Mini Project: 10 (+1) points

DESCRIPTION

In this (LAST!) mini project, you are going to use **Apache Pig** (ie. write a Pig Latin script) to count letters (not words!) in multiple input text files, using the HortonWorks Hadoop Sandbox setup running inside a VirtualBox VM.

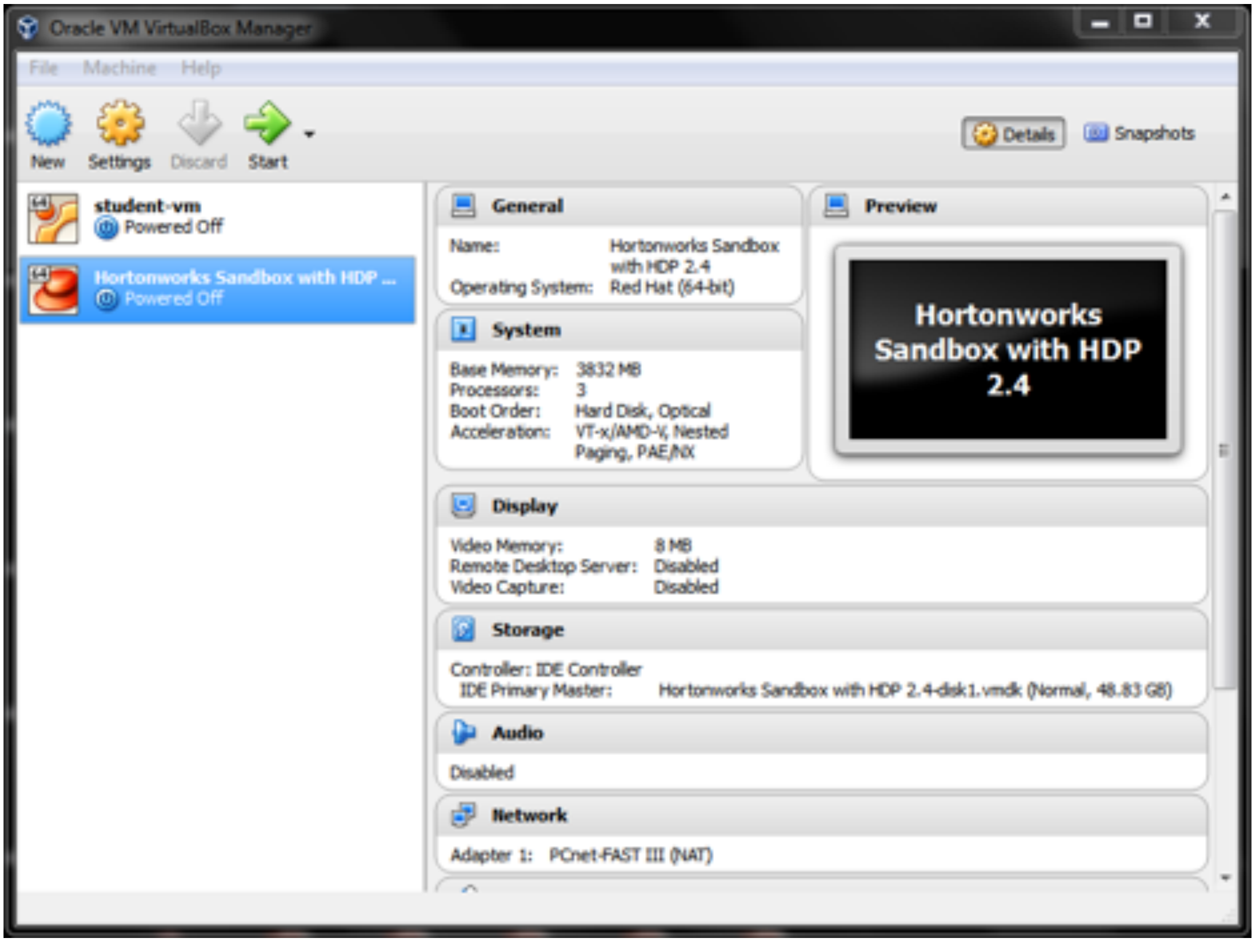
What you need to submit:

* charCount.pig - your Pig script that does the letter counting
* additional files, if any [eg. if you wrote a custom 'UDF', or if you created extra input files, submit those]

This is an extremely useful exercise, because it makes you learn the basics of Pig, a highly expressive and compact dataflow language. Using Pig, you can write SQL-like programs that get AUTOMATICALLY set up for mapping and reducing in a cluster! In other words, rather than write verbose mapper and reducer classes in Java and compiling them prior to running, you can write much smaller, equivalent Pig scripts (about 1/20th LOC!) that you can run without having to compile - this brings you the power of MapReduce, minus the heavy coding. Fun..

Following are the steps.

1. Install VirtualBox (NOT VMWare!).
2. Download the HortonWorks Hadoop Sandbox. If you are using a PC, download WinSCP as well.
3. Start the VM, and import the Hadoop Sandbox appliance:



General

Name: Hortonworks Sandbox with HDP 2.4
Operating System: Red Hat (64-bit)

System

Base Memory: 3832 MB
Processors: 3
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

Display

Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage

Controller: IDE Controller
IDE Primary Master: Hortonworks Sandbox with HDP 2.4-disk1.vmdk (Normal, 48.83 GB)

