Lathika P
210701131

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

Aim:

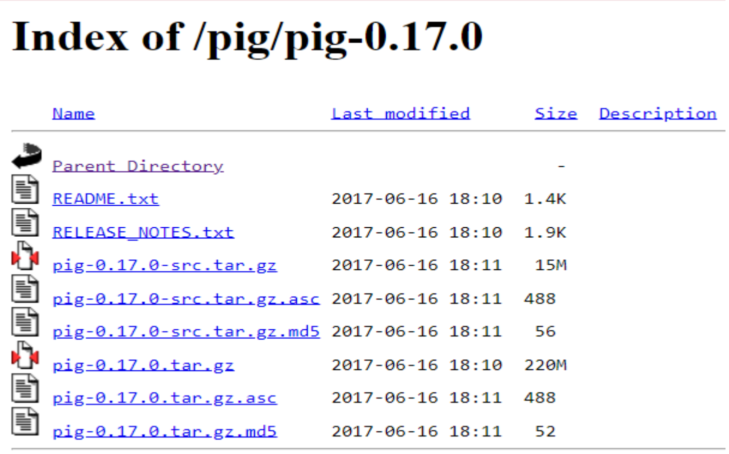
To create UDF in Apache Pig and execute it in MapReduce/HDFS mode.

Procedure:

1.Pig Installation:

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

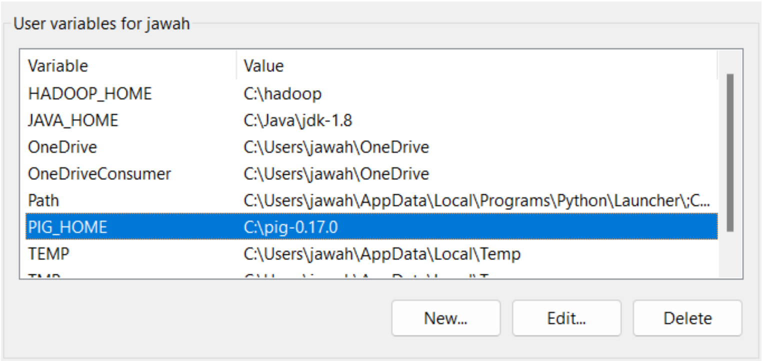
Click on pig-0.17.0.tar.gz and extract the downloaded files to c drive



Index of /pig/pig-0.17.0

Name	Last modified	Size	Description
Parent Directory		-	
README.txt	2017-06-16 18:10	1.4K	
RELEASE_NOTES.txt	2017-06-16 18:10	1.9K	
pig-0.17.0-src.tar.gz	2017-06-16 18:11	15M	
pig-0.17.0-src.tar.gz.asc	2017-06-16 18:11	488	
pig-0.17.0-src.tar.gz.md5	2017-06-16 18:11	56	
pig-0.17.0.tar.gz	2017-06-16 18:10	220M	
pig-0.17.0.tar.gz.asc	2017-06-16 18:11	488	
pig-0.17.0.tar.gz.md5	2017-06-16 18:11	52	

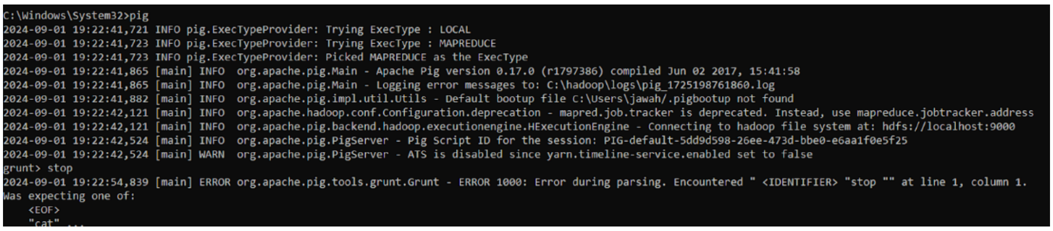
Step 2: Set up environment variables for PIG_HOME and set path in user variables to bin folder of pig



User variables for jawah

Variable	Value
HADOOP_HOME	C:\hadoop
JAVA_HOME	C:\Java\jdk-1.8
OneDrive	C:\Users\jawah\OneDrive
OneDriveConsumer	C:\Users\jawah\OneDrive
Path	C:\Users\jawah\AppData\Local\Programs\Python\Launcher\;C...
PIG_HOME	C:\pig-0.17.0
TEMP	C:\Users\jawah\AppData\Local\Temp

