

IE 6210 Data Management and Database Design

SKUNKWORKS: JOB DATABASE PROJECT

Domain: Data Scientist

Team Name: Deviants

Team Members:

Name 1: Harshil Mehta, NUID: 001470356

Name 2: Srikrishna Ram, NUID: 001415814

Project Code Link: <https://github.com/srikrishnaram1996/Database-Final-Project>

JOBS DATABASE PROJECT

The main objective of this project is to reduce job search and talent acquisition stress level which leads to less resume spamming. Matching the job seekers with the right employers and secondly, provide guidance to aspiring job seekers on the skills that are in demand so that they can build them to stay relevant in the job market.

We have researched on many Data Scientist job domain websites and found out most relevant skills required in a potential job seeker.

Objectives Achieved in this project

- ▶ We got the data from scrapping Indeed website
- ▶ Information related to 331 companies in Data Science domain
- ▶ The dataset includes Job Description, Salary, Location etc.
- ▶ Getting Social Media links of the companies
- ▶ Normalization of the database in order to reduce data redundancy
- ▶ Creating Use Cases that are practically useful
- ▶ Adding index to our database for increase in performance and creating functions which are practically useful
- ▶ Creating Stored procedures which are practically useful

Data Scarping Template

This template acts as a Robot for our project which extracts data related to a specific Location, Salary and Job Title.

<http://www.indeed.com/jobs?q=data+scientist+%2420%2C000&l={}&start={}>

There are different parameters we can alter to scrape the data:

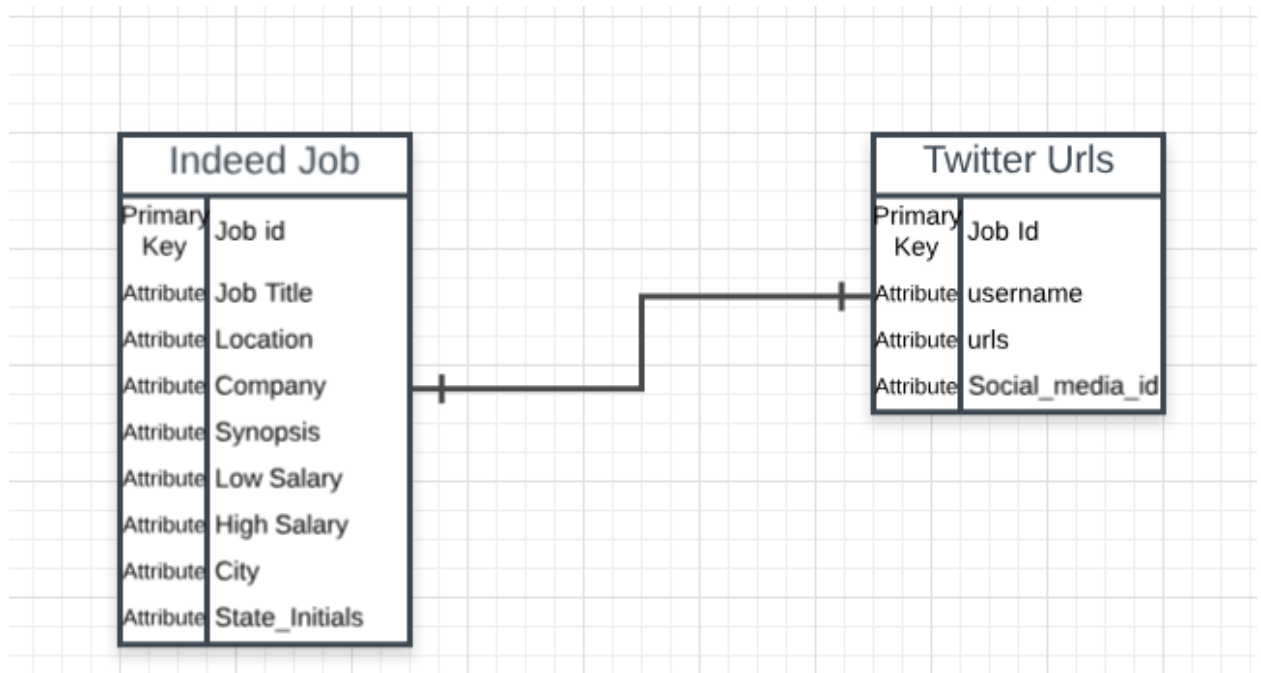
1: q for the job search

2: This is followed by "+20,000" to return results with salaries (or expected salaries >\$20,000)

