Final Project: Database Development for Career-Service Department in MSBA Program

Winter 422-002 Data Design and Representation

Yiying Wang, Kayla Zou, Lin Zhu, Alice Huang

3/15/2020

Table of Content

Executive Summary

In recent years, there has been an increasing demand for analytics professionals in the job market. Schools are launching new programs to fill in the gap between the supply and demand. In addition to delivering a well-crafted curriculum setting, schools have also been exploiting a great career-development service for students.

In order to gain a more comprehensive understanding of the job market in the area, we were reached out by UC Davis MSBA program, a newborn program in San Francisco, to help gather resources to lay a solid foundation of the career-service database. With the goal in mind, we decided to build a SQL database with job and company information in San Francisco/Bay Area. Information is all collected from Glassdoor, an online job-listing platform. Secondly, we generated a Tableau dashboard with summary statistics, which would be refreshed with the data tables once every week. Furthermore, we created an analyst-specific job search filter, which is grouped by job titles and displays company names with estimated salaries.

With the first week data, we built a database with a list of 1706 jobs in total from 899 companies. We can see that:

- San Francisco has the most job listings related to data analysts, taking 15% of the whole job market in northern California.
- More than 10% of the analytical jobs are from the IT industry.
- Around 60% of the companies are private companies.
- Uber ranked the top place as it posted the most positions, amounting to seven, that are relevant to data analytics.

We are aiming to refresh the data on a weekly basis to provide the most up-to-date data to our client and students. With the implementation of the database, our client can better serve the students and develop some better KPIs, such as the students’ placement rate, the average salary, etc., in the near future.

Background, Context and Domain Knowledge

Data scientists, data engineers and business analysts are among the most sought-after positions in America. According to IBM, the number of jobs for data professionals in the U.S will increase to 2,720,000 by 2020.

More and more schools have realized the demand of the market. Therefore, an increasing amount of new programs, such as Master Science of Business Analytics, Master Science of Data Science and etc. are launched in the market to better prepare people to get the qualified skills and the analytics-related jobs. Having the graduate placement as an important metric to stand out amongst all these new programs, the career service department of the schools are making a lot of efforts to provide the best service to the students. As an analytics consulting group, we were reached out by a career-service department from UC Davis MSBA for better resources to provide to students. We then decided to build a database with information from an online aggregated job-listing platform and to generate a summary statistic dashboard. We are aiming to update them on a weekly basis to enlarge their information source and thus better serve the students.

Database Development Process

Having the goal in mind, we decided to identify an online aggregated job-listing platform as our data source. We narrowed down to three websites, which are LinkedIn, Glassdoor, and Indeed. As we target to provide a dataset with a comprehensive presentation of not only the job description, but also the company information for further comparison and analysis, we eventually chose Glassdoor as our data source. We took job information, company information, company rating information and salary information into our database, defining companyID as the primary key to build a relational dataset.

After identifying our data source, we started our process of web-scraping. In order to initiate Glassdoor's search function, we need a location-related code and the specific job title to start. In this case, we took "San Francisco" and "San Jose" as our location, and "data analyst", "data scientist" and "business analyst" as our search terms. We first started a configure script, which is designed to aid in filling in the customized search terms for the search engine to execute. Then we embed these customized variables of the configure script to the URL, from which we used *get* request to retrieve all the related web pages and downloaded them. After getting the website html codes, we parsed them into Python by applying BeautifulSoup package and found all the job IDs from different pages. We stored them in one file and kept appending new IDs to it when there is any. Meanwhile, we made our algorithm sleep for three seconds after every retrieval, so that we would not be inspected and blocked by the data source. As of now, we have downloaded all the related html pages by applying our search terms, parsed them with BeautifulSoup and started a file with all the job listing IDs.

After ensuring the job IDs are unique, we embed them back into the URLs to get the json script of each job listing. JSON is a lightweight format that is used for data interchanging. It is built on two structures, which are a collection of name/pair values or an ordered list of values. With JSON, it is easier for us to retrieve the data from the name/pair values.

