Part 5:

- Following instructions from http://blog.cloudera.com/blog/2014/04/how-to-run-a-simple-apache-spark-app-in-cdh-5/
- I checkout sample repo from https://github.com/sryza/simplesparkapp
- I imported it into eclipse using File > import > maven project
- Eclipse then installed some plugins for scala and spark
- Used maven to build the package:
  mvn clean package
- Copied sample input file to hdfs
  sudo -u hdfs hadoop fs -copyFromLocal data/inputfile.txt /user/hive/warehouse/spark_samplewc
- Submit the job to spark using:
  spark-submit --class com.cloudera.sparkwordcount.SparkWordCount --master local target/sparkwordcount-0.0.1-SNAPSHOT.jar /user/hive/warehouse/spark_samplewc 2
- The output now contains:
  (e,6), (p,2), (a,4), (t,2), (i,1), (b,1), (u,1), (h,1), (o,2), (n,4), (f,1), (v,1), (r,2), (l,1), (c,1)

  Which is the expected output as per the tutorial.

Part 6: Analyze apache log file:

- Used CDH sample apache log file /opt/examples/log_files/access.log.2
- Moved it to HDFS
  sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/apache_access_logs
  sudo -u hdfs hadoop fs -copyFromLocal /opt/examples/log_files/access.log.2 /user/hive/warehouse/apache_access_logs/
- My idea is to find relative frequencies of products based on ip address of users, so if users usually buy product x after y,I would like to know and graph that.
-