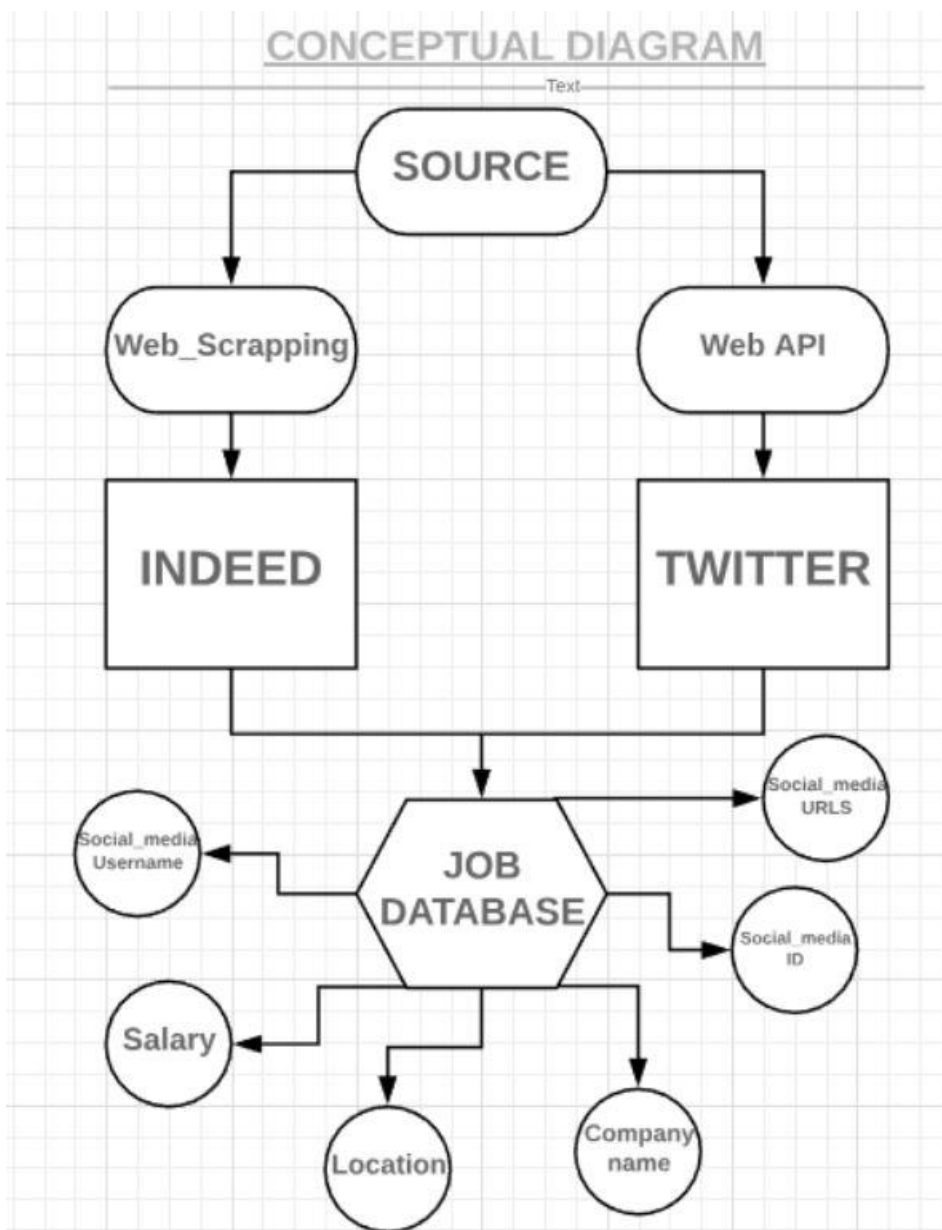
3: l for a location

4: start for what result number to start on

Entity Relationship Diagram:



Conceptual Diagram:



Data obtained from Indeed Web scraping:

