Query Processing using PIG

Steps:

1. **Download and set PATH for PIG**

Download pig from https://pig.apache.org/releases.html . Extract the zip file and save it in a folder.

To set path give below command

*export PATH=/home/hduser/pig-0.16.0/bin:$PATH*

1. **Load and check dataset in HDFS**

Make a directory in Hadoop dfs using below command.

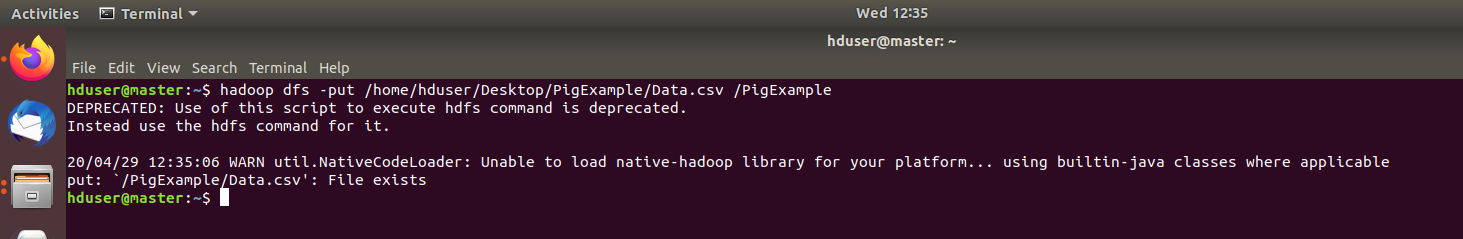
*hadoop fs -mkdir /<File\_directpry\_Name>*

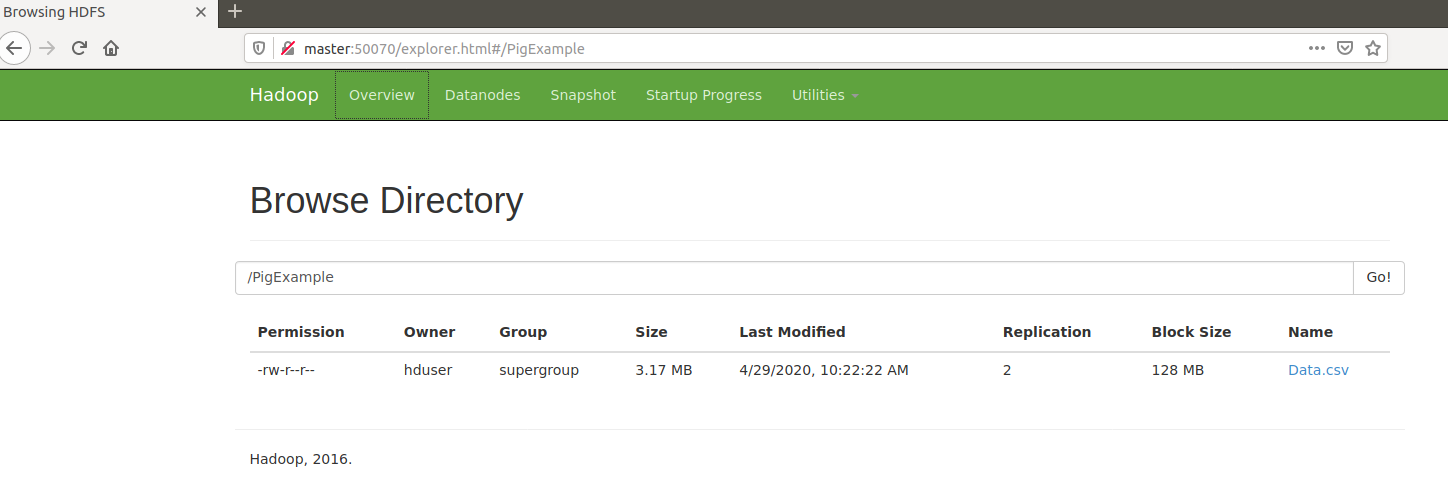
Upload the healthcare data in the hdfs directory using below command

*hadoop dfs -put <local\_directory\_loc> <hdfs\_directory\_loc>*

check the data by going to the hdfs directory using below address in browser.

