Data 607 Project 3: What Are the Most Valuable Data Science Skills?

Group members: Sean Connin, Ethan Haley, Daniel Moscoe, Zachary Safir.

Proposed Approach:

- Our project workflow includes the following:
    - Establish shared Github Repository for code, files, and related materials (complete)
    - Establish Slack channel and Google Drive to share information (complete) - we are meeting regularly via. Zoom to debrief and identify next steps as well as assignments
    - Develop R code to scrape job listings and descriptions from Indeed (complete)
    - Develop code to identify and extract “skills” identifiers from job descriptions (in progress)
    - Clean data and convert to Tidy format (in progress)
    - Develop and populate relational database (in progress)
    - Export local database to RDB if time permits
    - Prepare analyses, visualizations, and final report

Collaboration tools

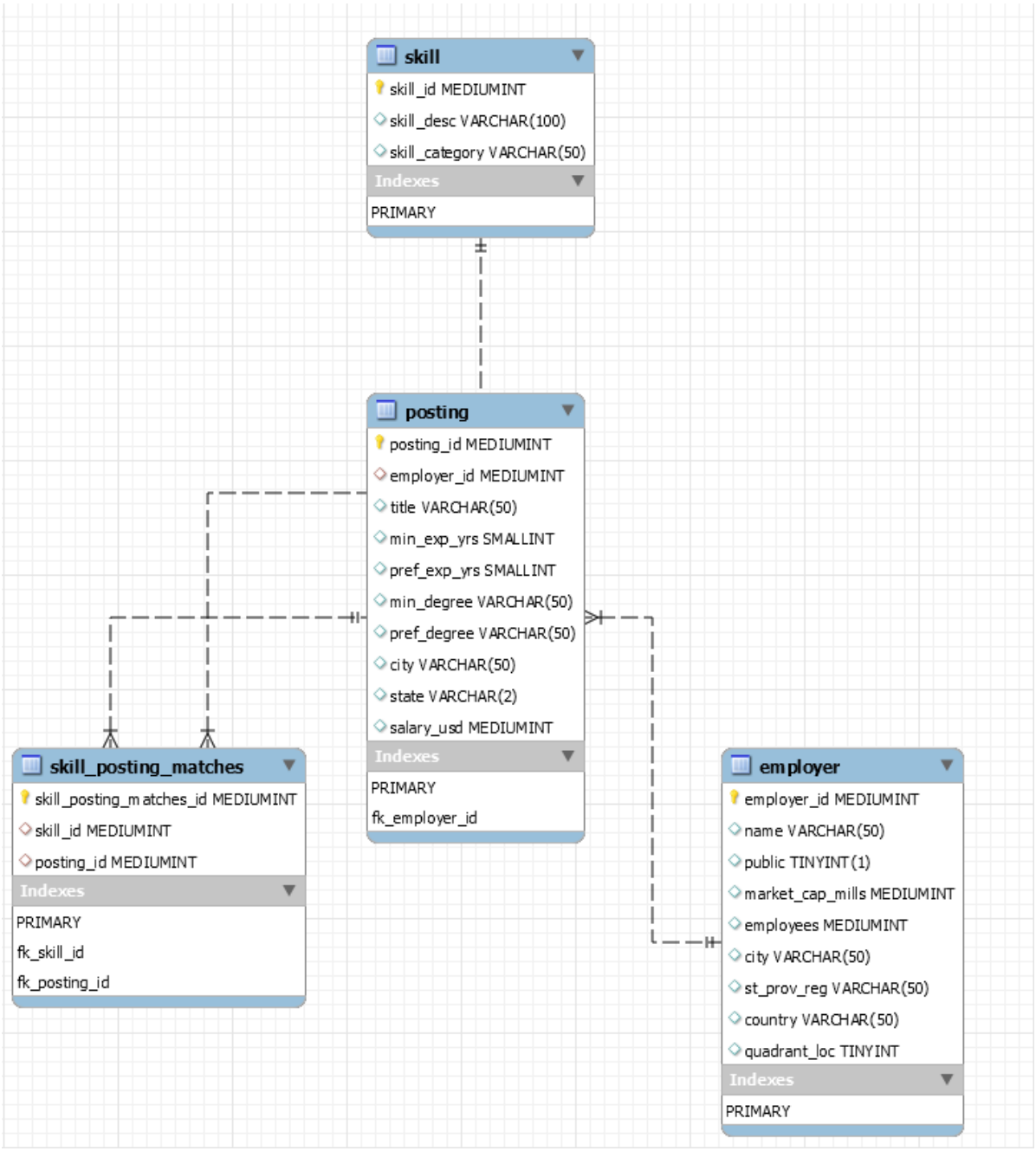
- MySQL database -- we are currently experimenting with various ways to share access during the development phase.
- Google Drive, for non-technical documents like this proposal;
- Slack, for asynchronous team communication;
- Zoom;
- GitHub, for shared R code and final product. We have already established a repository and individual branches for each collaborator.

Data sources

Job listings hosted by Indeed - we have a working script that returns search results as well as full job descriptions. While we have run into encoding errors associated with some search results, the data we have procured to date is sufficient to proceed.

Database design

Please see the following page for a tentative entity-relationship diagram. We expect to make modifications to this design as we progress. Time permitting, we will host the database on a remote server such as AWS.


skill
skill_id MEDIUMINT
skill_desc VARCHAR(100)
skill_category VARCHAR(50)
Indexes
PRIMARY
posting
posting_id MEDIUMINT
employer_id MEDIUMINT
title VARCHAR(50)
min_exp_yrs SMALLINT
pref_exp_yrs SMALLINT
min_degree VARCHAR(50)
pref_degree VARCHAR(50)
city VARCHAR(50)
state VARCHAR(2)
salary_usd MEDIUMINT
Indexes
PRIMARY
fk_employer_id
skill_posting_matches
skill_posting_matches_id MEDIUMINT
skill_id MEDIUMINT
posting_id MEDIUMINT
Indexes
PRIMARY
fk_skill_id
fk_posting_id
employer
employer_id MEDIUMINT
name VARCHAR(50)
public TINYINT(1)
market_cap_mills MEDIUMINT
employees MEDIUMINT
city VARCHAR(50)
st_prov_reg VARCHAR(50)
country VARCHAR(50)
quadrant_loc TINYINT
Indexes
PRIMARY