Lab 10: Getting started with Spark

1. Outline

In this lab, you will generate a density map with New York Taxi data (using the records of January, 2013).

2. Materials

The data and scripts are stored in: /gpfs\_scratch/geog479/lab10

Copy this folder to your home directory:

>> cp -r /gpfs\_scratch/geog479/lab10 ~/

3. Tasks

Task 1:

* Login to ROGER and copy data to your home directory:
  + >> ssh [NetID@roger-login.ncsa.illinois.edu](mailto:NetID@roger-login.ncsa.illinois.edu)
  + >> cp -r /gpfs\_scratch/geog479 ~/
* Login to cg-hm08
  + >> ssh cg-hm08
* Prepare data into HDFS
  + >>hdfs dfs -copyFromLocal trip\_data\_1.csv
* Find the commands in commands.txt
  + Before you do that, complete the code
* View the results
  + Copy the generated density map to your local computer and use ArcGIS to view the results
* Now, view the details in the script
* Play with different spark-submit parameters to see whether there are significant difference

Task 2:

* Make your own HTTP server with Python
* You can either do it with ROGER or you can use the lab computer as host
* To do it in ROGER, you will use remote display with Firefox
* To do it in lab computer, you need to copy the folder “d3\_flows” from ROGER to your own computer.
* The detailed command is in “commands.txt”
* Note that to do it in lab computer, you need to use ArcGIS’s python environment, since Python is not directly installed