|  |
| --- |
| **Lab Exercises**   1. Import the Hubway data into **Neo4j** and configure Neo4j. Then, answer the following questions using the Cypher Query Language:   a) List top 10 stations with most outbound trips (Show station name and number of trips)  b) List top 10 stations with most inbound trips (Show station name and number of trips)  c) List top 5 routes with most trips (Show starting station name, ending station name and number of trips) (4) List the hour number (for example 13 means 1pm -2pm) and number of trips which start from the station "B.U. Central"  d) List the hour number (for example 13 means 1pm -2pm) and number of trips which end at the station "B.U. Central"   1. The flight data can be found at <http://stat-computing.org/dataexpo/2009/thedata.html>   You need to download just one year and from there you can sample a subset of at least 10000 records. You can use the data from a full year if you want but we recommend using a smaller dataset for simplicity. Hint: If you need to unzip the data file, you can use the command: bzip2 –d datafile from a terminal. For example, for the 2008, you download the file and unzip it using: bzip2 -d 1987.csv.bz2. The airport data can be found at  <http://stat-computing.org/dataexpo/2009/supplemental-data.html>  1) Download the flight dataset and airport dataset.  (2) Clean the dataset (for example: remove columns you do not need, remove records with missing information, remove duplicate records and so on).  (3) Give the header to csv files  (4) Import the data into Neo4j.  (5) Write the queries to answer following questions:  (5.1) List top 10 airports with most outbound flights.  (5.2) List top 10 airports with most inbound flights.  (5.3) List top 5 routes with most flights in weekdays.  (5.4) List top 5 routes with most flights in weekends.  (5.5) List the hour number (for example 13 means 1pm -2pm) and number of flights, which depart from a specific airport in your data (e.g., Boston Logan Airport).  (5.6) List the hour number (for example 13 means 1pm -2pm) and number of flights, which arrive at specific airport in your data (e.g., Boston Logan Airport).  In your report, you should answer the following questions:  (a) List the year of the flights that you downloaded and prepared for this assignment. You can get a sample set from one-year data. However, the number of flights cannot be smaller than 10k.  (b) Describe how you clean the data (Which columns you remove and why? Which rows you remove and why?). Hint: You can clean your data by writing a small program in Java, Python, C, Matlab or any kind of programming language.  (c) Describe the header you give to the csv files.  (d) Write down the command for importing data.  (e) Write and execute the queries from step (5) above. |