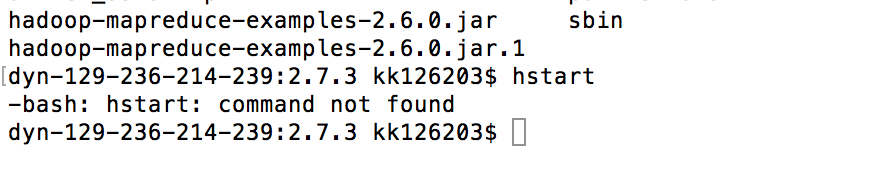
Big Data Analytics

UNI : kc3051

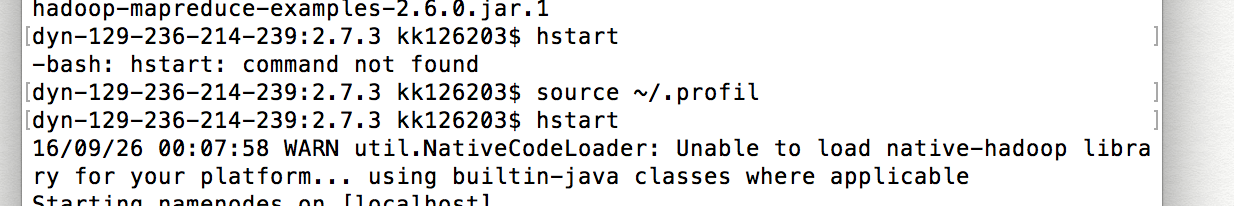
Name: Kuan-Sheng Chen

1. **Install Hadoop**

I followed the instructions step by step, except for the format problems resulted from my direct copy of the PDF (ex. -, whitespace), everything went smoothly. Since I am no very familiar with UNIX system, I spent a lot of time searching for the meanings of the commands, such as *alias, wget, ssh*.

I met the first annoying problem in step8 (start Hadoop). No matter how hard I tried the Hadoop just didn’t appear and the terminator keep showing “*unknown command*”.

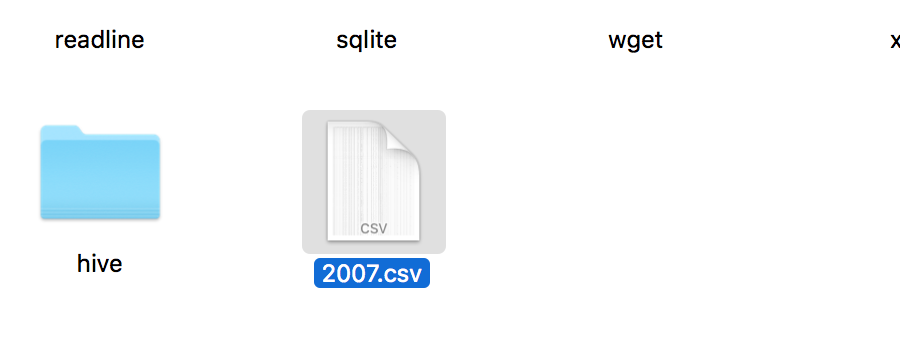
As soon as I went to the piazza, looking for other students’ questions, I found the solution is to type in “*source ~/.profil*”, and the hadoop just starts working!



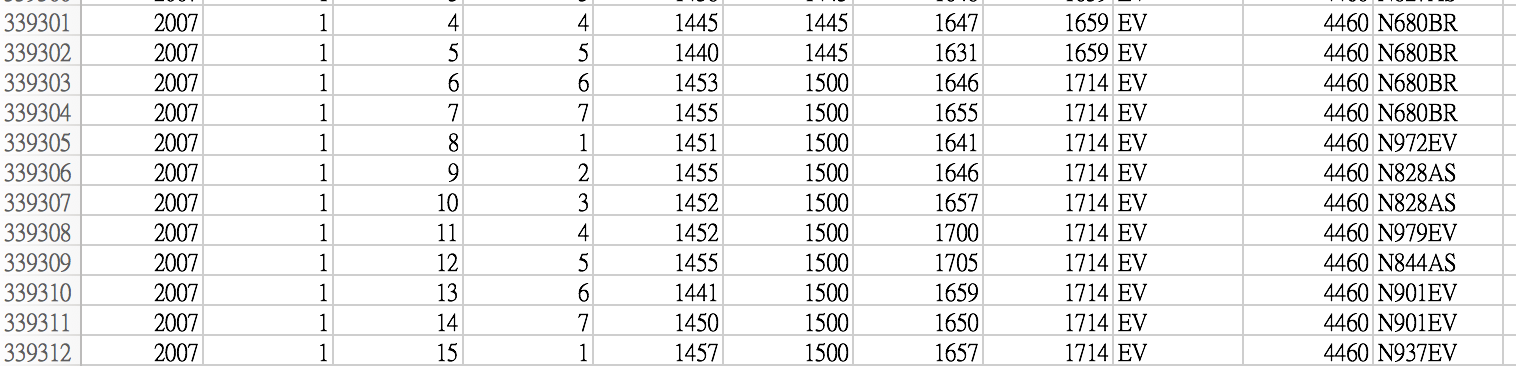
In addition, other problems I met when installing were also resolved after I looked for other students questions in piazza. For example, I couldn’t find the *NameNode* when I enter *jps* in Hadoop. The solution is to type in *hstop* and *hstart* to restart Hadoop (provided by professor). What’s more, my HDFS directory always disappeared when I needed it, and after typing in “*hdfs dfs -put /sourcefile /destination-file-location-on-hdfs*” (provided by the instructor), it was no longer a problem for me.

1. **Download Airline Data and one your own selected dataset from**

I downloaded the data of 2007 and put in the same directory of Hive.



and this in one part of the data of 2007’s AriLine



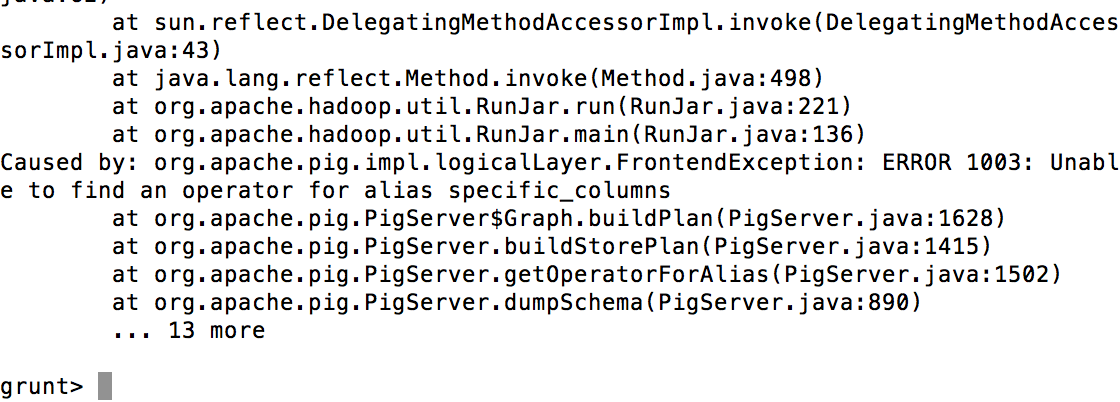
1. **Learn to use PIG. You can try the example in the reference.**

I installed PIG step by step and input the data in the tutorial, then describe its output. I picked three output data of them.

2b)



2c)



2e) The calculation performed using GROUP BY



1. Try HBase .Use your own example.
2. Try Hive. Use your own example.