Exp. No : 4

User Defined Function (UDF) in PIG

1. Create sample.txt
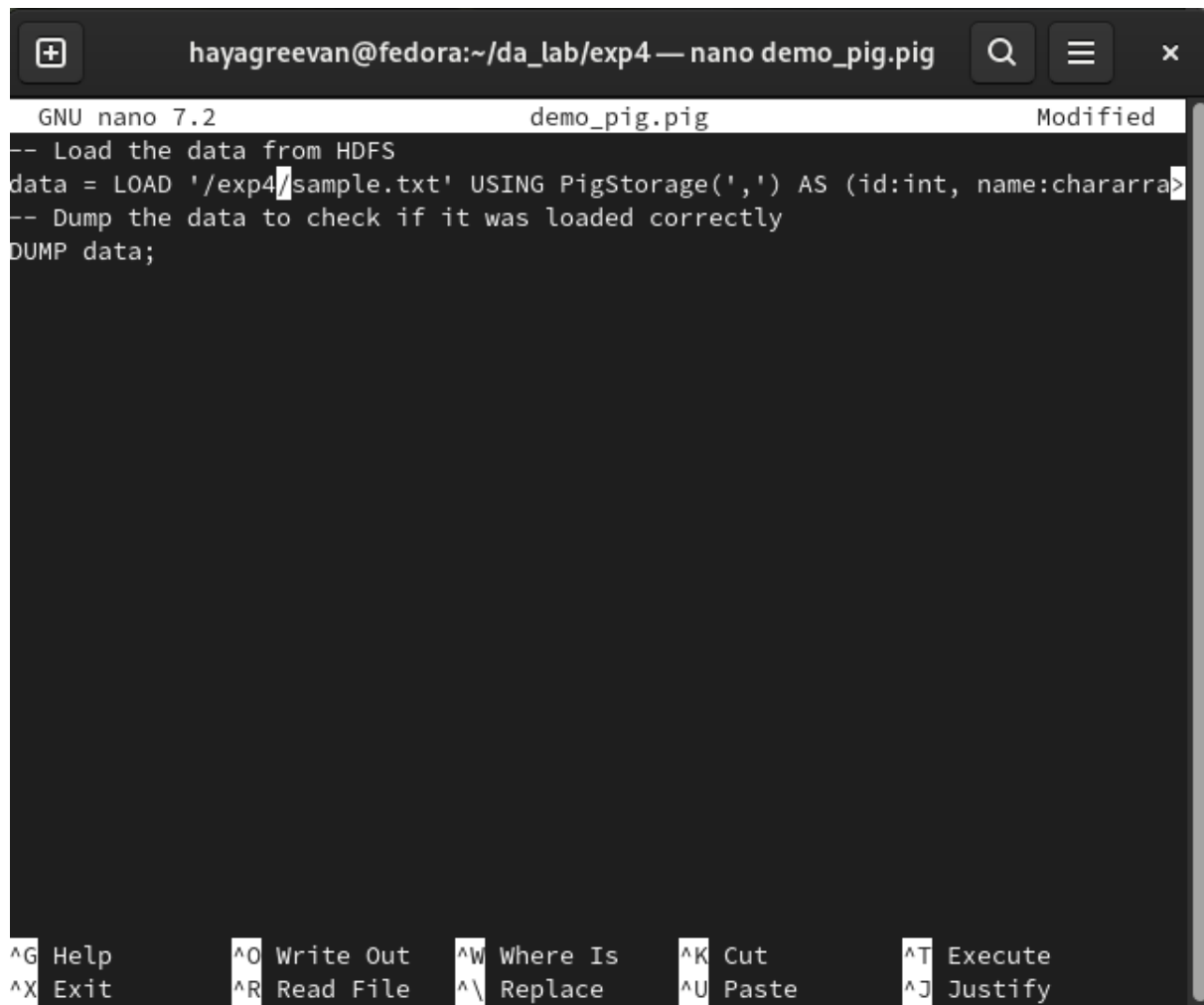

2. Upload sample.txt file to HDFS Storage.

```
hayagreevan@fedora:~/da_lab/exp4$ hdfs dfs -mkdir /exp4/
hayagreevan@fedora:~/da_lab/exp4$ hdfs dfs -put sample.txt /exp4/
```

3. Create demo_pig.pig file

```
GNU nano 7.2                    demo_pig.pig                    Modified
-- Load the data from HDFS
data = LOAD '/exp4/sample.txt' USING PigStorage(',') AS (id:int, name:chararra>
-- Dump the data to check if it was loaded correctly
DUMP data;




^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

4. Execute demo_pig.pig