We then wrote a program in python to connect with MySQL in the local server and created four tables in the dataset, which are "job", "company", "rating", and "salary". Having the tables ready, we need to define the tables with all the column names as well as the primary key(s). "JobID" and "CompanyID", which we got from the web pages, are defined as the primary key for table "job" and "company" respectively. "SalaryID" and "RatingID" are given by us with incremental numbers as the primary key for table "salary" and "rating" respectively. We then extracted the related information from the previous documents to insert them into the tables accordingly (relational schema is in Appendix I).

Until now, we finally developed our own dataset, consisting of four tables, with 1706 rows of unique JobIDs, 899 rows of unique companies, 6693 rows of salary information (one JobID may have different salary for different job title, pay period, and salary type), and 1589 rows of rating (one company may have multiple ratings towards different aspects).

Database Selection

We have rearranged the data we got from the aggregated online job listing platform into the Structured Query Language. We've built a relational database to store all the features that we thought would be helpful for students from the job market. The alternative database implementation could be NoSQL, but we preferred SQL.

On one hand, the business goal of us is to provide data-analyst related job market information for students in northern California, which can be updated on a weekly basis. With a specific job type that we are looking for, even if we enlarge our location scope to the east coast, the job listings are still going to be limited. We can expect an increase but the increase will not be exponential. If we just focus on the northern California area, then we will not anticipate a lot of changes or a huge growth of the total amount of the job listings. Therefore, there is no need to use NoSQL, since scalability is not a necessity in this case.

On the other hand, our clients and we will both be working with query and reports with our databases. With SQL we can build one script that retrieves and presents our data. However, NoSQL does not support relations between data types. In addition, when we deliver the database to the career-service department, they will be using this to assist students. An easy-to-learn and efficient query tool is more user-friendly to the less technical staff. Therefore, SQL is a preferable choice.

Business Value

On top of the database that we have developed, we also designed a Tableau dashboard with summary statistics of job market information and a more user-friendly job-selection filter to our client (Tableau dashboard link and screenshot can be found in Appendix II). With the dashboard, we could have a faster comprehension of what is happening in the job market. We will learn about where the most data-related jobs locate, which sectors need the most data professionals, which companies have the most relevant jobs posted, etc. With a list of 1706 jobs in total from 899 companies collected from the first week data, we can see that:

- San Francisco has the most job listings related to data analysts, taking 15% of the whole job market in northern California.
- More than 10% of the analytical jobs are from the IT industry.
- Around 60% of the companies are private companies.
- Uber ranked the top place as it posted the most positions, amounting to seven, that are relevant to data analytics.

With this information, the career-service department can sort out more specific suggestion.

The dashboard enables us to have some interaction with the data as well. If we click on San Francisco on the word cloud, all the information below will only demonstrate the statistics of San Francisco. Moreover, with the job selection filter we have built, students may easily locate the salary of their interested job titles in their interested companies. (The salary is calculated as the average of the range of salary provided by Glassdoor.)

