CAPP 3 Project Proposal

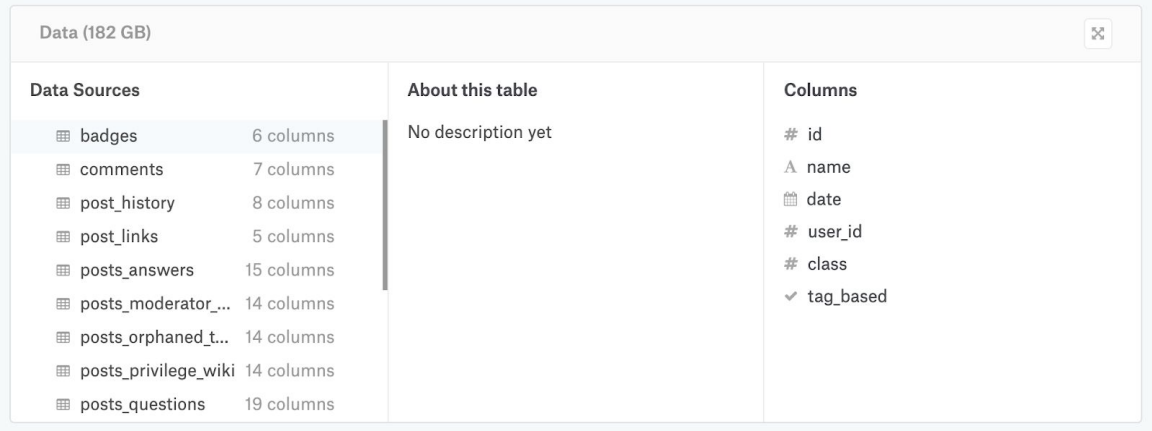
Group name: HackyStacks

Repository URL: https://github.com/liu431/Big-Data-Project

Group members: Adam Shelton, Dhruval Bhatt, Li Liu, Sanittawan Tan

Data sets: Stack Overflow Data

- The data set is available here: https://www.kaggle.com/stackoverflow/stackoverflow
- The size of the data set is 182 GB. It is constantly updated, so the size may be subject to change.
- Kaggle indicates that there are 16 data sources where we can query, merge, join, and retrieve a tabulated data set. The data can be obtained from Google BigQuery API. We can also query the data and write a subset of the data to CSV files and save them to local computers.



- Based on our quick analysis through the Kaggle kernel:
    - Data sources that we are interested in generally have millions of rows. For example, for “users” data source, there are 10,097,978 rows. “Badges” data source has 30,347,225 rows. “Stackoverflow_posts” data source has 31,017,889 rows. “Posts_answers” data source has 26,496,612 rows. “Posts_questions” has 17,278,709 rows. It is certainly possible that when we filter the dataset, some rows may be dropped out. However, the Stack Overflow data set is still fairly large.

Research Question and Hypotheses:

- What affects the number of responses a post receives?
    - The role of gender (from display name)
    - Profession listed (from “About Me” column)
    - Quality of writing on the post (by textstat package)

- Which is the most popular programming language?
    - Among professionals, students.
        - Does programming language usage change across professions?
    - Count across locations
- How much time does it take to get a response?
    - Based on location
    - Based on gender as determined by username
- How spread is the network of respondents?
    - Create a network of responders:
        - Using user id number
        - Locations
            - Do people from the same region respond more to others within the region?
- What is the trend in programming?
    - Buzz words among the most viewed posts (top 100 most viewed posts)
    - Find trend using tags identified
- What is the distribution of different participants on Stack Overflow?
    - Classify users into four categories based on their posting and answering behaviors
        - For example, (1) people who posted a lot but answered a few, (2) people who posted a few but answered a lot, (3) people who posted a lot and answered a lot, and (4) people who posted a few and answered a few

Algorithms and Tools:

- MapReduce/Hadoop
- Cloud Computing
- Efficient Sorting algorithms

Potential Issues:

- Missing values in the users table (Limited information)
- If we download the data, it's too large; if we query it every time, the answers would change.