```
hayagreevan@fedora:~/da_lab/exp4$ pig demo_pig.pig
2024-08-28 12:53:22,098 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-28 12:53:22,112 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-28 12:53:22,112 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the Exec
Type
2024-08-28 12:53:22,353 [main] INFO  org.apache.pig.Main - Apache Pig version 0
.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-08-28 12:53:22,354 [main] INFO  org.apache.pig.Main - Logging error messag
es to: /home/hayagreevan/da_lab/exp4/pig_1724829802330.log
2024-08-28 12:53:22,873 [main] INFO  org.apache.pig.impl.util.Utils - Default b
ootup file /home/hayagreevan/.pigbootup not found
2024-08-28 12:53:22,987 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2024-08-28 12:53:22,988 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-28 12:53:22,988 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:90
00
2024-08-28 12:53:24,194 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2024-08-28 12:53:24,194 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-08-28 12:53:24,196 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-28 12:53:24,261 [main] INFO  org.apache.pig.PigServer - Pig Script ID f
```
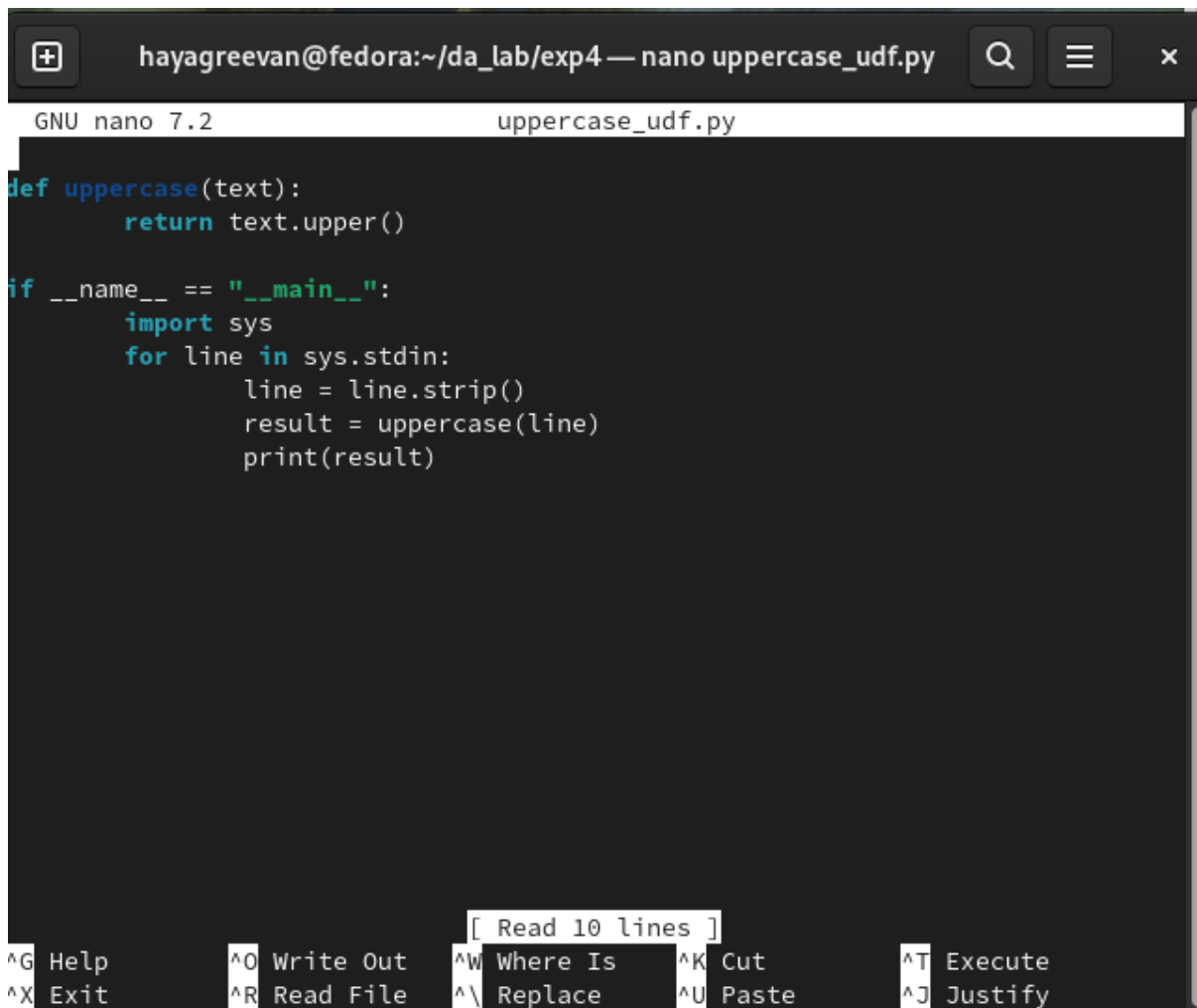