Audio

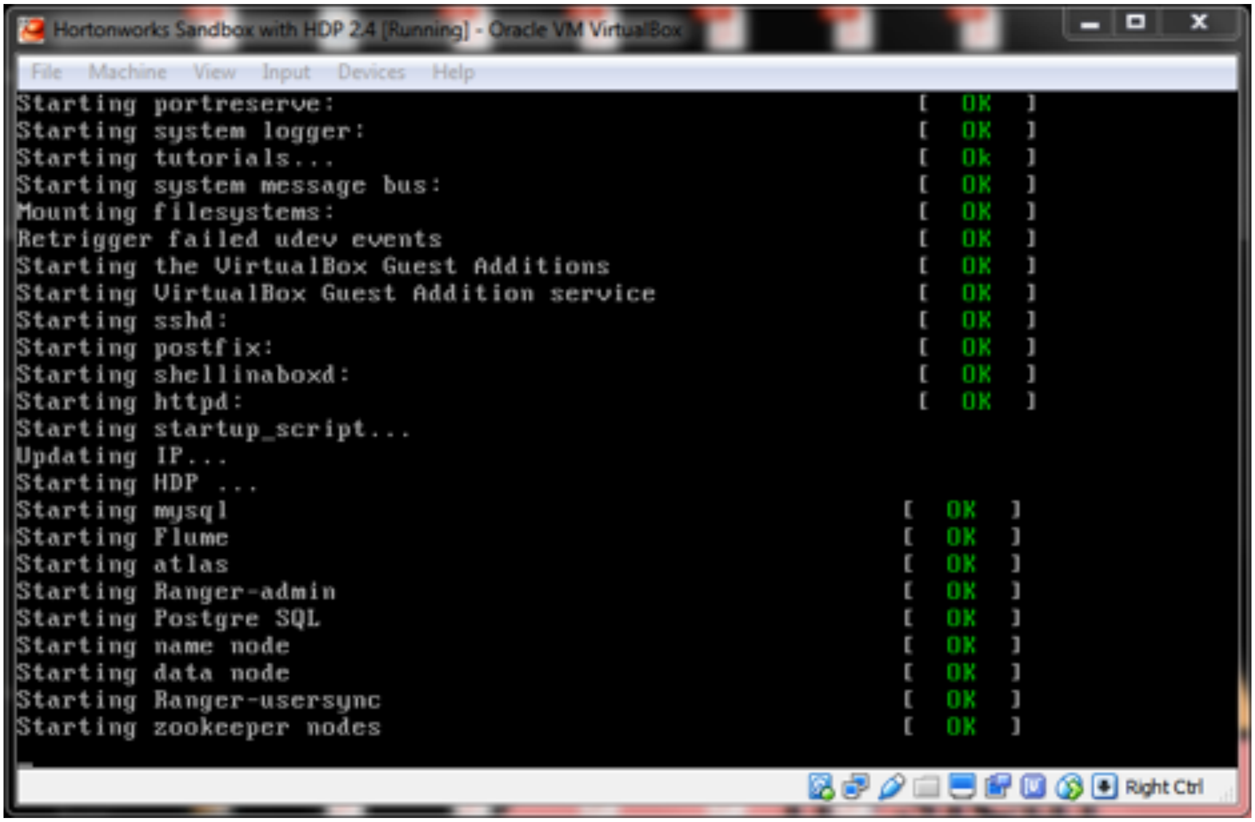
Disabled

Network

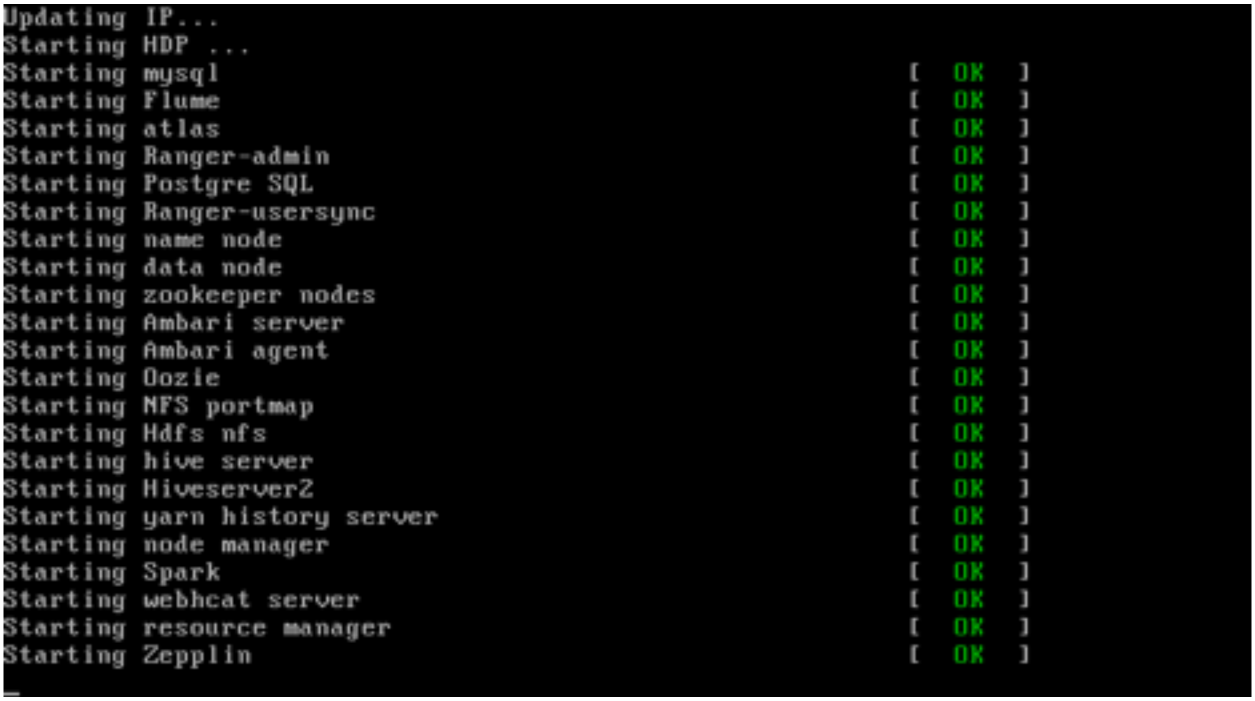
Adapter 1: PCnet-FAST III (NAT)

Here is a useful guide that will help with the setting up of the sandbox.

