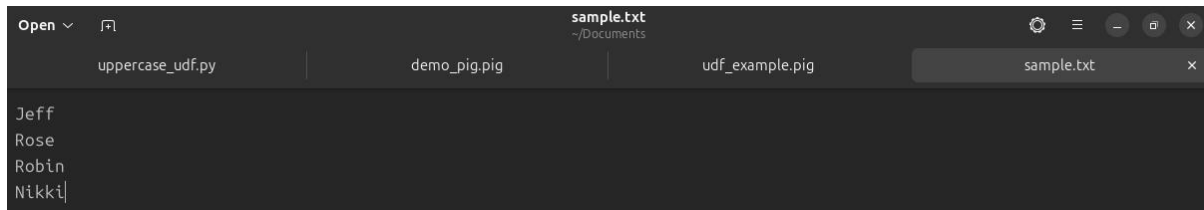


210701120

## User Defined Function(UDF) in PIG

Input(sample.txt):



```
sample.txt
~/Documents

uppercase_udf.py  demo_pig.pig  udf_example.pig  sample.txt x

Jeff
Rose
Robin
Nikki
```

demo\_pig.pig:

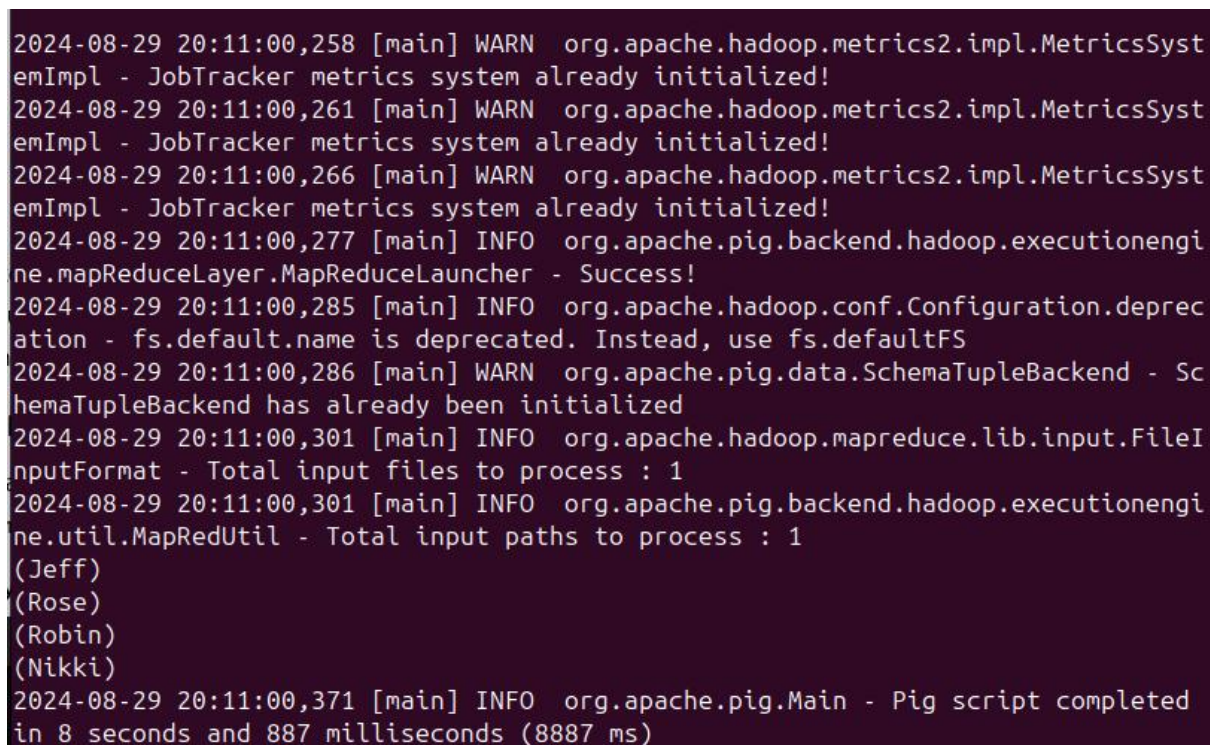


```
demo_pig.pig
~/Documents

uppercase_udf.py  demo_pig.pig x  udf_example.pig

data = LOAD '/piginput/sample.txt' USING PigStorage(',') AS (id:chararray);
DUMP data;
```

