Setting up your Hadoop environment in CentOS 7 (Hadoop for Dummies)

This guide is about installing Hadoop on CentOS 7 based on the Hadoop for Dummies book.

Downloading and installing Bigtop

Step 1 – Open terminal as root user.

su –

Enter password when prompted

Step 2 – Run the following command (the URL in the book is outdated)

wget -O /etc/yum.repos.d/bigtop.repo \

“http://www.apache.org/dist/bigtop/stable/repos/centos7/bigtop.repo&#8221”;

Step 3 – cd into “/etc/yum.repos.d/bigtop.repo” file and change the last line into

gpgkey=http://www.apache.org/dist/bigtop/stable/repos/GPG-KEY-bigtop

You’ll have to open the file as a root user:

3.1) In the terminal type

vi bigtop.repo

3.2) The cursor is controlled using the following four keys:

Key	Cursor Movement
h	left one space
j	down one line
k	up one line
l	right one space

3.3) Move the cursor to the last line and type **dd**. This should delete the last line.

3.4) Type **i**. This will let you insert something.

3.5) Type the following on the last line:

gpgkey=www.apache.org/dist/bigtop/stable/repos/GPG-KEY-bigtop

3.6) Press Esc to get out of insert mode.

Then type :wq (this will save and quit vi)

Step 4 – Use the yum installer to install components separately

yum install hadoop

yum install mahout

yum install oozie

yum install hbase

yum install hive

yum install hue

yum install pig

yum install zookeeper

Starting Hadoop (This section should work as is in the book)

Step 1 – Download and install Java:

yum install java-1.7.0-openjdk-devel.x86_64

Step 2 – Format the NameNode:

sudo /etc/init.d/hadoop-hdfs-namenode init

Step 3 – Start the Hadoop services for your pseudo distributed cluster:

for i in hadoop-hdfs-namenode hadoop-hdfs-datanode ; do sudo service $i start ; done

Step 4 – Create a subdirectory structure in HDFS:

sudo /usr/lib/hadoop/libexec/init-hdfs.sh

Step 5 – Start the YARN daemons:

sudo service hadoop-yarn-resourcemanager start

sudo service hadoop-yarn-nodemanager start

Download and copy the dataset to hdfs (No issues in this section either)

Step 1 – Download the 1987 file, extract it and save it to your home directory.

http://stat-computing.org/dataexpo/2009/the-data.html

Step 2 – hdfs dfs -copyFromLocal 1987.csv /user/root

This will load data to hdfs

Your First Hadoop program

Step 1 – Write the following in a file and save it in the home directory as totalmiles.pig

```
records = LOAD '1987.csv' USING
PigStorage(',')AS(Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTim
e,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDel
ay,Origin,Dest,Distance:int,TaxiIn,TaxiOut,Cancelled,CancellationCode,Diverted,CarrierDelay,Weat
herDelay,NASDelay,SecurityDelay,LateAircraftDelay);
milage_recs = GROUP records ALL;
tot_miles = FOREACH milage_recs GENERATE SUM(records.Distance);
STORE tot_miles INTO '/user/root/totalmiles';
```

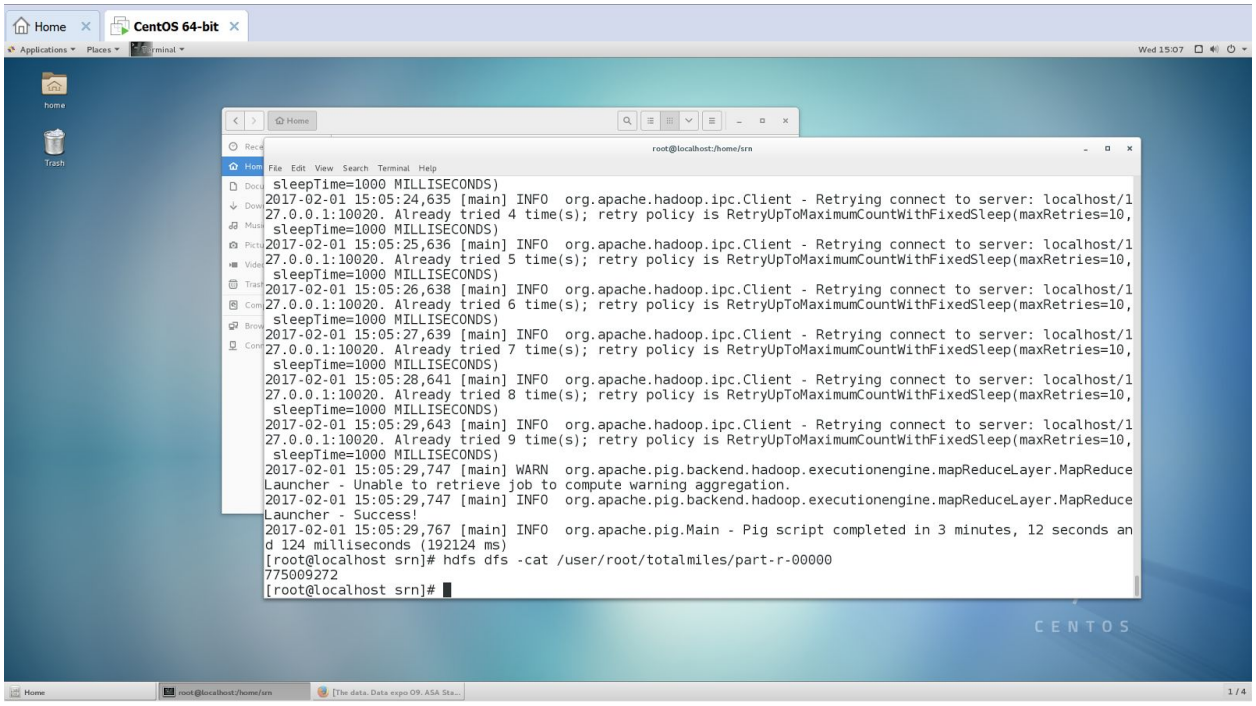
Note that there are no backslashes(\) in the code and I enclosed the path in the last line in single quotes('). And I changed '2013_subset.csv' in the book to '1987.csv' in my code. Also note that there are no unwanted spaces in my code.

Step 2 – pig totalmiles.pig

Step 3 – You should see the words success in the output. Type the following:

hdfs dfs -cat /user/root/totalmiles/part-r-00000

You should see the following in your terminal:



Copyright and License