210701141
K.MADHUMITHA

EXP 4: Create UDF (User Defined Functions) in Apache Pigand execute it in MapReduce / HDFS mode

AIM:

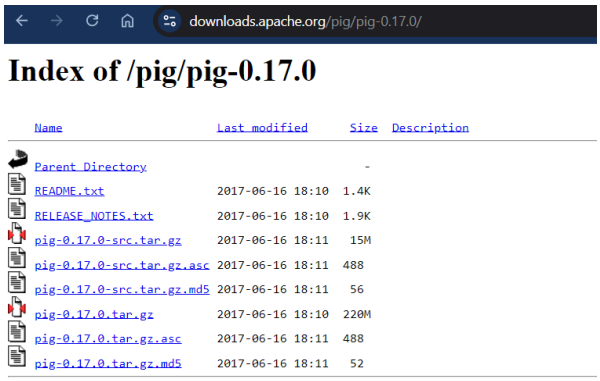
To create UDF in Apache Pig and execute it in MapReduce/HDFS mode.

PROCEDURE:

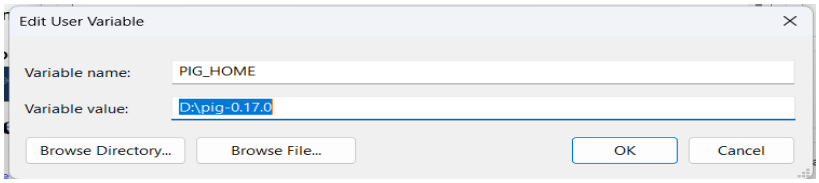
Pig Download and installation:

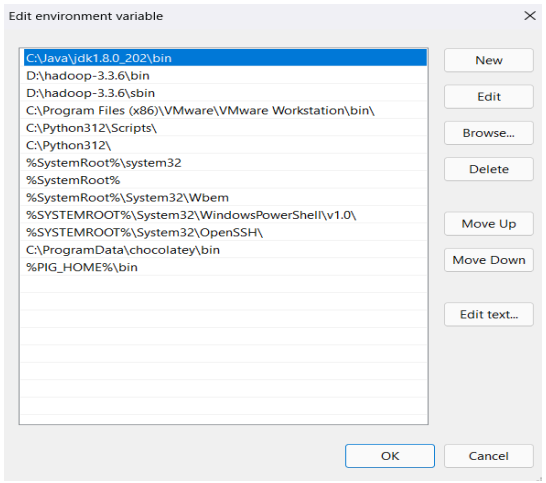
1. Download Pig:

Download Pig from "https://downloads.apache.org/pig/pig-0.17.0/"



