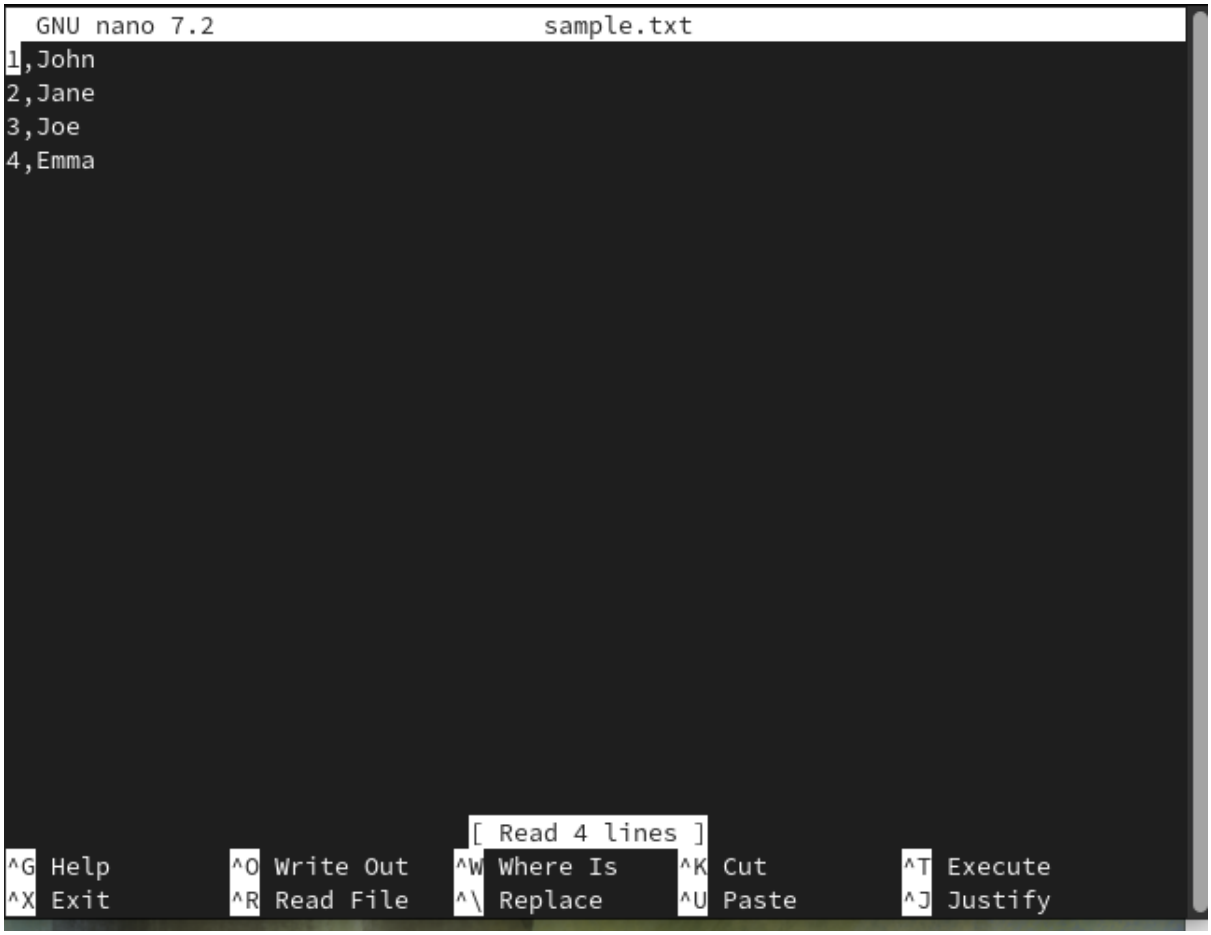
Exp. No : 4

User Defined Function (UDF) in PIG

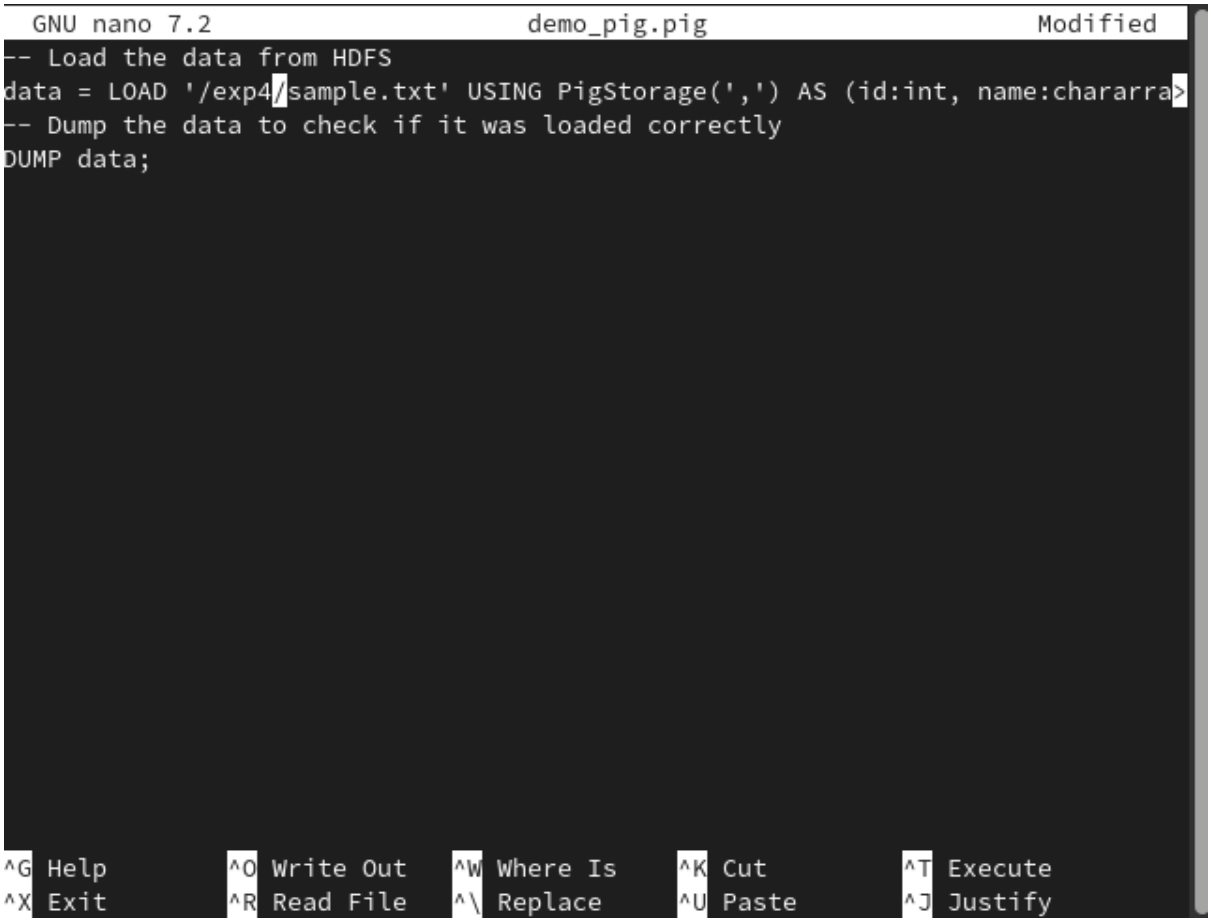
1. Create sample.txt



2. Upload sample.txt file to HDFS Storage.

```
harish@fedora:~$ hdfs dfs -ls /exp4
Found 3 items
drwxr-xr-x   - harish supergroup          0 2024-09-14 10:55 /exp4/output
-rw-r--r--   1 harish supergroup         27 2024-09-14 10:52 /exp4/sample.txt
-rw-r--r--   1 harish supergroup        172 2024-09-14 10:53 /exp4/uppercase_udf.py
harish@fedora:~$
```

