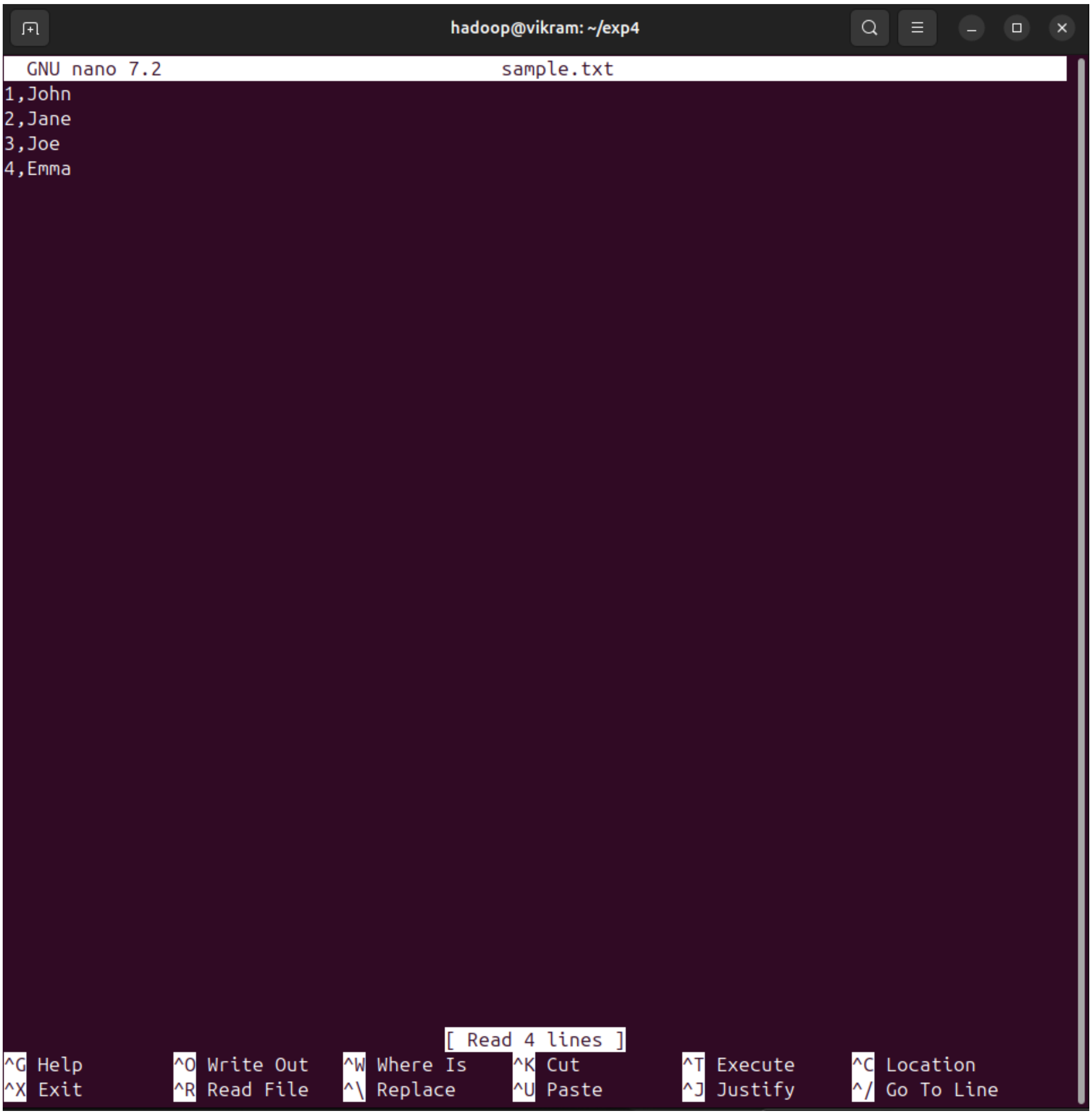
Exp. No : 4

User Defined Function (UDF) in PIG

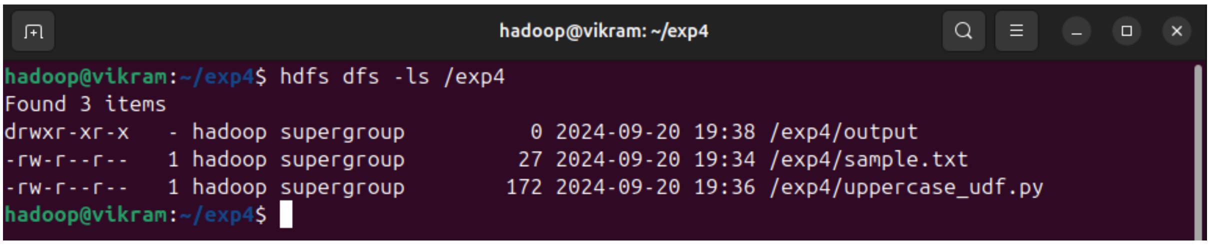
1. Create sample.txt



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

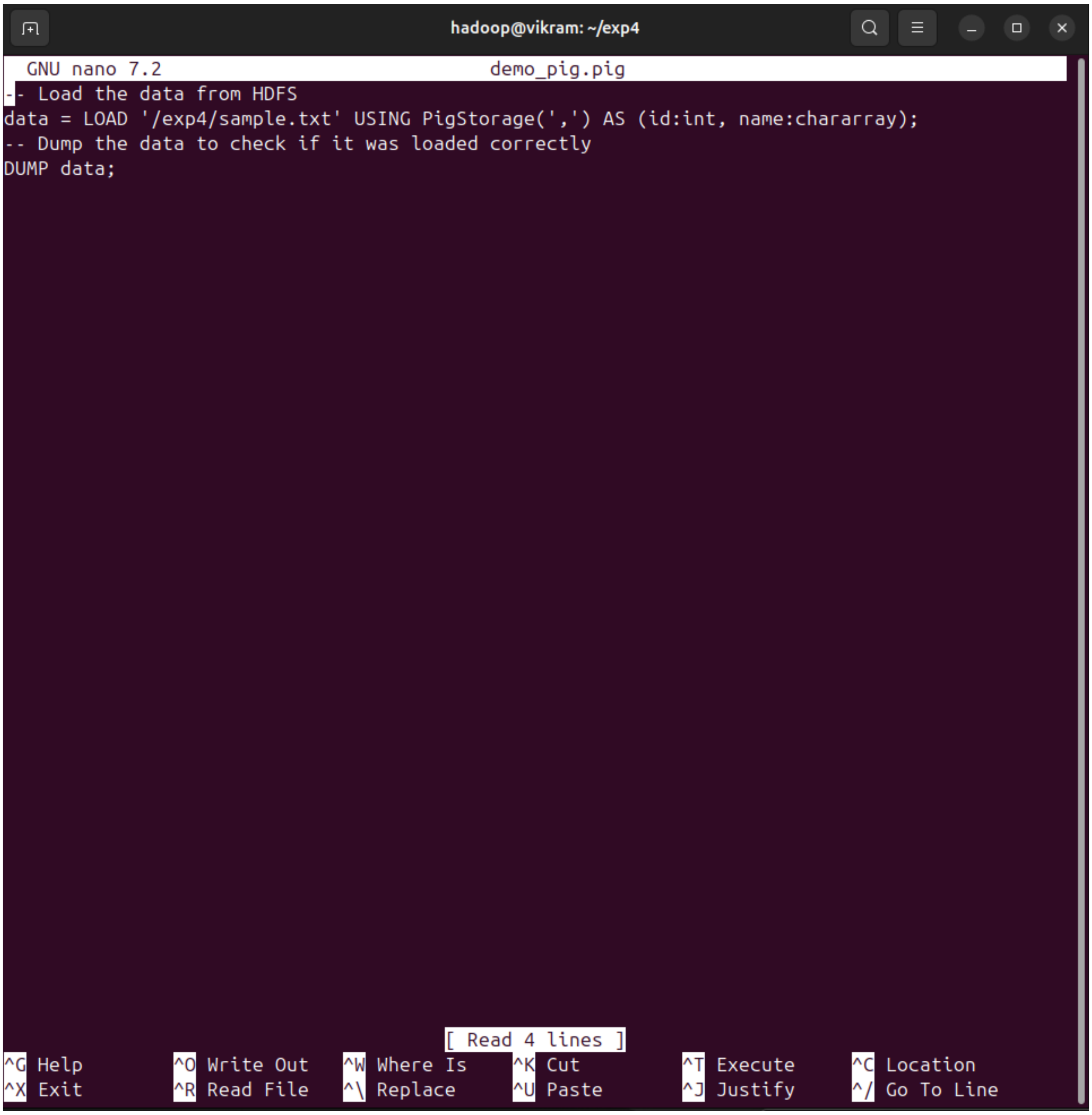
```
                                              [ Read 4 lines ]
^G Help          ^O Write Out     ^W Where Is      ^K Cut           ^T Execute       ^C Location
^X Exit          ^R Read File     ^\ Replace       ^U Paste         ^J Justify       ^/ Go To Line
```

