Exp No: 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

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=$PIG_HOME/confexport JAVA_HOME=/usr/lib/jvm/java-8-openjdkamd64export
PIG_CLASSPATH=$PIG_CONF_DIR:$PATH#PIG setting ends

```
export PIG_HOME=/home/haresh/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
```

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

```
haresh@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as haresh in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [fedora]
Starting resourcemanager
Starting nodemanagers
haresh@fedora:~$
```

Now you can launch pig by executing the following command:

$ pig

```
haresh@fedora:~$ pig
2024-09-13 09:46:24,963 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-13 09:46:24,964 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-13 09:46:24,964 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
ype
2024-09-13 09:46:25,012 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-13 09:46:25,012 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/haresh/pig_1726200985006.log
2024-09-13 09:46:25,056 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/haresh/.pigbootup not found
2024-09-13 09:46:25,375 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-13 09:46:25,375 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-13 09:46:25,375 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-13 09:46:26,058 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
```

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