How to Use Vi Editor??

* Vi <FileName> and Enter Button.
* Press I to write the content in file.
* Press ESC and type ZZ to save and exit.

**Horton Works Certification Tasks:**

**TASK 01**

**Question:**

Load the data from flights file where it was located at /user/horton/flights contains 24 fields with comma-separated. Fields required to the task

Column name

16 price

17 loc

Write a Pig Script which satisfies below criteria

1. Load the data from /user/Horton/flights
2. Compute the average of price column for each loc column
3. Save the output in /user/Horton/task1, with tab separated fields
4. Order the output in ascending alphabetical on column col
5. Save the script in task1 location was /home/horton/solutions

**Solution:**

**Input Data:** /home/user553158/user/horton/flights/ filight\_details.csv

1,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,200,Chennai,iuowyiu,nklk

2,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,100,Hyderabad,iuowyiu,nklk

3,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,1200,Banglore,iuowyiu,nklk

4,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,2000,Chennai,iuowyiu,nklk

5,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,2100,Chennai,iuowyiu,nklk

6,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,2200,Chennai,iuowyiu,nklk

7,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,2009,Hyderabad,iuowyiu,nklk

8,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,2008,Banglore,iuowyiu,nklk

9,abc,def,ghi,ijk,1234,678,987,asd34,678i,hsdfkjh,hkfhskfj,hfkhskf,weyiyw,iuoiuo,2007,Kolkata,iuowyiu,nklk

**Script :** /home/user553158/user/horton/tasks/ Task01

A = load '/home/user553158/user/horton/flights/\*.csv' using PigStorage(',');

B = foreach A generate $15,$16;

C = Group B by $1;

D = foreach C generate $0,AVG(B.$0);

E = order D by $0 asc;

store E into '/home/user553158/user/horton/solutions/task1' using PigStorage('\t');

**Execution Command: pig –x local Task01**

**Output:** /home/user553158/user/horton/solutions/task1/part-r-00000

Banglore 1604.0

Chennai 1625.0

Hyderabad 1054.5

Kolkata 2007.0

**TASK 02**

**Question:**

Load the data from players.txt file where it was located at /user/horton/basketball contains 42 fields with tab-separated. Fields required to the task

Column name

1 id

6 year

7 counts

Load the data from master.txt file where it was located at /user/horton/basketball contains 34 fields with tab-separated. Fields required to the task

Column name

1 id

6 first name

7 last name

Write a Pig Script which satisfies below criteria

1. Join both the datasets with id as common field
2. Output fields in prescribed order first name, last name, year, counts
3. Save the output in /user/Horton/task2, with comma-separated fields
4. Save the script in task2 location was /home/horton/solutions

**Solution:**

**Input Data:** /home/user553158/user/horton/basketball/players.txt

100 hjk uhh uhkjshf uuy 2001 555

101 zhkj kfhkj safjia jdkl 2002 324

102 zhkj kfhkj safjia jdkl 2003 1324

103 zhkj kfhkj safjia jdkl 2002 3124

104 zhkj kfhkj safjia jdkl 2001 3214

105 zhkj kfhkj safjia jdkl 2016 1051

/home/user553158/user/horton/basketball/master.txt

100 fskdjf fjhsdjkfh hkjh jlkj Viswa mandava

101 fskdjf fjhsdjkfh hkjh jlkj Viswa1 mandava

102 fskdjf fjhsdjkfh hkjh jlkj Viswa2 mandava

102 fskdjf fjhsdjkfh hkjh jlkj Viswa3 mandava

103 fskdjf fjhsdjkfh hkjh jlkj Viswa4 mandava

104 fskdjf fjhsdjkfh hkjh jlkj Viswa mandava

105 fskdjf fjhsdjkfh hkjh lkj Viswa5 mandava

106 fskdjf fjhsdjkfh hkjh jlkj Viswa6 mandava

**Script :** /home/user553158/user/horton/tasks/ Task02

A = load '/home/user553158/user/horton/basketball/players.txt' using PigStorage('\t');

B = foreach A generate $0,$5,$6;

C = load '/home/user553158/user/horton/basketball/master.txt' using PigStorage('\t');

D = foreach C generate $0,$5,$6;

E = join B by $0,D by $0;

F = foreach E generate $4,$5,$1,$2;

store F into '/home/user553158/user/horton/solutions/task02' using PigStorage(',');

**Execution Command: pig –x local Task02**

**Output:** **/home/user553158/user/horton/solutions/task02**

Viswa,mandava,2001,555

Viswa1,mandava,2002,324

Viswa3,mandava,2003,1324

Viswa2,mandava,2003,1324

Viswa4,mandava,2002,3124

Viswa,mandava,2001,3214

Viswa5,mandava,2016,1051

**TASK 03**

**Question:**

Notice the comma-separated values of the flightdelays files in HDFS contain historical data for airline flight delays. The columns in the files match the following schema:

