

Ex.No – 4

Roll No – 210701149

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

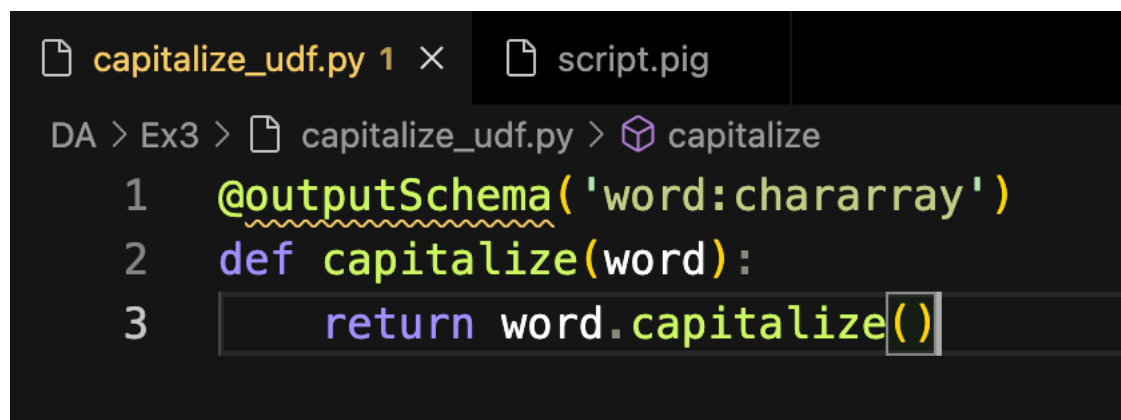
AIM:

To create user defined functions in Apache Pig and execute it in MapReduce / HDFS mode.

PROCEDURES:

1. **Write the Python UDF:** Created a Python function that reverses strings.
2. **Register the UDF in Pig:** Registered the Python script in the Pig script using the REGISTER command.
3. **Use the UDF in Pig Script:** Applied the UDF to the dataset using FOREACH ... GENERATE.
4. **Execute the Pig Script:** Ran the Pig script in MapReduce mode to process the data on HDFS.

OUTPUT:



```
capitalize_udf.py 1 × script.pig
DA > Ex3 > capitalize_udf.py > capitalize
1  @outputSchema('word:chararray')
2  def capitalize(word):
3      return word.capitalize()
```

script.pig x

DA > Ex3 > script.pig

```
1  -- Register the Python script containing the UDF
2  REGISTER 'capitalize_udf.py' USING jython AS myudf;
3
4  -- Load a sample dataset from HDFS
5  data = LOAD 'hdfs:///user/three/input.txt' USING PigStorage(',') AS (word:chararray);
6
7  -- Apply the UDF to each record
8  capitalized_data = FOREACH data GENERATE myudf.capitalize(word);
9
10 -- Store the result back to HDFS
11 STORE capitalized_data INTO 'hdfs:///user/three/output' USING PigStorage(',');
12
```

~/Documents/PYTHON/DA/Ex3

pig -x mapreduce

```
2024-08-20 20:19:06,581 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-08-20 20:19:06,582 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-08-20 20:19:06,582 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-08-20 20:19:06,607 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun
02 2017, 15:41:58
2024-08-20 20:19:06,607 [main] INFO org.apache.pig.Main - Logging error messages to: /Users/manoj/Documents/
PYTHON/DA/Ex3/pig_1724165346604.log
2024-08-20 20:19:06,618 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /Users/manoj/.pigbo
otup not found
2024-08-20 20:19:06,789 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop l
ibrary for your platform... using builtin-java classes where applicable
2024-08-20 20:19:06,798 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2024-08-20 20:19:06,798 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connect
ing to hadoop file system at: hdfs://127.0.0.1:9000
2024-08-20 20:19:07,130 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-c1
be8a42-4a25-4684-9bea-4f5329396c7a
2024-08-20 20:19:07,130 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.e
nabled set to false
grunt>
```

```
~/Documents/PYTHON/DA/Ex3 (1m 8.52s)

pig -x mapreduce

2024-08-20 20:19:45,948 [JobControl] INFO org.apache.hadoop.yarn.util.resource.ResourceUtils - Unable to find 'resource-types.xml'.
2024-08-20 20:19:46,013 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1724164650188_0002
2024-08-20 20:19:46,036 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8088/proxy/application_1724164650188_0002/
2024-08-20 20:19:46,036 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1724164650188_0002
2024-08-20 20:19:46,036 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases capitalized_data,data
2024-08-20 20:19:46,036 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: data[5,7],capitalized_data[-1,-1] C: R:
2024-08-20 20:19:46,041 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2024-08-20 20:19:46,041 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1724164650188_0002]
2024-08-20 20:19:58,175 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2024-08-20 20:19:58,176 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Running jobs are [job_1724164650188_0002]
2024-08-20 20:20:01,226 [main] INFO org.apache.hadoop.yarn.client.DefaultNoHARMFailoverProxyProvider - Connecting to ResourceManager at /127.0.0.1:8032
2024-08-20 20:20:01,236 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
```

```
~/Documents/PYTHON/DA/Ex3 (1.109s)

hadoop fs -cat /user/three/output/part-m-00000

2024-08-20 20:20:55,593 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Hello
Data
Analytics
From
Apache
Pig
```

RESULT:

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