EX.NO:4 **210701165**

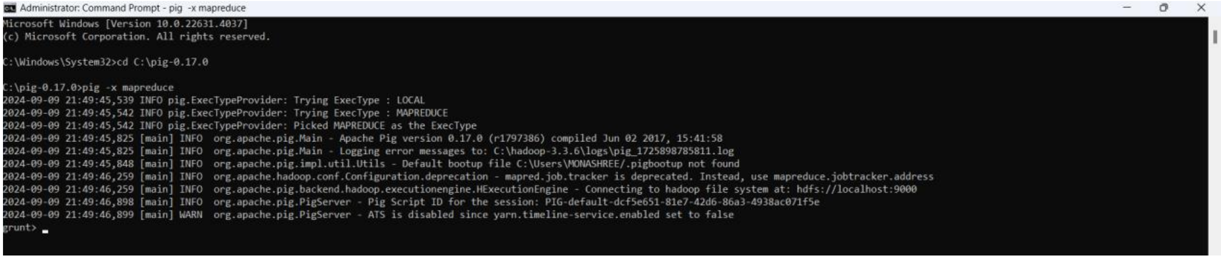
Create UDF (User Defined Functions) in Apache Pig and execute it inMapReduce/HDFS mode

AIM:

To create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode.

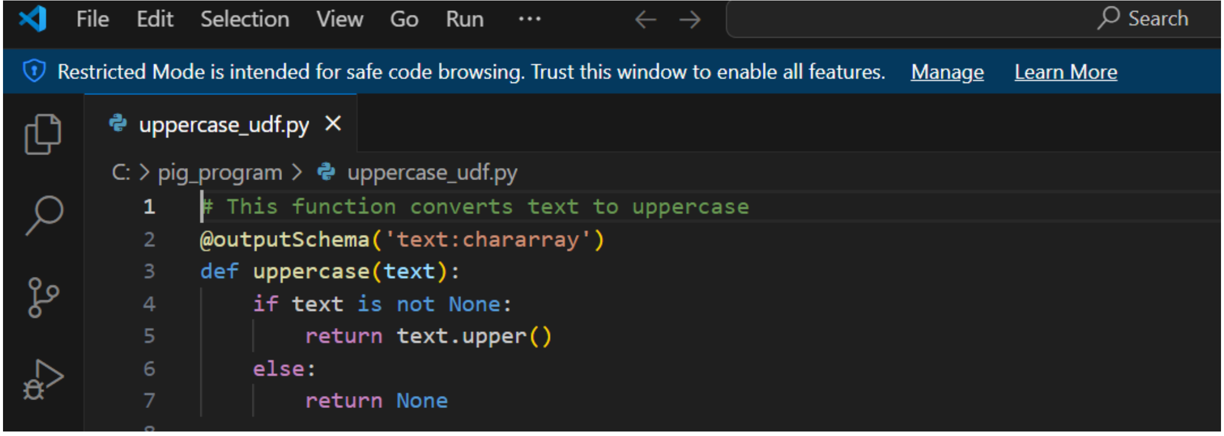
PROCEDURE:

1. Ensure that Apache Pig is installed and configured.



```
Microsoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd C:\pig-0.17.0

C:\pig-0.17.0>pig -x mapreduce
2024-09-09 21:49:45,539 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-09 21:49:45,542 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-09 21:49:45,542 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-09 21:49:45,825 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-09-09 21:49:45,825 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.6\logs\pig_1725898785811.log
2024-09-09 21:49:45,848 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\MONASHREE/.pigbootup not found
2024-09-09 21:49:46,259 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-09 21:49:46,259 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-09 21:49:46,898 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-dcf5e651-81e7-42d6-86a3-4938ac071f5e
2024-09-09 21:49:46,899 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

2. Create a python UDF (User Defined Functions).

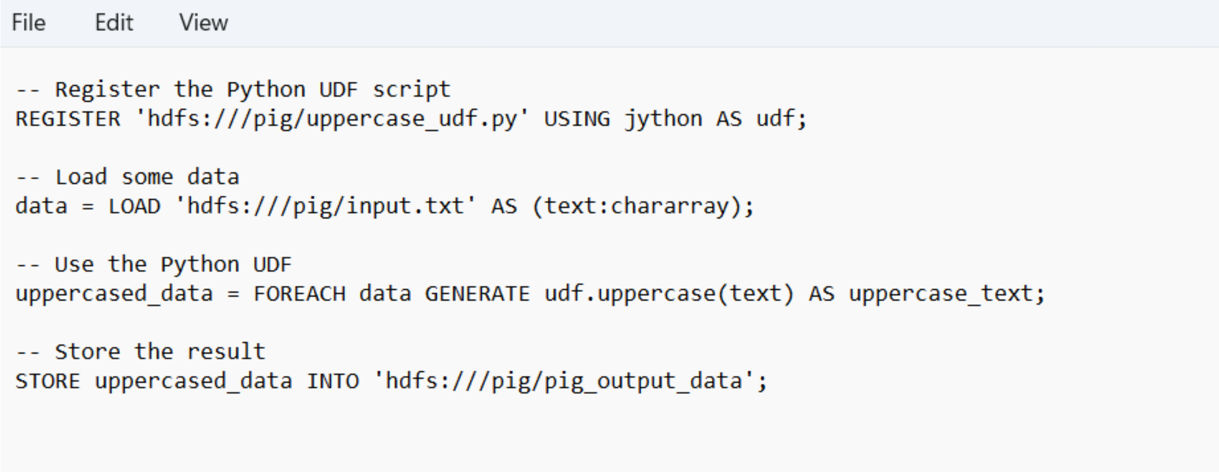


Restricted Mode is intended for safe code browsing. Trust this window to enable all features.

