

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

AIM:

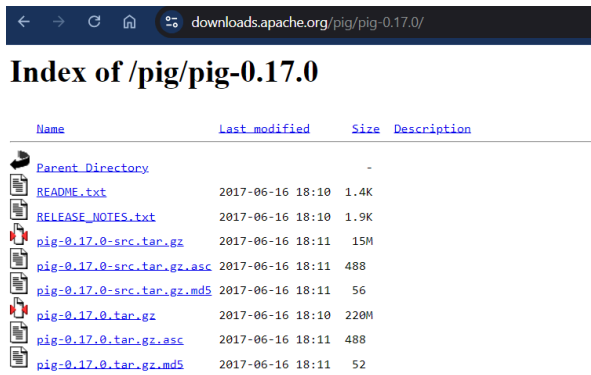
To create UDF in Apache Pig and execute it in MapReduce/HDFS mode.

PROCEDURE:

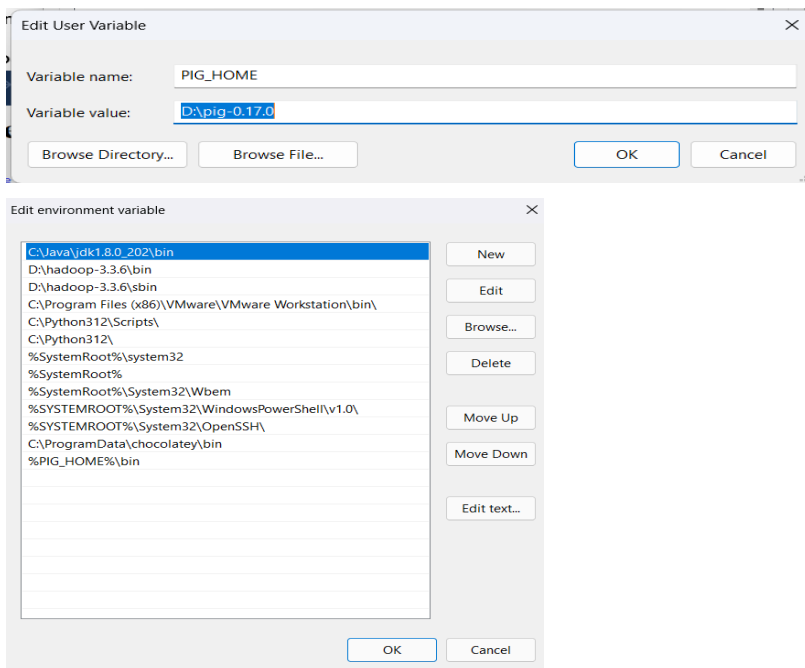
Pig Download and installation:

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”

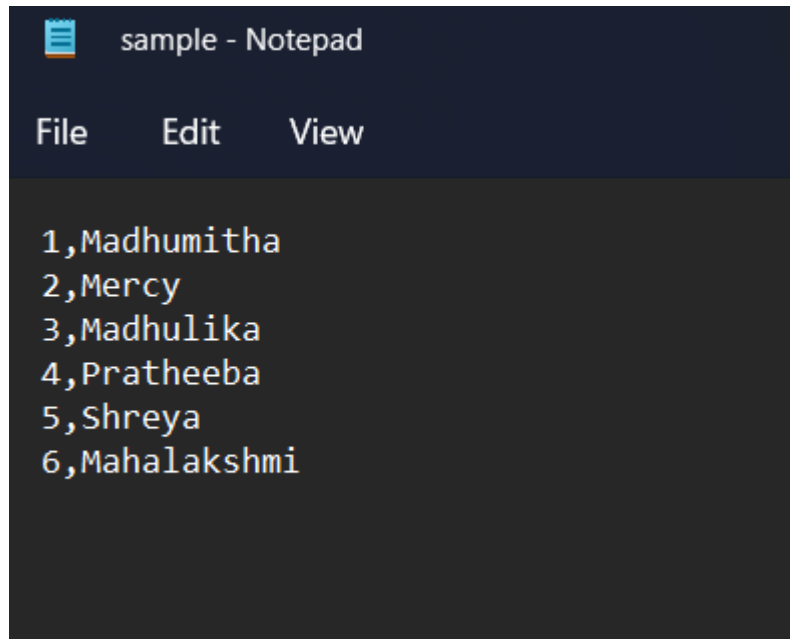
Started:	Tue Sep 17 18:50:24 +0530 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-0e30abef-07a2-416e-8d67-f8db310e9fd0
Block Pool ID:	BP-2145839966-192.168.72.1-17263060606534

Summary

Security is off.
Safemode is off.
152 files and directories, 94 blocks (94 replicated blocks, 0 erasure coded block groups) = 246 total filesystem object(s).
Heap Memory used 41.75 MB of 420.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 68.55 MB of 70 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	379.28 GB
Configured Remote Capacity:	0 B

