Ex No: 4 Roll No: 210701173

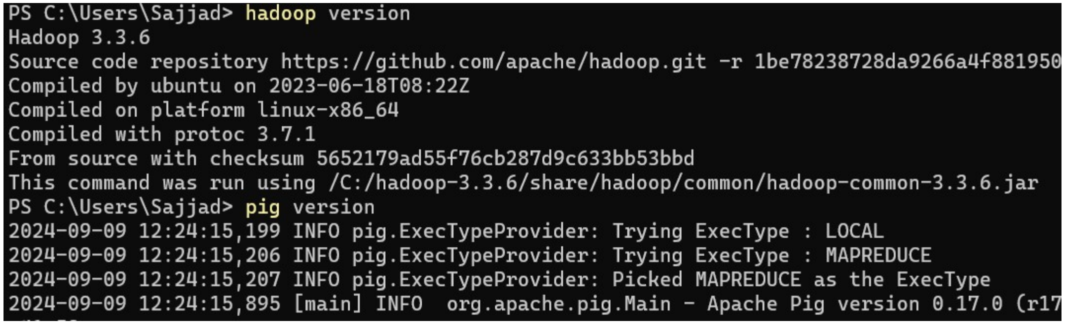
Create UDF (User Defined Functions) in Apache Pig and execute it in MapReduce/HDFS mode

AIM:

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

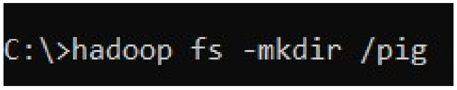
PROCEDURE:

1. Verify that Hadoop and Pig and installed successfully.



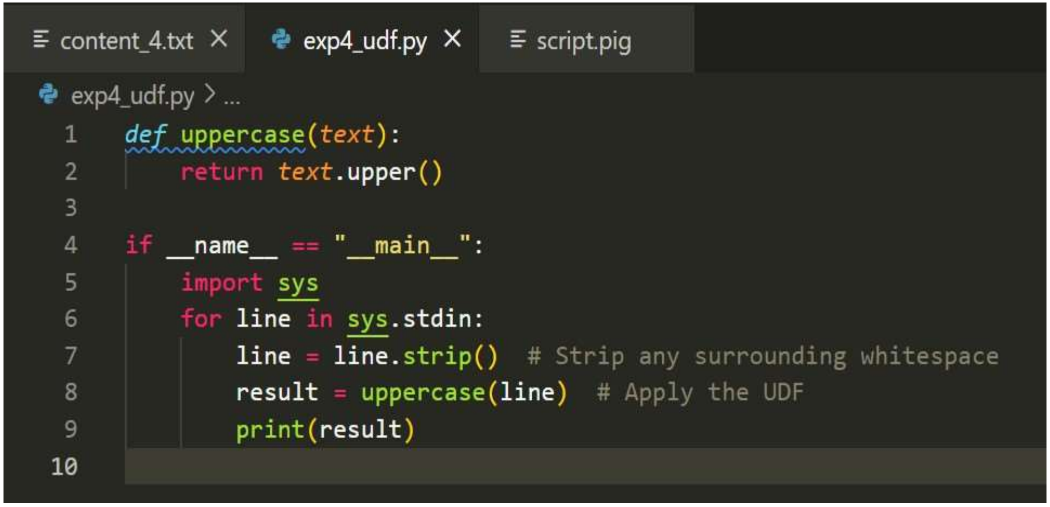
```
PS C:\Users\Sajjad> hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f881950
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /C:/hadoop-3.3.6/share/hadoop/common/hadoop-common-3.3.6.jar
PS C:\Users\Sajjad> pig version
2024-09-09 12:24:15,199 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-09 12:24:15,206 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-09 12:24:15,207 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-09 12:24:15,895 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r17
```

2. Start Hadoop using start-all.cmd command. This will enable HDFS.
3. Create a directory called pig in HDFS using the **-mkdir** command