```python
# This function converts text to uppercase
@outputSchema('text:chararray')
def uppercase(text):
    if text is not None:
        return text.upper()
    else:
        return None
```

3. Jython should be installed as Pig will use it to interpret the Python UDFs.

4. Create a Pig script that registers and uses the Python UDF.

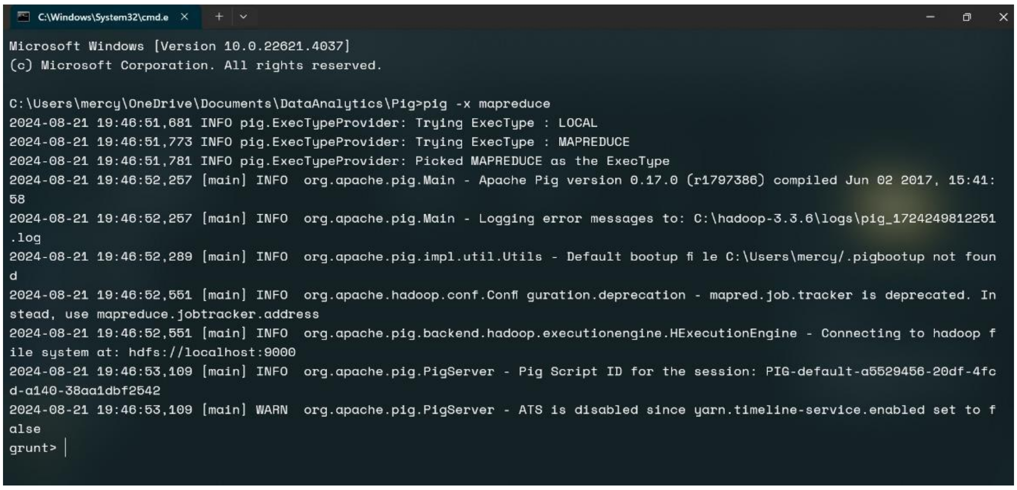


```
-- Register the Python UDF script
REGISTER 'hdfs:///pig/uppercase_udf.py' USING jython AS udf;

-- Load some data
data = LOAD 'hdfs:///pig/input.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:///pig/pig_output_data';
```

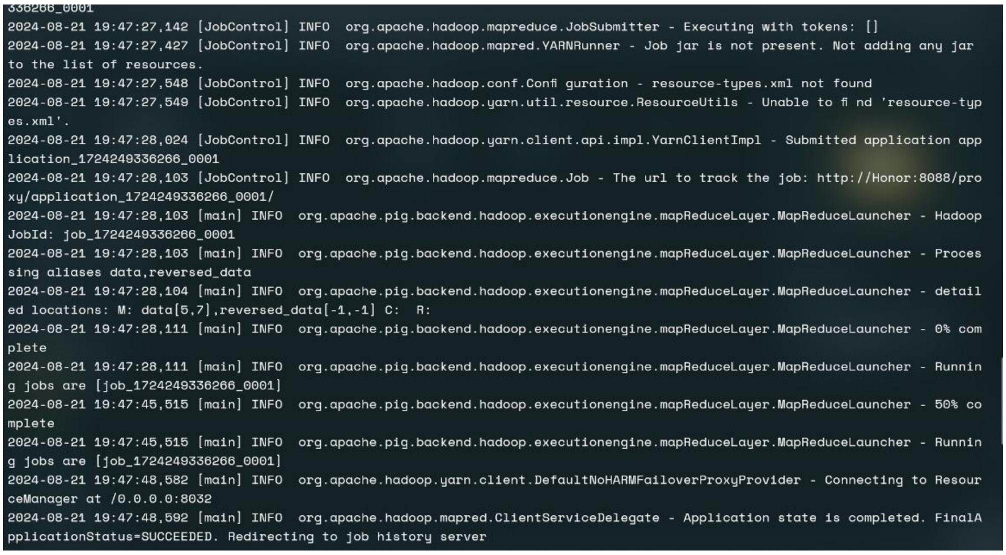
5. Execute the Pig Script in MapReduce Mode using the command:

