# INFO 6210 DATABASE MANAGEMENT AND DATABASE DESIGN

## JOBS DATABASE

**Group Members:**
Keval Shah
Akshay Kochhar
Rohit Chandramouli

### ABSTRACT:

The objective of this project is to build a job database by scrapping Glassdoor using python and calling for Social Media details using Twitter API. Database was successfully created in PostgreSQL. This project is focused on job domains such as supply chain analyst, data analyst and business analyst. Job listing and the respective details have been scrapped for 300 companies collectively and twitter specific details have been called for the 300 companies into the database.

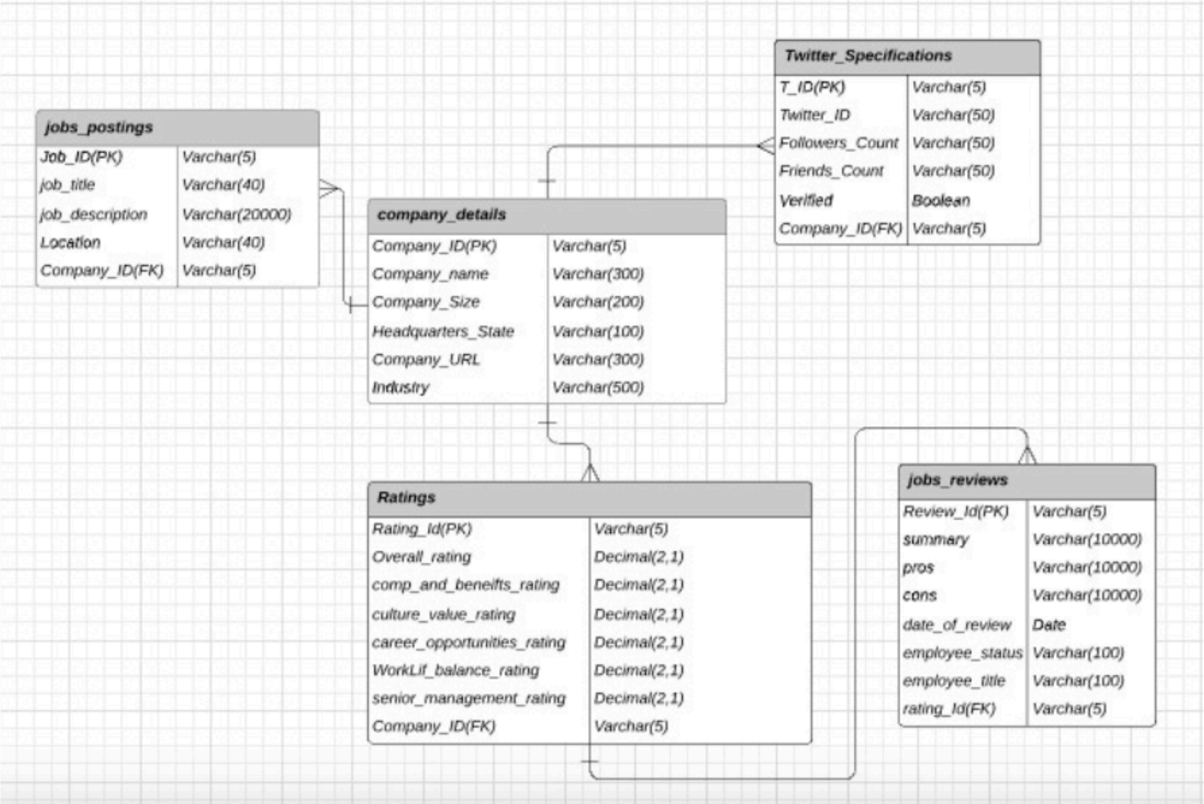
### DATA:

The job postings were specifically for United States and we have scraped about 100 jobs for each profile.
Some attributes about each job are:

1. Company Name
2. Job Description
3. Location
4. Company Reviews
5. Ratings

Apart from that, we have scraped twitter handles of these companies using Twitter API's which include their screen name, followers, friends and if profile is verified or not.
All this data was inserted into PostgreSQL database and use cases were built from this data.

The schema we came up with is:



| jobs_postings | |
|---|---|
| Job_ID(PK) | Varchar(5) |
| job_title | Varchar(40) |
| job_description | Varchar(20000) |
| Location | Varchar(40) |
| Company_ID(FK) | Varchar(5) |

| company_details | |
|---|---|
| Company_ID(PK) | Varchar(5) |
| Company_name | Varchar(300) |
| Company_Size | Varchar(200) |
| Headquarters_State | Varchar(100) |
| Company_URL | Varchar(300) |
| Industry | Varchar(500) |

| Twitter_Specifications | |
|---|---|
| T_ID(PK) | Varchar(5) |
| Twitter_ID | Varchar(50) |
| Followers_Count | Varchar(50) |
| Friends_Count | Varchar(50) |
| Verified | Boolean |
| Company_ID(FK) | Varchar(5) |

| Ratings | |
|---|---|
| Rating_Id(PK) | Varchar(5) |
| Overall_rating | Decimal(2,1) |
| comp_and_beneifts_rating | Decimal(2,1) |
| culture_value_rating | Decimal(2,1) |
| career_opportunities_rating | Decimal(2,1) |
| WorkLif_balance_rating | Decimal(2,1) |
| senior_management_rating | Decimal(2,1) |
| Company_ID(FK) | Varchar(5) |

| jobs_reviews | |
|---|---|
| Review_Id(PK) | Varchar(5) |
| summary | Varchar(10000) |
| pros | Varchar(10000) |
| cons | Varchar(10000) |
| date_of_review | Date |
| employee_status | Varchar(100) |
| employee_title | Varchar(100) |
| rating_Id(FK) | Varchar(5) |

