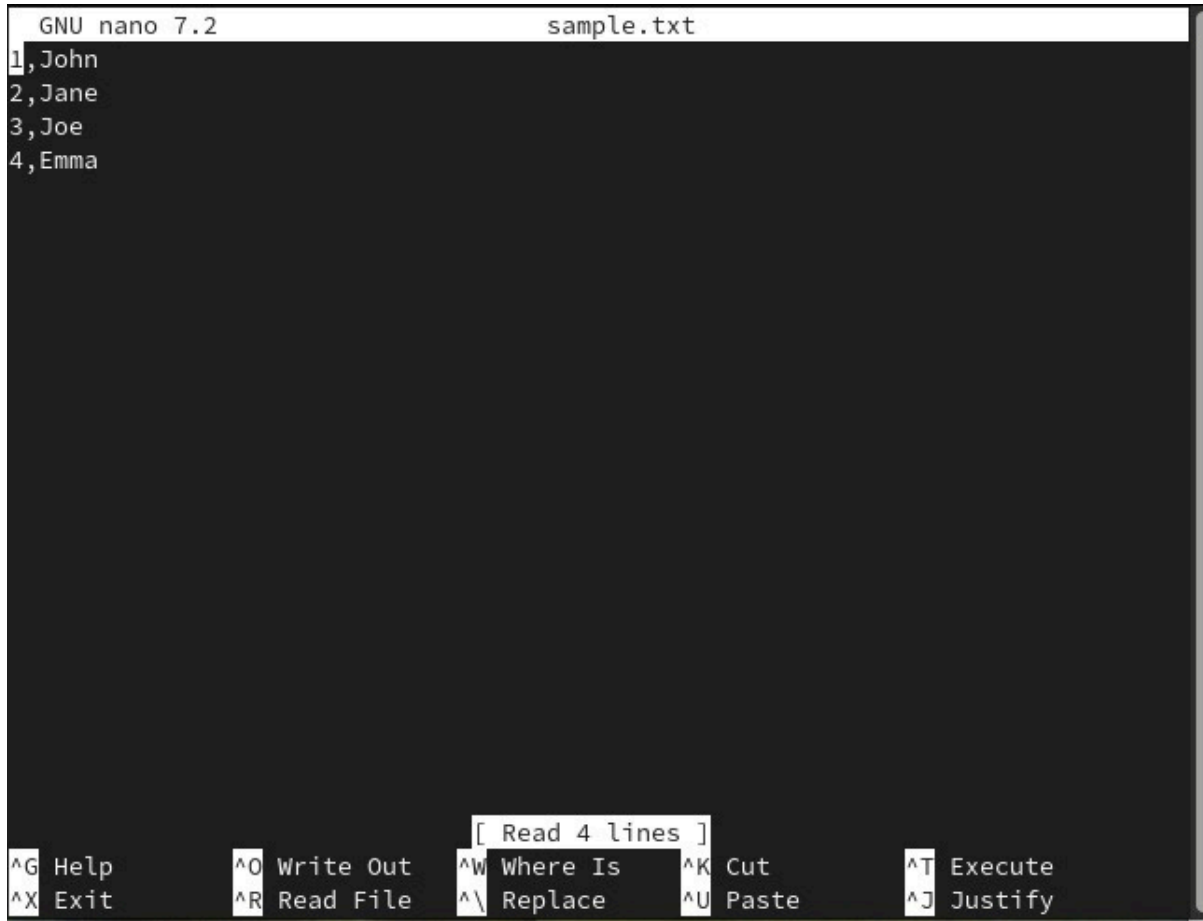


Exp. No : 4

User Defined Function (UDF) in PIG

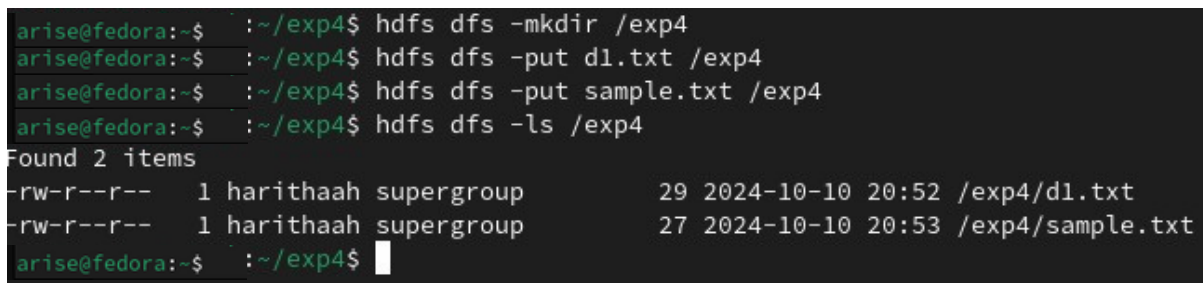
1. Create sample.txt



```
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
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```

2. Upload sample.txt file to HDFS Storage.