```
pig -x mapreduce script.pig
```

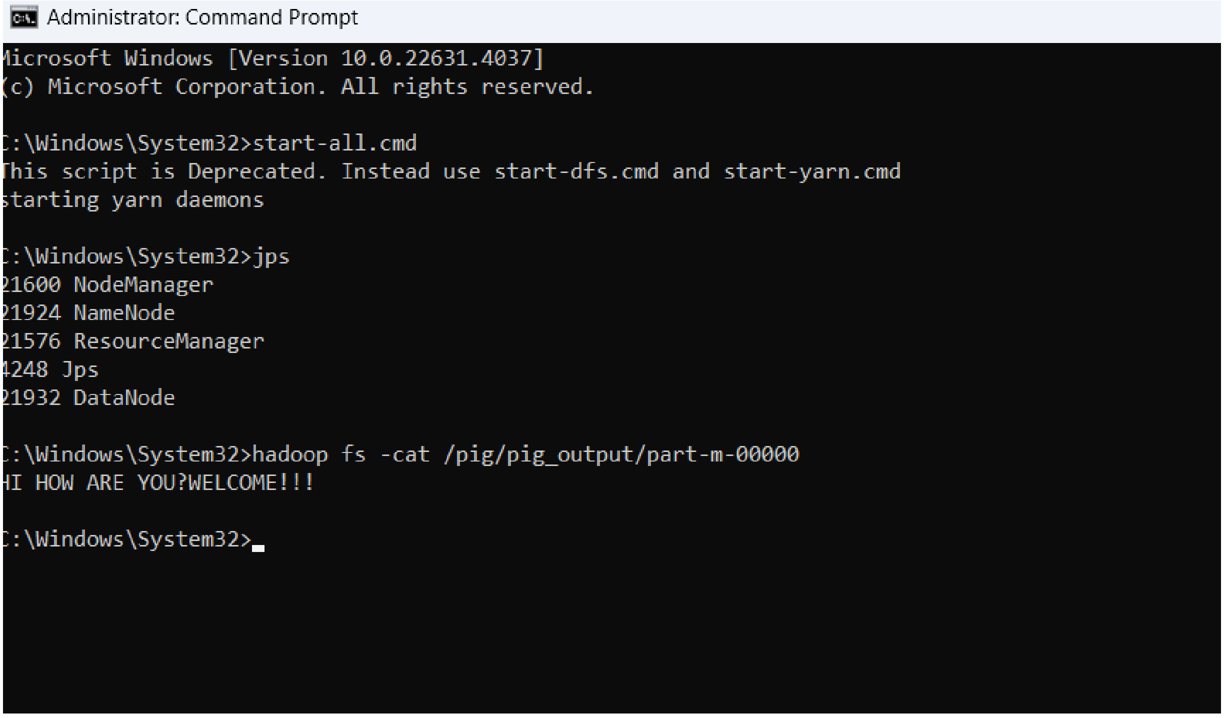
OUTPUT:



```
Microsoft Windows [Version 10.0.22621.4037]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mercy\OneDrive\Documents\DataAnalytics\Pig>pig -x mapreduce
2024-08-21 19:46:51,681 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-21 19:46:51,773 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-21 19:46:51,781 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-21 19:46:52,257 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:
58
2024-08-21 19:46:52,257 [main] INFO  org.apache.pig.Main - Logging error messages to: C:\hadoop-3.3.6\logs\pig_1724249812251
.log
2024-08-21 19:46:52,289 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file C:\Users\mercy/.pigbootup not foun
d
2024-08-21 19:46:52,551 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. In
stead, use mapreduce.jobtracker.address
2024-08-21 19:46:52,551 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
ile system at: hdfs://localhost:9000
2024-08-21 19:46:53,109 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-a5529456-20df-4fc
d-a140-38aa1dbf2542
2024-08-21 19:46:53,109 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to f
alse
grunt> |
```

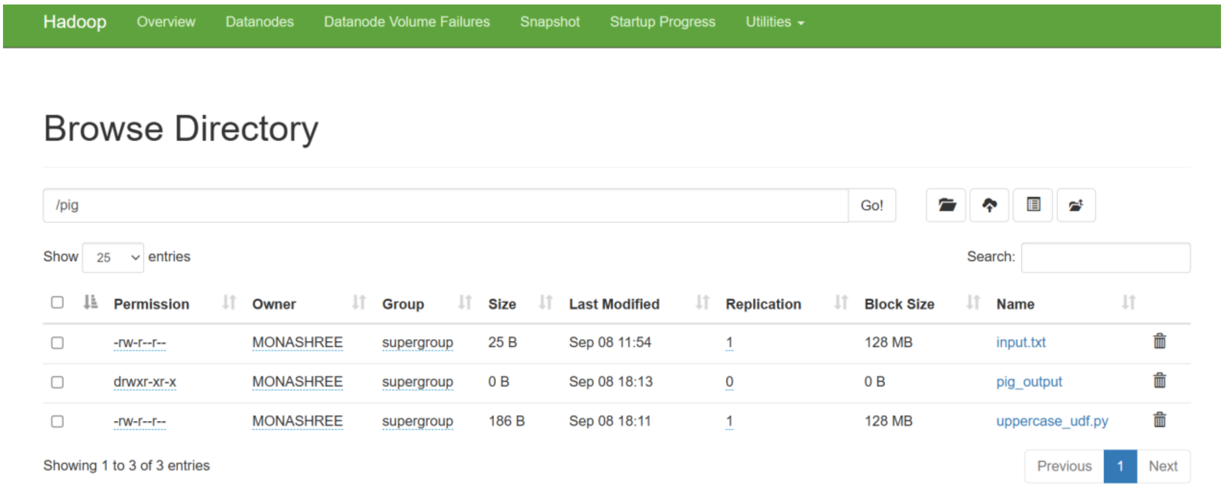


```
336266_0001
2024-08-21 19:47:27,142 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2024-08-21 19:47:27,427 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar
to the list of resources.
2024-08-21 19:47:27,548 [JobControl] INFO  org.apache.hadoop.conf.Configuration - resource-types.xml not found
2024-08-21 19:47:27,549 [JobControl] INFO  org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-typ
es.xml'.
2024-08-21 19:47:28,024 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application app
lication_1724249336266_0001
2024-08-21 19:47:28,103 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://Honor:8088/pro
xy/application_1724249336266_0001/
2024-08-21 19:47:28,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Hadoop
JobId: job_1724249336266_0001
2024-08-21 19:47:28,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Proces
sing aliases data,reversed_data
2024-08-21 19:47:28,104 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detail
ed locations: M: data[5,7],reversed_data[-1,-1] C:  R:
2024-08-21 19:47:28,111 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% com
plete
2024-08-21 19:47:28,111 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Runnin
g jobs are [job_1724249336266_0001]
2024-08-21 19:47:45,515 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% co
mplete
2024-08-21 19:47:45,515 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Runnin
g jobs are [job_1724249336266_0001]
2024-08-21 19:47:48,582 [main] INFO  org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to Resour
ceManager at /0.0.0.0:8032
2024-08-21 19:47:48,592 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalA
pplicationStatus=SUCCEEDED. Redirecting to job history server
```



Administrator: Command Prompt

```
licrosoft Windows [Version 10.0.22631.4037]
(c) Microsoft Corporation. All rights reserved.

:\Windows\System32>start-all.cmd
his script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
tarting yarn daemons

:\Windows\System32>jps
21600 NodeManager
21924 NameNode
21576 ResourceManager
1248 Jps
21932 DataNode

:\Windows\System32>hadoop fs -cat /pig/pig_output/part-m-00000
HI HOW ARE YOU?WELCOME!!!