Title	Location	Company	Salary	Synopsis	Low Salary	High Salary	City	State	State Initials
Data Scientist	Redlands, CA	Esri	104,000 – 154,000 ()	Experience dealing with massive da...	104000	154000 ()	Redlands	CA	CA
Spatial Data Scientist – Spatial Statistics	Redlands, CA	Esri	95,000 – 141,000 ()	Passion for storytelling using dat...	95000	141000 ()	Redlands	CA	CA
Environmental Compliance Specialist (Mid level)	Rancho Cucamonga, CA 91739	JE Group of Companies	55,000 – 65,000	Conducting site visits at client f...	55000	65000	Rancho Cucamonga	CA 91739	CA
Microfluidics Design Engineer	Irvine, CA 92618	FluxErgy	70,000 – 90,000	Communicate results effectively, g...	70000	90000	Irvine	CA 92618	CA
Data Scientist - LA	Los Angeles, CA	CruiTek	\$160,000	Perform data exploration and data ...	160000	None	Los Angeles	CA	CA

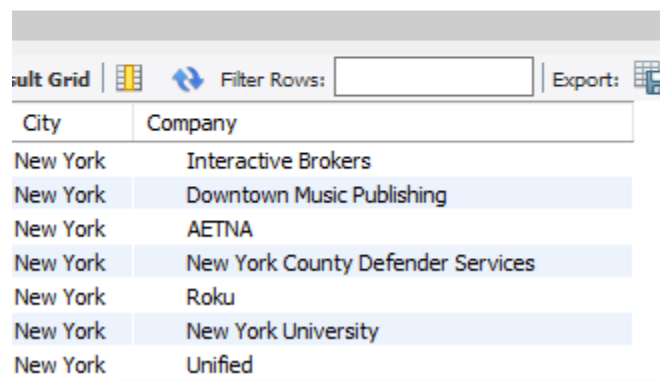
Data Obtained from Twitter API:

	user_name	social_media_id	company_url
1	VACULIVE	818510649139597312	https://t.co/PN9t3nogN8
2	VCU	156714051	http://t.co/a3ln7QjvFX
3	DSLCC	25579173	http://t.co/A7EtLSAS9A
4	FRC422	258023422	https://t.co/5bKlfSDvjN
5	AmyxInc	827209038786068481	https://t.co/1oluRoEPuY
6	csscorp09	85772994	http://t.co/sTEMfJU5WC
7	NewVAMajority	233630180	https://t.co/689Wo71shT
8	DSLCC	25579173	http://t.co/A7EtLSAS9A
9	CCASHMORE_BUYER	156568290	https://t.co/OCxt0dShFT
10	Covance	365005464	https://t.co/apKYTWy1Ny
11	ZotecPartners	188905742	https://t.co/1Mub9D0c9D
12	Cummins	87299367	http://t.co/vqpmj3Bha9
13	russellcrowe	133093395	https://t.co/OM7honTWnX
14	elementsCU	18975866	http://t.co/oNNgpU3bp4
15	weareoneamerica	46727968	http://t.co/olbQvFIRka

USE CASES:

- 1) Specific Location for example we consider New York

```
1 • SELECT * FROM jobs.v_case1;
```

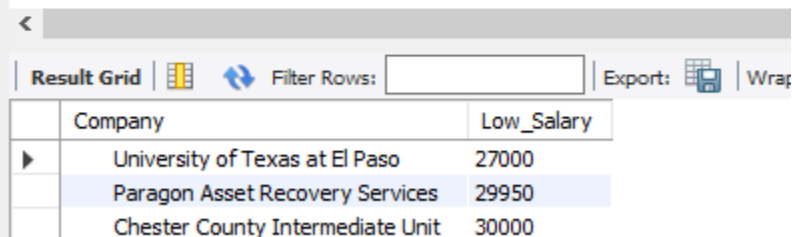


The screenshot shows a SQL query result grid with a toolbar at the top containing 'Result Grid', a grid icon, a refresh icon, a 'Filter Rows' input field, and an 'Export' button. The table has two columns: 'City' and 'Company'. The data rows are as follows:

City	Company
New York	Interactive Brokers
New York	Downtown Music Publishing
New York	AETNA
New York	New York County Defender Services
New York	Roku
New York	New York University
New York	Unified

- 2) Top 3 Lowest Salary paid by companies

```
1 • SELECT * FROM jobs.v_case3;
```