```
arise@fedora:~$ :~/exp4$ hdfs dfs -mkdir /exp4
arise@fedora:~$ :~/exp4$ hdfs dfs -put d1.txt /exp4
arise@fedora:~$ :~/exp4$ hdfs dfs -put sample.txt /exp4
arise@fedora:~$ :~/exp4$ hdfs dfs -ls /exp4
Found 2 items
-rw-r--r--  1 harithaah supergroup      29 2024-10-10 20:52 /exp4/d1.txt
-rw-r--r--  1 harithaah supergroup      27 2024-10-10 20:53 /exp4/sample.txt
arise@fedora:~$ :~/exp4$
```

3. Create demo_pig.pig file

```
GNU nano 7.2                                demo_pig.pig                                Modified
-- 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;
```

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

4. Execute demo_pig.pig

```

harise@fedora:~$ a:~/exp4$ pig demo.pig
2024-10-10 20:54:00,945 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-10 20:54:00,947 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-10 20:54:00,947 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
2024-10-10 20:54:01,026 [main] INFO org.apache.pig.Main - Apache Pig version 0.
2024-10-10 20:54:01,026 [main] INFO org.apache.pig.Main - Logging error message
2024-10-10 20:54:01,749 [main] INFO org.apache.pig.impl.util.Utils - Default bo
2024-10-10 20:54:01,818 [main] INFO org.apache.hadoop.conf.Configuration.deprec
2024-10-10 20:54:01,818 [main] INFO org.apache.hadoop.conf.Configuration.deprec
2024-10-10 20:54:01,818 [main] INFO org.apache.pig.backend.hadoop.executionengi
2024-10-10 20:54:03,352 [main] INFO org.apache.hadoop.conf.Configuration.deprec
2024-10-10 20:54:03,492 [main] INFO org.apache.pig.PigServer - Pig Script ID fo
2024-10-10 20:54:03,494 [main] WARN org.apache.pig.PigServer - ATS is disabled
2024-10-10 20:54:04,380 [main] INFO org.apache.hadoop.conf.Configuration.deprec
2024-10-10 20:54:05,010 [main] INFO org.apache.pig.tools.pigstats.ScriptState -
2024-10-10 20:54:05,172 [main] INFO org.apache.hadoop.conf.Configuration.deprec
2024-10-10 20:54:05,177 [main] INFO org.apache.pig.data.SchemaTupleBackend - Ke

```

```

2024-10-10 20:59:03,554 [main] INFO
2024-10-10 20:59:03,594 [main] INFO
e yarn.system-metrics-publisher.enabl
2024-10-10 20:59:03,596 [main] INFO
2024-10-10 20:59:03,610 [main] INFO
2024-10-10 20:59:04,259 [main] INFO
2024-10-10 20:59:04,260 [main] INFO
(1,Jane)
(2,John)
(3,Brian)
(4,Heka)
2024-10-10 20:59:04,620 [main] INFO
harithaah@fedora:~/exp4$
harithaah@fedora:~/exp4$
harithaah@fedora:~/exp4$

```

5. Create uppercase_udf.py