:\Windows\System32>_
```



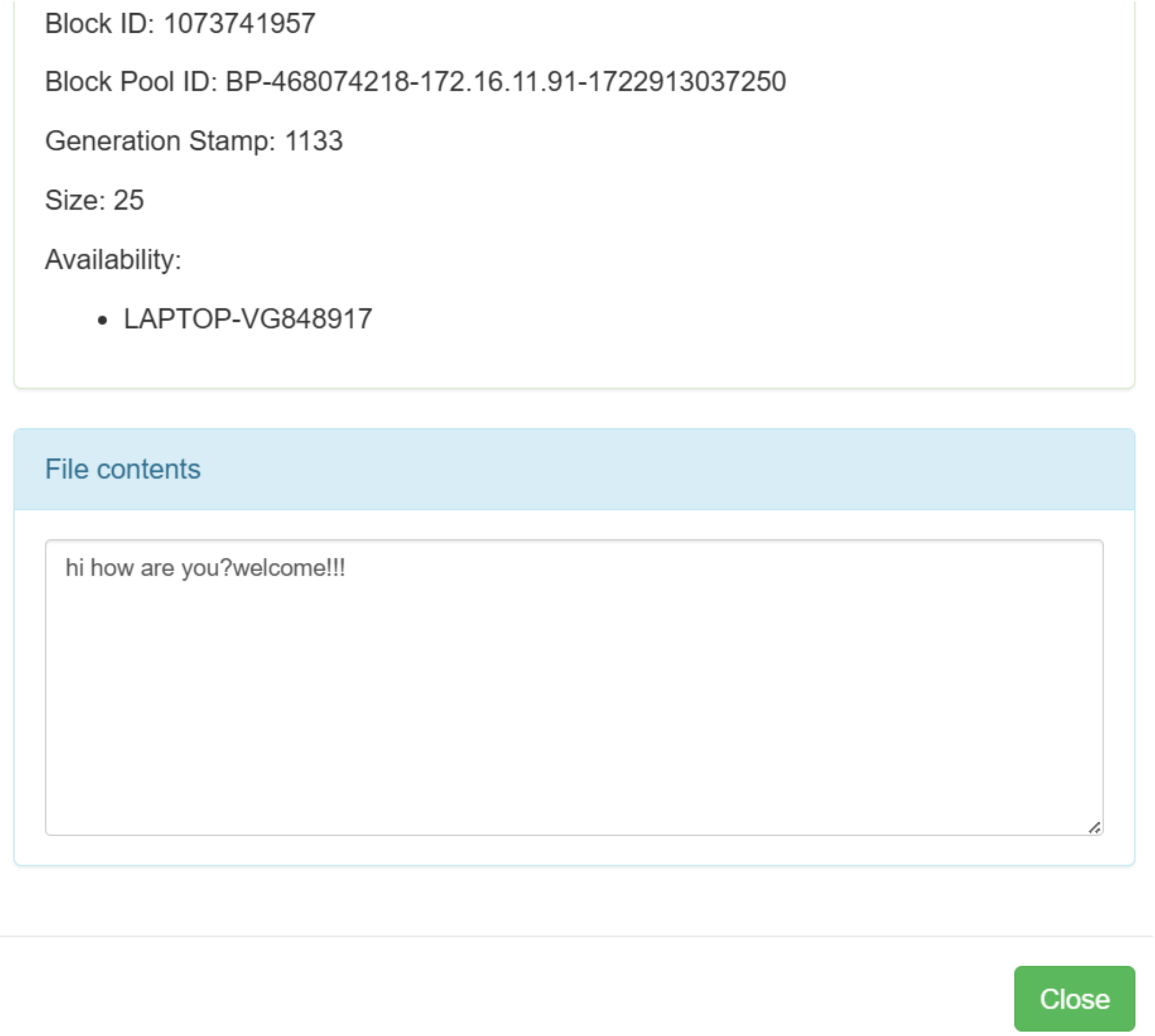
Browse Directory

/pig Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	MONASHREE	supergroup	25 B	Sep 08 11:54	1	128 MB	input.txt	🗑
□	drwxr-xr-x	MONASHREE	supergroup	0 B	Sep 08 18:13	0	0 B	pig_output	🗑
□	-rw-r--r--	MONASHREE	supergroup	186 B	Sep 08 18:11	1	128 MB	uppercase_udf.py	🗑

Showing 1 to 3 of 3 entries



Block ID: 1073741957

Block Pool ID: BP-468074218-172.16.11.91-1722913037250

Generation Stamp: 1133