2. Upload sample.txt file to HDFS Storage.



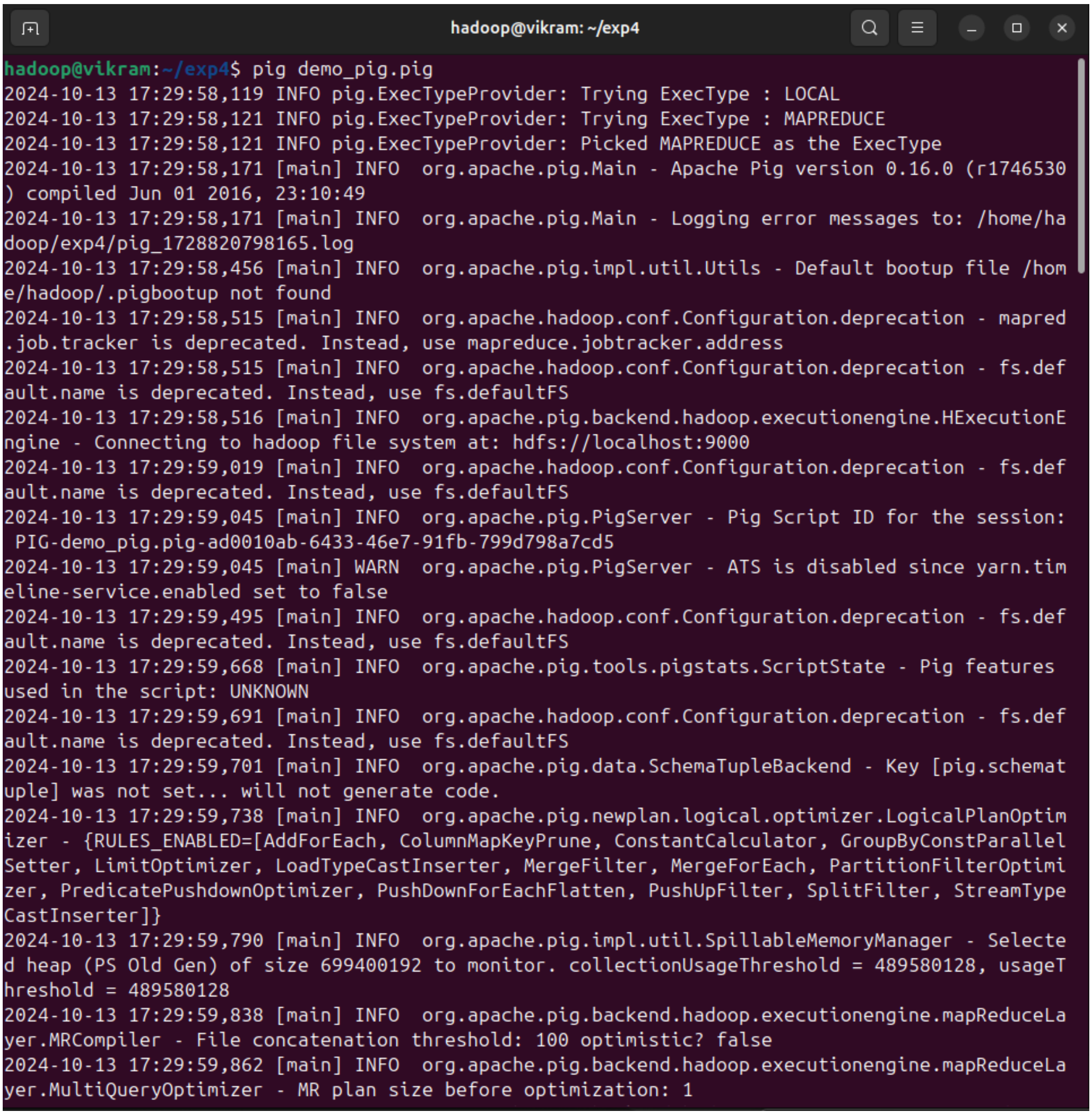
```
hadoop@vikram:~/exp4$ hdfs dfs -ls /exp4
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-20 19:38 /exp4/output
-rw-r--r--   1 hadoop supergroup         27 2024-09-20 19:34 /exp4/sample.txt
-rw-r--r--   1 hadoop supergroup        172 2024-09-20 19:36 /exp4/uppercase_udf.py
hadoop@vikram:~/exp4$
```