```
S)
2024-08-28 12:56:43,347 [main] WARN  org.apache.pig.backend.hadoop.executioneng
ine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warnin
g aggregation.
2024-08-28 12:56:43,348 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.mapReduceLayer.MapReduceLauncher - Success!
2024-08-28 12:56:43,690 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. I
nstead, use yarn.system-metrics-publisher.enabled
2024-08-28 12:56:43,690 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2024-08-28 12:56:43,691 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-28 12:56:43,692 [main] INFO  org.apache.pig.data.SchemaTupleBackend - K
ey [pig.schematuple] was not set... will not generate code.
2024-08-28 12:56:43,854 [main] INFO  org.apache.hadoop.mapreduce.lib.input.File
InputFormat - Total input files to process : 1
2024-08-28 12:56:43,855 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.util.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-08-28 12:56:44,251 [main] INFO  org.apache.pig.Main - Pig script completed
 in 3 minutes, 22 seconds and 263 milliseconds (202263 ms)
hayagreevan@fedora:~/da_lab/exp4$
```
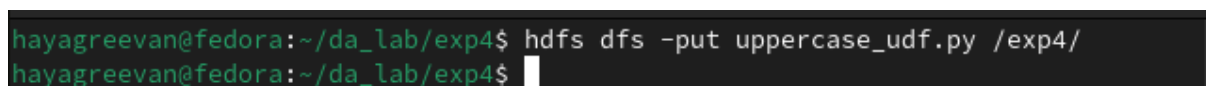
5. Create uppercase_udf.py

```
GNU nano 7.2                    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)




                              [ Read 10 lines ]
^G Help        ^O Write Out   ^W Where Is    ^K Cut        ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste      ^J Justify
```
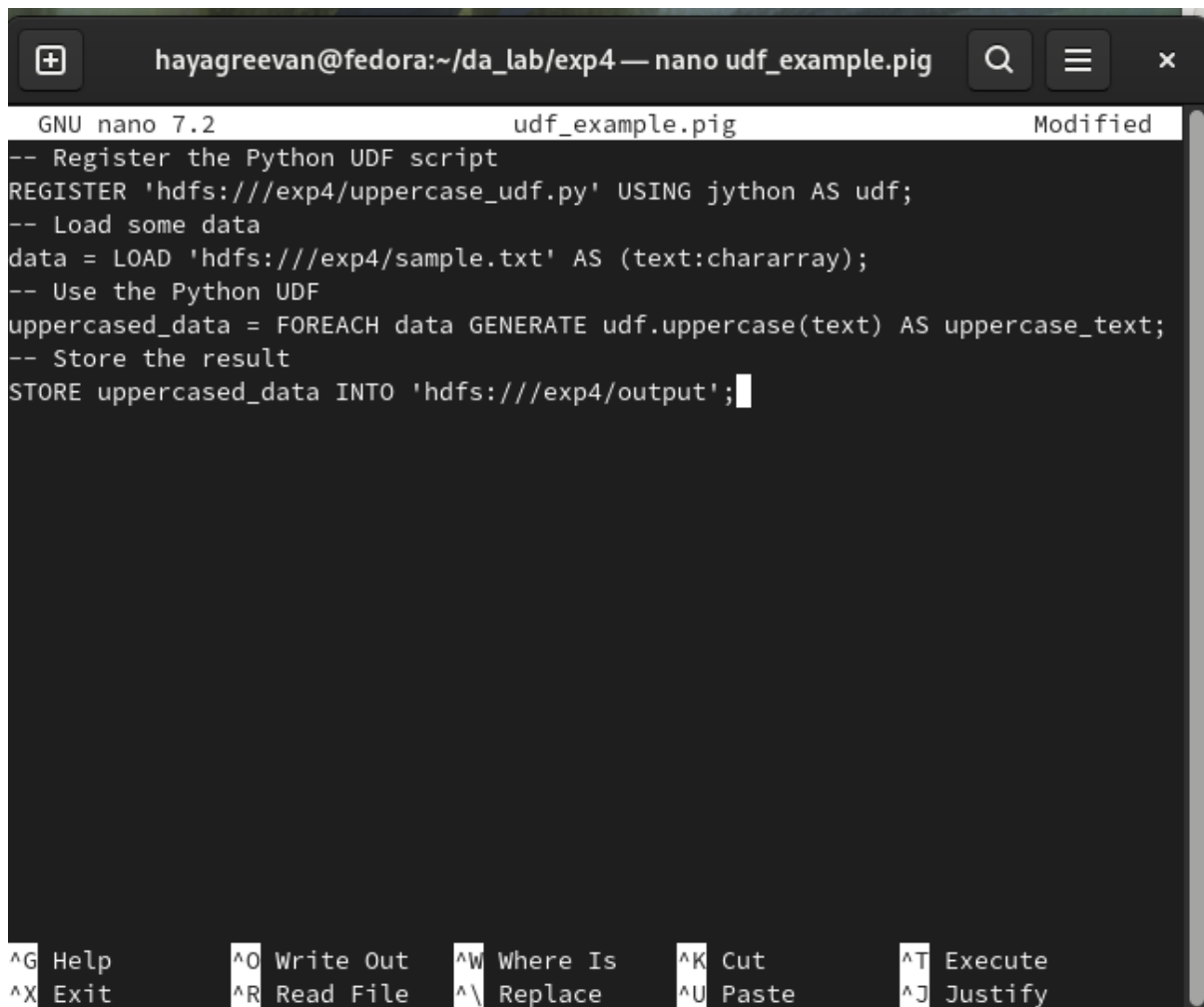
6. Upload uppercase_udf.py file to HDFS Storage.

