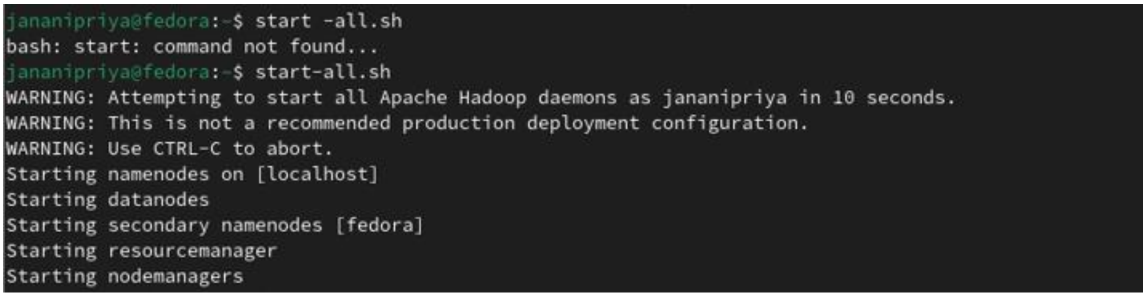
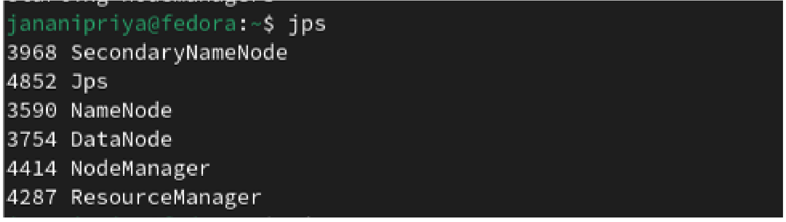
EXP NO: 4 **CREATE UDF IN PIG**

$start-all.sh



```
jananipriya@fedora:~$ start -all.sh
bash: start: command not found...
jananipriya@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as jananipriya in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [fedora]
Starting resourcemanager
Starting nodemanagers
```

$ jps

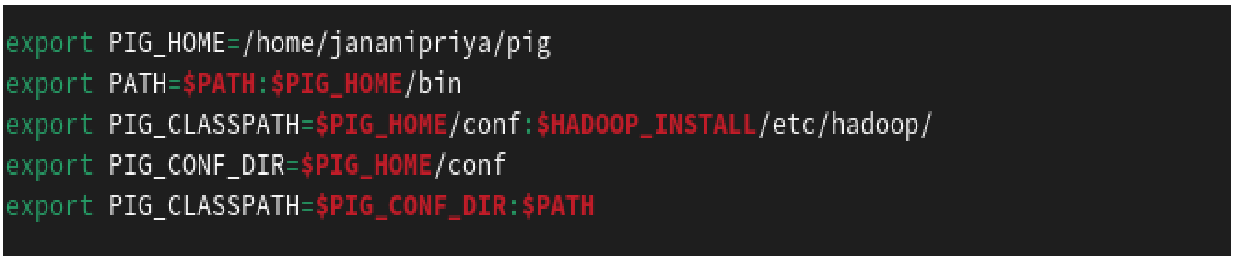


```
jananipriya@fedora:~$ jps
3968 SecondaryNameNode
4852 Jps
3590 NameNode
3754 DataNode
4414 NodeManager
4287 ResourceManager
```

$wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz

$ tar xvzf pig-0.16.0.tar.gz

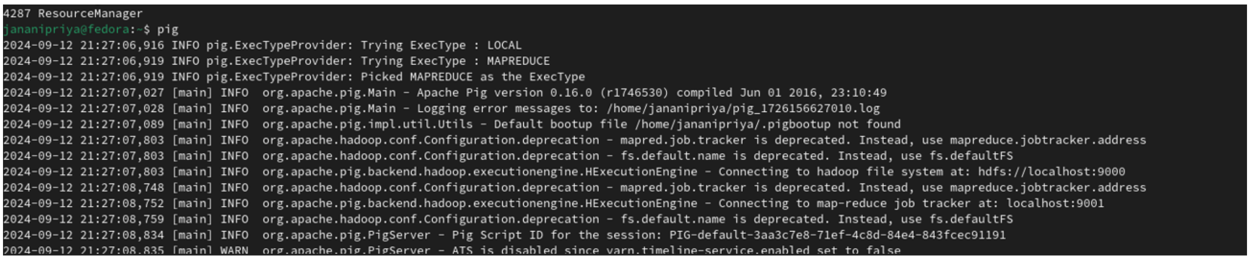
$nano ~/.bashrc



```
export PIG_HOME=/home/jananipriya/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
```

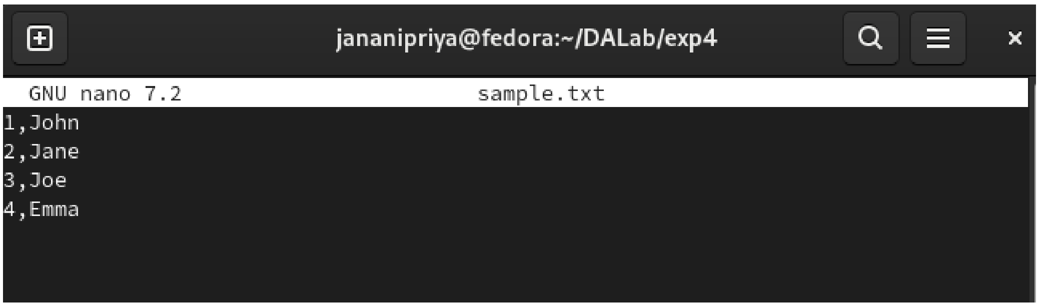
$mv pig-0.16.0 pig

$pig



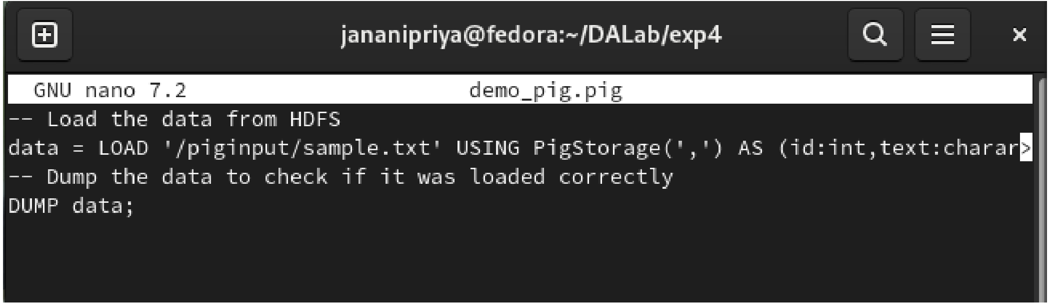
```
4287 ResourceManager
jananipriya@fedora:~$ pig
2024-09-12 21:27:06,916 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-12 21:27:06,919 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-12 21:27:06,919 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-12 21:27:07,027 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-12 21:27:07,028 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/jananipriya/pig_1726156627010.log
2024-09-12 21:27:07,089 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/jananipriya/.pigbootup not found
2024-09-12 21:27:07,803 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-12 21:27:07,803 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:27:07,803 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-12 21:27:08,748 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-12 21:27:08,752 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-09-12 21:27:08,759 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:27:08,834 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-3aa3c7e8-71ef-4c8d-84e4-843fcec91191
2024-09-12 21:27:08,835 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
```

$cd DA-Lab
$mkdir exp4
$cd exp4
$nano sample.txt



```
  GNU nano 7.2                          sample.txt
1,John
2,Jane
3,Joe
4,Emma
```