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

```
Administrator: Command Prompt

D:\madhumitha\DA>pig -x mapreduce udf.pig
2024-09-17 19:10:00,352 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-17 19:10:00,354 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-17 19:10:00,354 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-17 19:10:00,535 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compile
Jun 02 2017, 15:41:58
2024-09-17 19:10:00,535 [main] INFO org.apache.pig.Main - Logging error messages to: D:\hadoop-3.3.6\l
gs\pig_1726580400532.log
2024-09-17 19:10:00,684 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file C:\Users\ADMI
/.pigbootup not found
2024-09-17 19:10:00,727 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.trac
er is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-17 19:10:00,728 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - C
nnecting to hadoop file system at: hdfs://localhost:9000
2024-09-17 19:10:01,027 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf.
ig-205a9807-a668-400e-a2e9-19f7da6f661a
2024-09-17 19:10:01,027 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-ser
vice.enabled set to false
2024-09-17 19:10:01,245 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp p
thon.cachedir=C:\Users\ADMIN\AppData\Local\Temp\pig_jython_9133215464494257087
2024-09-17 19:10:03,134 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - Register scri
pting UDF: udf.uppercase
```

9. View the Output

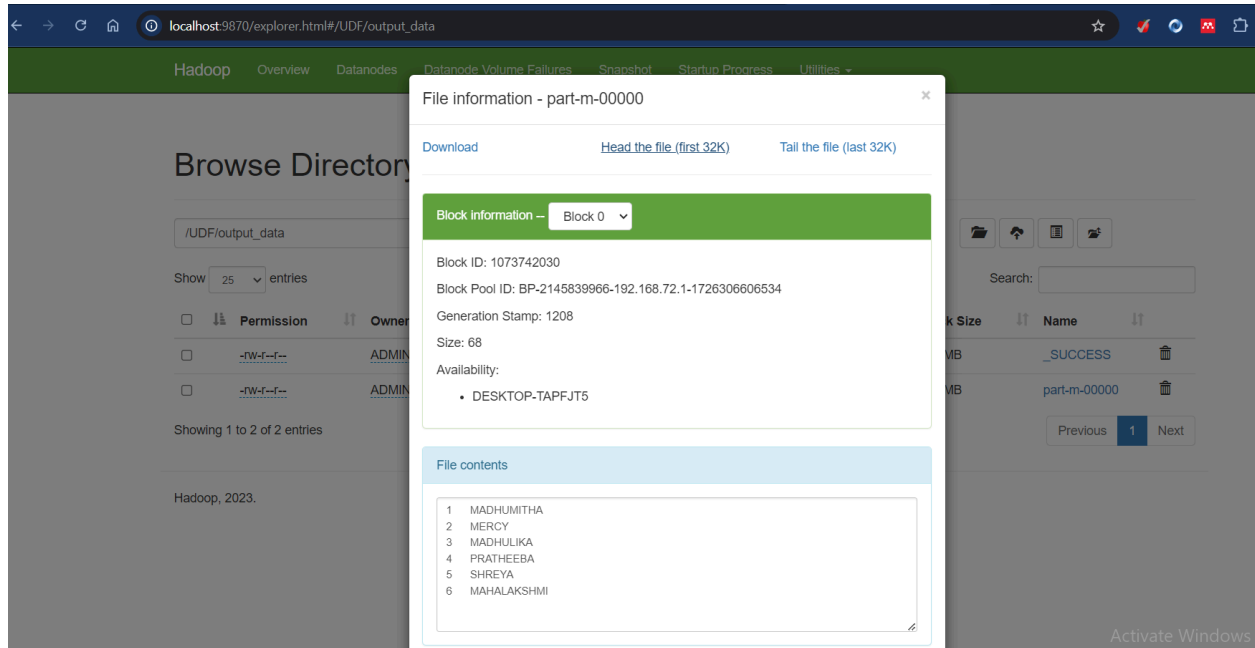
```
Administrator: Command Prompt

D:\madhumitha\DA>hdfs dfs -ls /UDF
Found 3 items
drwxr-xr-x - ADMIN supergroup 0 2024-09-17 19:11 /UDF/output_data
-rw-r--r-- 1 ADMIN supergroup 72 2024-09-17 18:53 /UDF/sample.txt
-rw-r--r-- 1 ADMIN supergroup 202 2024-09-17 19:09 /UDF/uppercase_udf.py

D:\madhumitha\DA>hdfs dfs -ls /UDF/output_data
Found 2 items
-rw-r--r-- 1 ADMIN supergroup 0 2024-09-17 19:11 /UDF/output_data/_SUCCESS
-rw-r--r-- 1 ADMIN supergroup 68 2024-09-17 19:11 /UDF/output_data/part-m-00000

D:\madhumitha\DA>hdfs dfs -cat /UDF/output_data/part-m-00000
1 MADHUMITHA
2 MERCY
3 MADHULIKA
4 PRATHEEBA
5 SHREYA
6 MAHALAKSHMI
```

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.