*localhost:50070*

**

**

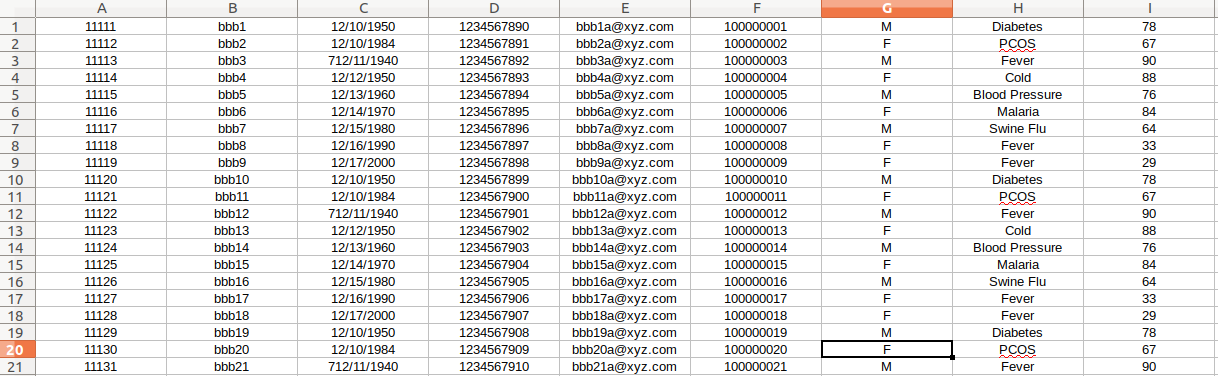
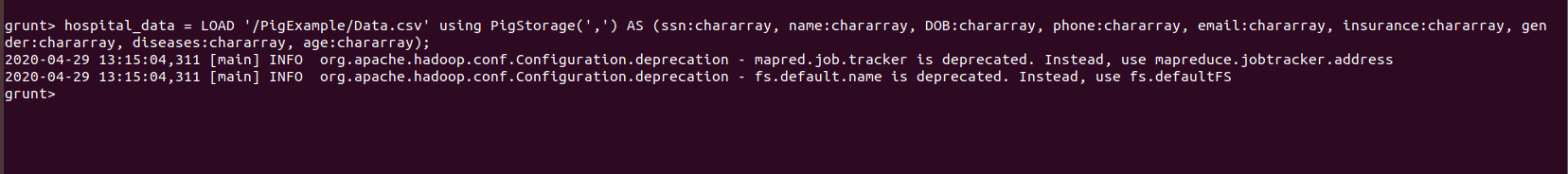
1. **Define Dataset**

Start pig using below command in terminal

*pig*

Now need to define column by column meaning of the uploaded healthcare data using below command.

> *hospital\_data = LOAD '/Uploaded\_FileName' using PigStorage(',') AS (ssn:chararray, name:chararray, DOB:chararray, phone:chararray, email:chararray,insurance: chararray, gender: chararray, diseases: chararray, age: chararray );*

> *dump employee;*

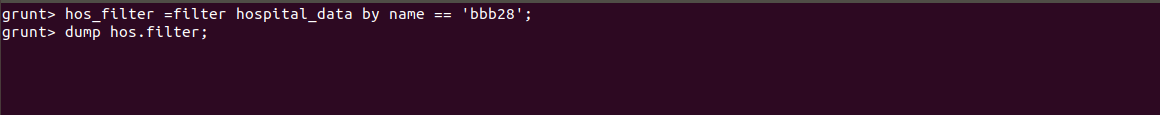
1. **Operations**

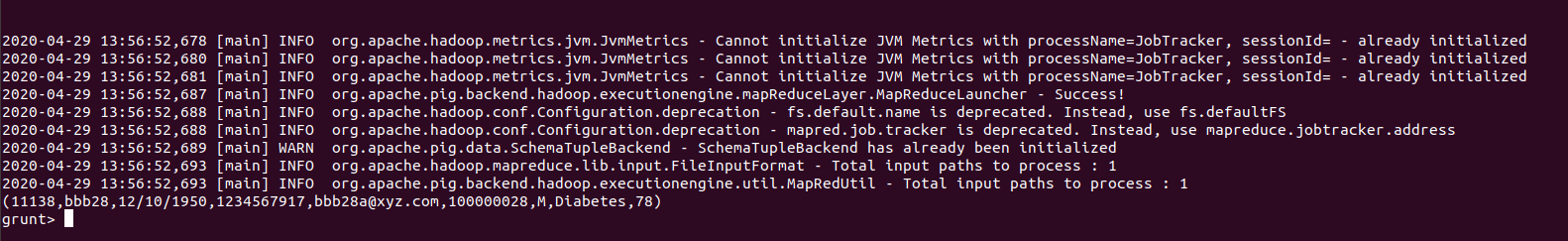
Filter:

Create a variable and use below filter command. Here filtering is used to get information using name of a specific person.