$nano demo_pig.pig

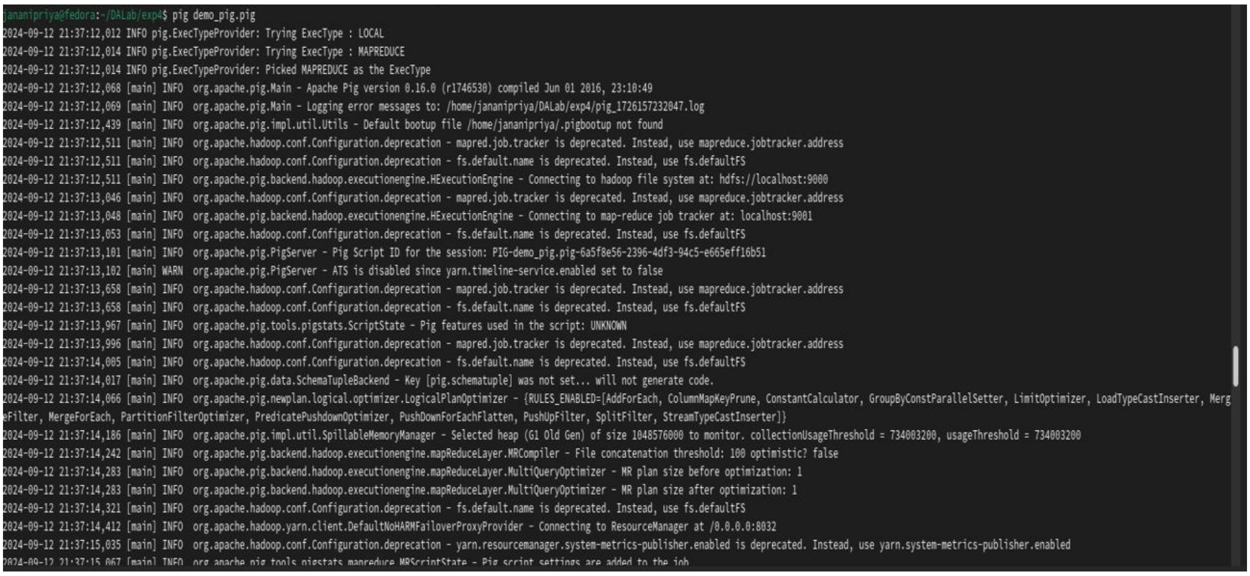


```
  GNU nano 7.2                          demo_pig.pig
-- Load the data from HDFS
data = LOAD '/piginput/sample.txt' USING PigStorage(',') AS (id:int,text:charar>
-- Dump the data to check if it was loaded correctly
DUMP data;
```

$hdfs dfs -mkdir /exp4

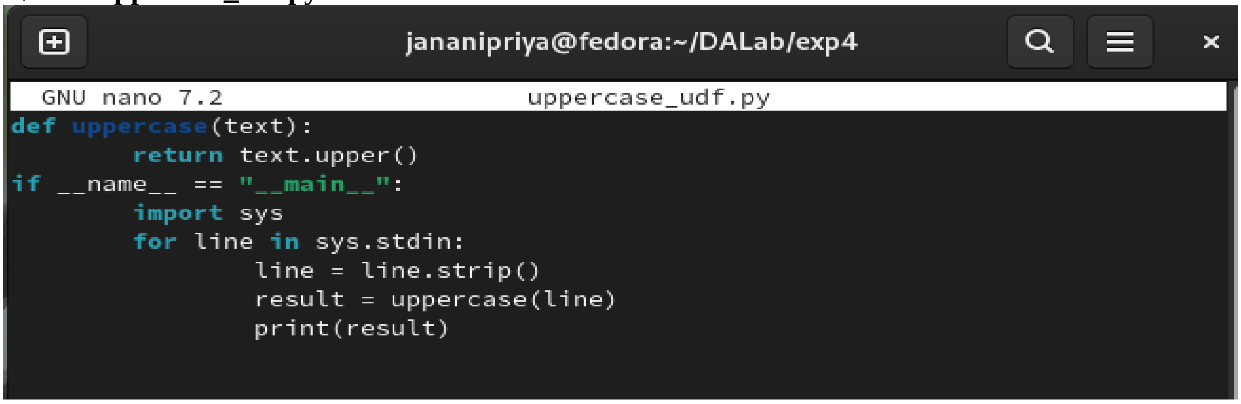
$hdfs dfs -copyFromLocal ~/DA-Lab/exp4/sample.txt /exp4

$pig demo_pig.pig



```
jananipriya@fedora:~/DALab/exp4$ pig demo_pig.pig
2024-09-12 21:37:12,012 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-12 21:37:12,014 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-12 21:37:12,014 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-12 21:37:12,068 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-12 21:37:12,069 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/jananipriya/DALab/exp4/pig_1726157232047.log
2024-09-12 21:37:12,439 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/jananipriya/.pigbootup not found
2024-09-12 21:37:12,511 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-12 21:37:12,511 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:37:12,511 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-12 21:37:13,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-12 21:37:13,048 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-09-12 21:37:13,053 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:37:13,101 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-6a5f8e56-2396-4df3-94c5-e665eff16b51
2024-09-12 21:37:13,102 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-12 21:37:13,658 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-12 21:37:13,658 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:37:13,967 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-12 21:37:13,996 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-12 21:37:14,005 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:37:14,017 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-12 21:37:14,066 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, Merg
eFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-09-12 21:37:14,186 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageThreshold = 734003200
2024-09-12 21:37:14,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-12 21:37:14,283 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-12 21:37:14,283 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-12 21:37:14,321 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-12 21:37:14,412 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /0.0.0.0:8032
2024-09-12 21:37:15,035 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-09-12 21:37:15,067 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
```

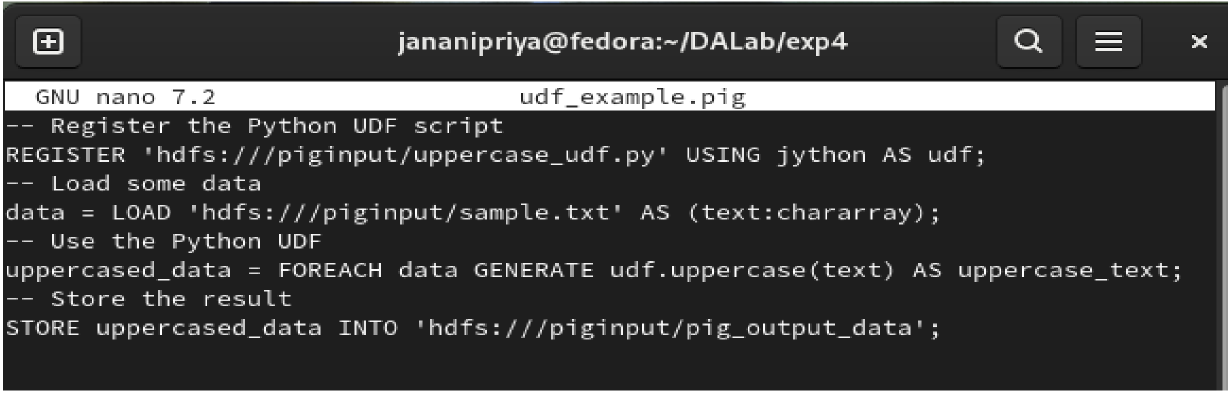
$nano uppercase_udf.py



```
  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)
```

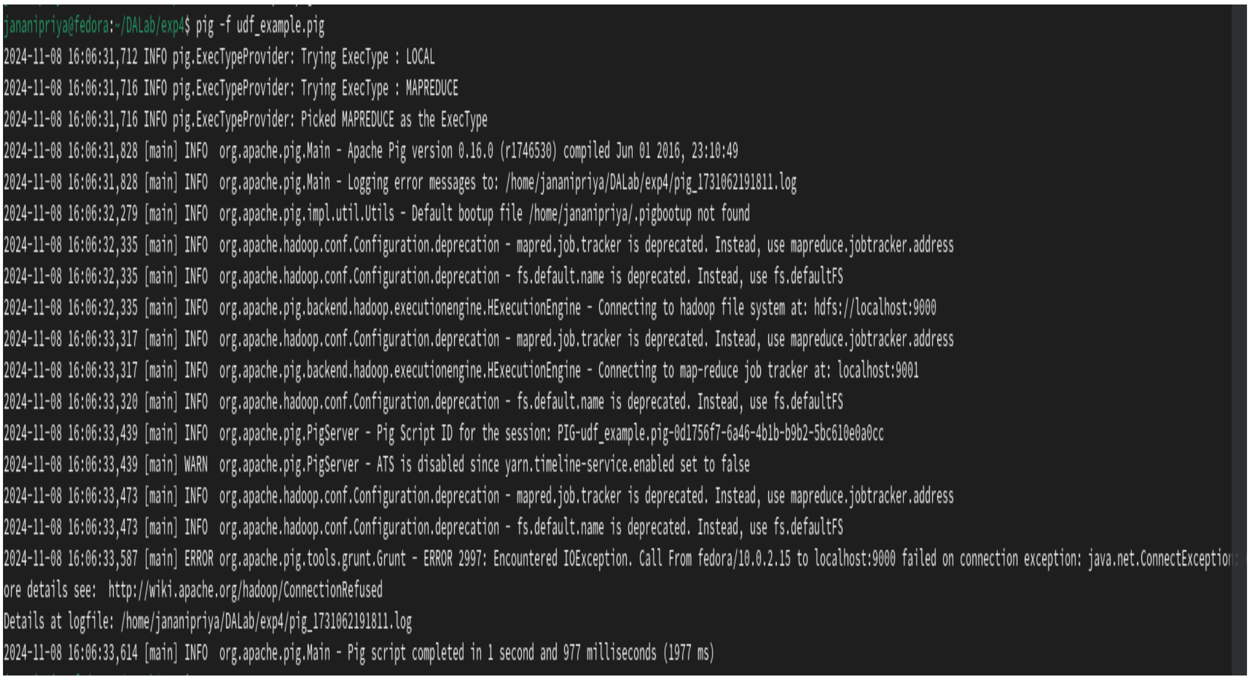
$hdfs dfs -copyFromLocal ~/DA-Lab/exp4/uppercase_udf.py /exp4

$nano udf_example.pig



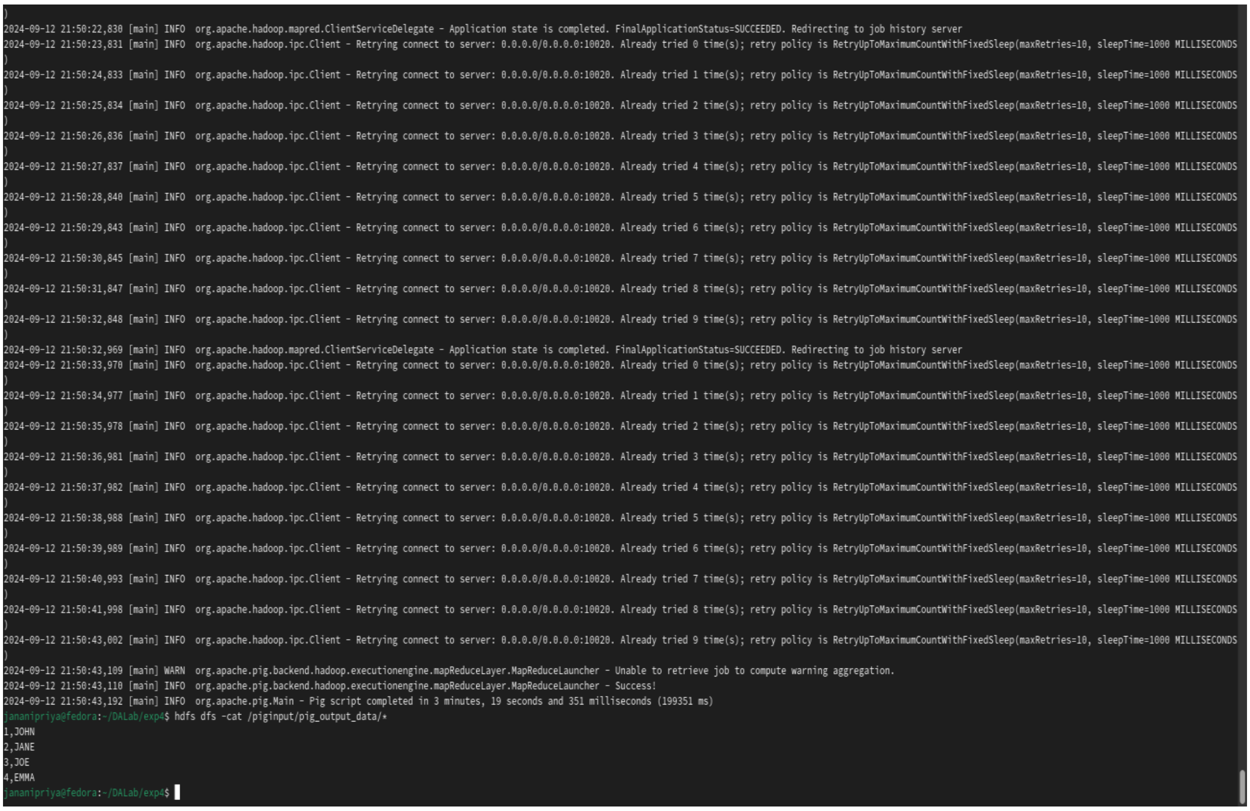
```
  GNU nano 7.2                          udf_example.pig
-- Register the Python UDF script
REGISTER 'hdfs:///piginput/uppercase_udf.py' USING jython AS udf;
-- Load some data
data = LOAD 'hdfs:///piginput/sample.txt' AS (text:chararray);
-- Use the Python UDF
uppercased_data = FOREACH data GENERATE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

$pig -f udf_example.pig



```
jananipriya@fedora:~/DALab/exp4$ pig -f udf_example.pig
2024-11-08 16:06:31,712 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-11-08 16:06:31,716 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-11-08 16:06:31,716 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-11-08 16:06:31,828 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-11-08 16:06:31,828 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/jananipriya/DALab/exp4/pig_1731062191811.log
2024-11-08 16:06:32,279 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/jananipriya/.pigbootup not found
2024-11-08 16:06:32,335 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-11-08 16:06:32,335 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-08 16:06:32,335 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-11-08 16:06:33,317 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-11-08 16:06:33,317 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-11-08 16:06:33,320 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-08 16:06:33,439 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-0d1756f7-6a46-4b1b-b9b2-5bc610e0a0cc
2024-11-08 16:06:33,439 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-11-08 16:06:33,473 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-11-08 16:06:33,473 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-11-08 16:06:33,587 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 2997: Encountered IOException. Call From fedora/10.0.2.15 to localhost:9000 failed on connection exception: java.net.ConnectException
ore details see:  http://wiki.apache.org/hadoop/ConnectionRefused
Details at logfile: /home/jananipriya/DALab/exp4/pig_1731062191811.log
2024-11-08 16:06:33,614 [main] INFO  org.apache.pig.Main - Pig script completed in 1 second and 977 milliseconds (1977 ms)
```

$hdfs dfs -cat /exp4/output/*



```
)
2024-09-12 21:50:22,830 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-09-12 21:50:23,831 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:24,833 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:25,834 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:26,836 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:27,837 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:28,840 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:29,843 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:30,845 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:31,847 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:32,848 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:32,969 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-09-12 21:50:33,970 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:34,977 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:35,978 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:36,981 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:37,982 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:38,988 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:39,989 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:40,993 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:41,998 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:43,002 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS
)
2024-09-12 21:50:43,109 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2024-09-12 21:50:43,110 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-12 21:50:43,192 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 19 seconds and 351 milliseconds (199351 ms)
jananipriya@fedora:~/DALab/exp4$ hdfs dfs -cat /piginput/pig_output_data/*
1,JOHN
2,JANE
3,JOE
4,EMMA
jananipriya@fedora:~/DALab/exp4$ 
```