: Create the directory in hadoop by the command
Hadoop fs -mkdir /user/pig

Step 2: Load the input file to that directory

hadoop fs -put C:\Users\jawah\OneDrive\Desktop\LathikaDA\ex4\pig_udf.txt /user/pig

//pig_udf.txt

1,hello

2,apache

3,pig

4,user

Step 3:Load the python file containing the user defined function into the hadoop directory

**hadoop fs -put
C:\Users\jawah\OneDrive\Desktop\LathikaDA\ex4\uppercase_udf.py/user/pig**

//uppercase_udf.py def
uppercase(text):   return
text.upper() if __name__ ==
"__main__":

   import sys   for line in
sys.stdin: line =
line.strip()   result =
uppercase(line)
print(result)

Step 4: Run the pig script using the command **pig -f
C:\Users\jawah\OneDrive\Desktop\LathikaDA\ex4\script.pig**

//script.pig

REGISTER 'hdfs://localhost:9000/user/pig/uppercase_udf.py' USING jython AS udf; data = LOAD 'hdfs://localhost:9000/user/pig/pig_udf_text.txt' AS (text:chararray); uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;

STORE uppercased_data INTO 'hdfs://localhost:9000/user/pig/output';

Step 8: View the output in the output directory of Hadoop

		Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
☐		drwxr-xr-x	jawah	supergroup	0 B	Sep 01 20:15	0	0 B	output	🗑
☐		-rw-r--r--	jawah	supergroup	32 B	Sep 01 19:23	3	128 MB	pig_udf_text.txt	🗑
☐		-rw-r--r--	jawah	supergroup	207 B	Sep 01 19:25	3	128 MB	uppercase_udf.py	🗑

Showing 1 to 3 of 3 entries

		Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
☐		-rw-r--r--	jawah	supergroup	0 B	Sep 01 20:15	3	128 MB	_SUCCESS	🗑
☐		-rw-r--r--	jawah	supergroup	30 B	Sep 01 20:15	3	128 MB	part-m-00000	🗑

hadoop fs -ls /user/pig/output

hadoop fs -cat /user/ex4/pig/part-m-00000

PIG SCRIPT IS RUN SUCCESSFULLY

Result:

Thus, UDF in Apache Pig has been created and executed in MapReduce/HDFS mode successful.

PIG SCRIPT IS RUN SUCCESSFULLY