*> hos\_filter =filter hospital\_data by name == 'bbb28';*

*> dump hos\_filter;*

**

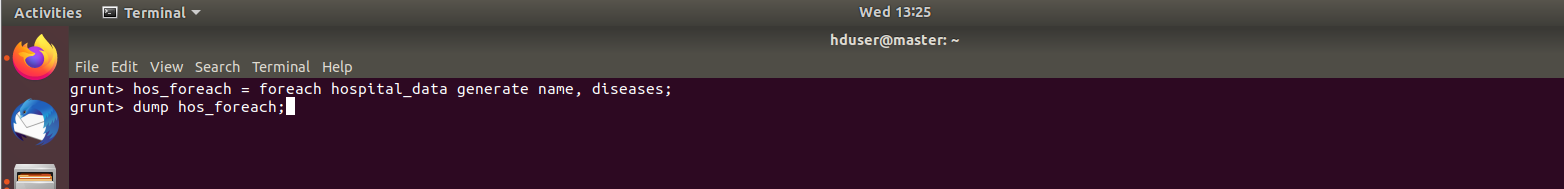
**

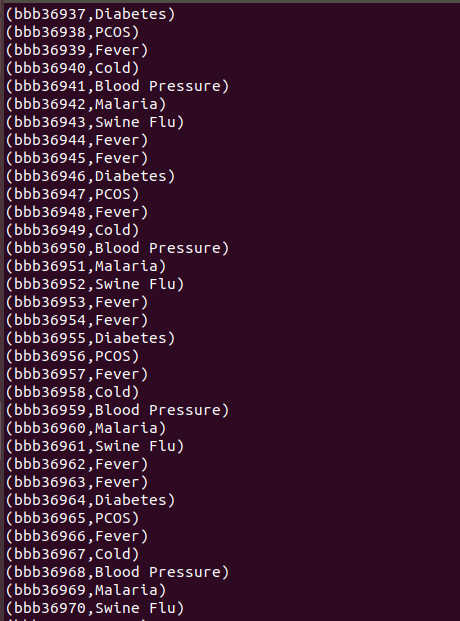
Printing only Specific information:

Again create a new variable and give below commands. In this particular example, it’ll print all names of the patient and their respective diseases.

*>hos\_foreach = foreach employee generate name, diseases;*

*>dump hos\_foreach;*

**

**

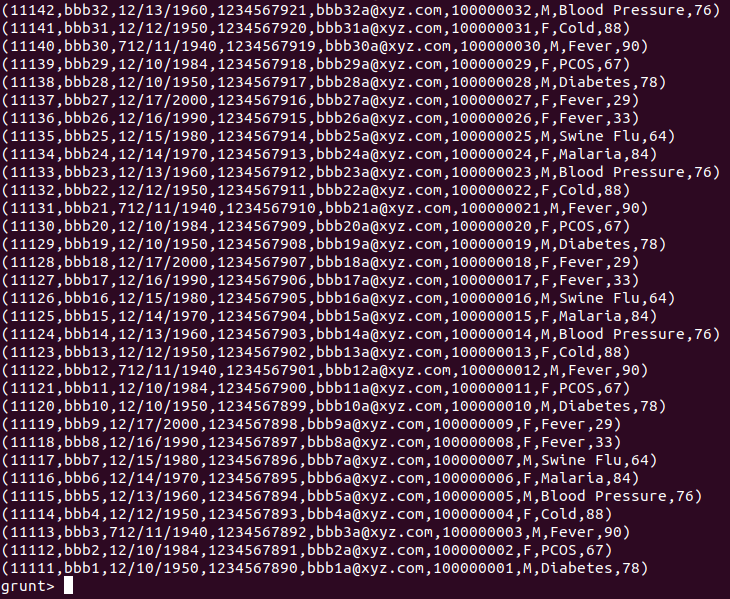
Ordering Data:

Again define another variable and give below commands in order to print data according descending order of ssn number.

