

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

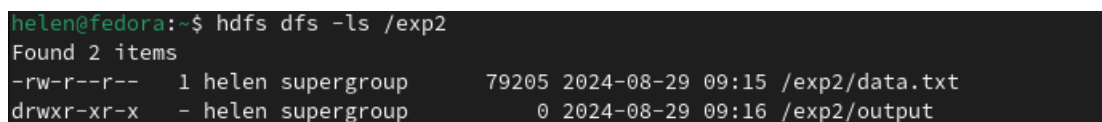
1. Create sample.txt



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

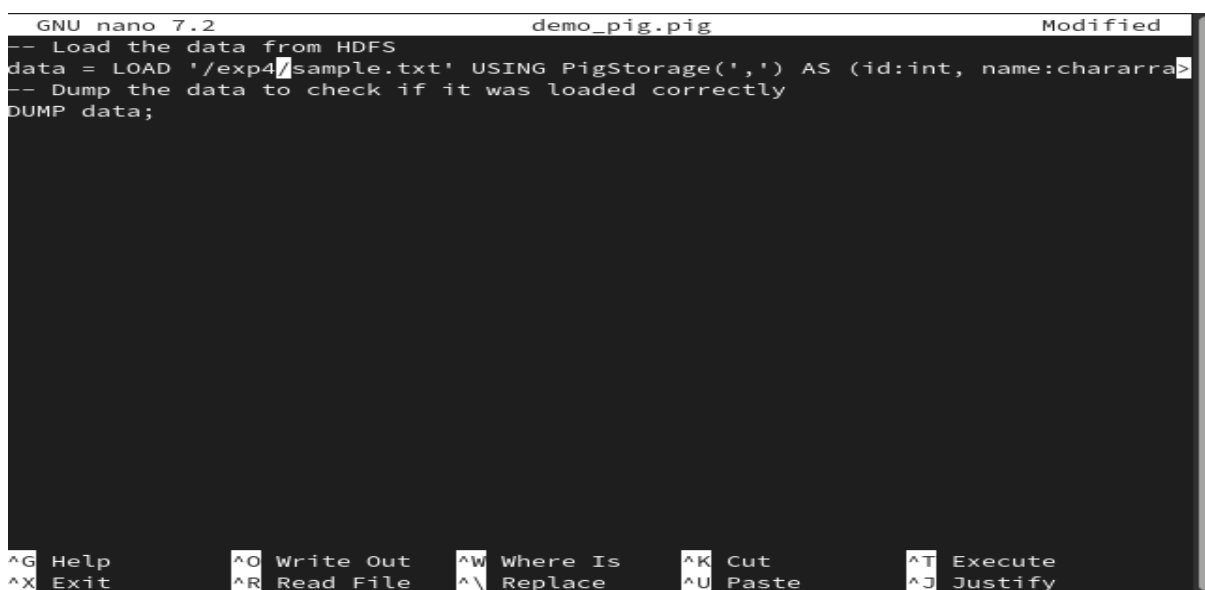
[ Read 4 lines ]
^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```

2. Upload sample.txt file to HDFS Storage.



```
helen@fedora:~$ hdfs dfs -ls /exp2
Found 2 items
-rw-r--r--  1 helen supergroup      79205 2024-08-29 09:15 /exp2/data.txt
drwxr-xr-x  - helen supergroup         0 2024-08-29 09:16 /exp2/output
```

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;
```

4. Execute demo_pig.pig

```

helen@fedora:~/exp3
helen@fedora:~$ cd exp3
helen@fedora:~/exp3$ pig demo.pig
2024-10-12 06:52:28,345 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 06:52:28,347 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 06:52:28,347 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-12 06:52:28,418 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0
(r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 06:52:28,419 [main] INFO org.apache.pig.Main - Logging error messages to:
/home/helen/exp3/pig_1728730348368.log
2024-10-12 06:52:28,747 [main] INFO org.apache.pig.impl.util.Utils - Default bootup
file /home/helen/.pigbootup not found
2024-10-12 06:52:28,809 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-12 06:52:28,809 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-12 06:52:28,809 [main] INFO org.apache.pig.backend.hadoop.executionengine.HE
xecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-10-12 06:52:29,189 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-12 06:52:29,213 [main] INFO org.apache.pig.PigServer - Pig Script ID for the
session: PIG-demo.pig-f04b0f94-310a-4be3-8b35-444348bc3cc2
2024-10-12 06:52:29,213 [main] WARN org.apache.pig.PigServer - ATS is disabled since
yarn.timeline-service.enabled set to false
2024-10-12 06:52:29,632 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-12 06:55:33,246 [main] INFO org.apache.pig.backend.hadoop.executionengine.ma
pReduceLayer.MapReduceLauncher - Success!
2024-10-12 06:55:33,273 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-12 06:55:33,277 [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 06:55:33,295 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pi
g.schematuple] was not set... will not generate code.
2024-10-12 06:55:33,369 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputF
ormat - Total input files to process : 1
2024-10-12 06:55:33,372 [main] INFO org.apache.pig.backend.hadoop.executionengine.ut
il.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-10-12 06:55:33,562 [main] INFO org.apache.pig.Main - Pig script completed in 3
minutes, 5 seconds and 241 milliseconds (185241 ms)
helen@fedora:~/exp3$

```

5. 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]

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

6. 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

7. Execute udf_example.pig

```
helen@fedora:~/exp3$ pig udf_example.pig
2024-10-12 07:11:56,056 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-12 07:11:56,060 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-10-12 07:11:56,060 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-10-12 07:11:56,149 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0
(r1746530) compiled Jun 01 2016, 23:10:49
2024-10-12 07:11:56,153 [main] INFO org.apache.pig.Main - Logging error messages to:
/home/helen/exp3/pig_1728731516078.log
2024-10-12 07:11:56,601 [main] ERROR org.apache.pig.Main - ERROR 2997: Encountered IO
Exception. File udf_example.pig does not exist
Details at logfile: /home/helen/exp3/pig_1728731516078.log
2024-10-12 07:11:56,704 [main] INFO org.apache.pig.Main - Pig script completed in 60
1 milliseconds (601 ms)
```

Output :

```
1,JOHN
2,JANE
3,JOE
4,EMMA
```