ROLL.NO: 210701271

EXP 4: Create UDF in PIG

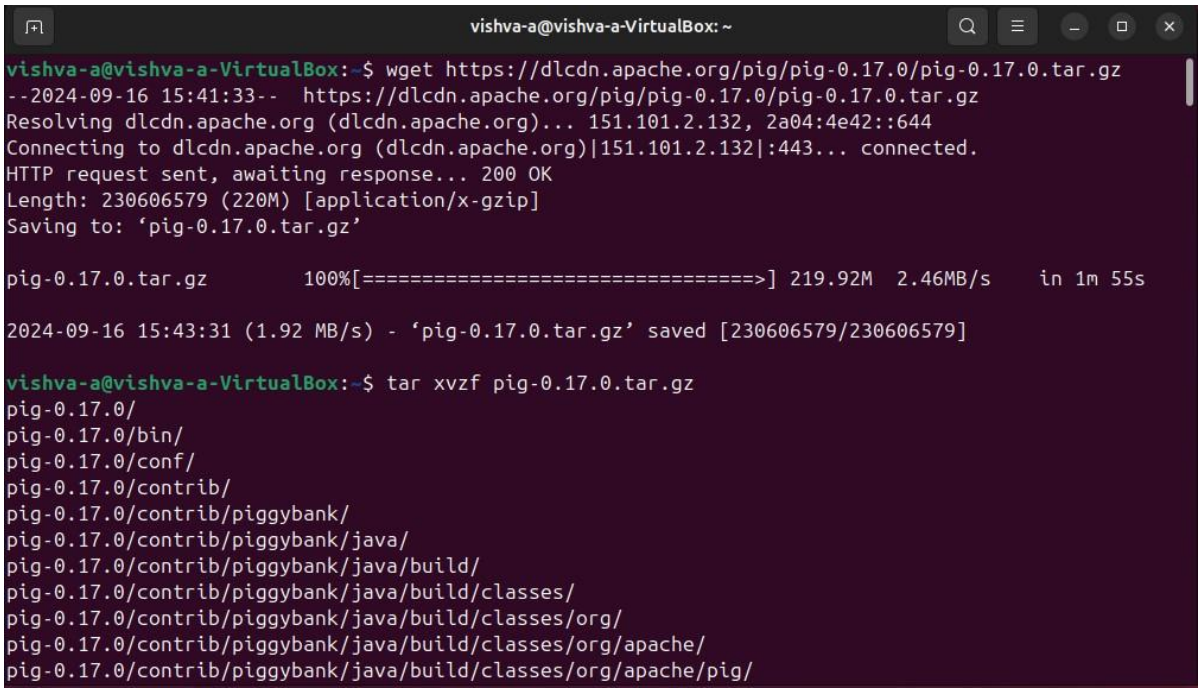
Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu Pre-requisite:

· Ubuntu 16.04 or higher version running (I have installed Ubuntu on Oracle VM (Virtual Machine) VirtualBox),

· Run Hadoop on ubuntu (I have installed Hadoop 3.2.1 on Ubuntu 16.04). You may refer to my blog "How to install Hadoop installation" click here for Hadoop installation).

Pig installation steps

Step 1: Login into Ubuntu



```
vishva-a@vishva-a-VirtualBox: ~
vishva-a@vishva-a-VirtualBox:~$ wget https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
--2024-09-16 15:41:33--  https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 230606579 (220M) [application/x-gzip]
Saving to: 'pig-0.17.0.tar.gz'

pig-0.17.0.tar.gz        100%[=================================>] 219.92M  2.46MB/s    in 1m 55s

2024-09-16 15:43:31 (1.92 MB/s) - 'pig-0.17.0.tar.gz' saved [230606579/230606579]

vishva-a@vishva-a-VirtualBox:~$ tar xvzf pig-0.17.0.tar.gz
pig-0.17.0/
pig-0.17.0/bin/
pig-0.17.0/conf/
pig-0.17.0/contrib/
pig-0.17.0/contrib/piggybank/
pig-0.17.0/contrib/piggybank/java/
pig-0.17.0/contrib/piggybank/java/build/
pig-0.17.0/contrib/piggybank/java/build/classes/
pig-0.17.0/contrib/piggybank/java/build/classes/org/
pig-0.17.0/contrib/piggybank/java/build/classes/org/apache/
pig-0.17.0/contrib/piggybank/java/build/classes/org/apache/pig/
```

