Project 2: Data Analytics

Due: Wednesday, Dec. 3rd at beginning of class

You will be required to hand in a report that contains all the code used as well as any required numeric or graphical output. There is no requirement to leave anything on the server, you can use whatever system you wish. Please be descriptive in both your report as well your code (remember, I will have to grade this!). This is an independent project and it is not allowed to be done in groups. Please do your own work.

1. Relational Theory:
   Design a 5th degree normalized database model for the following flat information. Use as many tables as you feel is required. Indicate all primary and foreign key constraints. Show all required SQL to create your database model.

   Fields in file solarareadata.txt:
   [SiteName, SiteAddress, SiteCity, PanelNum, PanelWatts, SensorNum, SensorTemp, SensorIrradiance, Date, Time]

   This is the output of solar-energy plants for a metro area. There are four plants. Each plant has sensors distributed throughout the facility measuring local conditions. Also, each PV panel measures the total watts it is producing at the time of the sample. The data was previously collected and assembled in a file by someone who didn't understand relational theory, so all the data has been placed into a single comma-delimited file.

2. Relational Implementation:
   Write a python program to upload the data for part 1 into the database model you designed in part 1. Write SQL statements to answer the following questions:
   1. What is the average monthly temperature for each of the 4 sites?
   2. What is the maximum monthly irradiance for each site? What time does this occur?
   3. What is the total wattage produced for each site per day? Per month?

   Show both the required SQL statements as well as a plot of each output data (you can decide the format for the plot but label all axis appropriately, use python for the plotting)

3. Map/Reduce: (use the testing technique discussed in class to implement, you will not have access to the hadoop cluster).
   Write a python map and reduce program to determine the average daily Extraterrestrial Radiation (ETR). Use the first column (YYY-MM-DD) and the column labeled (ETR (W/m^2)).
   - Download the solardata.csv file to your local hard drive (or to your directory on the MPI-main system, depending on where you want to run the program.
   - Show both the map and reduce program you used.

- Show your output, which should consist of two columns (“Day”, “Average ETR”)
- The data file contains two years worth of data. Data sampled each hour.