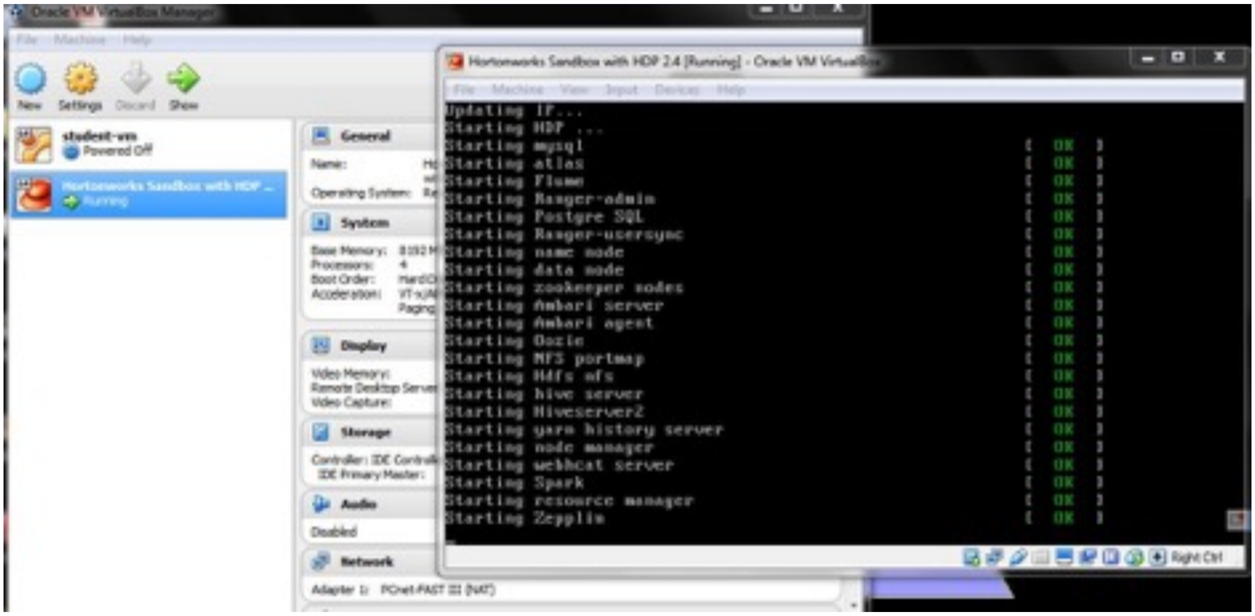
4. Bring up the sandbox (press the Start button on the VM) - this will take a few minutes:


Hortonworks Sandbox with HDP 2.4 [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Right Ctrl

```
Starting portreserve:                                       [  OK  ]
Starting system logger:                                     [  OK  ]
Starting tutorials...                                       [  Ok  ]
Starting system message bus:                                [  OK  ]
Mounting filesystems:                                       [  OK  ]
Retrigger failed udev events                                [  OK  ]
Starting the VirtualBox Guest Additions                     [  OK  ]
Starting VirtualBox Guest Addition service                  [  OK  ]
Starting sshd:                                              [  OK  ]
Starting postfix:                                           [  OK  ]
Starting shellinaboxd:                                      [  OK  ]
Starting httpd:                                             [  OK  ]
Starting startup_script...
Updating IP...
Starting HDP ...
Starting mysql                                             [  OK  ]
Starting Flume                                             [  OK  ]
Starting atlas                                             [  OK  ]
Starting Ranger-admin                                      [  OK  ]
Starting Postgre SQL                                       [  OK  ]
Starting name node                                         [  OK  ]
Starting data node                                         [  OK  ]
Starting Ranger-usersync                                   [  OK  ]
Starting zookeeper nodes                                   [  OK  ]
_
```



```
Updating IP...
Starting HDP ...
Starting mysql                                              [  OK  ]
Starting Flume                                              [  OK  ]
Starting atlas                                              [  OK  ]
Starting Ranger-admin                                       [  OK  ]
Starting Postgre SQL                                        [  OK  ]
Starting Ranger-usersync                                    [  OK  ]
Starting name node                                          [  OK  ]
Starting data node                                          [  OK  ]
Starting zookeeper nodes                                    [  OK  ]
Starting Ambari server                                      [  OK  ]
Starting Ambari agent                                       [  OK  ]
Starting Oozie                                              [  OK  ]
Starting NFS portmap                                        [  OK  ]
Starting Hdfs nfs                                           [  OK  ]
Starting hive server                                        [  OK  ]
Starting Hiveserver2                                        [  OK  ]
Starting yarn history server                                [  OK  ]
Starting node manager                                       [  OK  ]
Starting Spark                                              [  OK  ]
Starting webhcat server                                     [  OK  ]
Starting resource manager                                   [  OK  ]
Starting Zepplin                                            [  OK  ]
_
```