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 comment 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 xvzf 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:

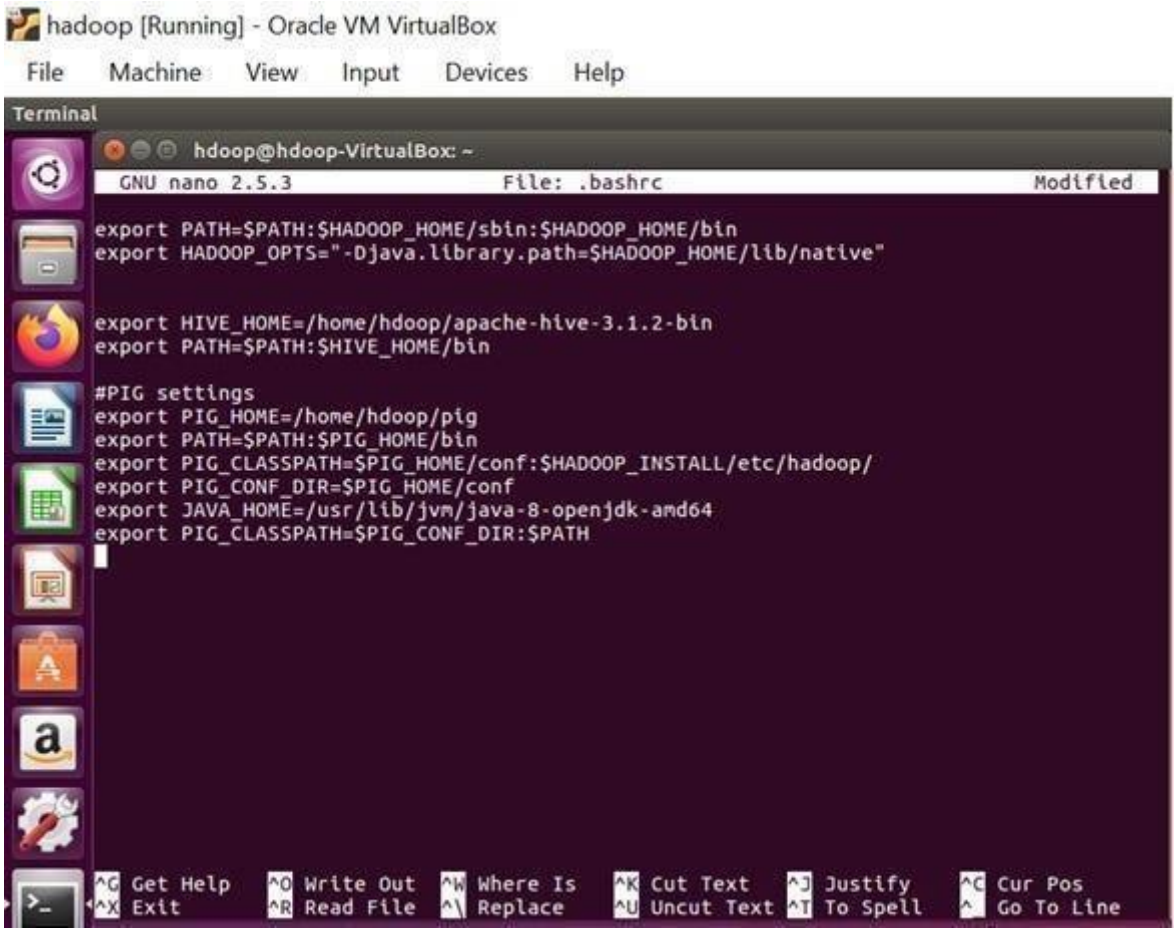
$ sudo mv /home/hdoop/pig-0.16.0 /home/hdoop/pig