3. Create demo_pig.pig file



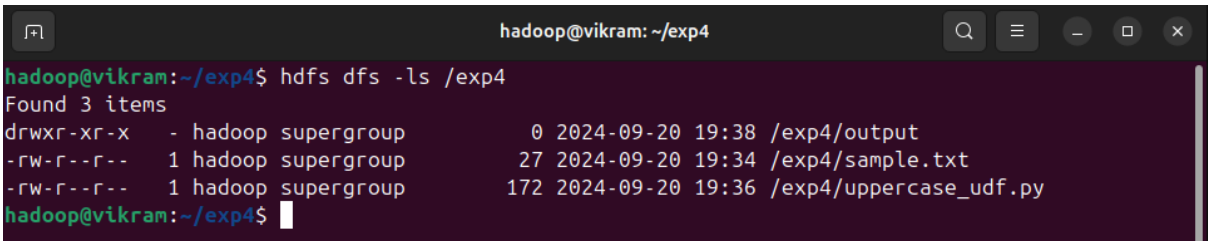
```
  GNU nano 7.2                          demo_pig.pig
-- Load the data from HDFS
data = LOAD '/exp4/sample.txt' USING PigStorage(',') AS (id:int, name:chararray);
-- Dump the data to check if it was loaded correctly
DUMP data;

                                      [ Read 4 lines ]
```

