# CS 410/510 Cloud and Cluster Data Management

# Spring 2017 Quarter

## Assignment Application Data Demands

## Due: Thursday, 13 April 2017 at the beginning of class

This assignment is to be completed by individuals or by teams of two students. If you work with a partner, he or she should be in the same section (410 or 510) and you should turn in one assignment paper, with both of your names on the paper. You should only talk to the instructors and your partner about this assignment. You may also post questions and comments to the course discussion list on Piazza.

Please turn in your completed assignments on paper. Put your last name, first name, the assignment number in that order in the first line of your assignment. List last name and first name for your partner, if you have one, on the second line of your assignment. (If you are working with a partner, turn in one assignment paper.)

Each question is worth 10 points. CS410 students answer Questions 1-9; CS510 students answer all questions.

### Part I: Data Sizes and Access

You should choose an online application that you are familiar with that has a large user base and maintains data *on a per-user basis*. Examples: Twitter, Snapchat, gmail, FaceBook, Minecraft, Healthcare.gov, Dropbox, Flickr, LinkedIn, Instagram, Ebay, Yelp, TripAdvisor, Quora.

Question 1: Estimate the total number of users with accounts for the application. Explain your estimation process. Cite any sources you consulted.

Question 2: List what you think are the three main kinds of data for the application that are maintained on a per-user basis.

Question 3: Estimate the average amount of data per user for each of the three kinds of data in Question 2. Explain the basis for your estimate.

Question 4: Use the answers to Questions 1 and 3 to estimate the total amount of user data managed by the application provider.

Question 5: Identify two common operations a user might do with the application that involves the data of one or more users. At least one operation should be an update. For each operation, say which kind(s) of data in Question 2 that it uses.

Question 6: For each operation in Question 5, estimate how much data is touched on average each time the operation is performed. Explain your estimation process.

Question 7: Estimate how many times *per day* an average user performs each operation in Question 5, and explain your estimation process.

Question 8: Use your answers to Questions 1, 6 and 7 to figure out how much data is accessed *per minute* across all users of the application.

Question 9: Use your answers to Questions 4 and 8 to estimate what percentage of the total user data is accessed each minute and each day. (*Note: In coming up with the daily percentage, you might want to take into account how much overlap there is in the data accessed during a day. Explain your adjustment if you make one.*)

## Part II: Synchronization (CS 510 only)

Question 10: To what extent do different instances of your operations need to see consistent data. Consider both two instances of the same operation (2 cases) and two instances of different operations (1 case). *You might think about this question in terms of whether serializable update, eventual consistency, best effort or no synchronization is required.*