3. Create demo_pig.pig file



4. Execute demo_pig.pig

```
harish@fedora:~/exp4$ pig demo.pig
2024-10-10 18:47:09,078 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-10 18:47:09,079 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-10 18:47:09,079 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-10 18:47:09,144 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-10-10 18:47:09,144 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/harish/exp4/pig_1728566229134.log
2024-10-10 18:47:09,503 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/harish/.pigbootup not found
2024-10-10 18:47:09,567 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-10 18:47:09,568 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-10 18:47:09,568 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-10-10 18:47:10,272 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-10 18:47:10,335 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo.pig-c3622257-a907-4452-bd4f-051624eb428b
2024-10-10 18:47:10,335 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-10-10 18:47:10,852 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-10 18:47:11,279 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-10-10 18:47:11,306 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-10 18:47:11,309 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-10-10 18:47:11,332 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatte
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-10-10 18:47:11,408 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold
 = 489580128, usageThreshold = 489580128
2024-10-10 18:47:11,447 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-10-10 18:47:11,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-10-10 18:47:11,487 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-10-10 18:47:11,514 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

```
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLI
2024-10-10 18:50:11,969 [main] INFO  org.apache.hadoop.ipc.Clie
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLI
2024-10-10 18:50:12,072 [main] WARN  org.apache.pig.backend.had
ion.
2024-10-10 18:50:12,072 [main] INFO  org.apache.pig.backend.had
2024-10-10 18:50:12,078 [main] INFO  org.apache.hadoop.conf.Con
e yarn.system-metrics-publisher.enabled
2024-10-10 18:50:12,078 [main] INFO  org.apache.hadoop.conf.Con
2024-10-10 18:50:12,079 [main] INFO  org.apache.pig.data.Schema
2024-10-10 18:50:12,163 [main] INFO  org.apache.hadoop.mapreduc
2024-10-10 18:50:12,164 [main] INFO  org.apache.pig.backend.had
(1,Jane)
(2,John)
(3,Brian)
(4,Heka)
2024-10-10 18:50:12,424 [main] INFO  org.apache.pig.Main - Pig
harish@fedora:~/exp4$
harish@fedora:~/exp4$
harish@fedora:~/exp4$
```

5. Create uppercase_udf.py

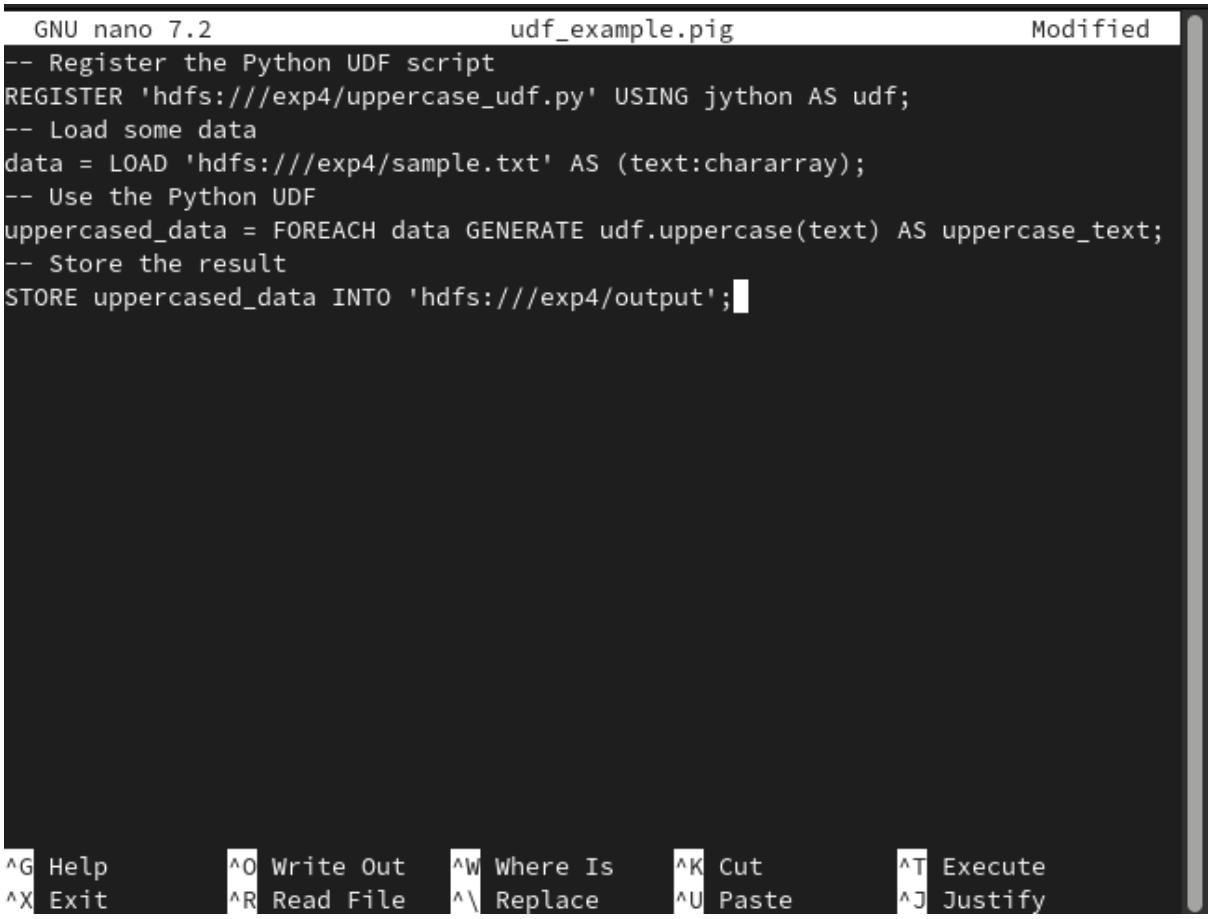
```
harish@fedora:~$ hdfs dfs -ls /exp4
Found 3 items
drwxr-xr-x   - harish supergroup          0 2024-09-14 10:55 /exp4/output
-rw-r--r--   1 harish supergroup         27 2024-09-14 10:52 /exp4/sample.txt
-rw-r--r--   1 harish supergroup        172 2024-09-14 10:53 /exp4/uppercase_udf.py
harish@fedora:~$
```

