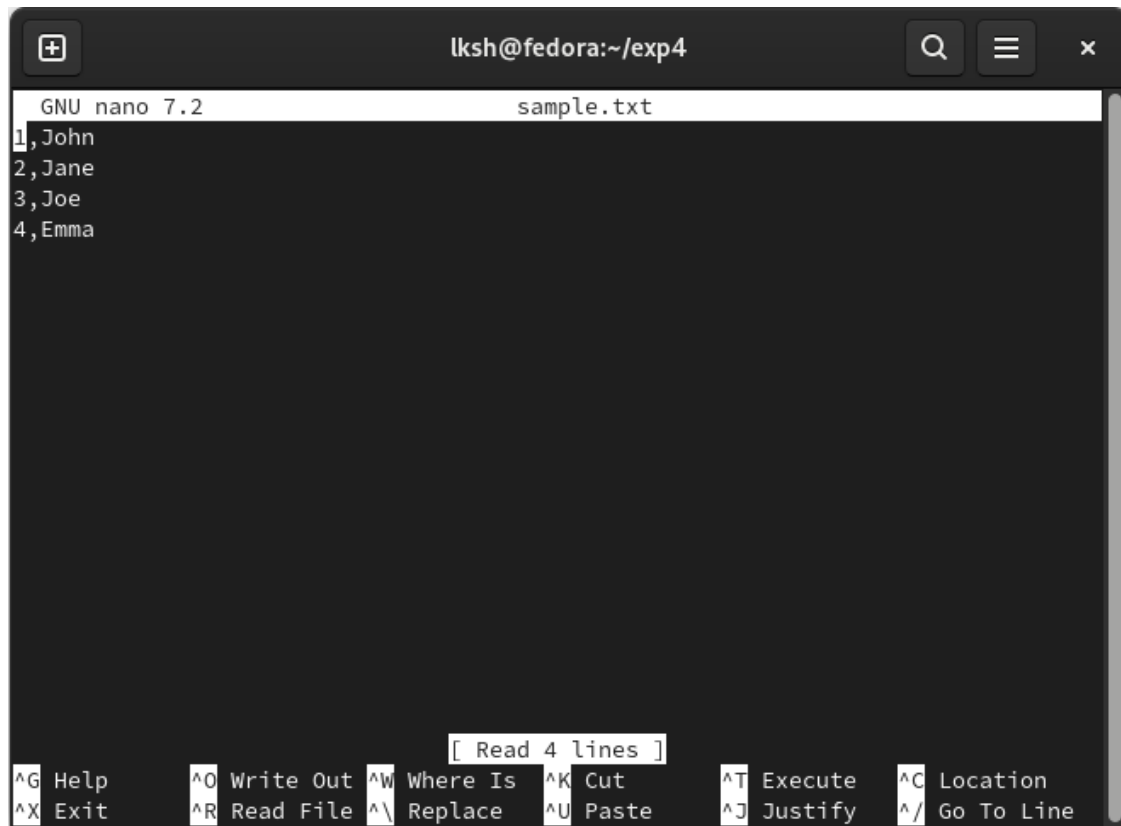


**Exp. No : 4****User Defined Function (UDF) in PIG**

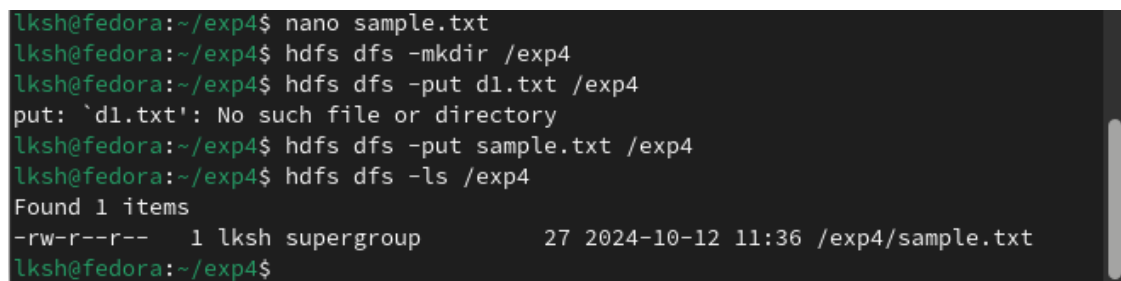
1. Create sample.txt



The screenshot shows a terminal window titled 'lksh@fedora:~/exp4'. Inside, the GNU nano 7.2 editor is open, editing a file named 'sample.txt'. The file contains four lines of text: '1, John', '2, Jane', '3, Joe', and '4, Emma'. The cursor is at the end of the fourth line. At the bottom of the editor, a status bar indicates '[ Read 4 lines ]'. Below the editor, a row of keyboard shortcuts is displayed: ^G Help, ^O Write Out, ^W Where Is, ^K Cut, ^T Execute, ^C Location, ^X Exit, ^R Read File, ^\ Replace, ^U Paste, ^J Justify, and ^\_ Go To Line.

```
lksh@fedora:~/exp4$ nano sample.txt
GNU nano 7.2 sample.txt
1, John
2, Jane
3, Joe
4, Emma
[ Read 4 lines ]
^G Help    ^O Write Out  ^W Where Is  ^K Cut       ^T Execute   ^C Location
^X Exit    ^R Read File  ^\ Replace   ^U Paste      ^J Justify   ^_ Go To Line
```

2. Upload sample.txt file to HDFS Storage.



The screenshot shows a terminal window with the following commands and output: 'nano sample.txt' is run, followed by 'hdfs dfs -mkdir /exp4', then 'hdfs dfs -put d1.txt /exp4' which results in an error 'put: `d1.txt': No such file or directory'. Then 'hdfs dfs -put sample.txt /exp4' is run successfully, followed by 'hdfs dfs -ls /exp4' which shows 'Found 1 items' and a file listing for '/exp4/sample.txt' with permissions '-rw-r--r--', owner 'lksh', group 'supergroup', size '27', and timestamp '2024-10-12 11:36'.

```
lksh@fedora:~/exp4$ nano sample.txt
lksh@fedora:~/exp4$ hdfs dfs -mkdir /exp4
lksh@fedora:~/exp4$ hdfs dfs -put d1.txt /exp4
put: `d1.txt': No such file or directory
lksh@fedora:~/exp4$ hdfs dfs -put sample.txt /exp4
lksh@fedora:~/exp4$ hdfs dfs -ls /exp4
Found 1 items
-rw-r--r--  1 lksh supergroup          27 2024-10-12 11:36 /exp4/sample.txt
lksh@fedora:~/exp4$
```