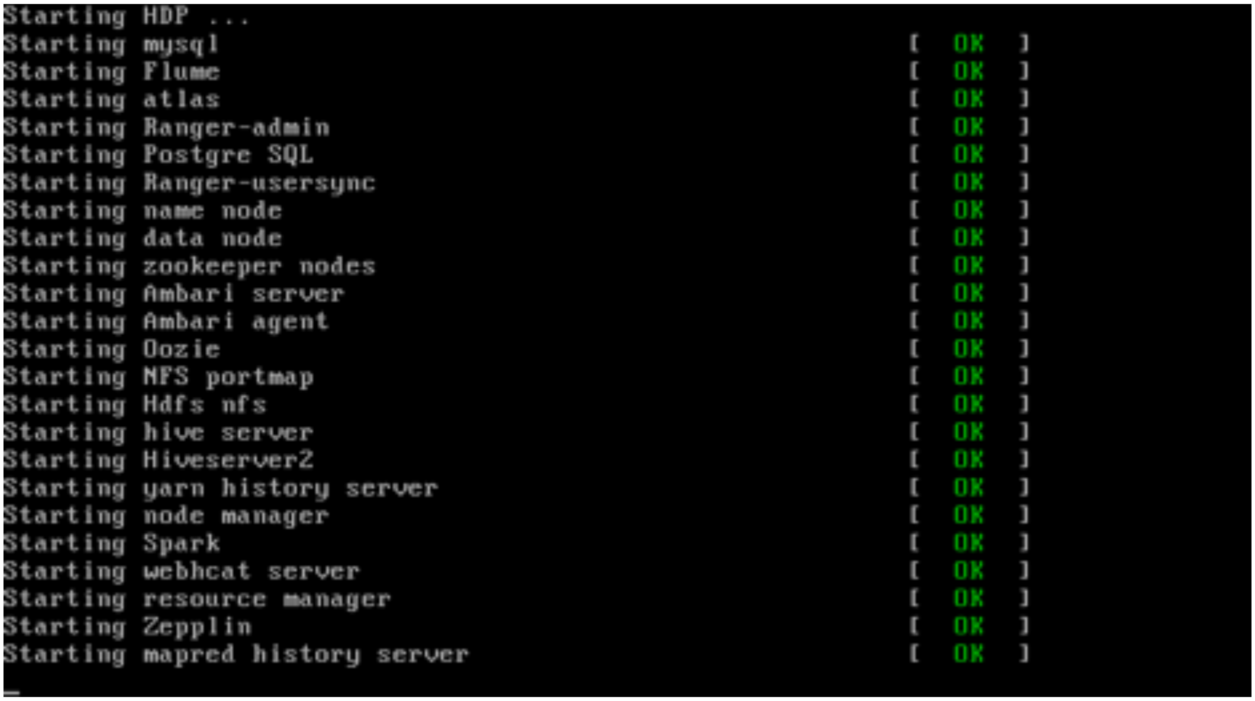
Base Memory: 8192 MB

Processors: 4

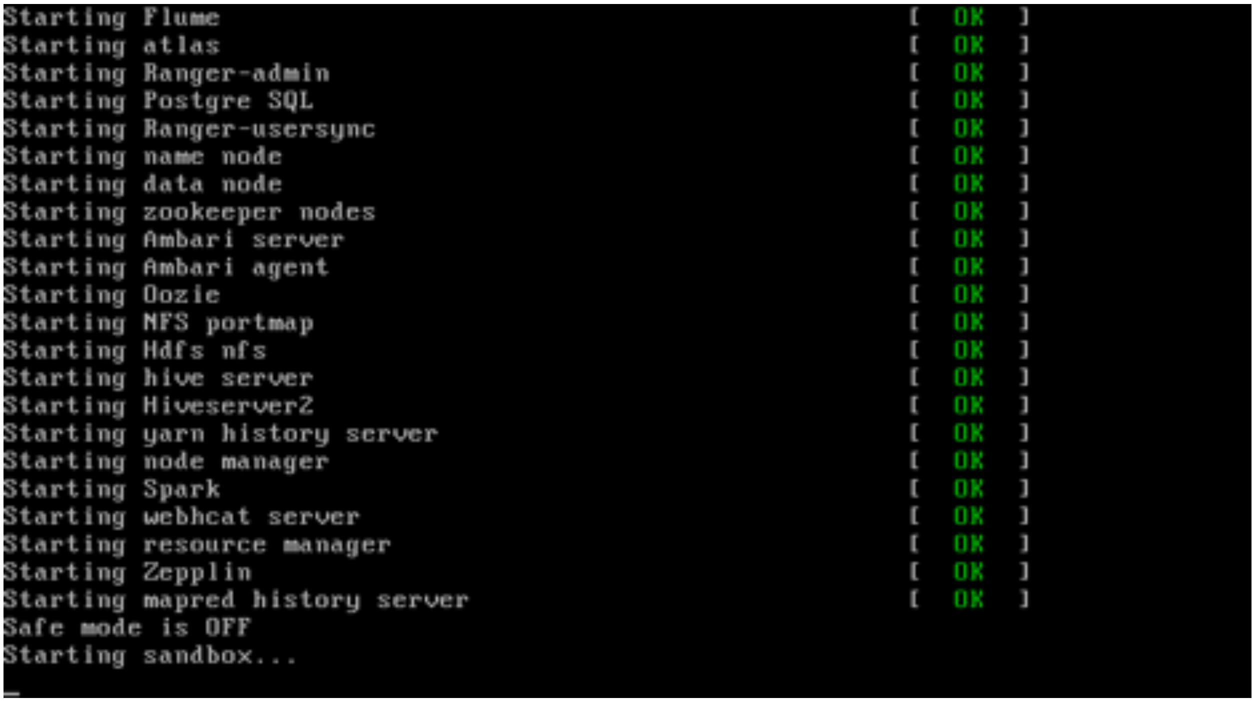
Audio: Disabled

Adapter 1: PCnet-FAST III (NAT)

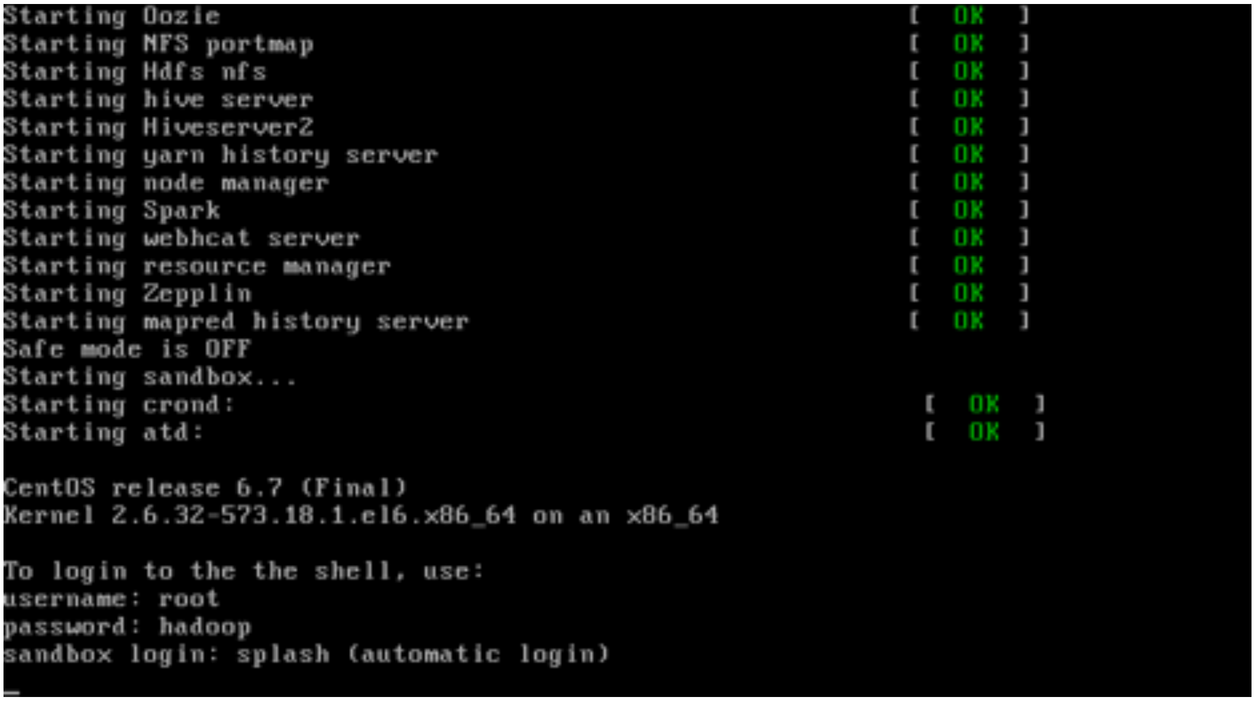
```
Updating IP...
Starting HDP ...
Starting mysql                                                    [  OK  ]
Starting atlas                                                    [  OK  ]
Starting Flume                                                    [  OK  ]
Starting Ranger-admin                                             [  OK  ]
Starting Postgre SQL                                              [  OK  ]
Starting Ranger-usersync                                          [  OK  ]
Starting name node                                                [  OK  ]
Starting data node                                                [  OK  ]
Starting zookeeper nodes                                          [  OK  ]
Starting Ambari server                                            [  OK  ]
Starting Ambari agent                                             [  OK  ]
Starting Oozie                                                    [  OK  ]
Starting NFS portmap                                              [  OK  ]
Starting Hdfs nfs                                                 [  OK  ]
Starting hive server                                              [  OK  ]
Starting Hiveserver2                                              [  OK  ]
Starting yarn history server                                      [  OK  ]
Starting node manager                                             [  OK  ]
Starting webhcat server                                           [  OK  ]
Starting Spark                                                    [  OK  ]
Starting resource manager                                         [  OK  ]
Starting Zepplin                                                  [  OK  ]
_
```



