DDS Phase 3 : Group name Bits Please

Team Members:

Ishita Pahwa
Palak Anmol
Sahil Sikka
Yatharth Sharma

Algorithm

The algorithm is divided into three parts:

a) Map reduce
b) Calculate getis
c) Sorting the treemap according to getis values in descending order.

1. Mapreduce Phase 1: Creation of cells
2. In mapToPair() parse the input file. As it is the map function we can get few of the fields (columns).
3. From the input file, the fifth column is the longitude, sixth column the latitude and first column the day.
4. Map step: Parse each row from the input file and pass the value to the reduce function.
5. Filter those cells which were outside a fixed geo-envelope (making sure that latitude between 40.5N to 40.9N, longitude between 73.7W to 74.25W).
6. Reduce step: Add up all the journeys in a cell to calculate the attribute value for the cell
7. Intermediate Calculations: After getting values from the reduced step we are passing the valid values to a hashmap where the key is a combination of day#latitude#longitude and the value is the second argument of the tuple.
8. Calculate the mean using mean() and standard deviation using calculateS() for every cell.

9. In calculateGetis(), calculate the getis values for every cell by adding the values of the adjacent cells using the formula

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{\left[n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j} \right)^2 \right]}{n-1}}}$$

[1]

Where x_j is the attribute value for cell j, $w_{i,j}$ is the spatial weight between cell *i* and *j,n* is equal to the total number of cells, and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$$

The total number of cells are also calculated in this function.

10. Top 50 values of getis score were put in a treemap and the corresponding latitude, longitude and day values were stored in the output file.