

Create User Define Function in Apache Pig and execute it on map reduce

Aim:

To create User Define Function in Apache Pig and execute it on map reduce

Procedure:

1. Firstly install PIG

Step 1: Login into Ubuntu

Step 2: Go to <https://pig.apache.org/releases.html> and copy the path of the latest version of pig that you want to install. Run the following command to download Apache Pig in Ubuntu:

\$ wget <https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz>

Step 3: To untar pig-0.16.0.tar.gz file run the following command:

\$ tar xvf pig-0.16.0.tar.gz

Step 4: To create a pig folder and move pig-0.16.0 to the pig folder, execute the following command:

\$ sudo mv /home/hadoop/pig-0.16.0 /home/hadoop/pig

Step 5: Now open the .bashrc file to edit the path and variables/settings for pig. Run the following command:

\$ sudo nano .bashrc

Add the below given to .bashrc file at the end and save the file.

```
# PIG settings
export
PIG_HOME=/home/hadoop/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
# PIG settings ends
```

Step 6: Run the following command to make the changes effective in the .bashrc file:

\$ source .bashrc

Step 7: To start all Hadoop daemons, navigate to the hadoop-3.2.1/sbin folder and run the following commands:

\$./start-dfs.sh \$./start-yarn.sh

Step 8: Now you can launch pig by executing the following command:

\$ pig

Step 9: Now you are in pig and can perform your desired tasks on pig. You can come out of the pig by the quit command:

> quit;

2. Create UDF in Pig

Create a sample text

```
filehadoop@Ubuntu:~/Documents$ nano
```

```
sample.txtPastethebelow content to
```

```
sample.txt
```

```
1,John2,Ja
```

```
ne3,Joe4,
```

```
Emma
```

```
hadoop@Ubuntu:~/Documents$hadoopfs-putsample.txt/home/hadoop/piginput/
```

CreatePIGFile

```
hadoop@Ubuntu:~/Documents$nanodemo_pig.pig
```

pastethe belowthe contentto demo_pig.pig

```
--LoadthedatafromHDFS
```

```
data=LOAD'/home/hadoop/piginput/sample.txt'USING PigStorage(',')AS(id:int>
```

```
-- Dump the data to check if it was loaded
```

```
correctlyDUMPdata;
```

Run the abovefile

```
hadoop@Ubuntu:~/Documents$pigdemo_pig.pig
```

```
2024-08-0712:13:08,791[main]INFO
```

```
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
```

- Total input paths to process :

1(1,John)

(2,Jane)

(3,Joe)

(4,Emma)

Createudffileansaveasuppercase_udf.py

uppercase_udf.py

```
def
uppercase(text):re
turntext.upper()

ifname__==
"main":import sys
forlinein sys.stdin:
line=line.strip()
result =
uppercase(line)print(result)
```

Createtheudfsfolderonhadoop

hadoop@Ubuntu:~/Documents\$ hadoopfs -mkdir/home/hadoop/udfs

puttheupppercase_udf.pyintotheabvfolder

hadoop@Ubuntu:~/Documents\$ hdfsdfs -putuppercase_udf.py/home/hadoop/udfs/

hadoop@Ubuntu:~/Documents\$ nano

udf_example.pigcopyand

pastethebelowcontentonudf_example.pig

--Register thePython UDFscript

```
REGISTER'hdfs:///home/hadoop/udfs/uppercase_udf.py'USINGjythonASudf;
```

```
--Load somedata
```

```
data=LOAD'hdfs:///home/hadoop/sample.txt'AS(text:chararray);
```

```
--Usethe Python UDF
```

```
uppercased_data=FOREACHdataGENERATEudf.uppercase(text)AS uppercase_text;
```

```
--Storetheresult
```

```
STOREuppercased_dataINTO'hdfs:///home/hadoop/pig_output_data';
```

placesample.txt file on hadoop

hadoop@Ubuntu:~/Documents\$hadoopfs -putsample.txt/home/hadoop/

ToRunthe pig file

hadoop@Ubuntu:~/Documents\$pig -fudf_example.pig

finally u

getSucces

s!

JobStats(timein seconds):

JobId Maps Reduces MaxMapTimeMinMapTime AvgMapTime

MedianMapTimeMaxReduceTimeMinReduceTime

AvgReduceTimeMedianReducetime

AliasFeatureOutputs

job_local1786848041_0001 1 0 n/a n/a n/a n/a 00 0

0data,uppercased_dataMAP_ONLYhdfs:///home/hadoop/pig_output_da

ta,

Input(s):

Successfullyread4records(42778068bytes)from:"hdfs:///home/hadoop/sample.txt"

Output(s):

Successfully stored 4 records (42777870 bytes) in: "hdfs:///home/hadoop/pig_output_data"

Counters:

Total records written: 4

Total bytes written: 42777870

Spillable Memory Manager spill count :

0 Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_local1786848041_0001

2024-08-07 13:33:04,631 [main] WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl

! -JobTracker metric system already initialized!