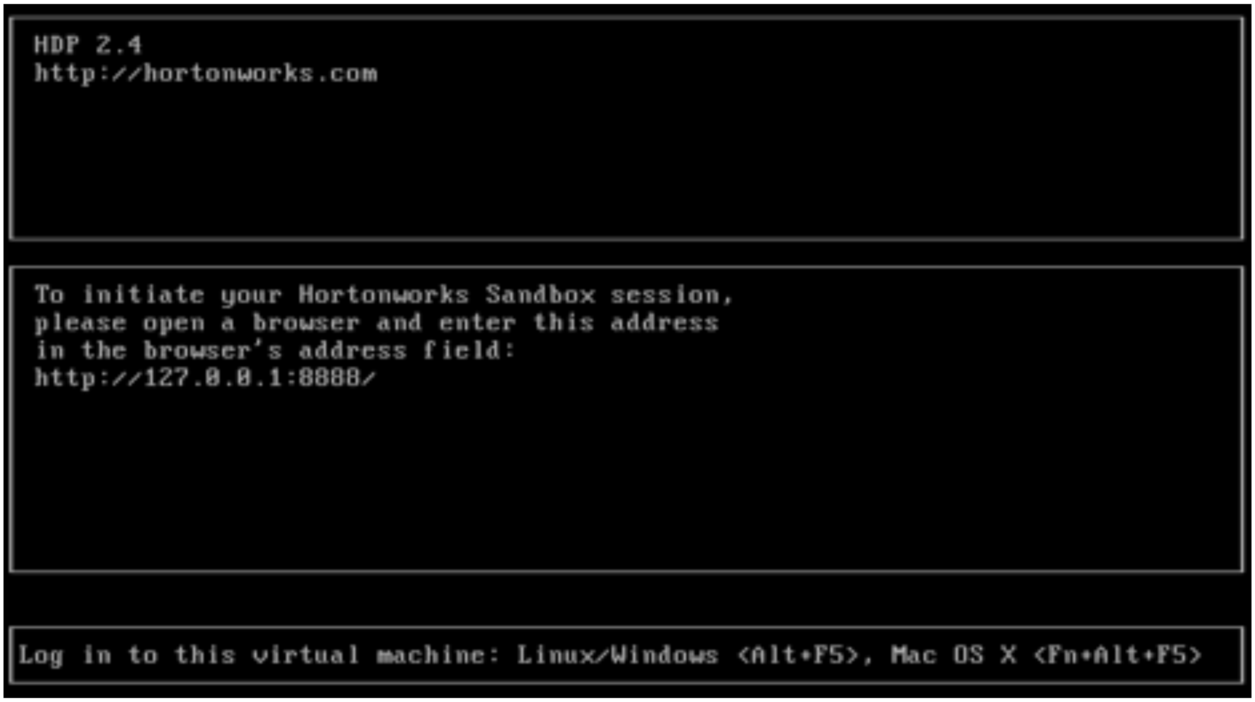
```
Starting HDP ...
Starting mysql                                                    [  OK  ]
Starting Flume                                                    [  OK  ]
Starting atlas                                                    [  OK  ]
Starting Ranger-admin                                             [  OK  ]
Starting Postgre SQL                                              [  OK  ]
Starting Ranger-usersync                                          [  OK  ]
Starting name node                                                [  OK  ]
Starting data node                                                [  OK  ]
Starting zookeeper nodes                                          [  OK  ]
Starting Ambari server                                            [  OK  ]
Starting Ambari agent                                             [  OK  ]
Starting Oozie                                                    [  OK  ]
Starting NFS portmap                                              [  OK  ]
Starting Hdfs nfs                                                 [  OK  ]
Starting hive server                                              [  OK  ]
Starting Hiveserver2                                              [  OK  ]
Starting yarn history server                                      [  OK  ]
Starting node manager                                             [  OK  ]
Starting Spark                                                    [  OK  ]
Starting webhcat server                                           [  OK  ]
Starting resource manager                                         [  OK  ]
Starting Zepplin                                                  [  OK  ]
Starting mapred history server                                    [  OK  ]
_
```