Year, Month, DayofMonth, DayOfWeek, DepTime, CRSDepTime, ArrTime, CRSArrTime,

UniqueCarrier, FlightNum, TailNum, ActualElapsedTime, CRSElapsedTime, AirTime,

ArrDelay, DepDelay, Origin, Dest, Distance, TaxiIn, TaxiOut, Cancelled,

CancellationCode, Diverted, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay,

LateAircraftDelay

Write a Pig script that satisfies all of the following criteria:

* Load all of the data in /user/user553158/flights/
* Remove all rows in the flightdelays data where the DepTime column equals the string "NA".
* The output should only contain the Year, Month, DayofMonth, DepTime, UniqueCarrier, FlightNum, ArrDelay, Origin and Dest
* Store the result as comma-separated records in a new directory in HDFS named /user/user553158/solutions/task03
* Save the script in a file named task04 and save it in the /home/horton/tasks directory on the local filesystem of the client machine.

**Solution:**

**Input Data:** /user/user553158/flights/flight\_delay.csv

2016,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2013,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2012,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2011,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2014,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2015,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2016,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2013,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2014,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2012,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2010,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

**Script :** /home/user553158/user/horton/tasks/ Task04

A = load '/user/user553158/flights/flight\_delay.csv' using PigStorage(',');

B = filter A by (chararray)$4!='NA';

C = foreach B generate $0,$1,$2,$3,$8,$9,$14,$16,$17;

store C into '/user/user553158/solutions/task03' using PigStorage(',');

**Execution Command: pig task04**

**Output:** **/user/user553158/solutions/task03**

2013,Aug,02,03,CTS,A6855,30,CHN,HYD

2011,Aug,02,03,CTS,A6855,30,CHN,HYD

2014,Aug,02,03,CTS,A6855,30,CHN,HYD

2015,Aug,02,03,CTS,A6855,30,CHN,HYD

2016,Aug,02,03,CTS,A6855,30,CHN,HYD

2013,Aug,02,03,CTS,A6855,30,CHN,HYD

2012,Aug,02,03,CTS,A6855,30,CHN,HYD

**TASK 04**

**Question:**

Write a Pig script saved on the client machine as /home/user553158/user/horton/tasks/ Task05.pig

that calculates the number of rows in the /user/user553158/flights/files in HDFS.

Store the output of your script in a new directory in HDFS named cleaned\_total.

The Dest column is the destination airport code where the flight ends.

Write a Pig script saved on the client machine as /home/horton/solutions/denver\_total.

pig that calculates the number of rows in the /user/horton/flightdelays\_clean data where

the Dest field equals the Denver airport code "CTS". Store the output of your script in

a new directory in HDFS named denver\_total.

The ArrDelay column is the number of minutes that a flight arrived late.

Write a Pig script saved on the client machine as /home/horton/solutions/denver\_late.pig that

counts the number of flights whose Dest is the "DEN" airport that arrived 60 or more mintues late.

Store the output of your script in a new directory in HDFS named denver\_late

//Year, Month, DayofMonth, DepTime, UniqueCarrier, FlightNum, ArrDelay, Origin,Dest;

A = load '/user/user553158/flights/flightdelays\_clean.csv' using PigStorage(',');

B = Group A all;

C = foreach B generate COUNT(A);

store C into '/user/user553158/solutions/task05' using PigStorage(',');

A1 = filter A by $8=='CTS';

B1 = group A1 all;

C1 = foreach B1 generate COUNT(A1);

store C1 into '/user/user553158/solutions/task06' using PigStorage(',');

A2 = filter A by $8=='CTS' and $15==60;

B2 = group A2 all;

C2 = foreach B2 generate COUNT(A2);

store C2 into '/user/user553158/solutions/task07' using PigStorage(',');

Input:

2016,Aug,02,03,NA,1:30,1:30,1:30,TCS,A6855,855,12:30,1:30,1:30,30,60,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2013,Aug,02,03,12:30,1:30,1:30,1:30,HCL,A6855,855,12:30,1:30,1:30,30,60,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2012,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,60,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2011,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,60,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2014,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,60,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2015,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2016,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2013,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2014,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2012,Aug,02,03,12:30,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30

2010,Aug,02,03,NA,1:30,1:30,1:30,CTS,A6855,855,12:30,1:30,1:30,30,30,CHN,HYD,300,520,650,N,NO,YES,30,30,30,30,30