4. Execute demo_pig.pig

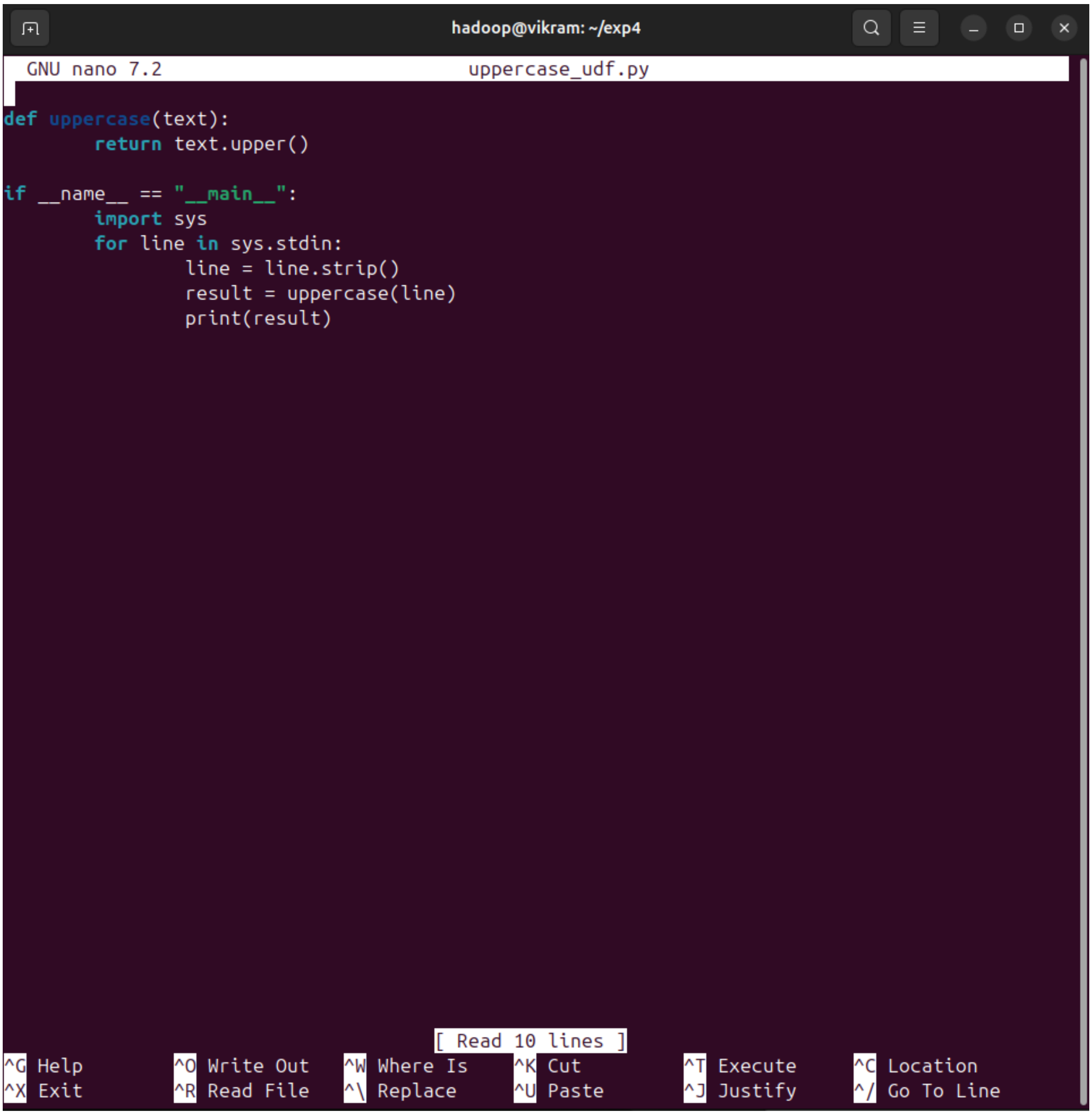


```
hadoop@vikram:~/exp4$ pig demo_pig.pig
2024-10-13 17:29:58,119 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-13 17:29:58,121 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-13 17:29:58,121 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-13 17:29:58,171 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530
) compiled Jun 01 2016, 23:10:49
2024-10-13 17:29:58,171 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ha
doop/exp4/pig_1728820798165.log
2024-10-13 17:29:58,456 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /hom
e/hadoop/.pigbootup not found
2024-10-13 17:29:58,515 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred
.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-13 17:29:58,515 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:29:58,516 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionE
ngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-10-13 17:29:59,019 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:29:59,045 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session:
 PIG-demo_pig.pig-ad0010ab-6433-46e7-91fb-799d798a7cd5
2024-10-13 17:29:59,045 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.tim
eline-service.enabled set to false
2024-10-13 17:29:59,495 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:29:59,668 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
used in the script: UNKNOWN
2024-10-13 17:29:59,691 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:29:59,701 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schemat
uple] was not set... will not generate code.
2024-10-13 17:29:59,738 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptim
izer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallel
Setter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimi
zer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamType
CastInserter]}
2024-10-13 17:29:59,790 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selecte
d heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageT
hreshold = 489580128
2024-10-13 17:29:59,838 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-10-13 17:29:59,862 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MultiQueryOptimizer - MR plan size before optimization: 1
```

