EX NO: 4 **Roll No: 210701159**

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

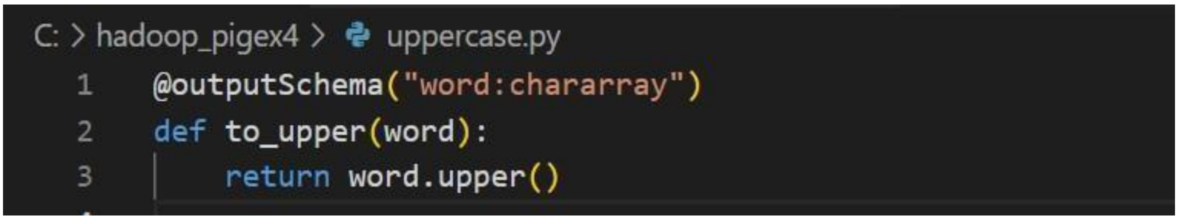
AIM:

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

PROCEDURE:

1. Install and Configure Apache Pig.

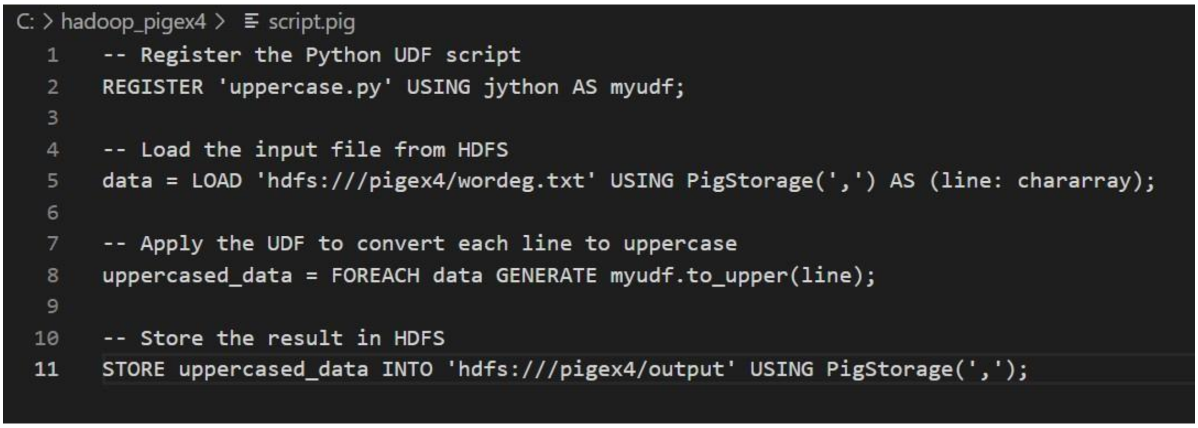
2. Create a python UDF (User Defined Functions).



```python
@outputSchema("word:chararray")
def to_upper(word):
    return word.upper()
```

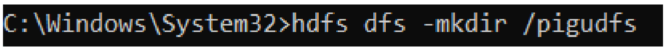
3. Install Jython because Pig will use it to interpret the Python UDFs.

4. Create a Pig script that registers and uses the Python UDF.



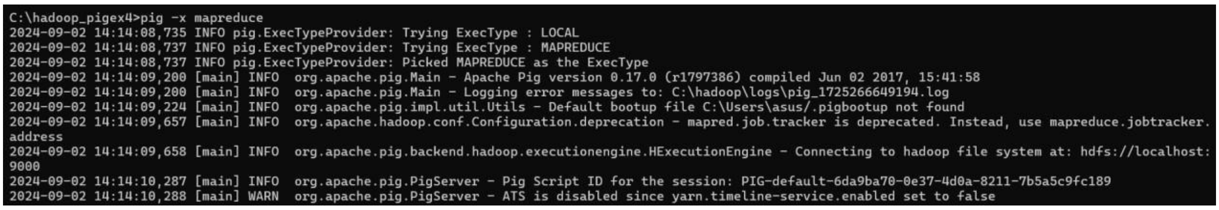
```
-- Register the Python UDF script
REGISTER 'uppercase.py' USING jython AS myudf;

-- Load the input file from HDFS
data = LOAD 'hdfs:///pigex4/wordeg.txt' USING PigStorage(',') AS (line: chararray);

-- Apply the UDF to convert each line to uppercase
uppercased_data = FOREACH data GENERATE myudf.to_upper(line);

-- Store the result in HDFS
STORE uppercased_data INTO 'hdfs:///pigex4/output' USING PigStorage(',');
```

5. Create Directory pigex4 and put the input files inside the created directory.



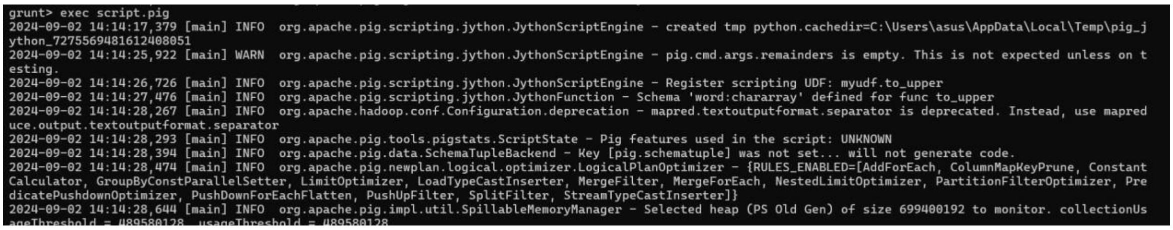
```
C:\Windows\System32>hdfs dfs -mkdir /pigudfs
```