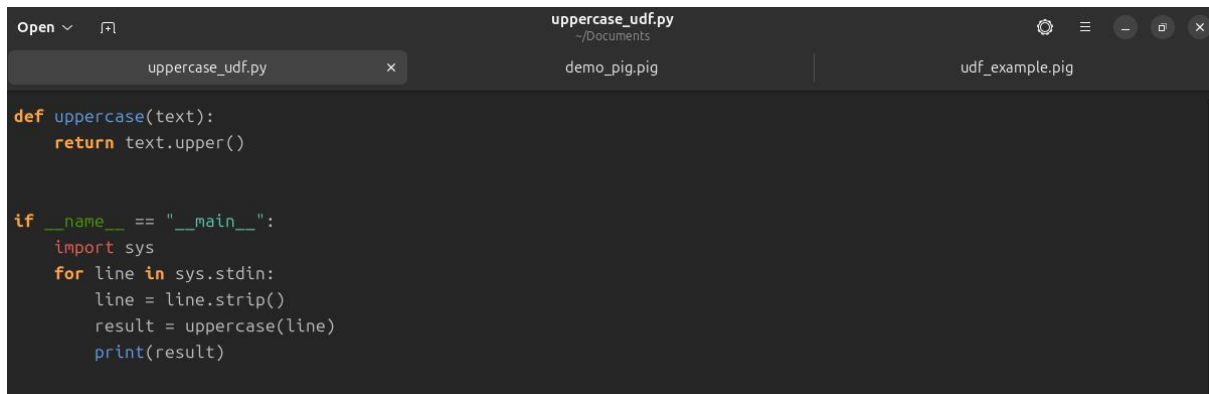
Output for demo\_pig.pig:



```
2024-08-29 20:11:00,258 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-08-29 20:11:00,261 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-08-29 20:11:00,266 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-08-29 20:11:00,277 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-08-29 20:11:00,285 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-08-29 20:11:00,286 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-08-29 20:11:00,301 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-08-29 20:11:00,301 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Jeff)
(Rose)
(Robin)
(Nikki)
2024-08-29 20:11:00,371 [main] INFO org.apache.pig.Main - Pig script completed in 8 seconds and 887 milliseconds (8887 ms)
```

uppercase\_udf.py:

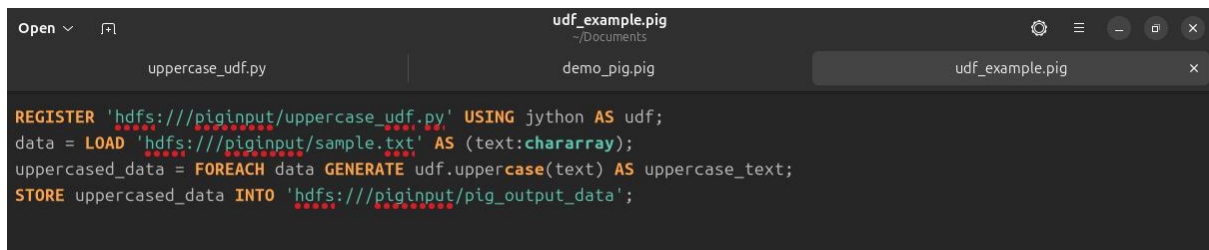
210701120



The screenshot shows a code editor with three tabs: 'uppercase\_udf.py', 'demo\_pig.pig', and 'udf\_example.pig'. The 'uppercase\_udf.py' tab is active, displaying the following Python code:

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

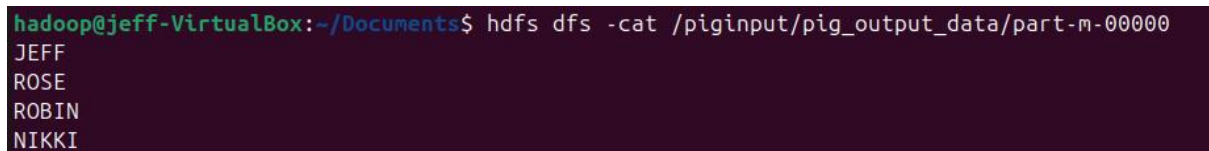
### udf\_example.pig:



The screenshot shows a code editor with three tabs: 'uppercase\_udf.py', 'demo\_pig.pig', and 'udf\_example.pig'. The 'udf\_example.pig' tab is active, displaying the following Pig Latin script:

```
REGISTER 'hdfs:///piginput/uppercase_udf.py' USING jython AS udf;  
data = LOAD 'hdfs:///piginput/sample.txt' AS (text:chararray);  
uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;  
STORE uppercased_data INTO 'hdfs:///piginput/pig_output_data';
```

### Output for udf\_example.pig:



The screenshot shows a terminal window with the following command and output:

```
hadoop@jeff-VirtualBox:~/Documents$ hdfs dfs -cat /piginput/pig_output_data/part-m-00000  
JEFF  
ROSE  
ROBIN  
NIKKI
```