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 settingsexport PIG_HOME=/home/hdoop/pigexport PATH=$PATH:$PIG_HOME/binexport PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/export PIG_CONF_DIR=$PIG_HOME/confexport JAVA_HOME=/usr/lib/jvm/java-8openjdkamd64export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH#PIG setting 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$ jps

```
vishva-a@vishva-a-VirtualBox: ~
WARNING: resourcemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
vishva-a@vishva-a-VirtualBox:~$ cd hadoop-3.3.6/sbin
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [vishva-a-VirtualBox]
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ jps
14884 NameNode
15686 Jps
15446 ResourceManager
15575 NodeManager
15180 SecondaryNameNode
15005 DataNode
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ cd
vishva-a@vishva-a-VirtualBox:~$ pig
2024-09-16 16:00:59,660 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-16 16:00:59,670 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-16 16:00:59,670 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-16 16:00:59,908 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) com
piled Jun 02 2017, 15:41:58
2024-09-16 16:00:59,908 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vishva-
```

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

```
vishva-a@vishva-a-VirtualBox: ~
15446 ResourceManager
15575 NodeManager
15180 SecondaryNameNode
15005 DataNode
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ cd
vishva-a@vishva-a-VirtualBox:~$ pig
2024-09-16 16:00:59,660 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-16 16:00:59,670 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-16 16:00:59,670 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-16 16:00:59,908 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) com
piled Jun 02 2017, 15:41:58
2024-09-16 16:00:59,908 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vishva-
a/pig_1726482659887.log
2024-09-16 16:01:00,102 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/vis
hva-a/.pigbootup not found
2024-09-16 16:01:02,722 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.
tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-16 16:01:02,725 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine
 - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-16 16:01:07,539 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-
default-2b684044-d43e-4d8f-b3cd-b5589e293acf
2024-09-16 16:01:07,539 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline
-service.enabled set to false
grunt>
```

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;

CREATE USER DEFINED FUNCTION(UDF)

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

Procedure:

Create a sample text file hadoop@Ubuntu:~/

nano sample.txt

Paste the below content to sample.txt

1,John

2,Jane

3,Joe

4,Emma

hadoop@Ubuntu:~/Documents$ hadoop fs -put sample.txt /home/hadoop/piginput/

Create PIG File hadoop@Ubuntu:~/Documents$

nano script.pig

paste the below the content to demo_pig.pig

-- Load the data from HDFS

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

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

DUMP data;

Create a user defined file named as expt_udf.py.

Run the following command, cat sample.txt |

python3 expt_udf.py

```
vboxuser@ubuntus: ~/pig-0.16.0
2024-09-21 05:26:31,661 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-09-21 05:26:31,680 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.4.0   0.16.0  vboxuser        2024-09-21 05:26:27     2024-09-21 05:26:31     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_local330257209_0001 1       0       n/a     n/a     n/a     n/a     0       0       0       0       data,uppercased MAP_ONLY        file:///home/vboxuser/pig-0.16.0/output,

Input(s):
Successfully read 3 records from: "file:///home/vboxuser/pig-0.16.0/data.txt"

Output(s):
Successfully stored 3 records in: "file:///home/vboxuser/pig-0.16.0/output"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local330257209_0001


2024-09-21 05:26:31,693 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-21 05:26:31,702 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-21 05:26:31,717 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-09-21 05:26:31,737 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-21 05:26:31,841 [main] INFO  org.apache.pig.Main - Pig script completed in 15 seconds and 804 milliseconds (15804 ms)
vboxuser@ubuntus:~/pig-0.16.0$ cat /home/vboxuser/pig-0.16.0/output/part-m-00000
SURENDHAR
VIMAL
VINOTH
vboxuser@ubuntus:~/pig-0.16.0$
```