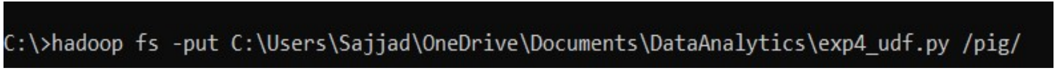
```
C:\>hadoop fs -mkdir /pig
```

4. Create a python file containing the function to perform uppercase.



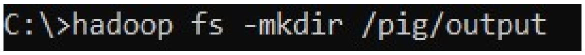
```python
def uppercase(text):
    return text.upper()

if __name__ == "__main__":
    import sys
    for line in sys.stdin:
        line = line.strip()  # Strip any surrounding whitespace
        result = uppercase(line)  # Apply the UDF
        print(result)

```

5. Upload the python file to HDFS using the **-put** command.



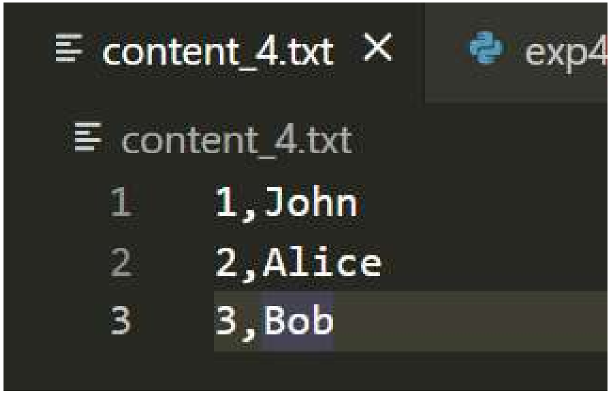
```
C:\>hadoop fs -put C:\Users\Sajjad\OneDrive\Documents\DataAnalytics\exp4_udf.py /pig/
```

6. Create a separate directory for output inside pig where the output will be stored.

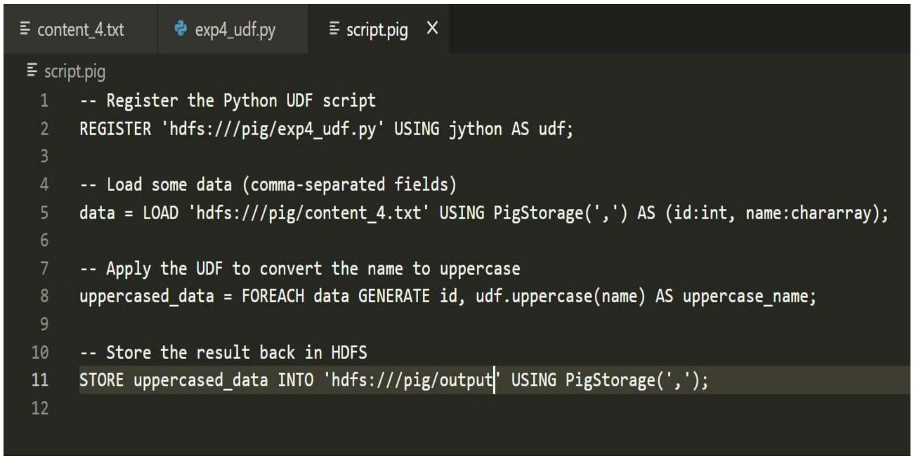


```
C:\>hadoop fs -mkdir /pig/output
```