## DESCRIPTION ABOUT TABLES:

- Job_Postings
- Company_Details
- Ratings
- Twitter_Specifications
- Reviews

1. **Job_Postings(Job Postings)**
   This table contains the following details:

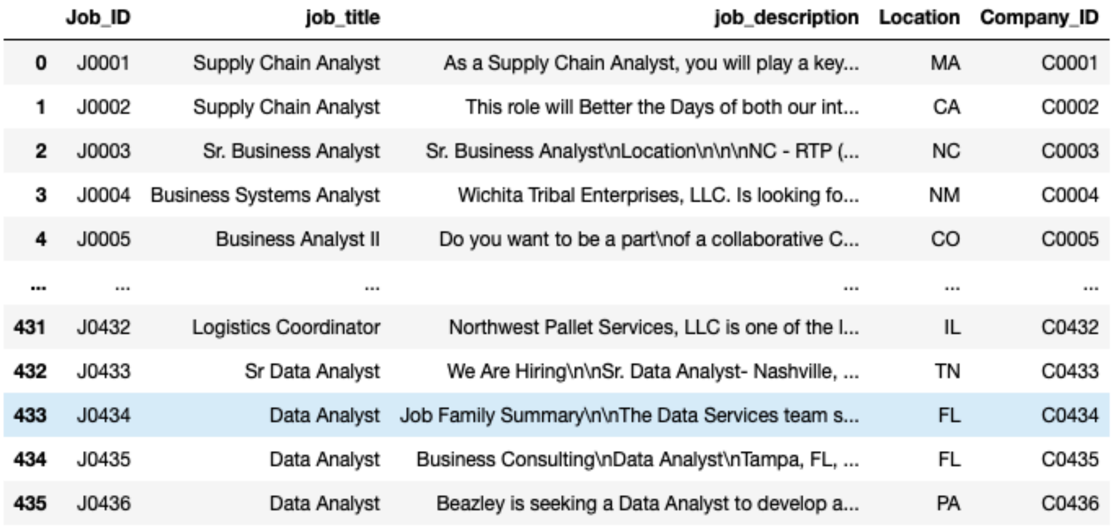
   **Job_ID(PK):** Since we collected data of 300 companies from Glassdoor, we created a unique ID for each company. This column is the unique ID of the table and each row can be uniquely identified by this ID.

   **Job_Title:** This column contains the position; the candidate is applying for. Ex: Business Analyst, Data Analyst

   **Job Description:** This column contains description about the job posting.

   **Location:** This column specifies the location where the position is situated in United States

   **CID(FK):** This column is a foreign key in the jobs table which is a primary key in the Company_Details table. This column is a unique identifier in the Company_Details table.



| | Job_ID | job_title | job_description | Location | Company_ID |
|---|---|---|---|---|---|
| 0 | J0001 | Supply Chain Analyst | As a Supply Chain Analyst, you will play a key... | MA | C0001 |
| 1 | J0002 | Supply Chain Analyst | This role will Better the Days of both our int... | CA | C0002 |
| 2 | J0003 | Sr. Business Analyst | Sr. Business Analyst\nLocation\n\n\nNC - RTP (... | NC | C0003 |
| 3 | J0004 | Business Systems Analyst | Wichita Tribal Enterprises, LLC. Is looking fo... | NM | C0004 |
| 4 | J0005 | Business Analyst II | Do you want to be a part\nof a collaborative C... | CO | C0005 |
| ... | ... | ... | ... | ... | ... |
| 431 | J0432 | Logistics Coordinator | Northwest Pallet Services, LLC is one of the l... | IL | C0432 |
| 432 | J0433 | Sr Data Analyst | We Are Hiring\n\nSr. Data Analyst- Nashville, ... | TN | C0433 |
| 433 | J0434 | Data Analyst | Job Family Summary\n\nThe Data Services team s... | FL | C0434 |
| 434 | J0435 | Data Analyst | Business Consulting\nData Analyst\nTampa, FL, ... | FL | C0435 |
| 435 | J0436 | Data Analyst | Beazley is seeking a Data Analyst to develop a... | PA | C0436 |

2. **Company_Details(Company Details)**
This table contains the following details:

**CID(PK):** Since we have data of about 300 companies, we have a unique identifier CID for every company. This column is also a foreign key in the Job_Postings table.

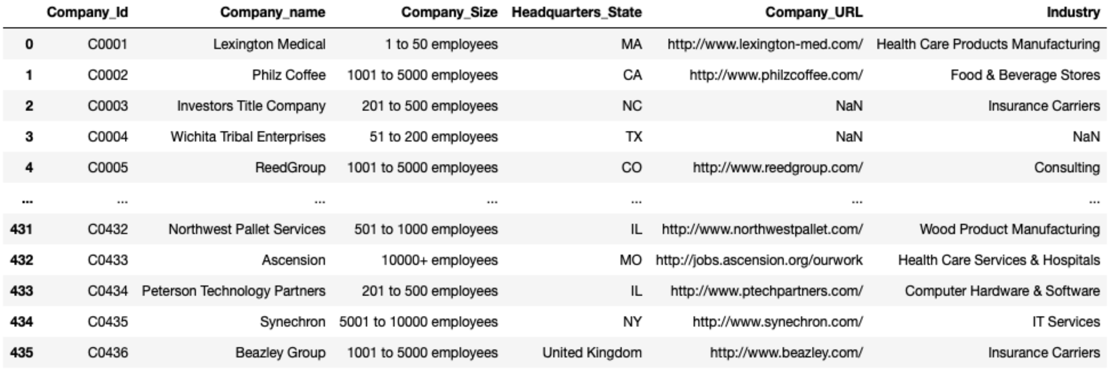
**Company_Name:** This attribute contains names of all companies whose jobs are posted.