2. Add the environment variable for Pig:



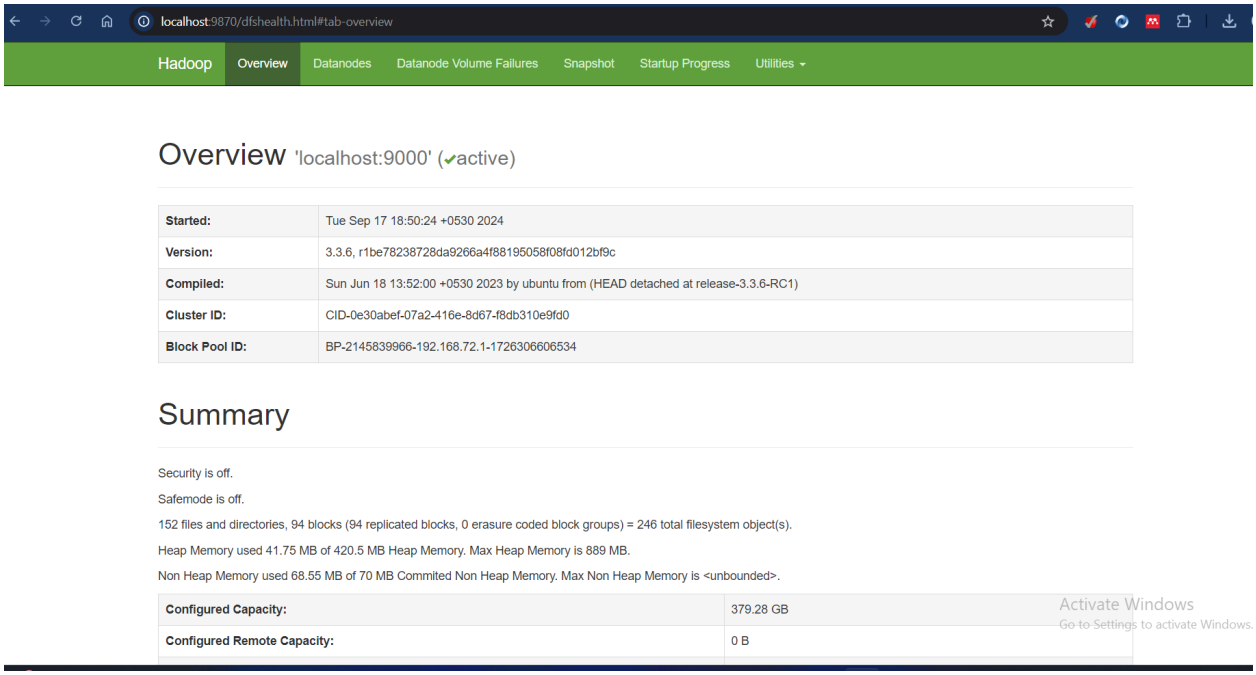


3. Go to D:\pig-0.17.0\bin and open pig (Windows Command Script)
4. Open Windows Powershell and type "pig –x local" and check whether pig grunt appears.

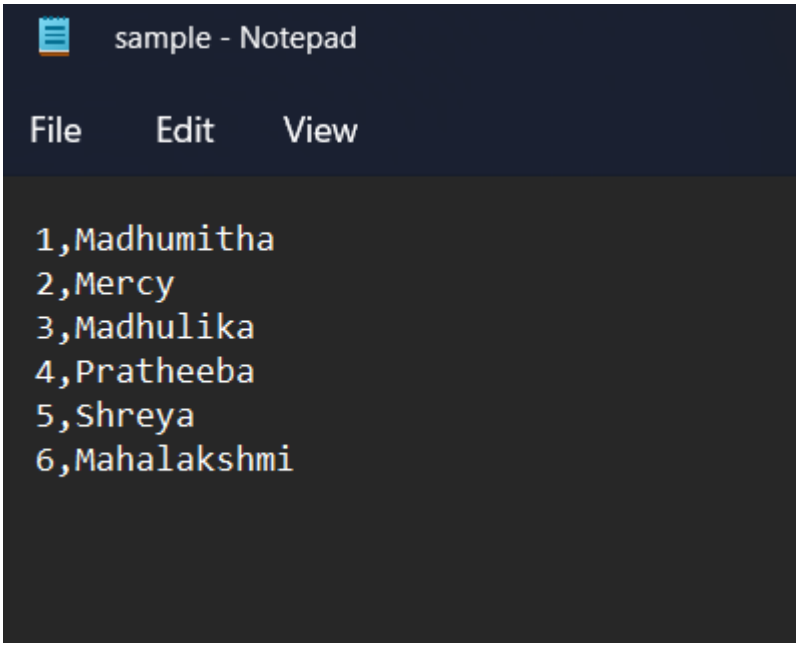
Pig is successfully installed.

Create UDF:

1. Start Hadoop services: Open command prompt as an administrator
   start-dfs.cmd
   start-yarn.cmd
2. Open the browser and go to the URL “localhost:9870”



3. Create a text file “sample.txt”:



4. Create a Directory in HDFS and copy the Input File to HDFS

```
C:\Windows\System32>hdfs dfs -mkdir /UDF

C:\Windows\System32>D:

D:\>cd madhumitha

D:\madhumitha>cd DA

D:\madhumitha\DA>hdfs dfs -put sample.txt /UDF
```

