CS 498: Cloud and Cluster Data Management
Spring 2022

Assignment 1: Big Data Intuition
Due: Tuesday, 1 February 2022, at Midnight

The purpose of this assignment is to acquaint you with the scale and properties associated with data in the Cloud.

This assignment is to be completed by individuals or by teams of two students. If you work with a partner, you should submit the assignment with both of your names included. You should only talk to the instructor, the TA and your partner about this assignment, except on course-associated discussion websites like Canvas.

Intuition and Scale: Introduction to Big Data

This assignment can be tested using our autograder at any point. Simply go to the autograder website we have provided you, and enter the URL of the Git repository you have created for this assignment.

I: Git Repo setup: Create a Git repository, and give access to our TAs and Professor Alawini. In this repository should be two files: **Answers.txt**, which consists of the answers to the below questions, and **Team.txt**, which consists of a space-separated list of your team's members.

II: Dataset Analysis: A dataset of Elon Musk tweets can be found at **https://www.kaggle.com/ayhmrba/elon-musk-tweets-2010-2021**. If you like, you can use Kaggle's notebook function to access and examine the data in your browser. Otherwise, you are free to download it and process it however you'd like. Try to get a grasp of the volume of data associated with even one social media user (albeit a famous one).

III: Questions: Calculate the values associated with the below questions, providing one answer on each line, rounded to the nearest whole number where appropriate. Submit a Python notebook of any code you write as **Code.ipynb**. You can run an IPython notebook on Kaggle's servers by clicking the "New Notebook" button towards the left of the page.

- ***Question 1:*** Estimate the average number of people that 'like' any given tweet by Elon Musk.

- ***Question 2:*** From the above, estimate how many people view one of his tweets, on average. Assume that engagement rate, calculated by summing likes, replies, and retweets, and dividing by views, is roughly .05 for any given tweet.

- ***Question 3:*** Estimate the average amount of data (in megabytes) stored for each Elon Musk tweet. Consider the sizes and types of any attached media.

- ***Question 4:*** Using the answers above, estimate the total data transfer involved in displaying any given Elon Musk tweet to Twitter users.

- ***Question 5:*** Estimate how many Elon tweets are viewed, per-minute, by the site's userbase.

- ***Question 6:*** Using the above, estimate how much data is accessed per-minute by Twitter in the process of displaying Elon Musk tweets.
  - A popular Twitter user likely has very active data, with more resources involved in sending it to others than in storing it. As a thought exercise, consider whether his old tweets are accessed as frequently as his new ones. Consider whether an average user's tweets are ever viewed at all – it may not be efficient to treat all data the same!

- ***Question 7:*** **You do not need to submit this answer in Answers.txt.** At the end of your Python notebook, choose a function that involves modifying a tweet (say, updating its 'likes' count, which you can assume is cached in a tweet's associated database row). How much data do you believe is touched by this operation on a monthly basis? Consider the amount of data affected, the frequency at which the operation is carried out, and so on.

IV: Submission: In the base directory of your Git repo, attach two files:

- **Team.txt** should be a space-separated list of your team member(s)' netIDs.
  - **Example content:** mweston3 mweston4

- **Answers.txt** should be a return character-separated list of numeric values, which correspond to the questions above.
  - **Example content:**
    1000
    890
    103000

  - **Code.ipynb** should include any code you used to compute the answers in **Answers.txt**. At the bottom should be your answer to question 7, along with any associated computations.

Be sure to run our autograder at least once, as described above.
This will upload the address of your repository, and allow us to calculate the score for your programming assignment.