```

harise@fedora:~$ a:~/exp4$ hdfs dfs -ls /exp4
Found 3 items
-rw-r--r-- 1 harithaah supergroup 29 2024-10-10 20:52 /exp4/d1.txt
-rw-r--r-- 1 harithaah supergroup 27 2024-10-10 20:53 /exp4/sample.txt
-rw-r--r-- 1 harithaah supergroup 172 2024-10-10 20:59 /exp4/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)
```

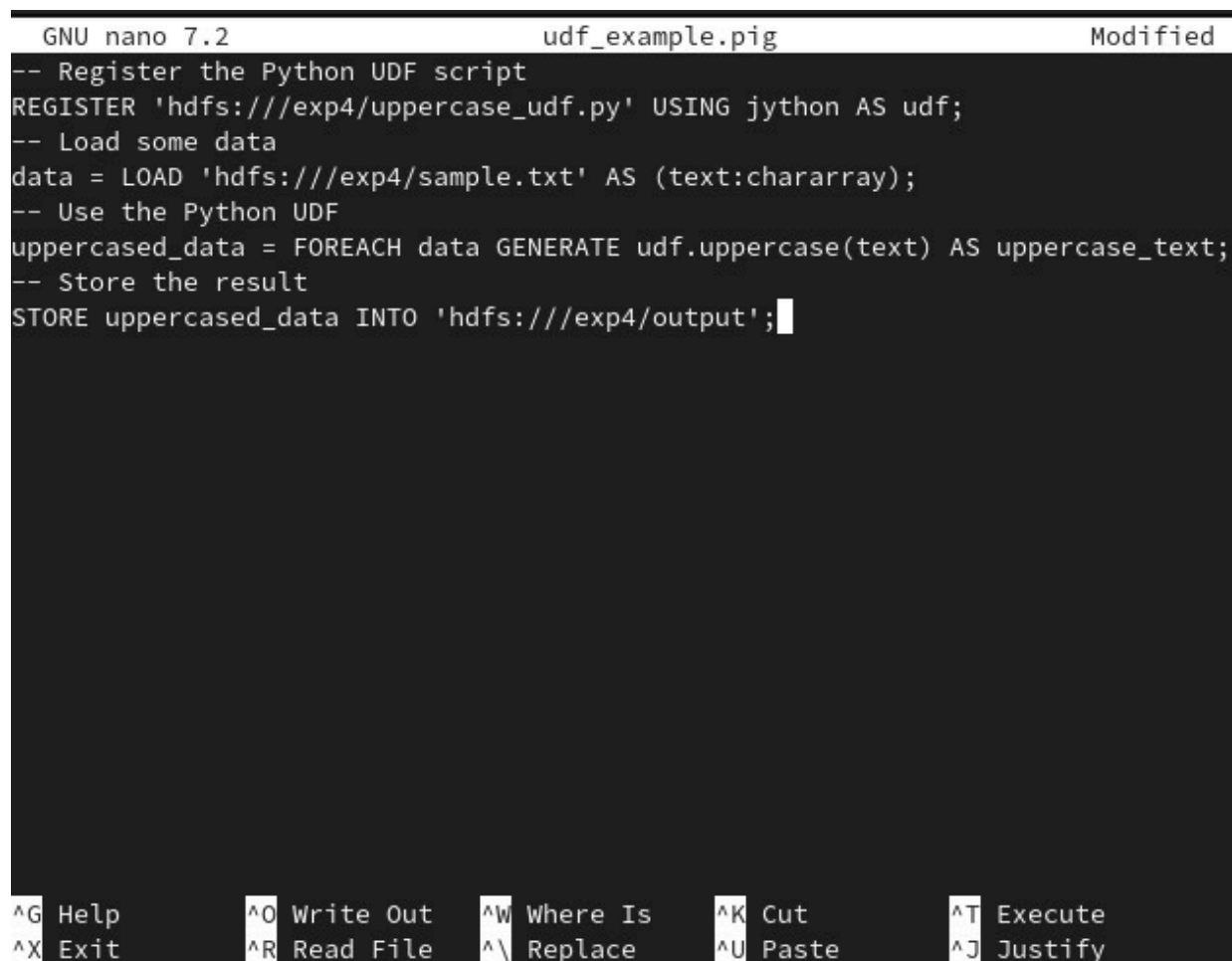
[Read 10 lines]

^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.

```
arise@fedora:~$ ./exp4$ hdfs dfs -ls /exp4
Found 3 items
-rw-r--r--  1 harithaah supergroup      29 2024-10-10 20:52 /exp4/d1.txt
-rw-r--r--  1 harithaah supergroup     27 2024-10-10 20:53 /exp4/sample.txt
-rw-r--r--  1 harithaah supergroup    172 2024-10-10 20:59 /exp4/uppercase_udf.py
```

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

```
arise@fedora:~$ cd ~/exp4$ pig udf_example.pig
2024-10-10 21:01:20,882 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-10 21:01:20,884 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-10 21:01:20,885 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-10 21:01:20,980 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0
2024-10-10 21:01:20,981 [main] INFO org.apache.pig.Main - Logging error messages to:
2024-10-10 21:01:21,555 [main] INFO org.apache.pig.impl.util.Utils - Default bootup
2024-10-10 21:01:21,626 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
2024-10-10 21:01:21,626 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
2024-10-10 21:01:21,626 [main] INFO org.apache.pig.backend.hadoop.executionengine.HE
2024-10-10 21:01:22,461 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
2024-10-10 21:01:22,482 [main] INFO org.apache.pig.PigServer - Pig Script ID for the
2024-10-10 21:01:22,483 [main] WARN org.apache.pig.PigServer - ATS is disabled since
2024-10-10 21:01:22,562 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
```

Output :

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