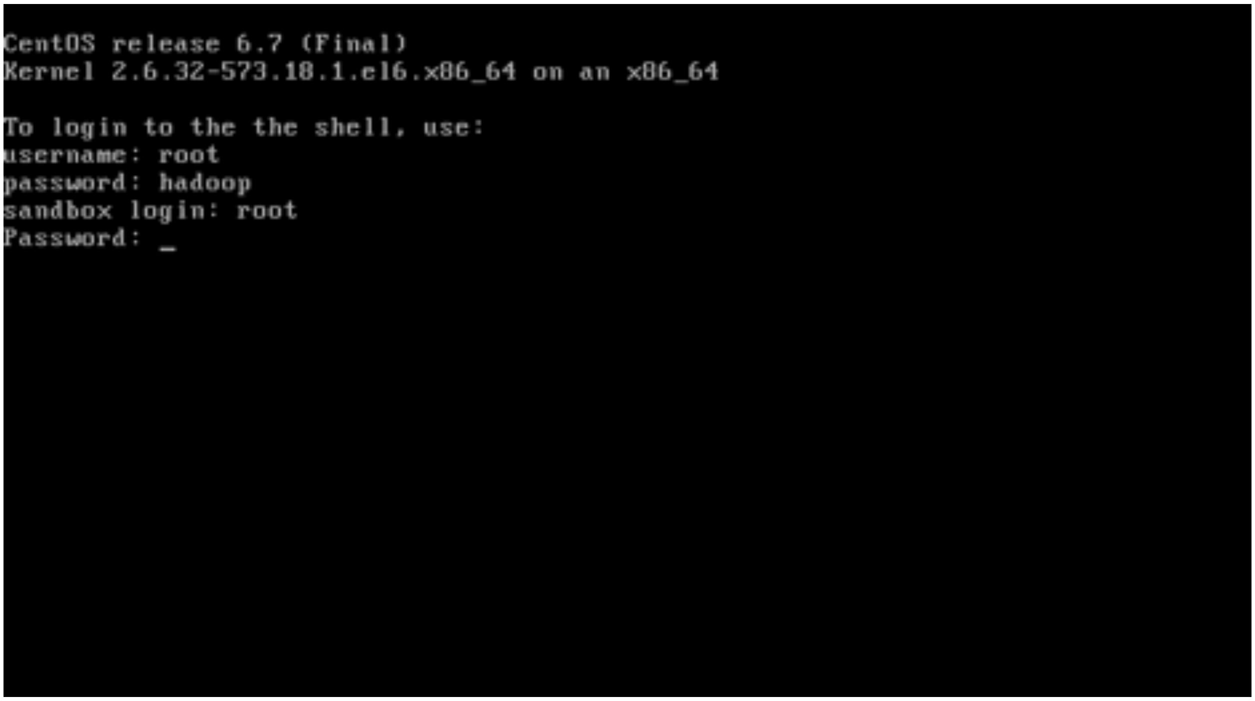
```
Starting Flume                                              [  OK  ]
Starting atlas                                              [  OK  ]
Starting Ranger-admin                                       [  OK  ]
Starting Postgre SQL                                        [  OK  ]
Starting Ranger-usersync                                    [  OK  ]
Starting name node                                          [  OK  ]
Starting data node                                          [  OK  ]
Starting zookeeper nodes                                    [  OK  ]
Starting Ambari server                                      [  OK  ]
Starting Ambari agent                                       [  OK  ]
Starting Oozie                                              [  OK  ]
Starting NFS portmap                                        [  OK  ]
Starting Hdfs nfs                                           [  OK  ]
Starting hive server                                        [  OK  ]
Starting Hiveserver2                                        [  OK  ]
Starting yarn history server                                [  OK  ]
Starting node manager                                       [  OK  ]
Starting Spark                                              [  OK  ]
Starting webhcat server                                     [  OK  ]
Starting resource manager                                   [  OK  ]
Starting Zepplin                                            [  OK  ]
Starting mapred history server                              [  OK  ]
Safe mode is OFF
Starting sandbox...
_
```



```
Starting Oozie                                              [  OK  ]
Starting NFS portmap                                        [  OK  ]
Starting Hdfs nfs                                           [  OK  ]
Starting hive server                                        [  OK  ]
Starting Hiveserver2                                        [  OK  ]
Starting yarn history server                                [  OK  ]
Starting node manager                                       [  OK  ]
Starting Spark                                              [  OK  ]
Starting webhcat server                                     [  OK  ]
Starting resource manager                                   [  OK  ]
Starting Zepplin                                            [  OK  ]
Starting mapred history server                              [  OK  ]
Safe mode is OFF
Starting sandbox...
Starting crond:                                              [  OK  ]
Starting atd:                                                [  OK  ]

CentOS release 6.7 (Final)
Kernel 2.6.32-573.18.1.el6.x86_64 on an x86_64

To login to the the shell, use:
username: root
password: hadoop
sandbox login: splash (automatic login)
_
```



```
HDP 2.4
http://hortonworks.com

To initiate your Hortonworks Sandbox session,
please open a browser and enter this address
in the browser's address field:
http://127.0.0.1:8888/

Log in to this virtual machine: Linux/Windows <Alt+F5>, Mac OS X <Fn+Alt+F5>
```

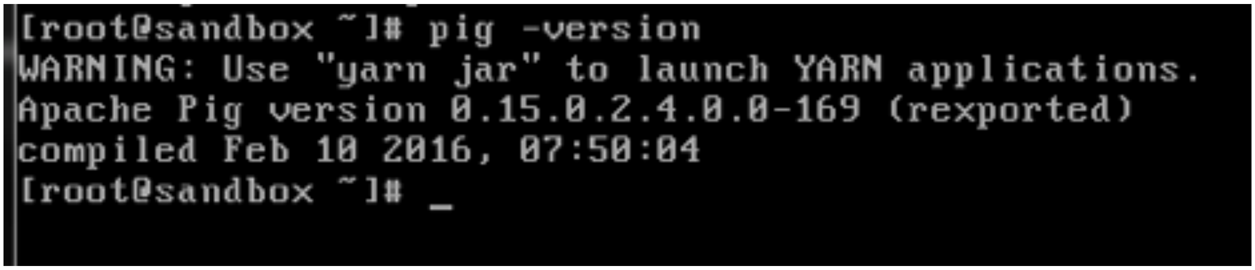


```
CentOS release 6.7 (Final)
Kernel 2.6.32-573.18.1.el6.x86_64 on an x86_64

To login to the the shell, use:
username: root
password: hadoop
sandbox login: root
Password: _
```

5. Once the sandbox shell (terminal) comes up, you can start to play! You can type in standard Unix commands (ls, rm, mkdir, more..), and use bash-like editing (Ctrl p,

Ctrl b, Ctrl d etc.). You can also run from this terminal, Hadoop map-reduce commands, Spark, Hive, Pig, etc. - LOTS of power!

