Cloud Computing: Lab Exercise 3

Look at the Quick Start for Spark
`http://spark.apache.org/docs/1.6.1/quick-start.html`

To install spark:

1. Start your Redhat 7 instance with HTTP, HTTPS, and SSH permissions.
2. Get the Cloudera repository:

```
sudo yum install  wget
wget https://archive.cloudera.com/cdh5/redhat/7/x86_64/cdh/cloudera-cdh5.repo
sudo mv cloudera-cdh5.repo /etc/yum.repos.d/
```

3. `sudo yum install hadoop-yarn-resourcemanager`
4. `sudo yum install hadoop-client`
5. `sudo yum install spark-core spark-history-server spark-python`
6. `sudo yum install java-1.8.0-openjdk`
7. Use *pyspark* to run spark.

Modify the `wordcounty.py` sample to obtain word counts on the literary texts from Shakespeare, Dickens and Wilde. Compare the relative frequency of words. `wordcount.py` can be found at `https://github.com/apache/spark/blob/branch-1.6/examples/src/main/python/wordcount.py`. To help understand how Spark can use SQL look at `https://github.com/apache/spark/blob/branch-1.6/examples/src/main/python/sql.py`. You will want to look into and understand the difference between Spark dataframes and RDD.

If you get an insufficient memory error from the Spark executor, you can either launch a different wirtual machine with more main memory, or edit the file
`/etc/spark/conf/spark-defaults.conf` adding the following

```
spark.driver.memory     640m
```

You can modify the value 640 to the largest possible size that still allows the application to run on your virtual machine.