2024-08-07 13:33:04,639 [main] WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl

! -JobTracker metric system already initialized!

2024-08-07 13:33:04,644 [main] WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl

! -JobTracker metric system already initialized!

2024-08-07 13:33:04,667 [main] INFO

org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -Success!

Note:

If any error check if python package is installed and check the path specified on the above steps are given correctly

To check the output file is created

hadoop@Ubuntu: ~/Documents\$ hdfs dfs -ls

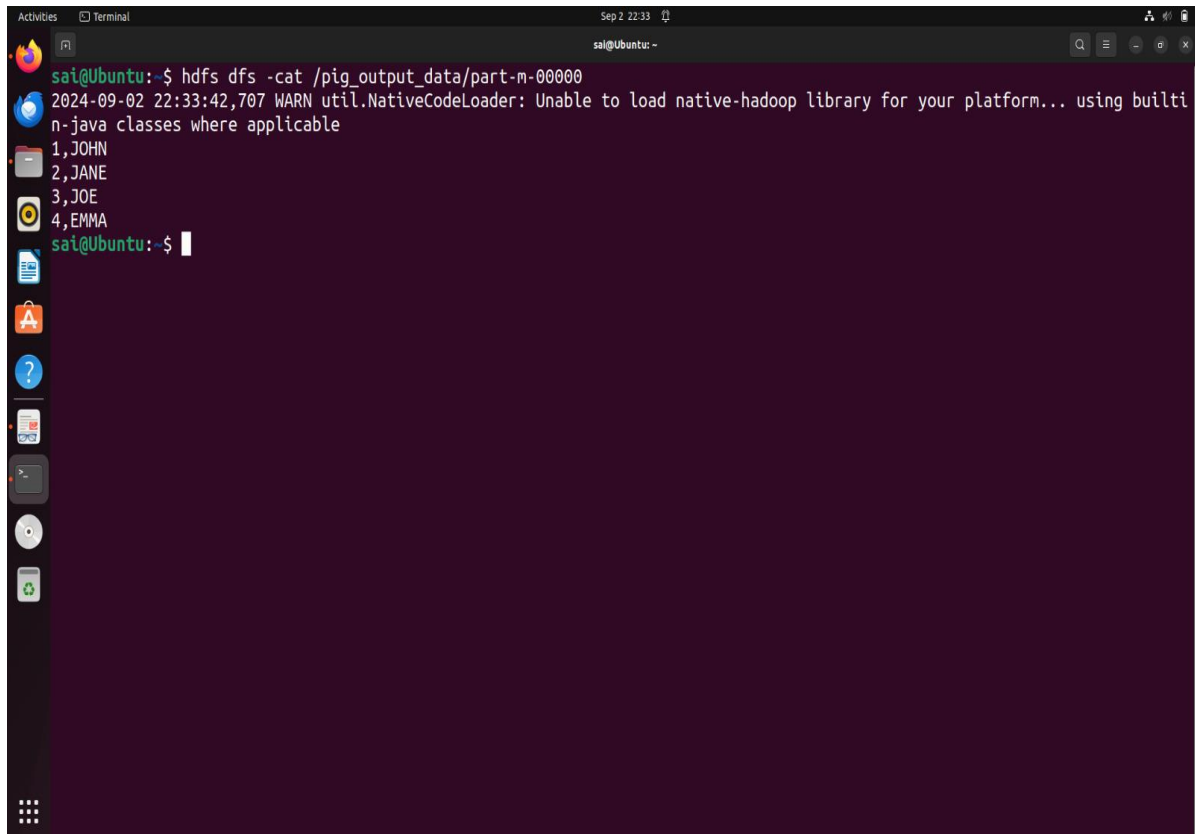
/home/hadoop/pig_output_data Found 2 items

If you need to examine the files in the output folder, use:

To view the output

```
hadoop@Ubuntu:~/Documents$ hdfs dfs -cat  
/home/hadoop/pig_output_data/part-m-00000
```

OUTPUT:

A screenshot of a Linux terminal window. The terminal title bar shows 'Activities', 'Terminal', and the date 'Sep 2 22:33'. The prompt is 'sai@Ubuntu: ~'. The command entered is 'hdfs dfs -cat /pig_output_data/part-m-00000'. The output shows a warning message: '2024-09-02 22:33:42,707 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable'. Below the warning, the output is a list of names: '1,JOHN', '2,JANE', '3,JOE', and '4,EMMA'. The prompt returns to 'sai@Ubuntu: ~\$'.

```
sai@Ubuntu:~$ hdfs dfs -cat /pig_output_data/part-m-00000  
2024-09-02 22:33:42,707 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built  
n-java classes where applicable  
1,JOHN  
2,JANE  
3,JOE  
4,EMMA  
sai@Ubuntu:~$
```

Result:

Thus the UserDefineFunction in Apache Pig and execute it on mapreduce is executed successfully.