**Headquarters:** This attribute specifies the headquarter location of the company.

**Company_Size:** This attribute specifies the employee strength of the company.

**Company_URL:** This attribute displays the company's website.

**Industry:** This attribute specifies which Industry the position belongs to. For eg: Banking, Manufacturing etc.



| | Company_Id | Company_name | Company_Size | Headquarters_State | Company_URL | Industry |
|---|---|---|---|---|---|---|
| 0 | C0001 | Lexington Medical | 1 to 50 employees | MA | http://www.lexington-med.com/ | Health Care Products Manufacturing |
| 1 | C0002 | Philz Coffee | 1001 to 5000 employees | CA | http://www.philzcoffee.com/ | Food & Beverage Stores |
| 2 | C0003 | Investors Title Company | 201 to 500 employees | NC | NaN | Insurance Carriers |
| 3 | C0004 | Wichita Tribal Enterprises | 51 to 200 employees | TX | NaN | NaN |
| 4 | C0005 | ReedGroup | 1001 to 5000 employees | CO | http://www.reedgroup.com/ | Consulting |
| ... | ... | ... | ... | ... | ... | ... |
| 431 | C0432 | Northwest Pallet Services | 501 to 1000 employees | IL | http://www.northwestpallet.com/ | Wood Product Manufacturing |
| 432 | C0433 | Ascension | 10000+ employees | MO | http://jobs.ascension.org/ourwork | Health Care Services & Hospitals |
| 433 | C0434 | Peterson Technology Partners | 201 to 500 employees | IL | http://www.ptechpartners.com/ | Computer Hardware & Software |
| 434 | C0435 | Synechron | 5001 to 10000 employees | NY | http://www.synechron.com/ | IT Services |
| 435 | C0436 | Beazley Group | 1001 to 5000 employees | United Kingdom | http://www.beazley.com/ | Insurance Carriers |

3. **Ratings**
This table contains the following details:

**RID(PK):** We have scraped data of 300 companies. This attribute is a unique identifier which specifies the rating of each company.

**Overall_ratings:** This attribute is an average of all the ratings of the company given by the employers.

**Comp_benefits_ratings:** This attribute specifies the ratings of compensation and benefits given by the company.

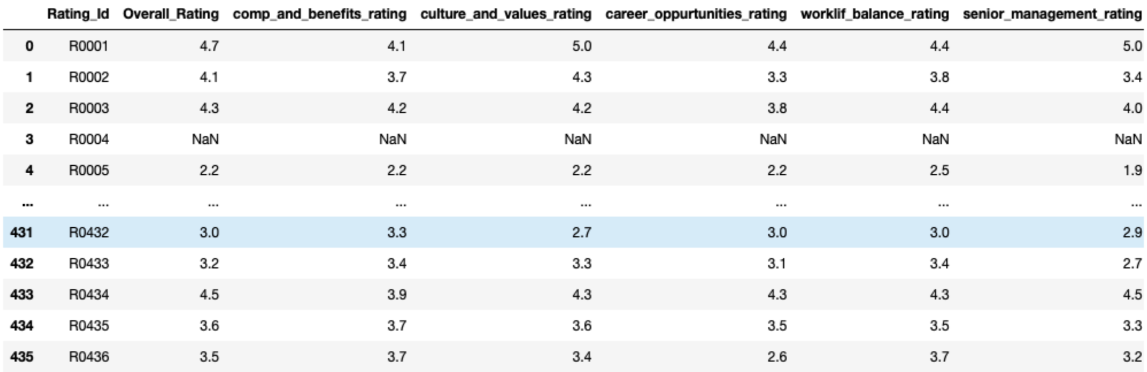
**Culture_value_ratings:** This attribute specifies the cultural value ratings of the company.

**Career_opportunities_ratings:** This attribute specifies the rating of career growth i.e how frequently an employer is promoted to higher positions.

**WorkLife_balance_ratings:** This attribute specifies the work life balance rating of the company.

**Senior_Management_Ratings:** This attribute specifies the ratings of the senior management of the company.

**CID(FK):** This attribute is a foreign key in the Ratings table which is a primary key in the Company_Details table. This column is a unique identifier in the Company_Details table.