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



```
C:\Windows\System32>pig
2024-09-01 19:22:41,721 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-01 19:22:41,723 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-01 19:22:41,723 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-01 19:22:41,865 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-01 19:22:41,865 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1725198761860.log
2024-09-01 19:22:41,882 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\jawah/.pigbootup not found
2024-09-01 19:22:42,121 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-01 19:22:42,121 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-01 19:22:42,524 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-5dd9d598-26ee-473d-bbe0-e6aa1f0e5f25
2024-09-01 19:22:42,524 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> stop
2024-09-01 19:22:54,839 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1000: Error during parsing. Encountered " <IDENTIFIER> "stop "" at line 1, column 1.
Was expecting one of:
    <EOF>
    "cat" ...
```

PIG INSTALLATION IS DONE

2.Create UDF

Step 1: Create the directory in hadoop by the command

Hadoop fs -mkdir /user/pig

```
C:\Windows\System32>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

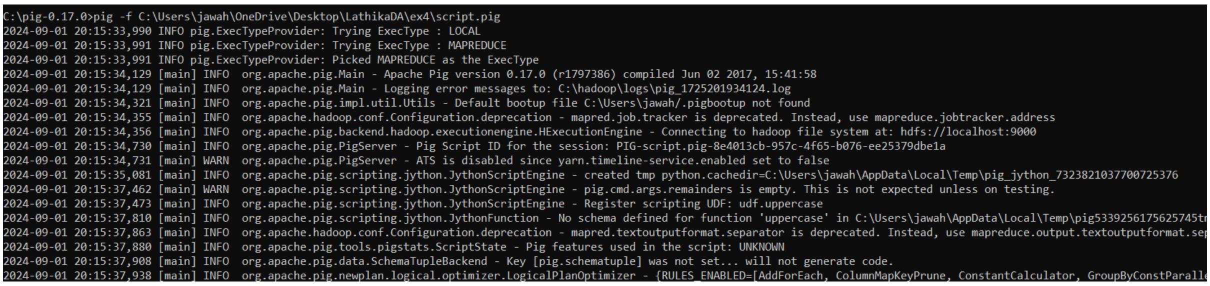
```
def uppercase(text):
   return text.upper()
if __name__ == "__main__":
   import sys
   for line in sys.stdin:
      line = line.strip()
      result = uppercase(line)
      print(result)
```

Step 4: Run the pig script using the command

pig -f C:\Users\jawah\OneDrive\Desktop\LathikaDA\ex4\script.pig

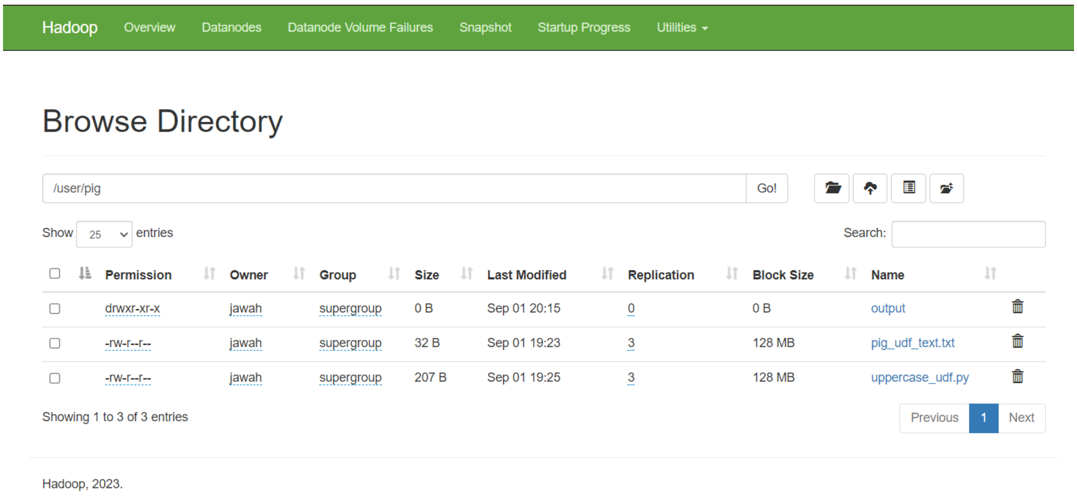
//script.pig