7. Create a text file and a script.pig file that contains the required .pig commands.

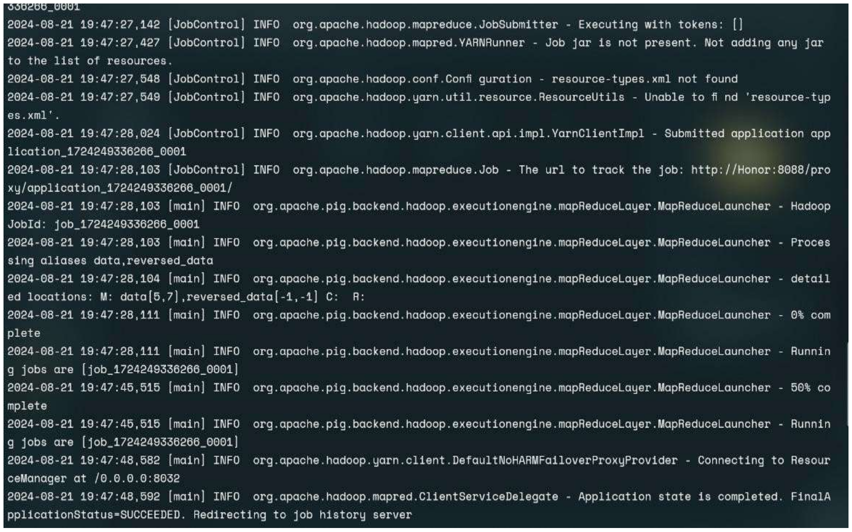


```
1,John
2,Alice
3,Bob
```



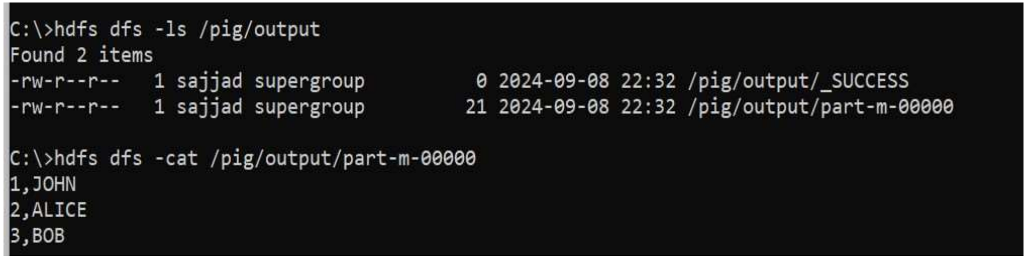
```
-- Register the Python UDF script
REGISTER 'hdfs:///pig/exp4_udf.py' USING jython AS udf;

-- Load some data (comma-separated fields)
data = LOAD 'hdfs:///pig/content_4.txt' USING PigStorage(',') AS (id:int, name:chararray);

-- Apply the UDF to convert the name to uppercase
uppercased_data = FOREACH data GENERATE id, udf.uppercase(name) AS uppercase_name;

-- Store the result back in HDFS
STORE uppercased_data INTO 'hdfs:///pig/output' USING PigStorage(',');
```

8. Upload the text file to HDFS using the **-put** command.
9. Execute the Pig Script in MapReduce mode using the command **pig -x mapreduce script.pig.**

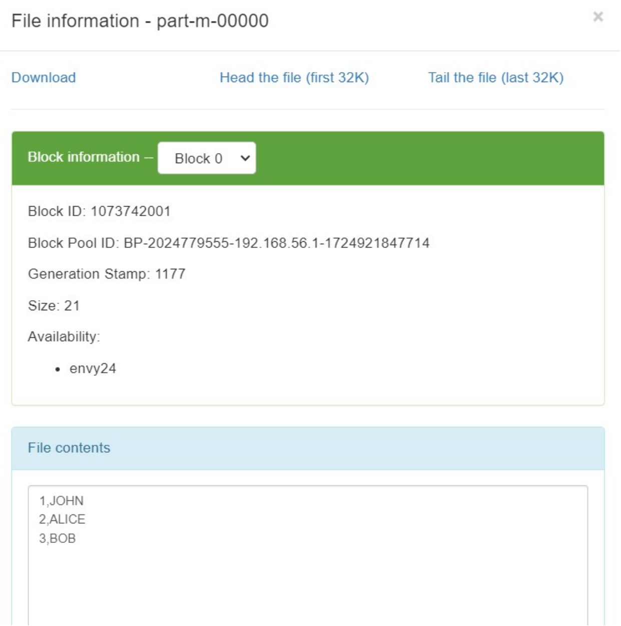


```
336266_0001
2024-08-21 19:47:27,142 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2024-08-21 19:47:27,427 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is not present. Not adding any jar
to the list of resources.
2024-08-21 19:47:27,548 [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
```

10. To check the output, use the -cat command



```
C:\>hdfs dfs -ls /pig/output
Found 2 items
-rw-r--r--   1 sajjad supergroup          0 2024-09-08 22:32 /pig/output/_SUCCESS
-rw-r--r--   1 sajjad supergroup         21 2024-09-08 22:32 /pig/output/part-m-00000

C:\>hdfs dfs -cat /pig/output/part-m-00000
1,JOHN
2,ALICE
3,BOB
```



File information - part-m-00000

Block information -- Block 0

Block ID: 1073742001

Block Pool ID: BP-2024779555-192.168.56.1-1724921847714

Generation Stamp: 1177

Size: 21

Availability:

- envy24

```
1,JOHN
2,ALICE
3,BOB
```

Result:

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