In this project, it was required to do spatial hot spot analysis. I had to complete two different hot spot analysis tasks.

1. Hot zone analysis

This task includes performing a range join operation on a rectangle dataset and a point dataset. For each rectangle, the number of points located within the rectangle is obtained. The hotter rectangle means that it includes more points. So, this task is to calculate the hotness of all the rectangles.

2. Hot cell analysis

Description

This task will focus on applying spatial statistics to spatial-temporal big data to identify statistically significant spatial hot spots using Apache Spark. The topic of this task is from ACM SIGSPATIAL GISCUP 2016.

The Problem Definition page is here: [http://sigspatial2016.sigspatial.org/giscup2016/problemLinks to an external site.](http://sigspatial2016.sigspatial.org/giscup2016/problem)

The Submit Format page is here: [http://sigspatial2016.sigspatial.org/giscup2016/submitLinks to an external site.](http://sigspatial2016.sigspatial.org/giscup2016/submit)

Special requirement (different from GIS CUP)

As stated in the Problem Definition page, in this task, you are asked to implement a Spark program to calculate the Getis-Ord statistic of NYC Taxi Trip datasets. We call it "**Hot cell analysis**”.

To reduce the computation power need, these following changes were made:

1. The input will be a monthly taxi trip dataset from 2009 - 2012. For example, "yellow\_tripdata\_2009-01\_point.csv", "yellow\_tripdata\_2010-02\_point.csv".
2. Each cell unit size is 0.01 \* 0.01 in terms of latitude and longitude degrees.
3. I had to use 1 day as the Time Step size. The first day of a month is step 1. Every month has 31 days.
4. I only had to consider Pick-up Location.