| | Rating_Id | Overall_Rating | comp_and_benefits_rating | culture_and_values_rating | career_oppurtunities_rating | worklif_balance_rating | senior_management_rating |
|---|---|---|---|---|---|---|---|
| 0 | R0001 | 4.7 | 4.1 | 5.0 | 4.4 | 4.4 | 5.0 |
| 1 | R0002 | 4.1 | 3.7 | 4.3 | 3.3 | 3.8 | 3.4 |
| 2 | R0003 | 4.3 | 4.2 | 4.2 | 3.8 | 4.4 | 4.0 |
| 3 | R0004 | NaN | NaN | NaN | NaN | NaN | NaN |
| 4 | R0005 | 2.2 | 2.2 | 2.2 | 2.2 | 2.5 | 1.9 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 431 | R0432 | 3.0 | 3.3 | 2.7 | 3.0 | 3.0 | 2.9 |
| 432 | R0433 | 3.2 | 3.4 | 3.3 | 3.1 | 3.4 | 2.7 |
| 433 | R0434 | 4.5 | 3.9 | 4.3 | 4.3 | 4.3 | 4.5 |
| 434 | R0435 | 3.6 | 3.7 | 3.6 | 3.5 | 3.5 | 3.3 |
| 435 | R0436 | 3.5 | 3.7 | 3.4 | 2.6 | 3.7 | 3.2 |

4. **Twitter_Specifications(Twitter_Specifications):**
This table contains the following details:

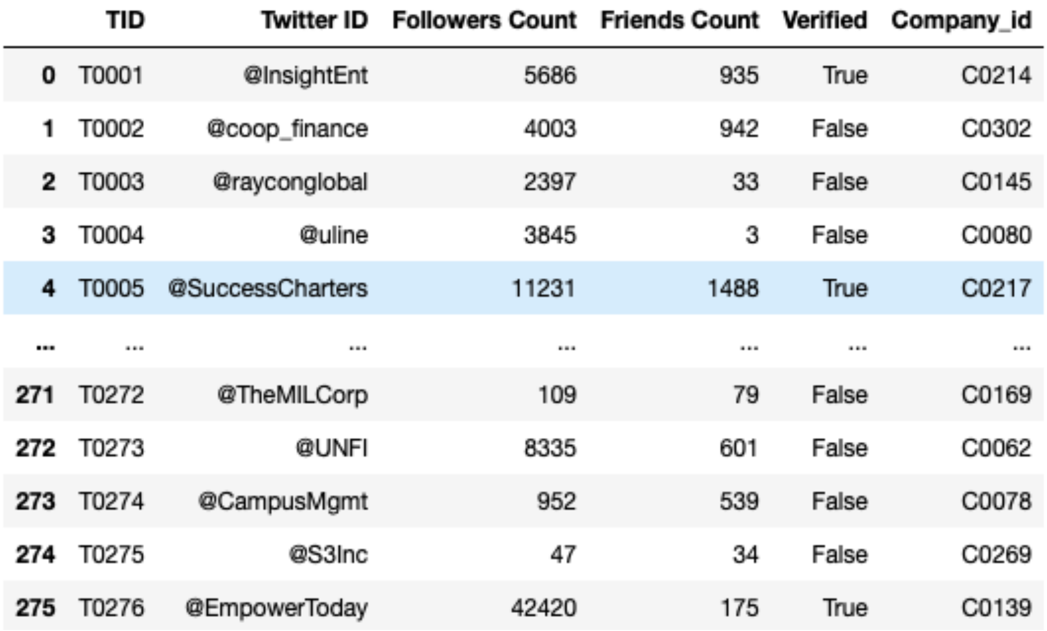
**Twitter_ID:** This attribute specifies the Twitter ID of all the companies. This ID is unique for every twitter handle.

**Followers_count:** This attributes specifies the Twitter followers of the company mentioned.

**Friends_count:** This attribute specifies which pages the company follows.

**Verified:** This attribute specifies if the company's twitter page is verified.

**CID(FK):** This attribute is a foreign key in the Twitter_Specifications table which is a primary key in the Company_Details table. This column is a unique identifier in the Company_Details table.



| | TID | Twitter ID | Followers Count | Friends Count | Verified | Company_id |
|---|---|---|---|---|---|---|
| **0** | T0001 | @InsightEnt | 5686 | 935 | True | C0214 |
| **1** | T0002 | @coop_finance | 4003 | 942 | False | C0302 |
| **2** | T0003 | @rayconglobal | 2397 | 33 | False | C0145 |
| **3** | T0004 | @uline | 3845 | 3 | False | C0080 |
| **4** | T0005 | @SuccessCharters | 11231 | 1488 | True | C0217 |
| **...** | ... | ... | ... | ... | ... | ... |
| **271** | T0272 | @TheMILCorp | 109 | 79 | False | C0169 |
| **272** | T0273 | @UNFI | 8335 | 601 | False | C0062 |
| **273** | T0274 | @CampusMgmt | 952 | 539 | False | C0078 |
| **274** | T0275 | @S3Inc | 47 | 34 | False | C0269 |
| **275** | T0276 | @EmpowerToday | 42420 | 175 | True | C0139 |

5. **Reviews:**
   This table contains the following details:

**ReviewID(PK):** This attribute is a primary key of the Reviews table. This attribute is the unique ID of the table and each row can be uniquely identified by this ID.

**Employee_Status:** This attribute specifies if the review posted by the employee is a current employee or a former employee.