Verify that Pig runs:



```
[root@sandbox ~]# pig -version
WARNING: Use "yarn jar" to launch YARN applications.
Apache Pig version 0.15.0.2.4.0.0-169 (rexported)
compiled Feb 10 2016, 07:50:04
[root@sandbox ~]# _
```

6. Learn (the basics of) Pig, start playing with it. You can directly run Pig commands in the shell, or bring up the Pig-specific 'grunt' shell and run commands inside it (I recommend not using 'grunt').

At the bottom of this page are PLENTY of resources for learning Pig - you'll become proficient by reading up, typing in commands, observing the results, reading, running, examining.. Be sure to allow time for this.

7. **The goal is to write a small script called countChars.pig** - you'll do all the

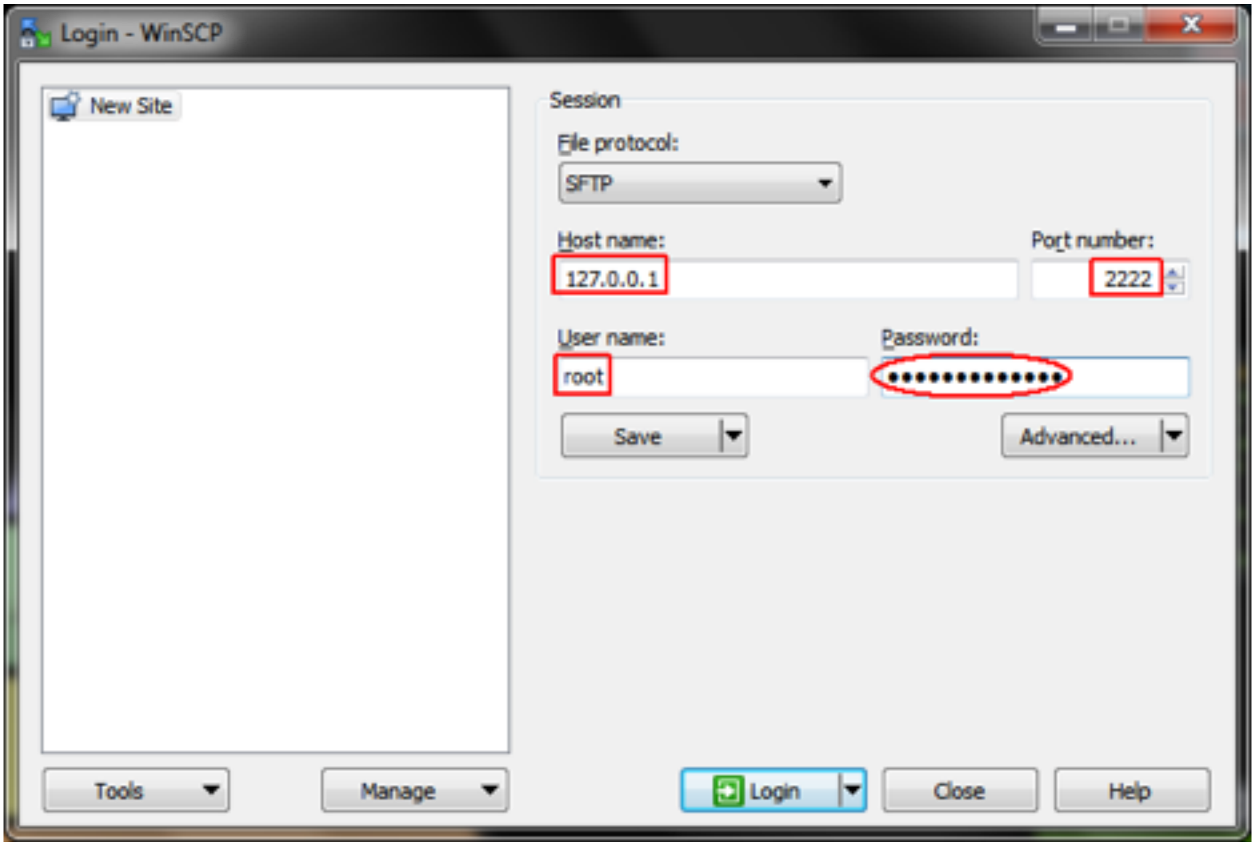
script typing on your own machine, transfer (upload) the script to the sandbox using 'scp' (Mac/Linux) or WinSCP (PC), and run it in the sandbox. Note: with WinSCP, you can drag and drop files and folders from your PC to the sandbox, and from the sandbox back to your PC.

Download the following 6 files (para1.txt through para6.txt), to use as 'official' inputs for your script - these are the same ones we used in class, for the MapReduce activity:

* para1.txt
* para2.txt
* para3.txt
* para4.txt
* para5.txt
* para6.txt

Transfer (scp) your .pig script plus the 6 input files, to the sandbox. The input files can be placed in an 'in' directory on the sandbox, to keep the sandbox clean (do 'mkdir in' in your sandbox, to create the

folder). On a PC, bring up WinSCP, and log on to the sandbox in order to transfer files:
* host: 127.0.0.1
* port: 2222
* user: root
* password: < whatever you picked when you were asked to change the default password 'hadoop' >



File protocol: SFTP
Host name: 127.0.0.1
Port number: 2222
User name: root
Password: ••••••••••••••

Run your program, debug, make changes to the script, upload, run, debug..

Below are a pair of clips that show my uploading and running countChars.pig, on four tiny input files named p1.txt, p2.txt, p3.txt, p4.txt - together these four text files contain 'The quick brown fox jumps over the lazy dog', which is a sentence that is special because it has all letters of the alphabet :)

download [right-click to save]

download [right-click to save]

Success! As you can tell, the output file contained counts for all 26 letters. Note that I had specified 'charcount' to be the output directory, in my .pig script. If you too specify a directory for output (recommended), make sure this directory **does not pre-exist** when you run your script! If it does, you'll get an error when your script runs.

Use 'rm -rf ' to remove the output directory and its contents, each time before running countChars.pig.

The letter counting should IGNORE case - so 'Tutti Frutti' for example would produce 5 for the 't' (or 'T') count, not 4.