## 3. Create demo\_pig.pig file

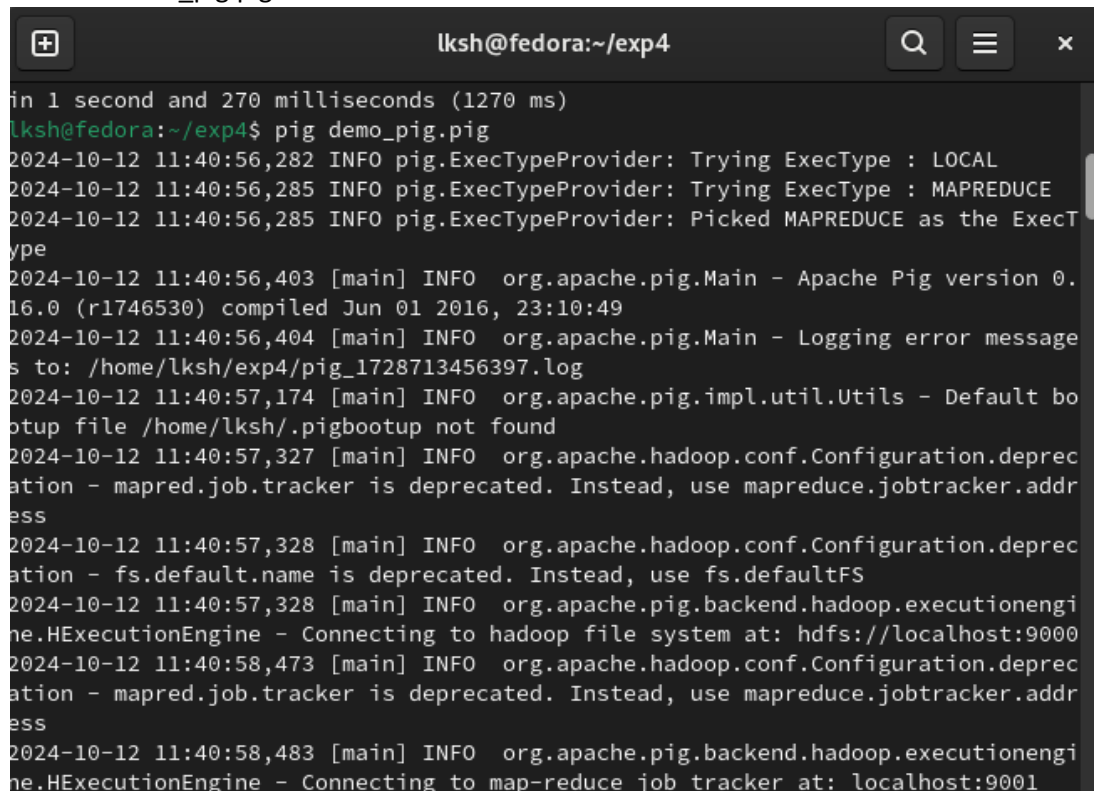


The screenshot shows a terminal window with the title bar 'lksh@fedora:~/exp4'. The window contains the GNU nano 7.2 editor editing a file named 'demo\_pig.pig'. The file content is as follows:

```
-- Load the data from HDFS
data = LOAD '/piginput/sample.txt' USING PigStorage(',') AS (id:int, name:chara>
-- Dump the data to check if it was loaded correctly
DUMP data;
```

The bottom status bar of the nano editor shows the following shortcuts: ^G Help, ^O Write Out, ^W Where Is, ^K Cut, ^T Execute, ^C Location, ^X Exit, ^R Read File, ^\ Replace, ^U Paste, ^J Justify, and ^\_ Go To Line. A status indicator '[ Read 4 lines ]' is also visible.

## 4. Execute demo\_pig.pig



The screenshot shows a terminal window with the title bar 'lksh@fedora:~/exp4'. The terminal output is as follows:

```
in 1 second and 270 milliseconds (1270 ms)
lksh@fedora:~/exp4$ pig demo_pig.pig
2024-10-12 11:40:56,282 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 11:40:56,285 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 11:40:56,285 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
ype
2024-10-12 11:40:56,403 [main] INFO org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 11:40:56,404 [main] INFO org.apache.pig.Main - Logging error message
s to: /home/lksh/exp4/pig_1728713456397.log
2024-10-12 11:40:57,174 [main] INFO org.apache.pig.impl.util.Utls - Default bo
otup file /home/lksh/.pigbootup not found
2024-10-12 11:40:57,327 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-10-12 11:40:57,328 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-12 11:40:57,328 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-10-12 11:40:58,473 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-10-12 11:40:58,483 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:9001
```

```

2024-10-12 11:44:09,422 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2024-10-12 11:44:09,423 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-12 11:44:09,427 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-12 11:44:09,433 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-10-12 11:44:09,502 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-10-12 11:44:09,509 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-10-12 11:44:09,714 [main] INFO org.apache.pig.Main - Pig script completed in 3 minutes, 13 seconds and 497 milliseconds (193497 ms)

```

##### 5. Create uppercase\_udf.py

```

lksh@fedora:~/exp4$ nano uppercase_udf.py
lksh@fedora:~/exp4$ hdfs dfs -ls /exp4
Found 1 items
-rw-r--r--  1 lksh supergroup          27 2024-10-12 11:36 /exp4/sample.txt
lksh@fedora:~/exp4$ hdfs dfs -put uppercase_udf.py /exp4

```

```

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.

```
lksh@fedora:~/exp4$ hdfs dfs -ls /exp4
Found 2 items
-rw-r--r--  1 lksh supergroup      27 2024-10-12 11:36 /exp4/sample.txt
-rw-r--r--  1 lksh supergroup    172 2024-10-12 11:47 /exp4/uppercase_udf.py
lksh@fedora:~/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

```
lksh@fedora:~/exp4$ pig udf_example.pig
2024-10-12 11:49:03,381 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 11:49:03,387 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 11:49:03,387 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-12 11:49:03,480 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 11:49:03,481 [main] INFO org.apache.pig.Main - Logging error messages to: /home/lksh/exp4/pig_1728713943478.log
2024-10-12 11:49:04,209 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/lksh/.pigbootstrap not found
2024-10-12 11:49:04,404 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```

## Output :

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