

CSE 512 Project Phase 1

Task:

Set up your running environment, including installing the Hadoop and Spark.

Load GeoSpark jar into Apache Spark Scala shell and execute the following operations using Scala:

1. Create GeoSpark SpatialRDD (PointRDD).
2. Spatial Range Query: Query the PointRDD using this query window [x1(-113.79), x2(-109.73), y1(35.08), y2(32.99)].
 - a. Query the PointRDD
 - b. Build R-Tree index on PointRDD then query this PointRDD.
3. Spatial KNN query: Query the PointRDD using this query point [x1(35.08),y1(-113.79)].
 - a. Query the PointRDD and find 5 Nearest Neighbors.
 - b. Build R-Tree index on PointRDD then query this PointRDD again.
4. Spatial Join query: Create a GeoSpark RectangleRDD and use it to join PointRDD
 - a. Join the PointRDD using Equal grid without R-Tree index.
 - b. Join the PointRDD using Equal grid with R-Tree index.
 - c. Join the PointRDD using R-Tree grid without R-Tree index.

Please refer to GeoSpark Java Program example in GeoSpark Github Repository:

<https://gist.github.com/jiayuasu/84029bc9877804b876d209936e5b757c>

Submission

A video demo around 5 - 10 minutes. Each group should put the video demo on YouTube. The submission will be a single text file including the video link and your running code.

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 have to pre-load data to HDFS and use Spark to read data from HDFS. Your video demo should clearly demonstrate this point.
3. You should use **zcta510** as the query window dataset (use GeoSpark RectangleRDD) and use **arealm** as the point datasets (use x1 and y1 column for GeoSpark PointRDD). The datasets are put into Blackboard.
4. GeoSpark JTS library uses [x1,x2,y1,y2] to define a rectangle (aka. Envelope) instead of using [x1,y1,x2,y2].
5. 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.
6. For the definition of Spatial Range, Join and KNN queries, please refer to the reference.

Reference:

Jia Yu, Jinxuan Wu, Mohamed Sarwat. "GeoSpark: A Cluster Computing Framework for Processing Large-Scale Spatial Data". (short paper) In Proceeding of the ACM International Conference on Advances in Geographic Information Systems ACM SIGSPATIAL GIS 2015, Seattle, WA, USA November 2015