5. Create uppercase_udf.py

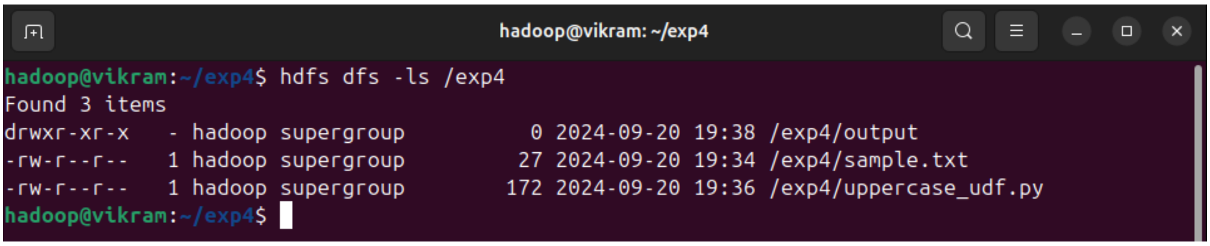


```
hadoop@vikram:~/exp4$ hdfs dfs -ls /exp4
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-20 19:38 /exp4/output
-rw-r--r--   1 hadoop supergroup         27 2024-09-20 19:34 /exp4/sample.txt
-rw-r--r--   1 hadoop supergroup        172 2024-09-20 19:36 /exp4/uppercase_udf.py
hadoop@vikram:~/exp4$ 
```



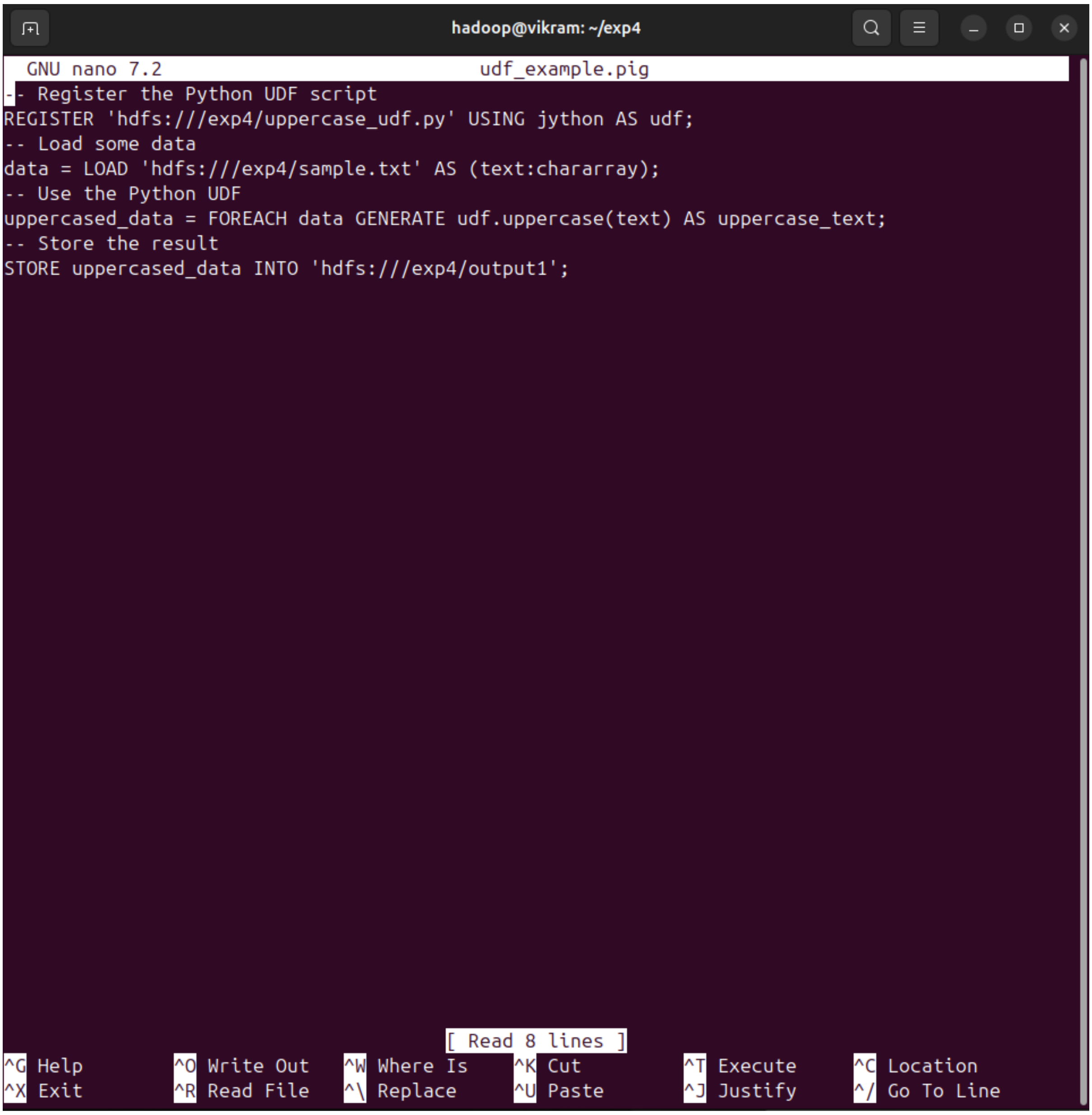
```
  GNU nano 7.2                          uppercase_udf.py

def uppercase(text):
        return text.upper()

if __name__ == "__main__":
        import sys
        for line in sys.stdin:
                line = line.strip()
                result = uppercase(line)
                print(result)
```