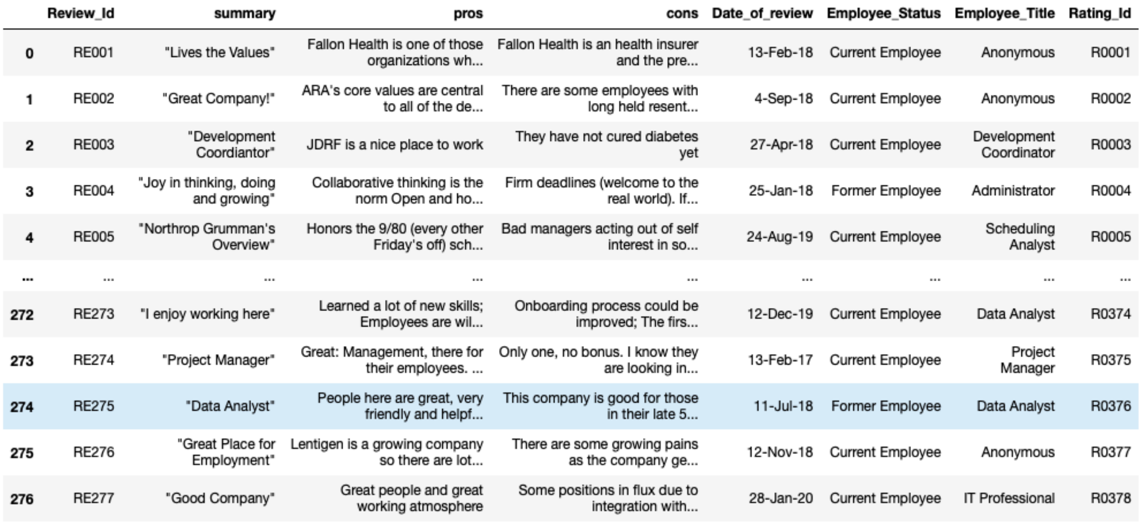
**Employee_title:** This attribute specifies the position of the employee who has posted the review.

**Latest_review:** This attribute specifies the most recent review about the company

**Pros:** This attribute specifies the pros of working in the company on basis of review posted by the employees.

**Cons:** This attribute specifies the cons of working in the company on basis of review posted by the employees.

**RID(FK):** This attribute is a foreign key in the reviews table. This attribute is a primary key of the ratings table and is a unique identifier in the ratings table.



| | Review_Id | summary | pros | cons | Date_of_review | Employee_Status | Employee_Title | Rating_Id |
|---|---|---|---|---|---|---|---|---|
| **0** | RE001 | "Lives the Values" | Fallon Health is one of those organizations wh... | Fallon Health is an health insurer and the pre... | 13-Feb-18 | Current Employee | Anonymous | R0001 |
| **1** | RE002 | "Great Company!" | ARA's core values are central to all of the de... | There are some employees with long held resent... | 4-Sep-18 | Current Employee | Anonymous | R0002 |
| **2** | RE003 | "Development Coordiantor" | JDRF is a nice place to work | They have not cured diabetes yet | 27-Apr-18 | Current Employee | Development Coordinator | R0003 |
| **3** | RE004 | "Joy in thinking, doing and growing" | Collaborative thinking is the norm Open and ho... | Firm deadlines (welcome to the real world). If... | 25-Jan-18 | Former Employee | Administrator | R0004 |
| **4** | RE005 | "Northrop Grumman's Overview" | Honors the 9/80 (every other Friday's off) sch... | Bad managers acting out of self interest in so... | 24-Aug-19 | Current Employee | Scheduling Analyst | R0005 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **272** | RE273 | "I enjoy working here" | Learned a lot of new skills; Employees are wil... | Onboarding process could be improved; The firs... | 12-Dec-19 | Current Employee | Data Analyst | R0374 |
| **273** | RE274 | "Project Manager" | Great: Management, there for their employees. ... | Only one, no bonus. I know they are looking in... | 13-Feb-17 | Current Employee | Project Manager | R0375 |
| **274** | RE275 | "Data Analyst" | People here are great, very friendly and helpf... | This company is good for those in their late 5... | 11-Jul-18 | Former Employee | Data Analyst | R0376 |
| **275** | RE276 | "Great Place for Employment" | Lentigen is a growing company so there are lot... | There are some growing pains as the company ge... | 12-Nov-18 | Current Employee | Anonymous | R0377 |
| **276** | RE277 | "Good Company" | Great people and great working atmosphere | Some positions in flux due to integration with... | 28-Jan-20 | Current Employee | IT Professional | R0378 |

## NORMALIZATION:

Once the tables have been created, the next step is to normalize the database. Normalization is done to reduce data redundancy. Data redundancy cannot be eliminated completely but it can be reduced by dividing the repeating columns in particular table and generate a unique ID to that table. Now we can give a unique ID to this table instead of repeating the columns and this unique ID can act as a link between these tables.

## 1ST NORMAL FORM:

A table is in first normal form if it is atomic and have no repeating rows and columns. All the tables in this database are in 1NF as they satisfy each requirement of first Normalization form.

## 2nd NORMAL FORM:

For a table to be in second normal form, two conditions must be satisfied.

1. It should be in 1NF
2. There should be no partial dependency, which means that no value in the table should be dependent on a part of primary key.

Our tables are in 2NF as they satisfy every requirement of second Normalization form.

## 3rd NORMAL FORM:

For a table to be in third normal form, two conditions must be satisfied.