5. Create a Python file “uppercase_udf.py”:

```
# 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)
```

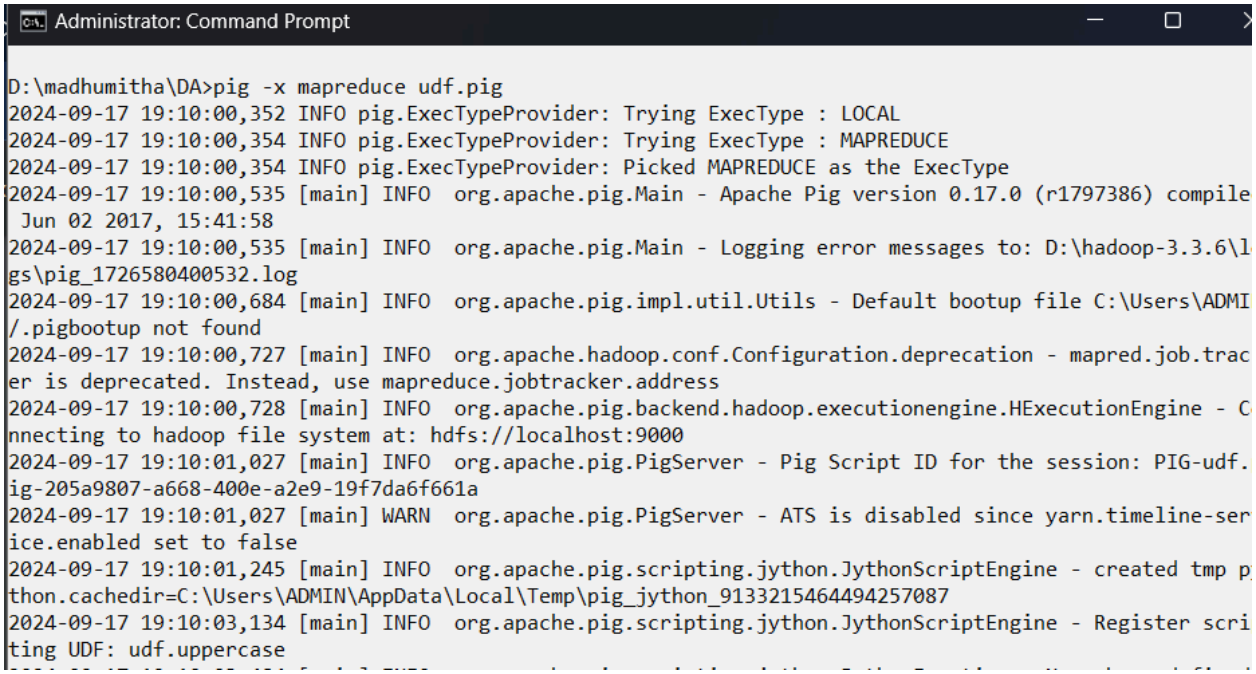
6. Create a Directory in HDFS and copy the Input File to HDFS

D:\madhumitha\DA>hdfs dfs -put uppercase_udf.py /UDF

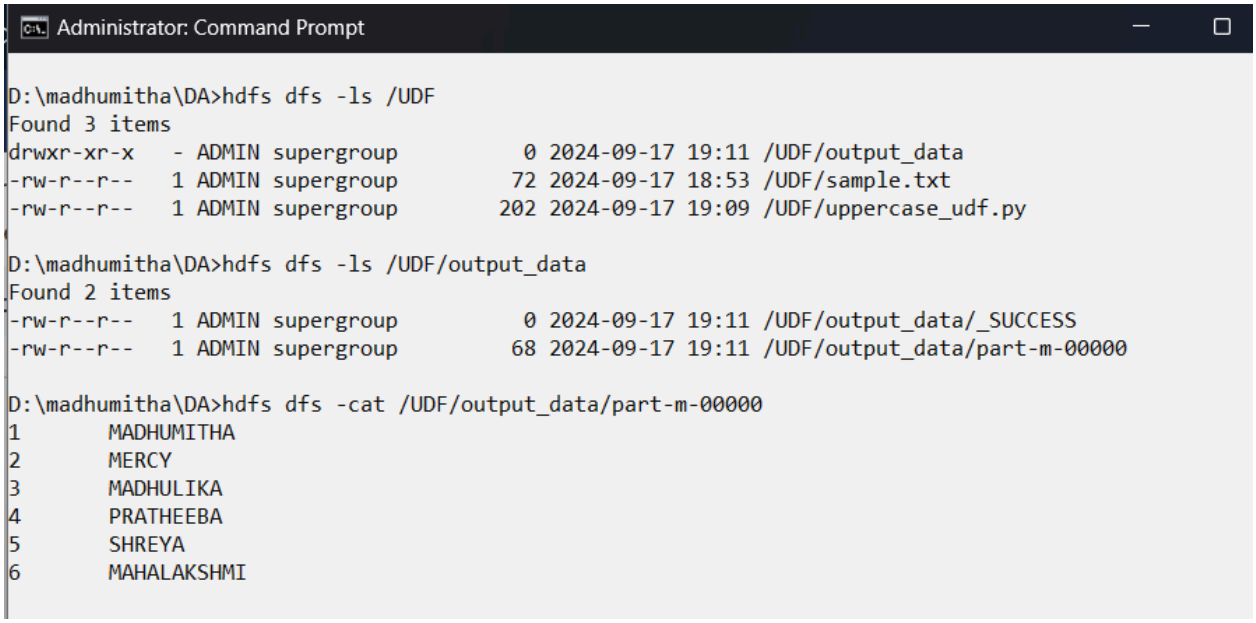
7. Create pig file “UDF.pig”:

```
-- Register the UDF
REGISTER 'hdfs://localhost:9000/UDF/uppercase_udf.py' USING jython AS udf;

-- Load the data
data = LOAD 'hdfs://localhost:9000/UDF/sample.txt'
      USING PigStorage(',')
      AS (id:int, name:chararray);

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

-- Store the result
STORE uppercased_data INTO 'hdfs://localhost:9000/UDF/output_data';
```