6. Upload uppercase_udf.py file to HDFS Storage.



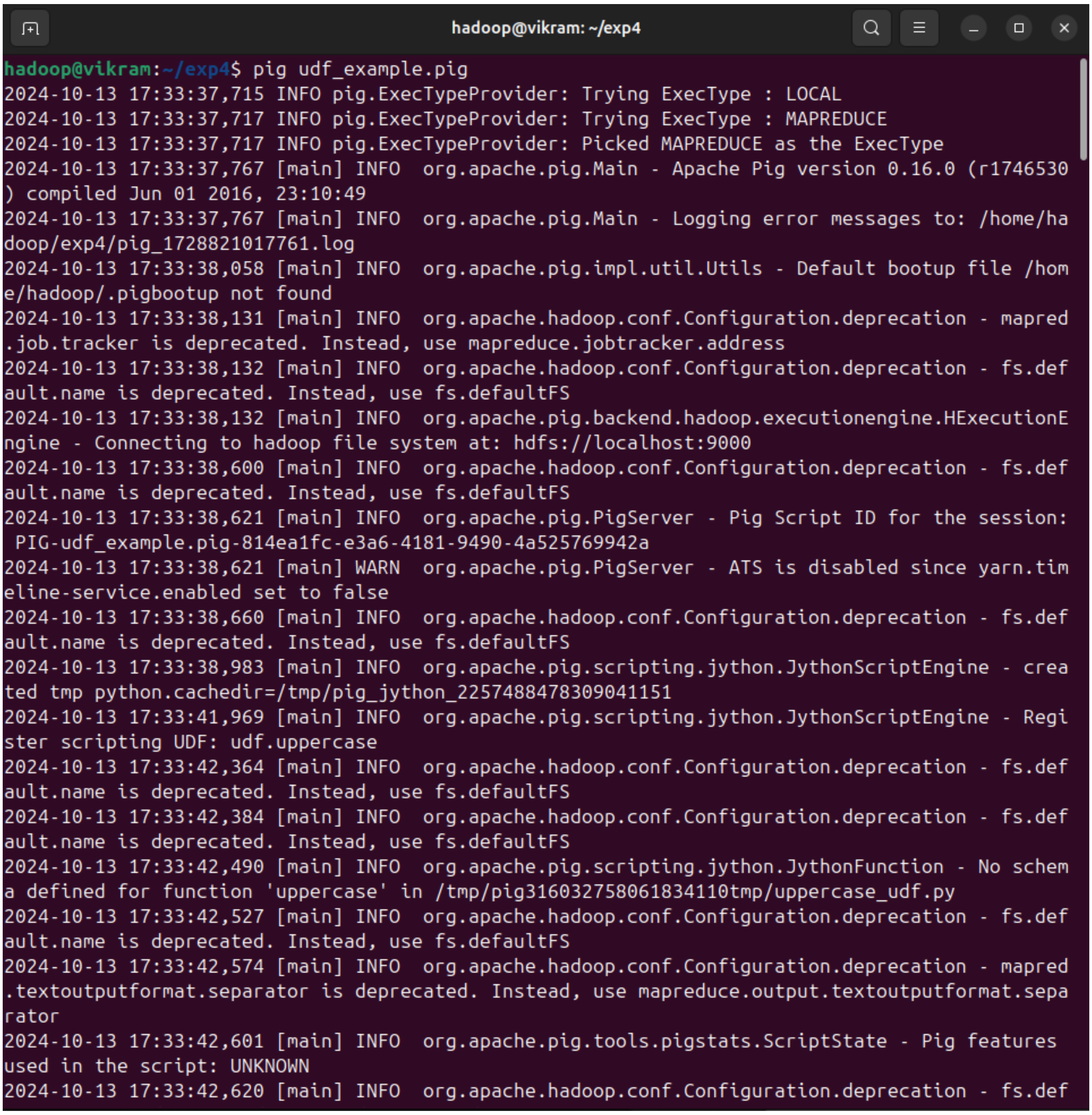
```
hadoop@vikram:~/exp4$ hdfs dfs -ls /exp4
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-20 19:38 /exp4/output
-rw-r--r--   1 hadoop supergroup         27 2024-09-20 19:34 /exp4/sample.txt
-rw-r--r--   1 hadoop supergroup        172 2024-09-20 19:36 /exp4/uppercase_udf.py
hadoop@vikram:~/exp4$ 
```

