CSE 512 Project Phase 1

Task:

1. Set up your running environment, including installing the Hadoop and Spark. Record a video to show that you have already successfully set up the cluster as required. The video should show the following:
    - You can connect from master to at least one worker by password-less ssh using both public ip and private ip.
    - Open the "master_ public_ip:50070", there should be **three** live nodes.
    - Open the "master_public_ip:8080", there should be **three** workers.

2. You need to write two User Defined Functions ST_Contains and ST_Within in SparkSQL and use them to perform some spatial queries. Detailed instruction is explained in the README in the following Github repository: https://github.com/YuhanSun/CSE512-Project-Phase1-Template.

Submission

1. A video demo less than 5 minutes. Each group should put the video demo on YouTube. The submission will be a single text file including the video link.
2. As required on the Github readme, you need to submit a .zip file, which includes the whole project folder and a .jar file which can be directly run by using ./bin/spark-submit.

Notes:

1. You have to run Apache Spark on a cluster. This means you should have at least three machines or Virtual Machines (One master and two workers). The master should be able to communicate with workers using bi-directional Password-less SSH. Your video demo should clearly demonstrate this point.
2. You can use Amazon EC2 to run your experiment. It provides free trial for new users. The free trial can be valid for a year. But we are not responsible for any fees on EC2.