6. Use the command pig -x mapreduce is used to run Apache Pig scripts in MapReduce mode



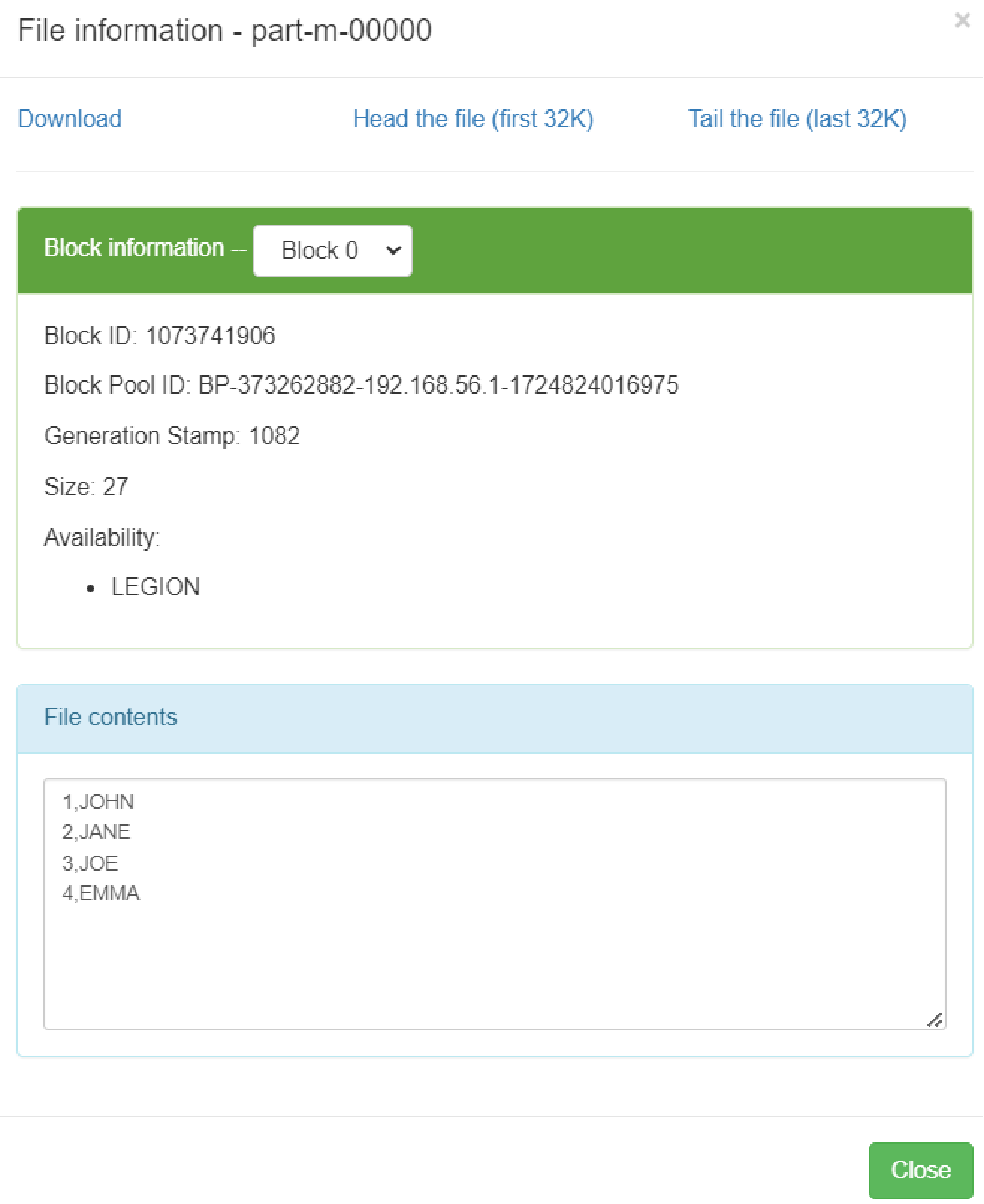
```
C:\hadoop_pigex4>pig -x mapreduce
2024-09-02 14:14:08,735 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-02 14:14:08,737 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-02 14:14:08,737 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-02 14:14:09,200 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-02 14:14:09,200 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1725266649194.log
2024-09-02 14:14:09,224 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\asus/.pigbootup not found
2024-09-02 14:14:09,657 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.
address
2024-09-02 14:14:09,658 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:
9000
2024-09-02 14:14:10,287 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-6da9ba70-0e37-4d0a-8211-7b5a5c9fc189
2024-09-02 14:14:10,288 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
```

7. After executing the above command you will enter the grunt shell. Here we can execute the script.pig



```
grunt> exec script.pig
2024-09-02 14:14:17,379 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=C:\Users\asus\AppData\Local\Temp\pig_j
ython_7275569481612408051
2024-09-02 14:14:25,922 [main] WARN  org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.remainders is empty. This is not expected unless on t
esting.
2024-09-02 14:14:26,726 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: myudf.to_upper
2024-09-02 14:14:27,476 [main] INFO  org.apache.pig.scripting.jython.JythonFunction - Schema 'word:chararray' defined for func to_upper
2024-09-02 14:14:28,267 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapred
uce.output.textoutputformat.separator
2024-09-02 14:14:28,293 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-02 14:14:28,394 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-02 14:14:28,474 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, Constant
Calculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Pre
dicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-09-02 14:14:28,644 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUs
ageThreshold = 489580128, usageThreshold = 489580128
```

OUTPUT:



File information - part-m-00000

Block information -- Block 0

Block ID: 1073741906

Block Pool ID: BP-373262882-192.168.56.1-1724824016975

Generation Stamp: 1082

Size: 27

Availability:

- LEGION

File contents

```
1,JOHN
2,JANE
3,JOE
4,EMMA
```

RESULT:

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