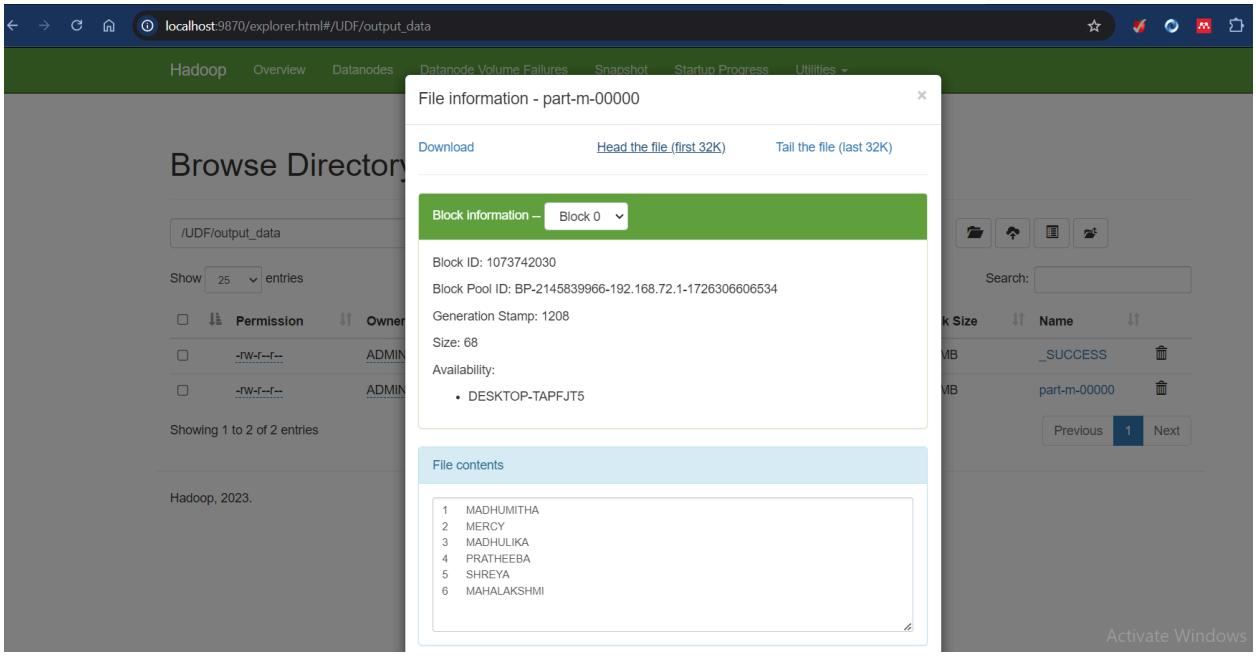
8. Execute Pig file



9. View the Output



10. Once the map reduce operations are performed successfully, the output will be present in the specified directory.



RESULT:

Thus, UDF in Apache Pig has been created and executed in MapReduce/HDFS mode successfully.