*>hos\_order = order hospital\_data by ssn desc;*

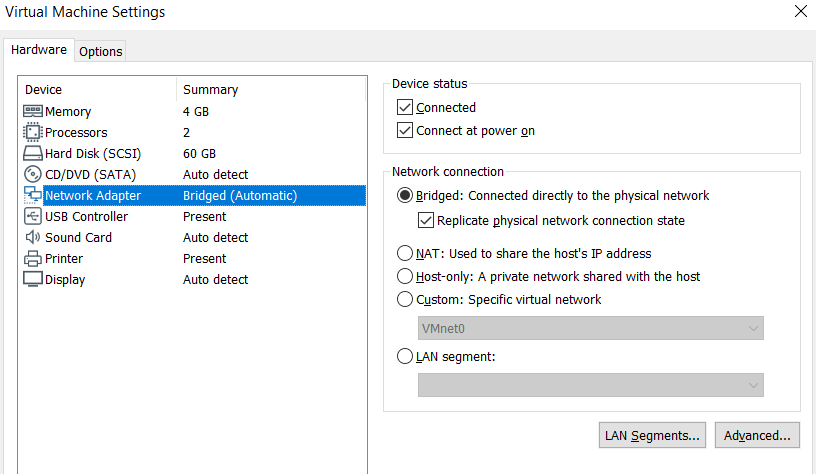
*>dump hos\_order;*





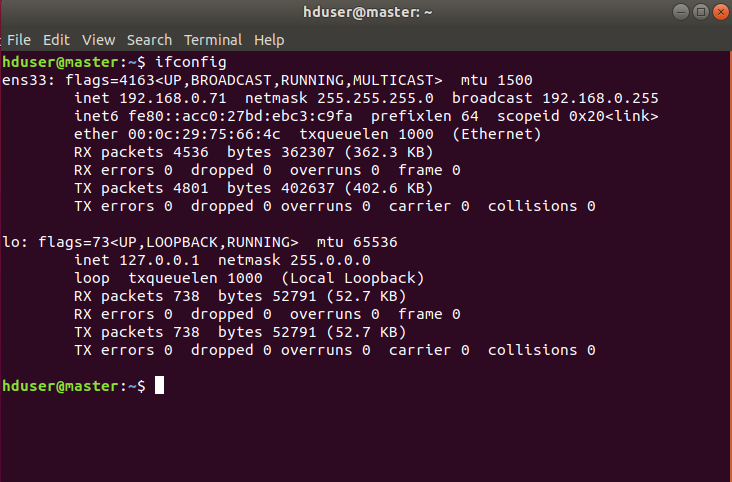
Connecting nodes from different physical devices

Select every configured Hadoop node and go to *settings>Network Adapter* and select ***bridged*** Network connection.



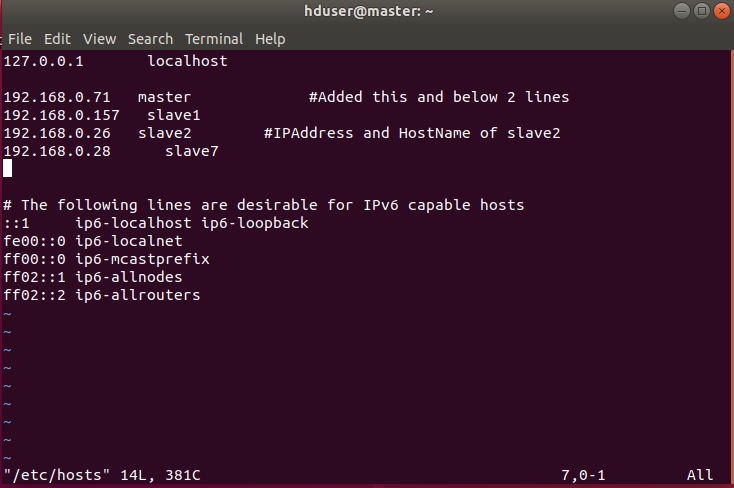
Go to terminal of each node and get all the ip addresses by using below command

*$ ifconfig*



Go to every node , give below command and put all nodes ip addresses and host name there.

*$ sudo vim /etc/hosts*



Save it and try to ping every node. If ping is ok then start the dfs and yarn.