CSE 512 Spring 2017
Project Phase 2 Requirement

(Deadline: 19 March)

Task A

Write a Java function to do a Spatial Join Query using the simple Cartesian Product algorithm. Specifically, for each rectangle from the query window dataset, check this rectangle against the point datasets using the regular GeoSpark Spatial Range Query: RangeQuery.SpatialRangeQuery(objectRDD, queryEnvelope, 0, boolean useIndex)

1. This function should be integrated into GeoSpark source code. You should fork or download GeoSpark Github master branch **LATEST** source code and practice. You can use the latest version of Spark and Hadoop.
2. The result format should be same as GeoSpark spatial join query: JavaPairRDD<Envelope, HashSet<Point>>.
3. A piece of pseudo code is available here: https://gist.github.com/aniquetahir/acb2a781a55cf76a5d2a32d4f0a4d5d6
4. Your Scala API usage should be like that in this link: https://gist.github.com/aniquetahir/7cf09b31aa6906b2e80b139510dc55eb

Task B

Write a report:

1. For the tasks in Phase 1, compare the execution time / average memory / average CPU utilization of the entire cluster and explain differences of the following operations if any:
    1) For Task 2, difference between (a) and (b)
    2) For Task 3, difference between (a) and (b)
    3) For Task 4, difference between (a) and (b); difference between (a) and (c).
2. For Phase 1 Task 4 (a)(b)(c) and Phase 2 Task A, compare the execution time / average memory / average CPU utilization of the entire cluster and explain the differences if any.
3. The comparison above should be done in Spark Scala shell. You might need to run each program many times to get an accurate result.
4. Clearly mention your group number, member name, 10 digits Student ID and ASU Email in the report.

Submission

1. Your modified GeoSpark source code and compiled binary Jar. The source code folder name should be like this: “geospark_group1” and the binary Jar name should be like this “geospark_group1.jar”. This binary jar should be able to work

in Scala shell directly and support regular GeoSpark operations.

2. The report mentioned in Task B. The report should be in PDF format and named as “group1_report”.

1 and 2 should be put in a folder together (cse512_group1_phase2). This folder should be compressed to a zip or tar file (e.g., cse512_group1_phase2.zip) and submitted to Blackboard.

Notes:

1. You have to run Apache Spark on a cluster. This means you should have at least three machines (One master and two workers). The master should be able to communicate with workers using bi-directional Password-less SSH.
2. You have to pre-load data to HDFS and use Spark to read data from HDFS.
3. You should use zcta510 as the query window datasets (use GeoSpark RectangleRDD) and use arealm as the point datasets (use GeoSpark PointRDD). The datasets are put into Blackboard. This time the PointRDD in all Tasks should have 10 RDD partitions.
4. You have to use Apache Spark 2.1.0.
5. You’d better use some cluster monitoring software such as Ganglia to monitor the resources of your cluster. Single machine monitoring software like dstat is OK but it may take you lots of time to collect and process data from the machines in your cluster. You need to explain what software you are using and how you collect and process statistical data. You also need to provide screenshots of the software which is running on your cluster such as Figure 1.

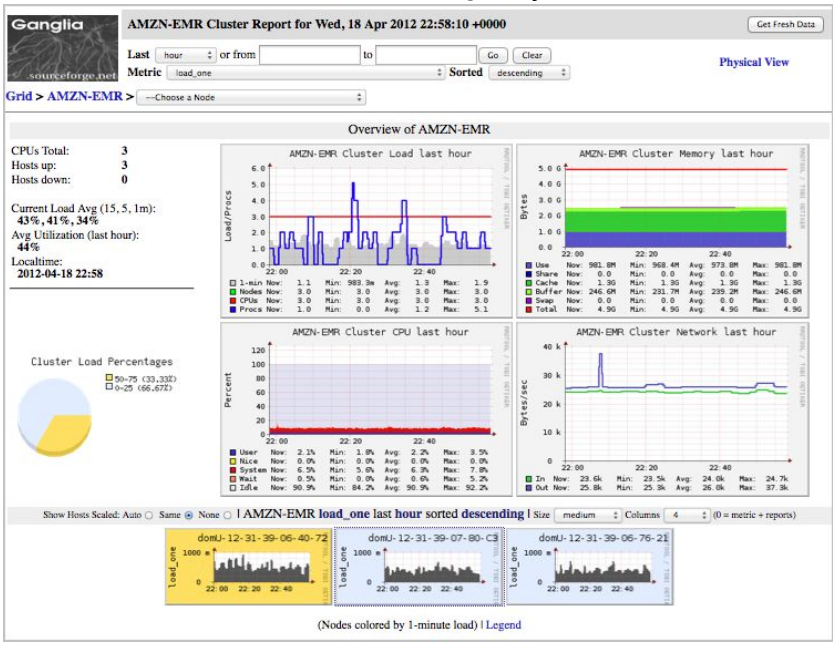

Figure 1 Ganglia Web Interface

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