Ira A. Fulton Schools of Engineering
Arizona State University

CSE 511 - Data Processing at Scale

Project Hotspot Analysis

Requirement

In this part of the project, you are required to do spatial hot spot analysis. In particular,you need to complete two different hot spot analysis tasks.

Hot zone analysis

This task will need to perform a range join operation on a rectangle datasets and a pointdataset. For each rectangle, the number of points located within the rectangle will be obtained. The hotter rectangle means that it includes more points. So this task is to calculate the hotness of all the rectangles.

Hot cell analysis

Description

This task will focus on applying spatial statistics to spatio-temporal big data in order 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/problem

The Submit Format page is here: http://sigspatial2016.sigspatial.org/giscup2016/

submitSpecial requirement (different from GIS CUP)

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

To reduce the computation power need, we made the following changes:

1. The input will be "yellow_trip_sample_100000.csv" dataset
2. Each cell unit size is 0.01 * 0.01 in terms of latitude and longitude degrees.
3. You only need to consider Pick-up Location.
4. We don't use Jaccard similarity to check your answer. However, you don't need toworry about how to decide the cell coordinates because the code template generated cell coordinates. You just need to write the rest of the task.

Ira A. Fulton Schools of Engineering
Arizona State University

Coding template specification

Input parameters

1. Output path (Mandatory)
2. Task name: "hotzoneanalysis" or "hotcellanalysis"
3. Task parameters: (1) Hot zone (2 parameters): nyc taxi data path, zone path(2) Hotcell (1 parameter): nyc taxi data path

Example

< / > Generic code block | Make it interactive

```
test/output hotzoneanalysis src/resources/point-hotzone.csv
src/resources/zone-hotzone.csv hotcellanalysis
src/resources/yellow_trip_sample_100000.csv
```

Note:

1. The number/order of tasks do not matter.
2. But, the first 7 of our final test cases will be hot zone analysis, the last 8 will be hotcell analysis.

Input data format

The main function/entrace is "cse512.Entrance" scala file.

1. Point data: the input point dataset is the pickup point of New York Taxi trip datasets. The data format of this phase is the original format of NYC taxi trip whichis similar from Phase 2. But the coding template already parsed it for you. Find thedata in the .zip file below.

yellow_trip_sample_100000
ZIP File

2. Zone data (only for hot zone analysis): at "src/resources/zone-hotzone" of thetemplate

Hot zone analysis

The input point data can be any small subset of NYC taxi dataset.

Hot cell analysis

The input point data is a sample dataset like “yellow_trip_sample_100000.csv"

Output data format

Hot zone analysis

All zones with their count, sorted by "rectangle" string in an ascending order.

< / > Generic code block | Make it interactive

```
"-73.795658,40.743334,-73.753772,40.779114",1
"-73.797297,40.738291,-73.775740,40.770411",1
"-73.832707,40.620010,-73.746541,40.665414",20
```

Hot cell analysis

The coordinates of top 50 hotest cells sorted by their G score in a descending order. Note,DO NOT OUTPUT G score.

< / > Generic code block | Make it interactive

```
-7399,4075,15
-7399,4075,29
-7399,4075,22
```

Example answers

An example input and answer are put in "testcase" folder of the coding template

Ira A. Fulton Schools of Engineering
Arizona State University

Where you need to change

DO NOT DELETE any existing code in the coding template unless you see this "YOU NEED TO CHANGE THIS PART"

Hot zone analysis

In the code template,

1. You need to change "HotzoneAnalysis.scala and HotzoneUtils.scala".
2. The coding template has loaded the data and wrote the first step, range join query,for you. Please finish the rest of the task.
3. The output DataFrame should be sorted by you according to "rectangle" string.

Hot cell analysis

In the code template,

1. You need to change "HotcellAnalysis.scala and HotcellUtils.scala".
2. The coding template has loaded the data and decided the cell coordinate, x, y, zand their min and max. Please finish the rest of the task.
3. The output DataFrame should be sorted by you according to G-score. The codingtemplate will take the first 50 to output. DO NOT OUTPUT G-score.

Submission

Submission files

1. Submit your project jar package in My Submission Tab.
2. Note: You need to make sure your code can compile and package by entering sbt clean assembly. We will run the compiled package on our cluster directly using "spark-submit" with parameters. If your code cannot compile and package, you willnot receive any points.

IMPORTANT: Grading will take about 10~20 minutes.

Tips (Optional)

This section is same as that in Phase 2.

How to debug your code in IDE

If you are using the Scala template,

1. Use IntelliJ Idea with Scala plug-in or any other Scala IDE.
2. Replace the logic of User Defined Functions ST_Contains and ST_Within inSpatialQuery.scala.
3. Append .master("local[*]") after .config("spark.some.config.option", "some-value")to tell IDE the master IP is localhost.
4. In some cases, you may need to go to "build.sbt" file and change % "provided" to % "compile" in order to debug your code in IDE
5. Run your code in IDE
6. You must revert Step 3 and 4 above and recompile your code before use of spark-submit!!!

How to submit your code to Spark

If you are using the Scala template

1. Go to project root folder,
2. Run sbt clean assembly. You may need to install sbt in order to run this command.
3. Find the packaged jar in "./target/scala-2.11/CSE511-Project-Hotspot-Analysis-Template-assembly-0.1.0.jar"
4. Submit the jar to Spark using Spark command "./bin/spark-submit". A pseudo codeexample: ./bin/spark-submit ~/GitHub/CSE511-Project-Hotspot-Analysis-Template/target/scala-2.11/CSE511-Project-Hotspot-Analysis-Template-assembly-0.1.0.jar test/output hotzoneanalysis src/resources/point-hotzone.csv src/resources/zone- hotzone.csv hotcellanalysis src/resources/ yellow_trip_sample_100000.csv