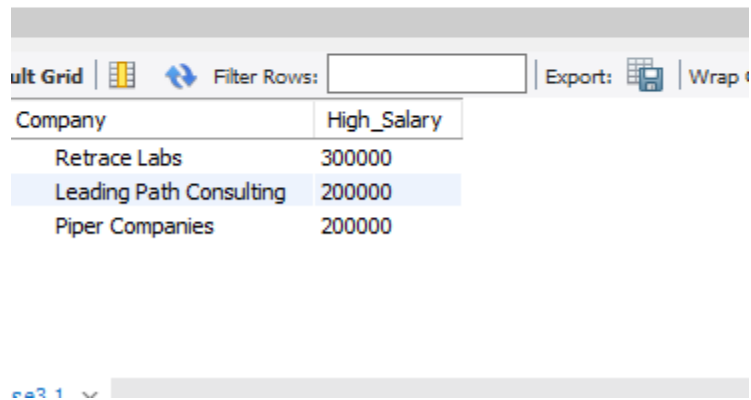


The screenshot shows a SQL query result grid with a toolbar at the top containing '<', 'Result Grid', a grid icon, a refresh icon, a 'Filter Rows' input field, an 'Export' button, and a 'Wrap' button. The table has two columns: 'Company' and 'Low_Salary'. The data rows are as follows:

Company	Low_Salary
University of Texas at El Paso	27000
Paragon Asset Recovery Services	29950
Chester County Intermediate Unit	30000

3) Top 3 Highest Salary paid by companies

```
1 • SELECT * FROM jobs.v_case3;
```

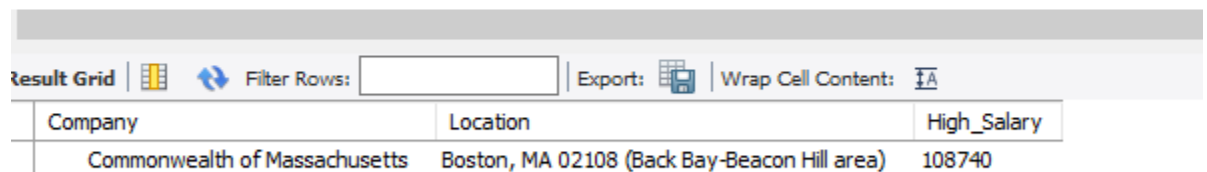


The screenshot shows a SQL query result grid with the following data:

Company	High_Salary
Retrace Labs	300000
Leading Path Consulting	200000
Piper Companies	200000

4) Salary greater than 90000 in Boston

```
1 • SELECT * FROM jobs.v_case4;
```



The screenshot shows a SQL query result grid with the following data:

Company	Location	High_Salary
Commonwealth of Massachusetts	Boston, MA 02108 (Back Bay-Beacon Hill area)	108740

5) Job title is Data Analyst

1 • `SELECT * FROM jobs.v_case5;`

Result Grid	
Filter Rows:	Export: Wrap Cell Content:
Company	Title
Lennon Wright Associates	Data Analyst - FinTech Company
Logic20/20	Data Analyst / Jr. Data Scientist
Johns Hopkins University	Research Data Analyst
Legalmation	Legal Research/Data Analyst
Shapestone Inc.	Data Analyst
Ambry Genetics	Intern, Financial Revenue Data Analyst/...
Urban Health Collaborative- Drexel University	Senior Data Analyst

case5 1 x

Output

References:

- <https://github.com/aakashtandel/Web-Scraping-Indeed/blob/master/Code/Scratch%20Notebooks/project-3-aakash-version4.ipynb>
- <https://nycdatascience.com/blog/student-works/project-3-web-scraping-company-data-from-indeed-com-and-dice-com/>
- <https://stackoverflow.com/questions/42225364/getting-whole-user-timeline-of-a-twitter-user> <http://docs.tweepy.org/en/v3.5.0/api.html>
- <https://stackoverflow.com/questions/22469713/managing-tweepy-api-search>
- <https://labsblog.f-secure.com/2018/01/26/how-to-get-tweets-from-a-twitter-account-using-python-andtweepy/>
- <https://stackoverflow.com/questions/8282553/removing-character-in-list-of-strings> <https://stackoverflow.com/questions/42225364/getting-whole-user-timeline-of-a-twitter-user>