Apache Pig

Steps to install Ant

1. Install latest ant version

wgethttp://www.apache.org/dist/ant/binaries/apache-ant-1.9.4-bin.tar.gz
tar xzf apache-ant-1.9.4-bin.tar.gz

2. setup ANT_HOME and update PATH in .bash_profile

export ANT_HOME=/home/ec2-user/apache-ant-1.9.4

Steps to install Pig

1. Run the below command to download Pig-0.12.1to home directory

wget http://mirror.metrocast.net/apache/pig/pig-0.12.1/pig-0.12.1-src.tar.gz

2. Extract the downloaded tar file and move it to /user/local

tar xzf pig-0.12.1-src.tar.gz

3. Compile Pig: (Apache Pig 0.12.1 expects an older version of Hadoop by default. **You must recompile Pig for Hadoop 2.4.1**)

cd pig-0.12.1-src
ant clean jar-all -Dhadoopversion=23

4. Update Pig executable PATH so that you can run Pig programs without issuing the full path and by just issuing the pig command.

vi ~/.bash_profile
#Paste the below content to the file
PIG binary paths
export PIG_HOME=/home/ec2-user/pig-0.12.1-src
PATH=$PATH:$JAVA_HOME/bin:$ANT_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PIG_HOME/bin
source .bash_profile

5. If you want to create a dedicated directory for pig log files, create one and update the directory in pig.properties file

under home dir :
mkdir pig_logs

vi ~/pig-0.12.1-src/conf/pig.properties
#updated in pig.properties file

pig.logfile=/home/ec2-user/pig_logs/

Pig Setup Verification

#Verify the version of Pig installed
pig version

#To run the pig in local mode, run the below command
pig –x local

#To run the pig in mapreduce mode, both the below commands will work, run any one of these
pig –x mapreduce
or
pig

Running a Pig script

An Apache Pig script works in two modes:

Local Mode: In 'local mode', you can execute the pig script in local file system. In this case you don't need to store the data in Hadoop HDFS file system, instead you can work with the data stored in local file system itself.
HDFS Mode: In 'HDFS mode', the data needs to be stored in HDFS file system and you can process the data with the help of pig script.

Running in localmode:

1. Get word count program dir from local machine to EC2 instance

scp -i ~/aws/aws-key.pem -r pig-wordcount ec2-user@54.183.19.28:

2. To run word count pig script

go to the directory pig-wordcount
cd pig-wordcount

pig -x local wordcount.pig

#Output directory is wordcount
cd wordcount

```
[ec2-user@ip-172-31-5-251 pig-wordcount]$ cd wordcount
[ec2-user@ip-172-31-5-251 wordcount]$ ls
part-r-00000  _SUCCESS
[ec2-user@ip-172-31-5-251 wordcount]$ cat part-r-00000
2       in
1       for
2       pig
2       2012
1       word
1       count
2       school
2       summer
1       indiana
2       tutorial
```

Running in Hadoop mode:

1. Go to directory pig-wordcount
   hdfs dfs -mkdir /input1
   hdfs dfs -copyFromLocal input.txt /input1
   hdfs dfs -cat /input1/inputpig.txt

2. In wordcount.pig , make changes to reflect input file and output file

vi wordcount.pig
A = load '/input1/inputpig.txt';
B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
C = group B by word;
D = foreach C generate COUNT(B), group;
store D into '/wordcount1';

3. Run pig script in hadoop mode
pig wordcount.pig

4. verify the output directory wordcount1:
hdfs dfs -ls /wordcount1

```
[ec2-user@ip-172-31-5-251 pig-wordcount]$ hdfs dfs -cat /wordcount1/part-r-00000
14/07/20 16:45:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library
icable
2       in
1       for
2       pig
2       2012
1       word
1       count
2       school
2       summer
1       indiana
2       tutorial
```

Apache Hive

Installation of hive:

1. Download Apache hive

wget http://mirrors.advancedhosters.com/apache/hive/hive-0.13.1/apache-hive-0.13.1-bin.tar.gz

2. Untar the archive

tar xzf apache-hive-0.13.1-bin.tar.gz

3. Create a soft link

ln -s apache-hive-0.13.1-bin hive

4. Setting HIVE_HOME in bash_profile

vi ~/.bash_profile
#copy below line
exportHIVE_HOME=/home/ec2-user/hive

#Set path to HIVE_HOME
PATH=$PATH:$JAVA_HOME/bin:$ANT_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PIG_HOME/bin:**$HIVE_HOME/bin**

Demo

hive> CREATE TABLE PRODUCT (productid INT, productname STRING,price FLOAT, category STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

hive> describe product;

##Insert data into the table.
hive> load data local inpath '/home/ec2-user/hive/inputhive.txt' into table product;

##Retrieve data
hive> select * from product;

```
hive>
```

```
hive> CREATE TABLE PRODUCT (
    > productid INT, productname STRING,price FLOAT, category STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.931 seconds
hive> describe product;
OK
productid               int
productname             string
price                   float
category                string
Time taken: 0.718 seconds, Fetched: 4 row(s)
hive>
    >
    > load data local inpath '/home/ec2-user/hive/input.txt' into table product;
Copying data from file:/home/ec2-user/hive/input.txt
Copying file: file:/home/ec2-user/hive/input.txt
Loading data to table default.product
Table default.product stats: [numFiles=1, numRows=0, totalSize=103, rawDataSize=0]
OK
Time taken: 1.157 seconds
hive> select * from product;
OK
1       Books   25.0    Stationery
2       Pens    10.0    Stationery
3       Sugar   40.05   House HOld Items
4       Furniture       130000.0        Interiors
Time taken: 0.542 seconds, Fetched: 4 row(s)
hive>
```