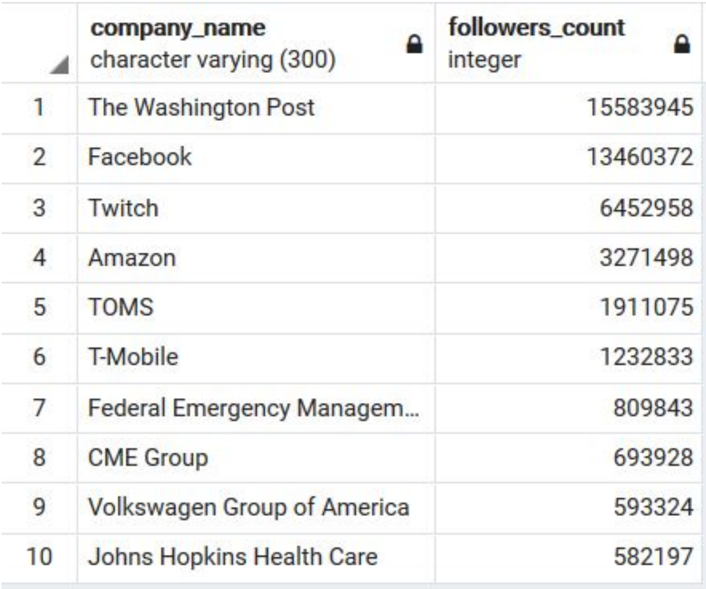
1. It should be in 2NF
2. No non primary attribute in the table should be dependent on other non-primary attribute in the table

Our tables are in 3NF as they satisfy every requirement of third Normalization form.

## USE-CASES:

### Use case-1: Querying the top 10 Popular Companies

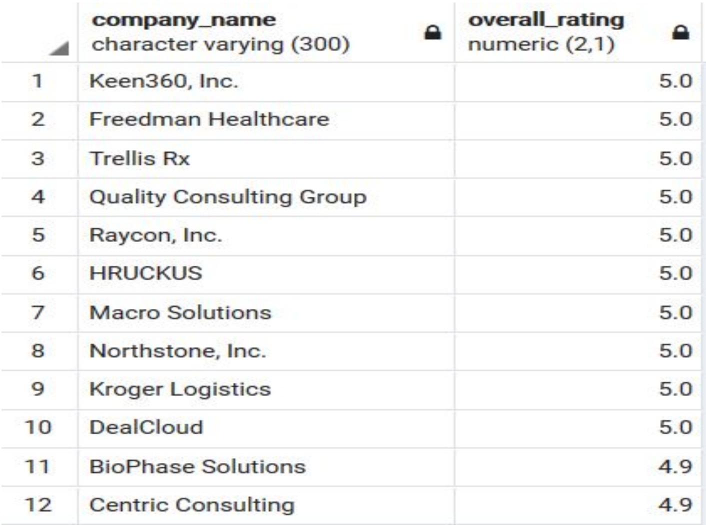
```
SELECT c.company_name, t.followers_count
FROM twitter_specifications t
LEFT JOIN company_details c
ON t.company_id = c.company_id
ORDER BY t.followers_count DESC
LIMIT 10;
```



| | company_name<br>character varying (300) | followers_count<br>integer |
|---|---|---|
| 1 | The Washington Post | 15583945 |
| 2 | Facebook | 13460372 |
| 3 | Twitch | 6452958 |
| 4 | Amazon | 3271498 |
| 5 | TOMS | 1911075 |
| 6 | T-Mobile | 1232833 |
| 7 | Federal Emergency Managem... | 809843 |
| 8 | CME Group | 693928 |
| 9 | Volkswagen Group of America | 593324 |
| 10 | Johns Hopkins Health Care | 582197 |

### Use case-2: Querying for top 50 companies having best overall ratings

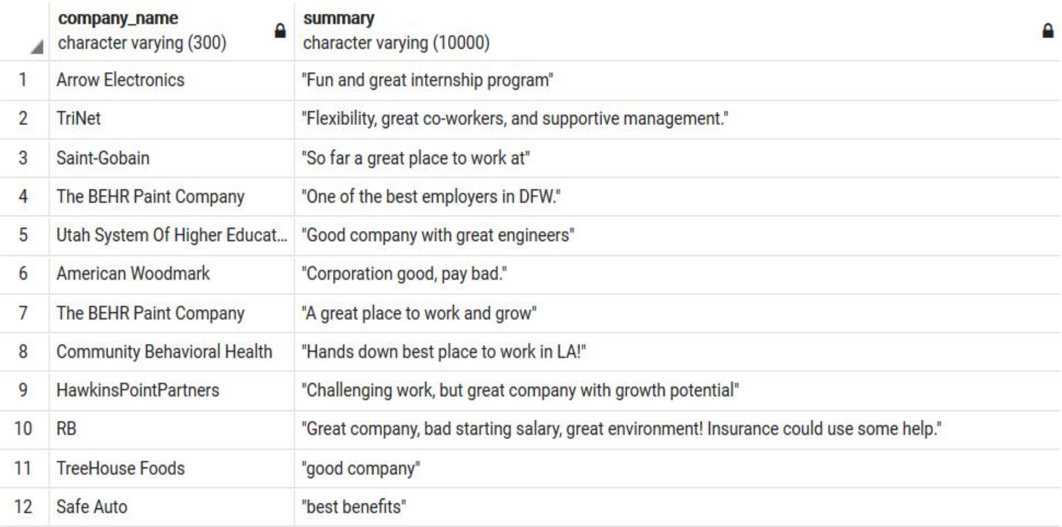
```
SELECT c.company_name, r.overall_rating
FROM company_details c
LEFT JOIN ratings r
ON c.company_id = r.company_id
WHERE r.overall_rating IS NOT NULL
ORDER BY r.overall_rating DESC
LIMIT 50;
```



| | company_name<br>character varying (300) | overall_rating<br>numeric (2,1) |
|---|---|---|
| 1 | Keen360, Inc. | 5.0 |
| 2 | Freedman Healthcare | 5.0 |
| 3 | Trellis Rx | 5.0 |
| 4 | Quality Consulting Group | 5.0 |
| 5 | Raycon, Inc. | 5.0 |
| 6 | HRUCKUS | 5.0 |
| 7 | Macro Solutions | 5.0 |
| 8 | Northstone, Inc. | 5.0 |
| 9 | Kroger Logistics | 5.0 |
| 10 | DealCloud | 5.0 |
| 11 | BioPhase Solutions | 4.9 |
| 12 | Centric Consulting | 4.9 |