```
hayagreevan@fedora:~/da_lab/exp4$ hdfs dfs -put uppercase_udf.py /exp4/
hayagreevan@fedora:~/da_lab/exp4$
```

7. Create udf_example.pig

```
GNU nano 7.2                    udf_example.pig                    Modified
-- Register the Python UDF script
REGISTER 'hdfs:///exp4/uppercase_udf.py' USING jython AS udf;
-- Load some data
data = LOAD 'hdfs:///exp4/sample.txt' AS (text:chararray);
-- Use the Python UDF
uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;
-- Store the result
STORE uppercased_data INTO 'hdfs:///exp4/output';

^G Help       ^O Write Out   ^W Where Is   ^K Cut      ^T Execute
^X Exit       ^R Read File   ^\ Replace    ^U Paste    ^J Justify
```

8. Execute udf_example.pig

```
hayagreevan@fedora:~/da_lab/exp4$ pig -f udf_example.pig
2024-08-28 13:03:16,716 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-28 13:03:16,718 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-28 13:03:16,718 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the Exec
Type
2024-08-28 13:03:16,930 [main] INFO  org.apache.pig.Main - Apache Pig version 0
.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-08-28 13:03:16,930 [main] INFO  org.apache.pig.Main - Logging error messag
es to: /home/hayagreevan/da_lab/exp4/pig_1724830396914.log
2024-08-28 13:03:17,383 [main] INFO  org.apache.pig.impl.util.Utils - Default b
ootup file /home/hayagreevan/.pigbootup not found
2024-08-28 13:03:17,507 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2024-08-28 13:03:17,507 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-28 13:03:17,507 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:90
00
2024-08-28 13:03:19,755 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2024-08-28 13:03:19,756 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-08-28 13:03:19,759 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-28 13:03:19,865 [main] INFO  org.apache.pig.PigServer - Pig Script ID f
```

```
nnect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy i
s RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECOND
S)
2024-08-28 13:07:07,297 [main] INFO  org.apache.hadoop.ipc.Client - Retrying co
nnect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy i
s RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECOND
S)
2024-08-28 13:07:08,300 [main] INFO  org.apache.hadoop.ipc.Client - Retrying co
nnect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy i
s RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECOND
S)
2024-08-28 13:07:09,308 [main] INFO  org.apache.hadoop.ipc.Client - Retrying co
nnect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy i
s RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECOND
S)
2024-08-28 13:07:10,328 [main] INFO  org.apache.hadoop.ipc.Client - Retrying co
nnect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy i
s RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECOND
S)
2024-08-28 13:07:10,527 [main] WARN  org.apache.pig.backend.hadoop.executioneng
ine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warnin
g aggregation.
2024-08-28 13:07:10,528 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.mapReduceLayer.MapReduceLauncher - Success!
2024-08-28 13:07:11,124 [main] INFO  org.apache.pig.Main - Pig script completed
 in 3 minutes, 54 seconds and 309 milliseconds (234309 ms)
hayagreevan@fedora:~/da_lab/exp4$
```

Output :

```
hayagreevan@fedora:~/da_lab/exp4$ hdfs dfs -cat /exp4/output/*
1,JOHN
2,JANE
3,JOE
4,EMMA
```