That was a lot of output from the running processes! The underlying YARN manager takes our .pig script, parses it, and automagically spawns a series of mappers and reducers to run the Pig commands, where possible in parallel. Cool!

Note the syntax for executing a Pig script: 'pig -x local countChars.pig'. Local execution (ie. in the sandbox) is simpler than executing in the HDFS (Hadoop file system), something which you can learn later (you need to copy the inputs and your .pig script to HDFS, then run the script using 'pig -x mapreduce').

Next, I upload the 'official' para?.txt files and a slightly modified countChars.pig that points to para?.txt, and execute the script:

download [right-click to save]

download [right-click to save]

download [right-click to save]

Again, success! You can see the total letter counts for all the letters in the six input paragraphs.

Note - your submittable (.pig) would be as small as just 6 lines (one or two more lines if you attempt the bonus question below)!! Do allocate enough time, though, to experiment with Pig commands and data flow, that is how you will arrive at the solution. Translation: do not put off working

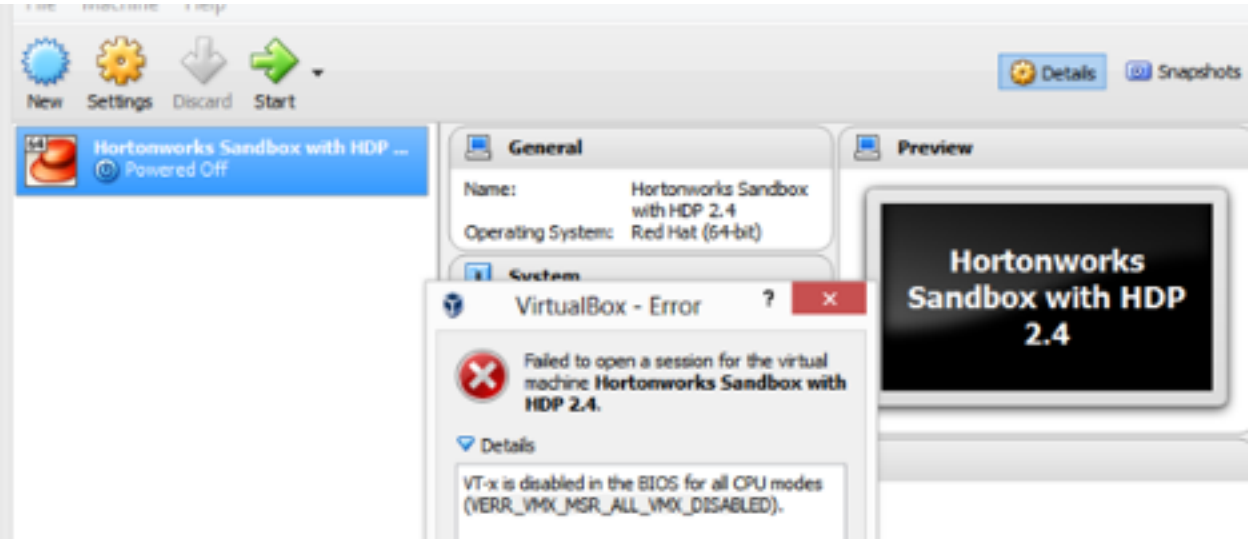
on this because there doesn't seem much to type in.
Tip: 'dump' and 'describe' are VERY useful Pig commands to add to your code, they are great debugging aids.

BONUS QUESTION [1 point]
Modify countChars.pig so that it only outputs totals for the vowels, ie. for a, e, i, o, u; submit it as countChars_Bonus.pig. Feel free to create any extra files you might need for this - if you do so, submit these extra files as well.

RESOURCES [15 links!]
Below are links to official Pig documentation, a paper, a tutorial, and to several 'word count' Pig scripts as well - even though you won't be writing code for counting words in this assignment, these will help you get famililar with Pig commands, syntax and workflow.
* a paper on Pig - explains the overall idea and syntax

* a tutorial - good way to get your feet wet
* open dir containing docs for various Pig releases - eg. look in https://pig.apache.org/docs/r0.15.0/
* Pig Cookbook
* Pig functions - VERY useful!!
* 'wordcount' links:
** http://salsahpc.indiana.edu/ScienceCloud/pig_word_count_tutorial.htm
** http://www.folkstalk.com/2013/09/word-count-example-pig-script.html
** https://gist.github.com/tomgullo/186460
** https://amalgjose.wordpress.com/2014/05/31/simple-word-count-using-apache-pig/
** http://www.hadooplessons.info/2015/01/word-count-in-pig-latin.html
** http://stackoverflow.com/questions/17951375/what-exactly-am-i-doing-wrong-with-my-wordcount-program-pig
** https://www.pluralsight.com/blog/tutorials/pig-vs-java-mapreduce
** https://groups.google.com/forum/#!topic/seattle-scale/fbv61LXqO3c
** https://www.ibm.com/support/knowledgecenter/SSGSMK_7.1.1/mapreduce_integration/map_reduce_pig_apps.dita
** https://blog.pivotal.io/pivotal/products/hadoop-101-programming-mapreduce-with-native-libraries-hive-pig-and-cascading

UPDATE, 4/20: If you want you can use a UDF [written in the language of your choosing] to do the assignment - both for the regular part and the extra credit part.
UPDATE, 4/21: If you are unable to run the HortonWorks Sandbox, as an alternative, you can run your .pig scripts inside Amazon's AWS cloud; note that it involves signing up for an Amazon AWS account (free) - please be sure to TERMINATE any running jobs before you log out of AWS!
UPDATE, 4/24: Are you getting the following 'Failed to open...' error on a PC, with the HortonWorks Sandbox on VirtualBox?



Name: Hortonworks Sandbox with HDP 2.4
Operating System: Red Hat (64-bit)

Failed to open a session for the virtual machine **Hortonworks Sandbox with HDP 2.4**.

VT-x is disabled in the BIOS for all CPU modes (VERR_VMX_MSR_ALL_VMX_DISABLED).

Go to https://support.lenovo.com/us/en/documents/ht081446, look for the virtualization setting called "virtualization", "VT-x" or "AMD-V" and enable it - you need to turn on virtualization at the hardware (BIOS) level [the error is on account of this setting being 'off'].