```
  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.

```
harish@fedora:~$ hdfs dfs -ls /exp4
Found 3 items
drwxr-xr-x   - harish supergroup          0 2024-09-14 10:55 /exp4/output
-rw-r--r--   1 harish supergroup         27 2024-09-14 10:52 /exp4/sample.txt
-rw-r--r--   1 harish supergroup        172 2024-09-14 10:53 /exp4/uppercase_udf.py
harish@fedora:~$
```

7. Create udf_example.pig



8. Execute udf_example.pig

```
harish@fedora:~/exp4$ pig udf_example.pig
2024-10-10 18:50:11,073 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-10 18:50:11,074 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-10 18:50:11,075 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-10 18:50:11,134 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r
2024-10-10 18:50:11,134 [main] INFO  org.apache.pig.Main - Logging error messages to: /
2024-10-10 18:50:11,513 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup fi
2024-10-10 18:50:11,590 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:11,591 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:11,591 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExe
2024-10-10 18:50:12,536 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:12,556 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the s
2024-10-10 18:50:12,556 [main] WARN  org.apache.pig.PigServer - ATS is disabled since y
2024-10-10 18:50:12,603 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:13,073 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine
2024-10-10 18:50:15,393 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine
2024-10-10 18:50:15,678 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:15,692 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:15,806 [main] INFO  org.apache.pig.scripting.jython.JythonFunction - N
case_udf.py
2024-10-10 18:50:15,839 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:15,871 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
extoutputformat.separator
2024-10-10 18:50:15,891 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig fe
2024-10-10 18:50:15,904 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
2024-10-10 18:50:15,906 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
```

```
2024-10-10 18:53:31,202 [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
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-10 18:53:32,216 [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
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-10 18:53:33,252 [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
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-10 18:53:34,288 [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
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-10 18:53:35,294 [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
oMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-10-10 18:53:35,413 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job
ion.
2024-10-10 18:53:35,413 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-10-10 18:53:35,489 [main] INFO  org.apache.pig.Main - Pig script completed in 3 minutes, 24 seconds and 421 milliseconds (204421 ms)
harish@fedora:~/exp4$
```

Output :

```
harish@fedora:~/exp4$ hdfs dfs -cat /exp4/output1/*
1,JOHN
2,JANE
3,JOE
4,EMMA
```