## Use Case-3: Finding the companies which has good reviews

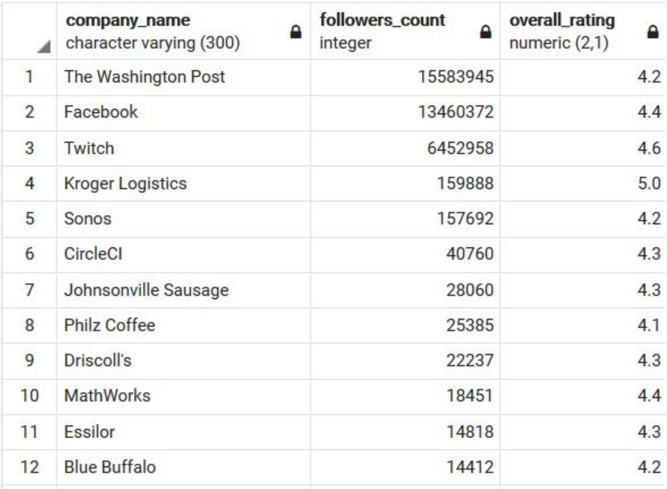
```
SELECT c.company_name, re.summary
FROM ratings r
LEFT JOIN company_details c ON r.company_id = c.company_id
LEFT JOIN Jobs_review re ON r.rating_id = re.rating_id
WHERE re.summary LIKE '%good%'
OR re.summary LIKE '%best%'

OR re.summary LIKE '%healthy%'
OR re.summary LIKE '%great%'
OR re.summary LIKE '%joy%'
OR re.summary LIKE '%supportive%'
```



| | company_name<br>character varying (300) | summary<br>character varying (10000) |
|---|---|---|
| 1 | Arrow Electronics | "Fun and great internship program" |
| 2 | TriNet | "Flexibility, great co-workers, and supportive management." |
| 3 | Saint-Gobain | "So far a great place to work at" |
| 4 | The BEHR Paint Company | "One of the best employers in DFW." |
| 5 | Utah System Of Higher Educat... | "Good company with great engineers" |
| 6 | American Woodmark | "Corporation good, pay bad." |
| 7 | The BEHR Paint Company | "A great place to work and grow" |
| 8 | Community Behavioral Health | "Hands down best place to work in LA!" |
| 9 | HawkinsPointPartners | "Challenging work, but great company with growth potential" |
| 10 | RB | "Great company, bad starting salary, great environment! Insurance could use some help." |
| 11 | TreeHouse Foods | "good company" |
| 12 | Safe Auto | "best benefits" |

## Use Case-4: Finding the companies having overall rating greater than 4 ordered by their popularity

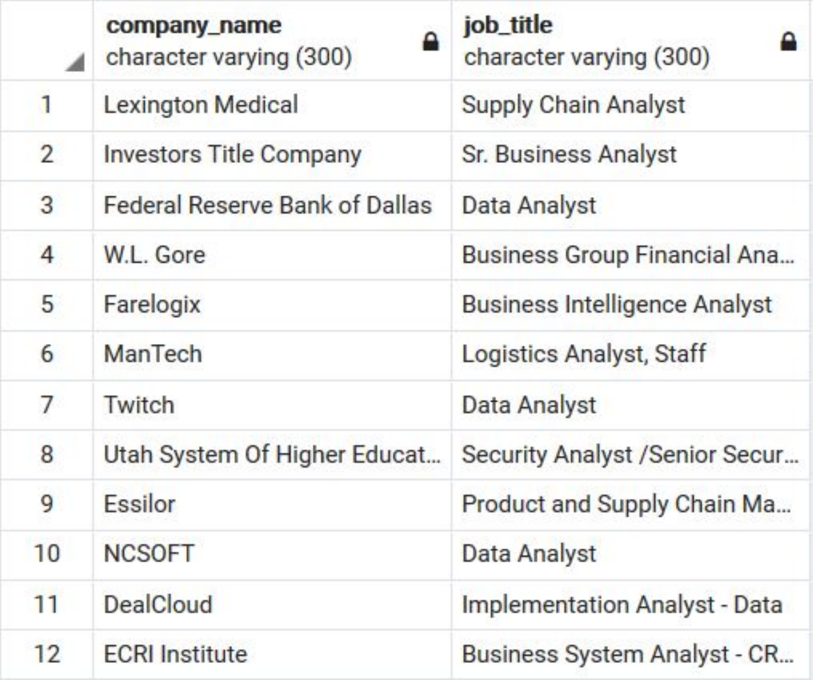
```
SELECT c.company_name, t.followers_count, r.overall_rating
FROM company_details c
LEFT JOIN ratings r ON r.company_id = c.company_id
LEFT JOIN twitter_specifications t ON t.company_id = c.company_id
WHERE t.followers_count IS NOT NULL AND r.overall_rating > 4
ORDER BY t.followers_count DESC
```