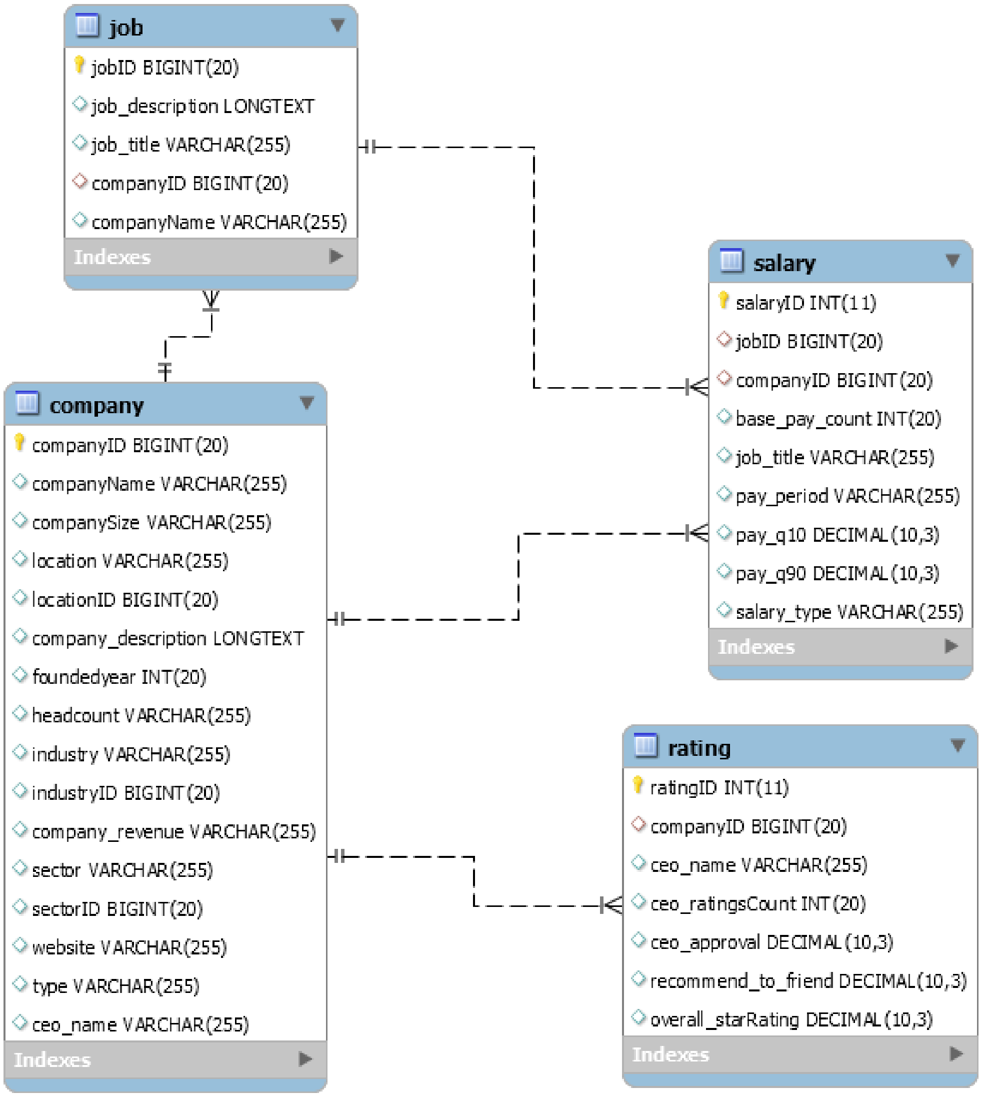
Since the career-service department in school is not profit-oriented, the KPIs of our project will be student-oriented. With our work implemented, we are expecting an increase of the students' placement rate, an increase of the average salary, as well as a decrease of students' average time of getting a job, etc.

