Contemporary Database Technologies

Project Specifications

Objectives
- To allow students to apply database technologies in solving Big Data problems
- To be able to analyze a data source and build database structures to accommodate solving big data problems

Instructions
1) Download the file relatively_large_gps_data.json from moodle. Each object contains data related to a vehicle (device_code), coordinates (lon, lat), when it was logged (logged_at). The data spans 3 months (from august to october) and comes from several vehicles.
2) We'd like to perform a comprehensive analysis on what the data can tell us about the traffic situation in Metro Manila. To do this, parse the datasets and store it in a MongoDB instance. Decide on what the structure of the collection will be (it's only one collection anyway) and write a script/program that will do this automatically for you.
3) Create another collection that would serve as a lookup table for cities within metro manila. Each city should have a set of bounds that define it (Example: { city_name: "Makati", bounds: [{ lat: x, lon: y }, { lat: x, lon: y }, { lat: x, lon: y }, { lat: x, lon: y }] }). You may research on an external tool to help you get this data (example: http://www.rubygeocoder.com/).
4) Use Map Reduce to be able to answer the following questions:
   a) Count the number of times a vehicle travels within a city at every hour of the day by taking the average number of times a vehicle entered that city for a given hour.
   b) Count how many times vehicle travels within a city on a given day by taking the average of number of times a vehicle is found to be in a given city during that day (i.e. Monday, Device 100, 35.5).
   c) Based on the results of a) and b), determine the city that vehicles travel the most. To do this, perform a map reduce on the collections of the the results above to be stored in another collection. You will be graded according to how your documents in this collection will be structured.
5) (*grad students only*) Create a paper that first discusses what Map Reduce is. Cite 3 other existing applications that use Map Reduce to determine traffic situations in a city and how it compares to this current methodology. Discuss this project's methodology and how Map Reduce is applied. Present the results of your data in table format (or any other form of diagram) and how long it took to produce the data (i.e. How long in milliseconds it took to crunch the numbers for each scenario). Please use IEEE format for this paper (http://www.ieee.org/conferences_events/conferences/publishing/templates.html).

Things to Submit:
1) Parsing program
2) README.txt on how to run the parsing program (so it can be repeated locally)
3) Data on City bounds (a json file will do unless you use an external service)
4) Map Reduce functions for Scenario A, Scenario B, Scenario C (it should be repeatable based on your initial document structure)
5) IEEE formated paper (*grad students only*)