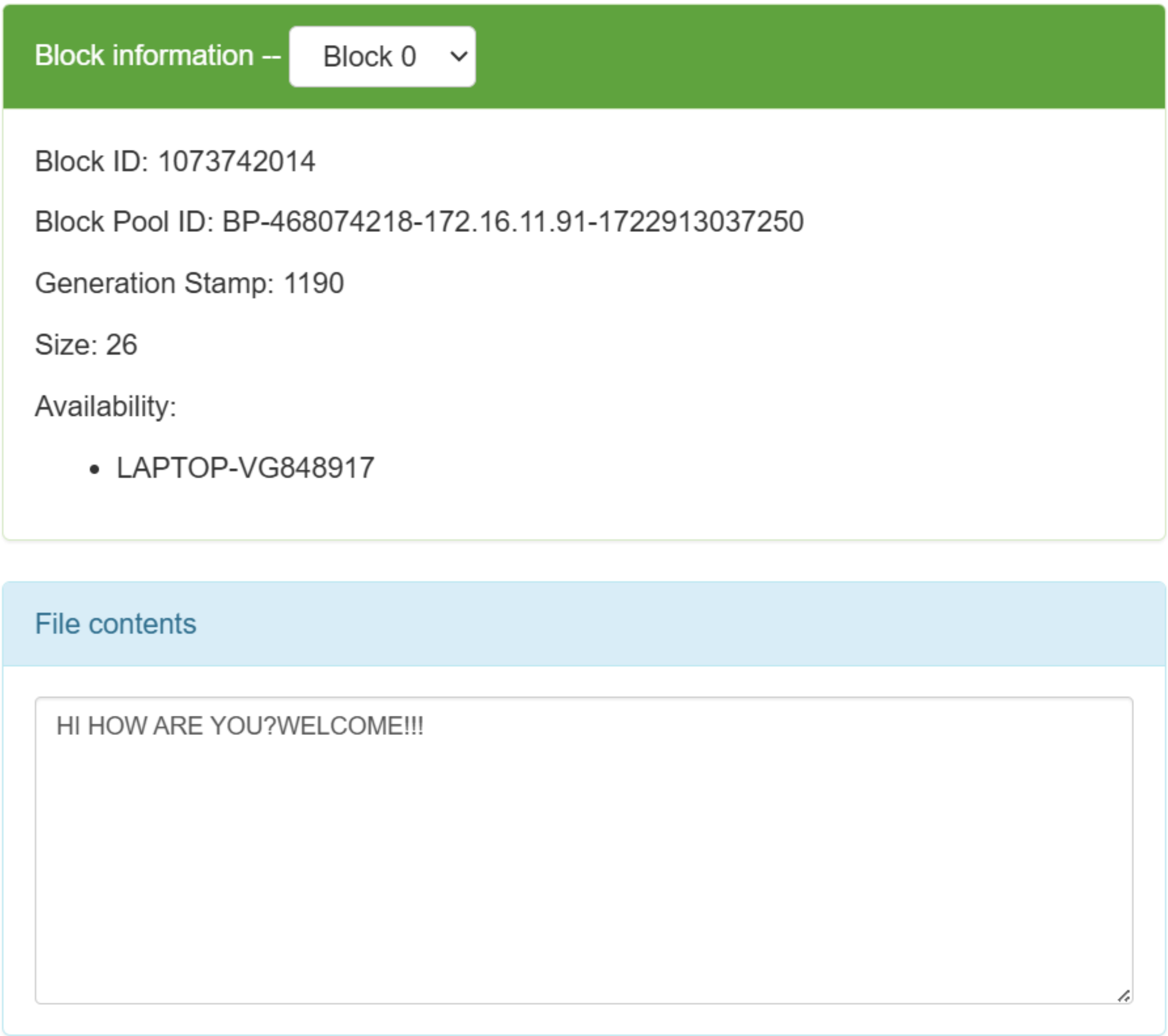
Size: 25

Availability:

- LAPTOP-VG848917

File contents

```
hi how are you?welcome!!!
```



Block information -- Block 0

Block ID: 1073742014

Block Pool ID: BP-468074218-172.16.11.91-1722913037250

Generation Stamp: 1190

Size: 26

Availability:

- LAPTOP-VG848917

File contents

```
HI HOW ARE YOU?WELCOME!!!
```

RESULT:

Thus, to create a UDF in Apache Pig and execute in MapReduce mode has been executed successfully