Summary and Conclusions

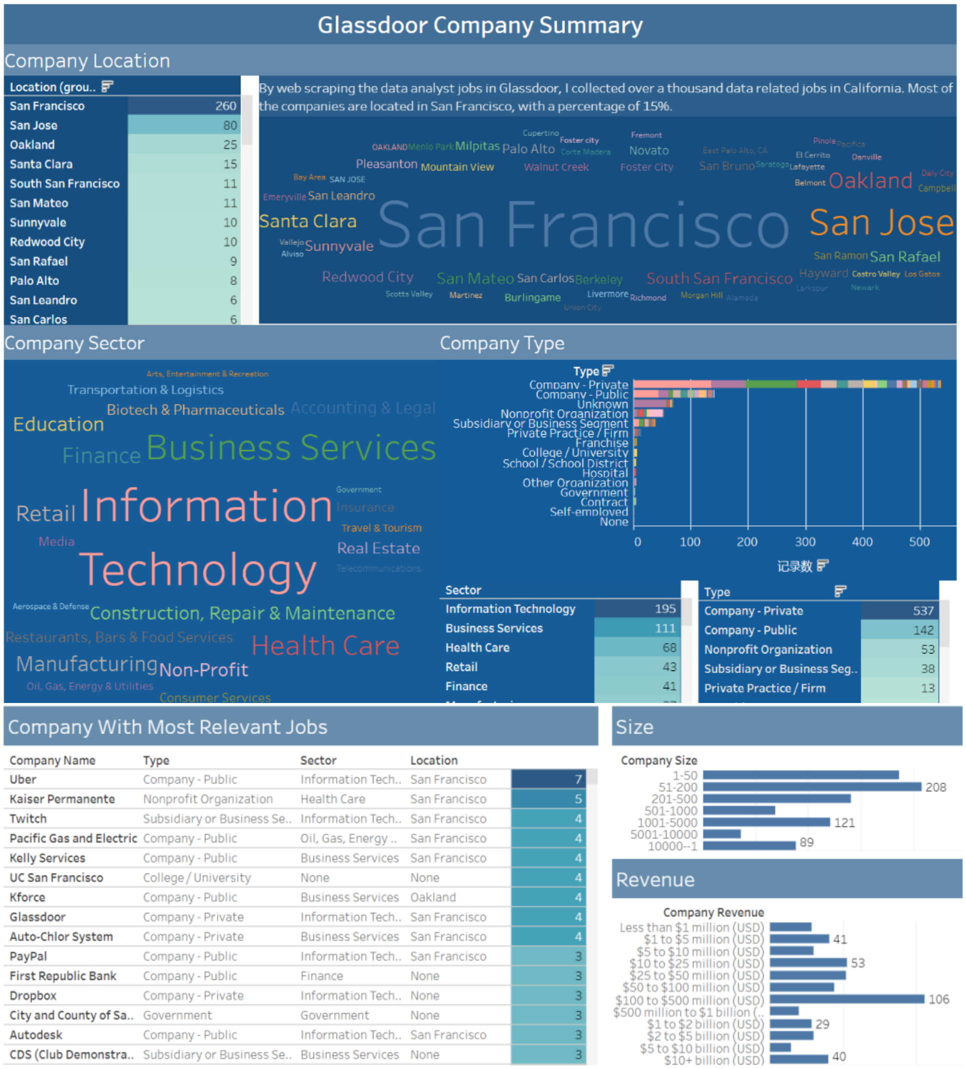
Above all, we have developed a SQL dataset containing four tables of companies and jobs related information retrieved from an online job listing platform, Glassdoor. Furthermore, we developed a Tableau dashboard consisting of some summary statistics and a job-selection filter from our database. We are aiming at updating our database on a weekly basis, in order to provide the most up-to-date job market information to our client, the career service department of the school. With the easy-to-query database as well as the easy-to-look-up dashboard, our client will be more resourceful when serving students, and eventually assist the students to understand better of the job market and get the best offer as they graduate from school.

Appendix

I. Relational Schema of our database



II. Tableau dashboard https://public.tableau.com/profile/yiying7787#!/vizhome/ddr_15838219621930/Dashboard12



Glassdoor Company Summary

Company Location

Location (grou..	
San Francisco	260
San Jose	80
Oakland	25
Santa Clara	15
South San Francisco	11
San Mateo	11
Sunnyvale	10
Redwood City	10
San Rafael	9
Palo Alto	8
San Leandro	6
San Carlos	6

By web scraping the data analyst jobs in Glassdoor, I collected over a thousand data related jobs in California. Most of the companies are located in San Francisco, with a percentage of 15%.

Company Sector

Sector	
Information Technology	195
Business Services	111
Health Care	68
Retail	43
Finance	41