| | company_name<br>character varying (300) | followers_count<br>integer | overall_rating<br>numeric (2,1) |
|---|---|---|---|
| 1 | The Washington Post | 15583945 | 4.2 |
| 2 | Facebook | 13460372 | 4.4 |
| 3 | Twitch | 6452958 | 4.6 |
| 4 | Kroger Logistics | 159888 | 5.0 |
| 5 | Sonos | 157692 | 4.2 |
| 6 | CircleCI | 40760 | 4.3 |
| 7 | Johnsonville Sausage | 28060 | 4.3 |
| 8 | Philz Coffee | 25385 | 4.1 |
| 9 | Driscoll's | 22237 | 4.3 |
| 10 | MathWorks | 18451 | 4.4 |
| 11 | Essilor | 14818 | 4.3 |
| 12 | Blue Buffalo | 14412 | 4.2 |

Use Case-5: Finding the companies and their respective job title that has cultural values rating greater than 3.5 and worklife balance rating greater than 4

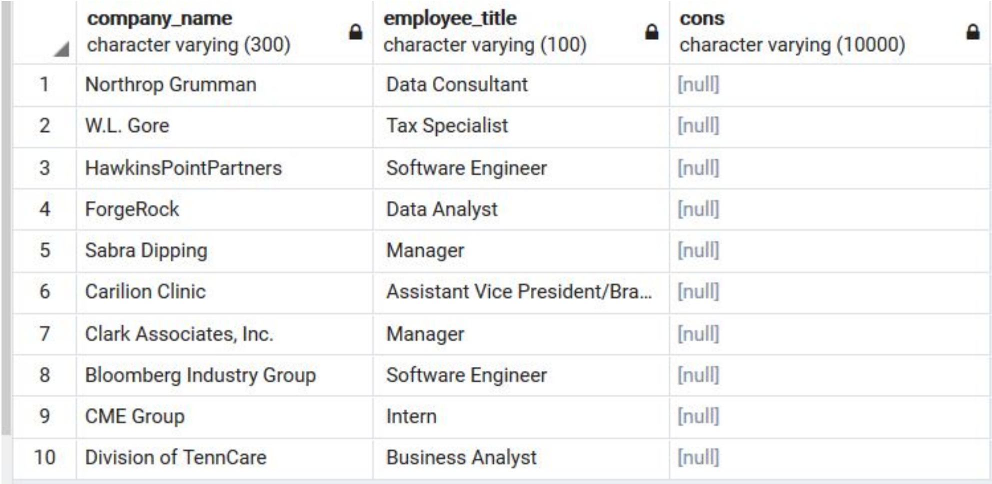
```
SELECT company_name, job_title
FROM job_postings j
LEFT JOIN ratings r ON j.company_id = r.company_id
LEFT JOIN company_details c on j.company_id = c.company_id
WHERE culture_values_rating > 3.5 AND worklif_balance_rating > 4
```



|  | company_name<br>character varying (300) | job_title<br>character varying (300) |
|---|---|---|
| 1 | Lexington Medical | Supply Chain Analyst |
| 2 | Investors Title Company | Sr. Business Analyst |
| 3 | Federal Reserve Bank of Dallas | Data Analyst |
| 4 | W.L. Gore | Business Group Financial Ana... |
| 5 | Farelogix | Business Intelligence Analyst |
| 6 | ManTech | Logistics Analyst, Staff |
| 7 | Twitch | Data Analyst |
| 8 | Utah System Of Higher Educat... | Security Analyst /Senior Secur... |
| 9 | Essilor | Product and Supply Chain Ma... |
| 10 | NCSOFT | Data Analyst |
| 11 | DealCloud | Implementation Analyst - Data |
| 12 | ECRI Institute | Business System Analyst - CR... |

Use Case-6: Querying for companies and respective job title that has no negative reviews

```
SELECT company_name, employee_title, cons
FROM ratings r
FULL OUTER JOIN jobs_review j ON j.rating_id = r.rating_id
FULL OUTER JOIN company_details c on c.company_id = r.company_id
WHERE cons IS NULL AND employee_title IS NOT NULL;
```



|  | company_name<br>character varying (300) | employee_title<br>character varying (100) | cons<br>character varying (10000) |
|---|---|---|---|
| 1 | Northrop Grumman | Data Consultant | [null] |
| 2 | W.L. Gore | Tax Specialist | [null] |
| 3 | HawkinsPointPartners | Software Engineer | [null] |
| 4 | ForgeRock | Data Analyst | [null] |
| 5 | Sabra Dipping | Manager | [null] |
| 6 | Carilion Clinic | Assistant Vice President/Bra... | [null] |
| 7 | Clark Associates, Inc. | Manager | [null] |
| 8 | Bloomberg Industry Group | Software Engineer | [null] |
| 9 | CME Group | Intern | [null] |
| 10 | Division of TennCare | Business Analyst | [null] |

## CONCLUSION:

Our primary focus of this project was to build a structured database for the selected job domains and its social media activities. We have successfully built the database using PostgreSQL and normalized tables without any redundancy.

## CITATION:

https://stackoverflow.com/

https://www.w3schools.com/sql

https://www.lucidchart.com

https://github.com/arapfaik/scraping-glassdoor-selenium/blob/master/glassdoor%20scraping.ipynb

## LICENSE: