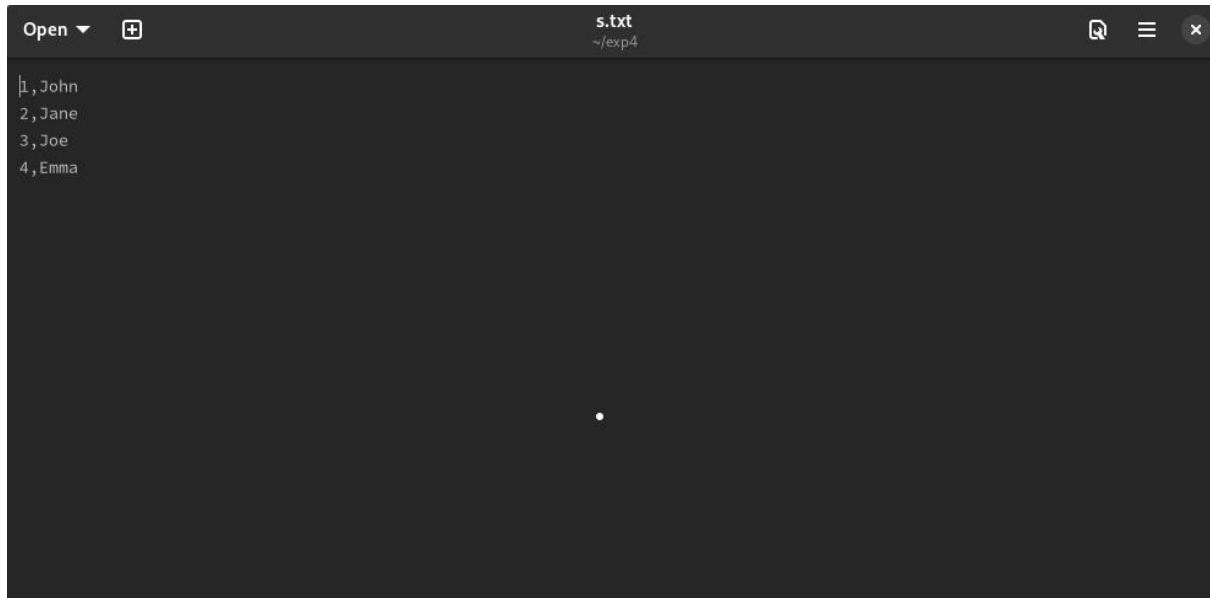
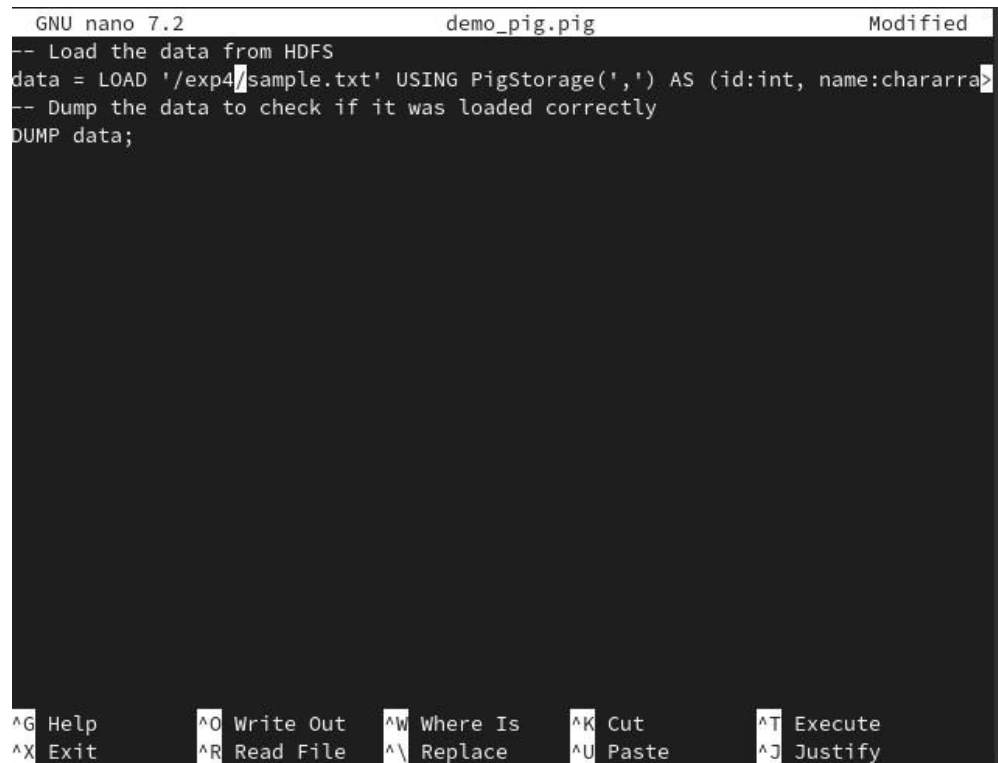


**Exp. No : 4****User Defined Function (UDF) in PIG****1. Create sample.txt**A screenshot of a text editor window titled 's.txt' with a path of '~/.exp4'. The window contains four lines of text: '1, John', '2, Jane', '3, Joe', and '4, Emma'. The editor has a dark theme and standard window controls.**2. Create demo\_pig.pig file**A screenshot of a GNU nano 7.2 text editor window titled 'demo\_pig.pig'. The window contains the following Pig script: 

```
-- 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;
```

 The bottom of the window shows a status bar with various keyboard shortcuts like ^G Help, ^O Write Out, ^W Where Is, ^K Cut, ^T Execute, ^X Exit, ^R Read File, ^\ Replace, ^U Paste, and ^J Justify.

## 3. Execute demo\_pig.pig

```

keerthi@fedora:~/exp4
keerthi@fedora:~/exp4  x  keerthi@fedora:~/exp4 — /usr/lib/jvm/java-8-openjdk/bin/java -Dpro...  x
keerthi@fedora:~/exp4$ pig demo_pig.pig
2024-10-20 12:20:54,013 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-20 12:20:54,014 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-20 12:20:54,014 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-20 12:20:54,068 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-20 12:20:54,071 [main] INFO org.apache.pig.Main - Logging error messages to: /home/keerthi/exp4/pig_1729441254067.log
2024-10-20 12:20:54,330 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/keerthi/.pigbootup not found
2024-10-20 12:20:54,385 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:20:54,386 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:20:54,386 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at : hdfs://localhost:9000
2024-10-20 12:20:54,834 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:20:54,834 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-10-20 12:20:54,835 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:20:54,847 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-a94672f3-3cf6-405b-892b-cb56133c2658
2024-10-20 12:20:54,848 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-10-20 12:20:55,199 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:20:55,200 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:20:55,356 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-10-20 12:20:55,384 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:20:55,388 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:20:55,399 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.

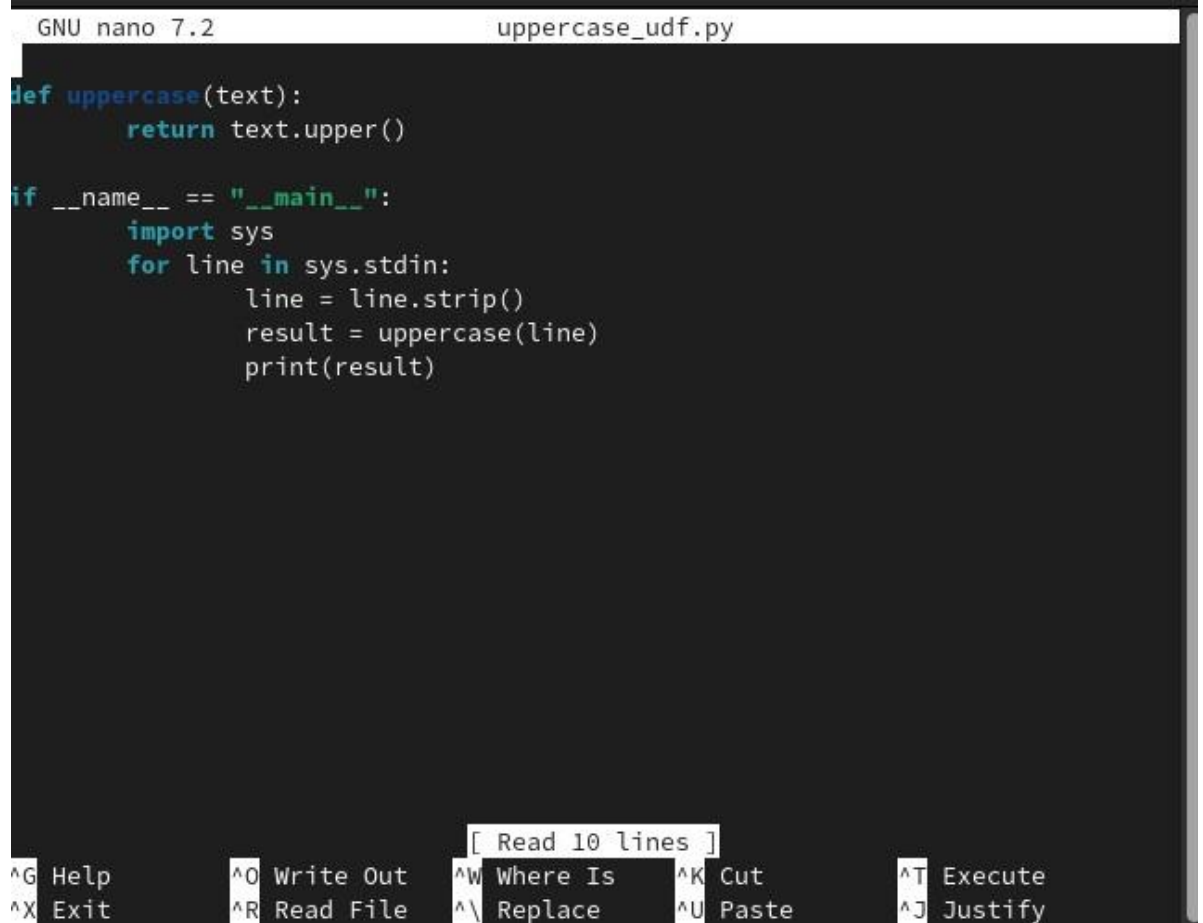
```

```

keerthi@fedora:~/exp4
keerthi@fedora:~/exp4  x  keerthi@fedora:~/exp4 — /usr/lib/jvm/java-8-openjdk/bin/java -Dpro...  x
2024-10-20 12:42:37,546 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:38,548 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:39,550 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:40,550 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:41,550 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:42,552 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:43,555 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 times(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-20 12:42:43,657 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2024-10-20 12:42:43,657 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-10-20 12:42:43,662 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-10-20 12:42:43,662 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:42:43,663 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:42:43,665 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code
.
2024-10-20 12:42:43,841 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-10-20 12:42:43,843 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-10-20 12:42:44,177 [main] INFO org.apache.pig.Main - Pig script completed in 3 minutes, 31 seconds and 447 milliseconds (211447 ms)
keerthi@fedora:~/exp4$

```

## 4. 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 ]

<b>^G</b> Help	<b>^O</b> Write Out	<b>^W</b> Where Is	<b>^K</b> Cut	<b>^T</b> Execute
<b>^X</b> Exit	<b>^R</b> Read File	<b>^_</b> Replace	<b>^U</b> Paste	<b>^J</b> Justify

## 5. 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';

```

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

## 5. Execute udf\_example.pig

```

keerthi@fedora:~/exp4$ pig udf_example.pig
2024-10-20 12:39:47,253 INFO keerthi@fedora:~/exp4$ Trying ExecType : LOCAL
2024-10-20 12:39:47,255 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-20 12:39:47,255 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-20 12:39:47,618 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-20 12:39:47,640 [main] INFO org.apache.pig.Main - Logging error messages to: /home/keerthi/exp4/pig_1729442387609.log
2024-10-20 12:39:48,533 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/keerthi/.pigbootup not found
2024-10-20 12:39:48,717 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:39:48,717 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:39:48,717 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at : hdfs://localhost:9000
2024-10-20 12:39:52,010 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:39:52,010 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
2024-10-20 12:39:52,012 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:39:52,039 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-cb93d312-312a-48de-93af-eafcc7d9ca0e
2024-10-20 12:39:52,047 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-10-20 12:39:52,183 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:39:52,186 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 12:39:53,259 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=/tmp/pig_jython_6524728859344915343
2024-10-20 12:40:01,795 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: udf.uppercase
2024-10-20 12:40:02,154 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-20 12:40:02,154 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

```

Output :

```
keerthi@fedora:~/exp4$ hdfs dfs -cat /exp4/output/part-m-00000  
1,JOHN  
2,JANE  
3,JOE  
4,EMMA
```