Company Type

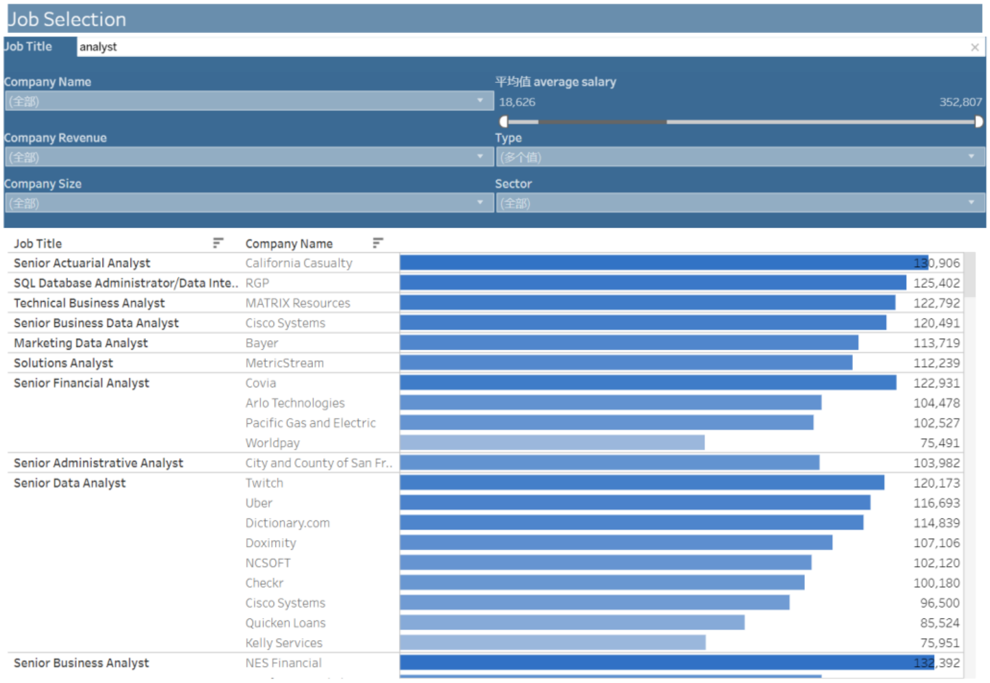
Type	
Company - Private	537
Company - Public	142
Nonprofit Organization	53
Subsidiary or Business Seg..	38
Private Practice / Firm	13

Company With Most Relevant Jobs

Company Name	Type	Sector	Location	
Uber	Company - Public	Information Tech..	San Francisco	7
Kaiser Permanente	Nonprofit Organization	Health Care	San Francisco	5
Twitch	Subsidiary or Business Se..	Information Tech..	San Francisco	4
Pacific Gas and Electric	Company - Public	Oil, Gas, Energy ..	San Francisco	4
Kelly Services	Company - Public	Business Services	San Francisco	4
UC San Francisco	College / University	None	None	4
Kforce	Company - Public	Business Services	Oakland	4
Glassdoor	Company - Private	Information Tech..	San Francisco	4
Auto-Chlor System	Company - Private	Business Services	San Francisco	4
PayPal	Company - Public	Information Tech..	San Francisco	3
First Republic Bank	Company - Public	Finance	None	3
Dropbox	Company - Private	Information Tech..	None	3
City and County of Sa..	Government	Government	None	3
Autodesk	Company - Public	Information Tech..	San Francisco	3
CDS (Club Demonstra..	Subsidiary or Business Se..	Business Services	None	3

Size

Revenue



Job Selection

Job Title: analyst

Company Name: (全部)

平均值 average salary: 18,626 – 352,807

Company Revenue: (全部)

Type: (多个值)

Company Size: (全部)

Sector: (全部)

Job Title	Company Name	
Senior Actuarial Analyst	California Casualty	130,906
SQL Database Administrator/Data Inte..	RGP	125,402
Technical Business Analyst	MATRIX Resources	122,792
Senior Business Data Analyst	Cisco Systems	120,491
Marketing Data Analyst	Bayer	113,719
Solutions Analyst	MetricStream	112,239
Senior Financial Analyst	Covia	122,931
	Arlo Technologies	104,478
	Pacific Gas and Electric	102,527
	Worldpay	75,491
Senior Administrative Analyst	City and County of San Fr..	103,982
Senior Data Analyst	Twitch	120,173
	Uber	116,693
	Dictionary.com	114,839
	Doximity	107,106
	NCSOFT	102,120
	Checkr	100,180
	Cisco Systems	96,500
	Quicken Loans	85,524
	Kelly Services	75,951
Senior Business Analyst	NES Financial	132,392