7. Create udf_example.pig



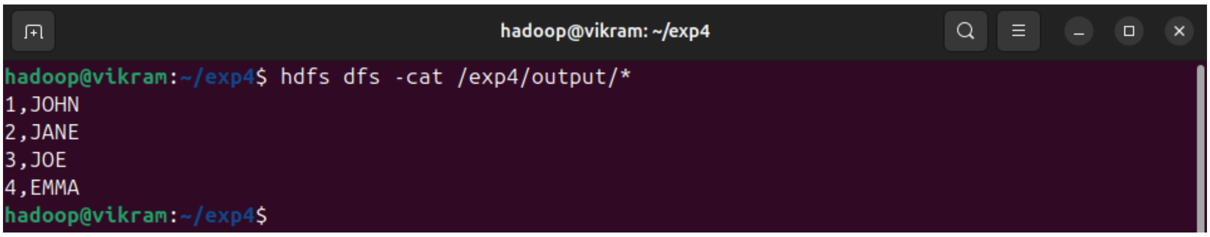
```
  GNU nano 7.2                                udf_example.pig
-- Register the Python UDF script
REGISTER 'hdfs:///exp4/uppercase_udf.py' USING jython AS udf;
-- Load some data
data = LOAD 'hdfs:///exp4/sample.txt' AS (text:chararray);
-- Use the Python UDF
uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;
-- Store the result
STORE uppercased_data INTO 'hdfs:///exp4/output1';

                                      [ Read 8 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
```

8. Execute udf_example.pig



```
hadoop@vikram:~/exp4$ pig udf_example.pig
2024-10-13 17:33:37,715 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-13 17:33:37,717 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-13 17:33:37,717 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-13 17:33:37,767 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530
) compiled Jun 01 2016, 23:10:49
2024-10-13 17:33:37,767 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/ha
doop/exp4/pig_1728821017761.log
2024-10-13 17:33:38,058 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /hom
e/hadoop/.pigbootup not found
2024-10-13 17:33:38,131 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred
.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-13 17:33:38,132 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:33:38,132 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionE
ngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-10-13 17:33:38,600 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:33:38,621 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session:
 PIG-udf_example.pig-814ea1fc-e3a6-4181-9490-4a525769942a
2024-10-13 17:33:38,621 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.tim
eline-service.enabled set to false
2024-10-13 17:33:38,660 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:33:38,983 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - crea
ted tmp python.cachedir=/tmp/pig_jython_2257488478309041151
2024-10-13 17:33:41,969 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - Regi
ster scripting UDF: udf.uppercase
2024-10-13 17:33:42,364 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:33:42,384 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:33:42,490 [main] INFO  org.apache.pig.scripting.jython.JythonFunction - No schem
a defined for function 'uppercase' in /tmp/pig316032758061834110tmp/uppercase_udf.py
2024-10-13 17:33:42,527 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
ault.name is deprecated. Instead, use fs.defaultFS
2024-10-13 17:33:42,574 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred
.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.sepa
rator
2024-10-13 17:33:42,601 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
used in the script: UNKNOWN
2024-10-13 17:33:42,620 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.def
```

Output :



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