```
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';
```



```
C:\pig-0.17.0>pig -f C:\Users\jawah\OneDrive\Desktop\LathikaDA\ex4\script.pig
2024-09-01 20:15:33,990 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-01 20:15:33,991 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-01 20:15:33,991 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-01 20:15:34,129 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-01 20:15:34,129 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1725201934124.log
2024-09-01 20:15:34,321 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\jawah/.pigbootup not found
2024-09-01 20:15:34,355 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-01 20:15:34,356 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-01 20:15:34,730 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-script.pig-8e4013cb-957c-4f65-b076-ee25379dbe1a
2024-09-01 20:15:34,731 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-01 20:15:35,081 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=C:\Users\jawah\AppData\Local\Temp\pig_jython_7323821037700725376
2024-09-01 20:15:37,462 [main] WARN  org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.remainders is empty. This is not expected unless on testing.
2024-09-01 20:15:37,473 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: udf.uppercase
2024-09-01 20:15:37,810 [main] INFO  org.apache.pig.scripting.jython.JythonFunction - No schema defined for function 'uppercase' in C:\Users\jawah\AppData\Local\Temp\pig5339256175625745tı
2024-09-01 20:15:37,863 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.sep
2024-09-01 20:15:37,880 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-01 20:15:37,908 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-01 20:15:37,938 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParalle
```

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



Browse Directory

/user/pig Go!

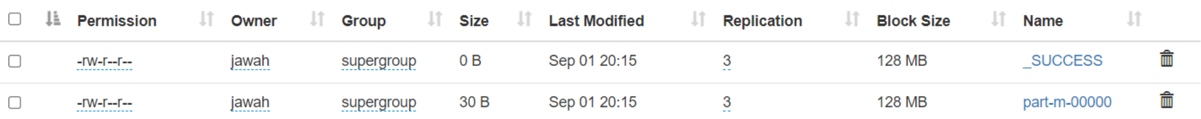
Show 25 entries Search:

□	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

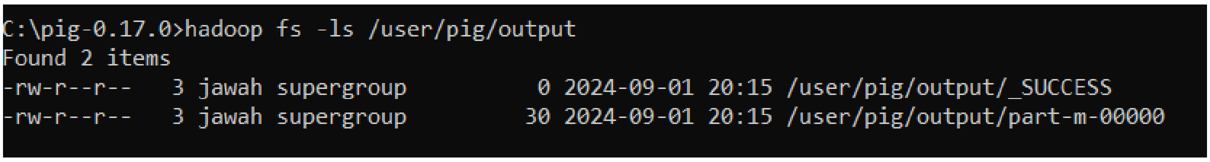
Previous 1 Next

Hadoop, 2023.



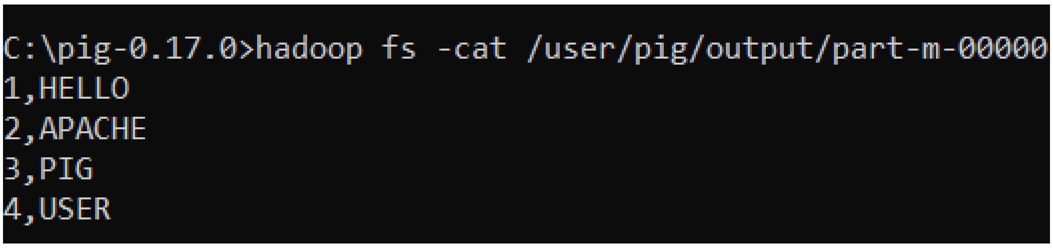
□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	jawah	supergroup	0 B	Sep 01 20:15	3	128 MB	_SUCCESS	🗑
□	-rw-r--r--	jawah	supergroup	30 B	Sep 01 20:15	3	128 MB	part-m-00000	🗑

hadoop fs -ls /user/pig/output



```
C:\pig-0.17.0>hadoop fs -ls /user/pig/output
Found 2 items
-rw-r--r--   3 jawah supergroup          0 2024-09-01 20:15 /user/pig/output/_SUCCESS
-rw-r--r--   3 jawah supergroup         30 2024-09-01 20:15 /user/pig/output/part-m-00000
```

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



```
C:\pig-0.17.0>hadoop fs -cat /user/pig/output/part-m-00000
1,HELLO
2,APACHE
3,PIG
4,USER
```

PIG SCRIPT IS RUN SUCCESSFULLY

Result:

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

Lathika P
210701131