INFO 6210 Project JOBS DATABASE

By

- Meda Saikanth Reddy Neuid: 001389103
- Naga vuyyuru Neuid: 001344678
- Vennela Reddy Neuid: 001081643

ABSTRACT

In this project we will contribute to building a jobs database using mysql and python with proper documentation. We will each focus on job domain Software Engineering and Data Science. The database is focused around an employer like a company and research lab

PROJECT

You have been given the beginnings of a jobs database but the previous developer was let go because the organization of the previous project is a mess. The developers all worked independently and the tables and schemas didn't match. There are some working scripts but they sometimes replicate and some are missing. Some scripts have bugs and may be in python 2.7. Data is in .csv, json, sqlite and mysql Your job is to salvage what data, schemas and scripts that you can and add what is missing.
Everything must be done in mysql using the InnoDB engine. All python scripst must be in python 3 or above and using Google Python Style Guide

Tasks

- Build the list of companies in some domain.
- Web scrape the data for each of the companies for job details.
- Automate the scraping.
- Get additional relevant data from sites like Glassdoor, LinkedIn (one per person)
- Get data from social media sites – Twitter, YouTube, Instagram, Steam, Twitter, etc.. (One per person)
- Tag the social media posts, including synonyms for the tags
- Clean and integrate data.
- Build an ER diagram and model the db.
- Build the dB schema and insert the data
- Generate use cases.
- Optimize the database.
- Properly document that database
- Professionalism (Licensing, code style, file naming, README. Etc.)

2. Web scrape the data for each of the companies for job details

In [7]:

```python
# imports
import numpy as np
import pandas as pd
import seaborn as sns
import requests
from bs4 import BeautifulSoup
```

In [8]:

```python
def indeed_jobs_scrapper(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')
    return soup

soup = indeed_jobs_scrapper("https://www.indeed.com/jobs?q=data+scientist&l=U
```

In [9]:

```python
def scrape_jobtitle(soup):
    jobs = []
    for div in soup.find_all(name="div", attrs={"class":"row"}):
        for a in div.find_all(name="a", attrs={"data-tn-element":"jobTitle"}
            jobs.append(a["title"])
    return(jobs)

scrape_jobtitle(soup)
```

Out[9]:

```
['Data Scientist',
 'Data Scientist',
 'Data Scientist, Office of Data Science',
 'Data Scientist',
 'Principal Data Scientist',
 'Actuarial Services + Data Science Intern',
 'Data Scientist / Data Analytics',
 'Data Scientist/Machine Learning Engineer',
 'Data Scientist Entry Level - Pathrise Recruiting Partners',
 'Data Scientist',
 'Data Scientist Analyst',
 'Data Scientist',
 'Data Scientist - 68924BR',
 'Content Data Scientist',
 'Data Scientist',
 'Jr. Data Scientist',
 'Junior Data Scientist',
 'Analyst II, Data Science',
 'Data Scientist, Medical Diagnostics']
```

In [10]:

```python
def extract_company_from_result(soup):
    companies = []
    for div in soup.find_all(name="div", attrs={"class":"row"}):
        company = div.find_all(name="span", attrs={"class":"company"})
        if len(company) > 0:
            for b in company:
                companies.append(b.text.strip())
        else:
            sec_try = div.find_all(name="span", attrs={"class": "result-link
            for span in sec_try:
                companies.append(span.text.strip())
    return(companies)

extract_company_from_result(soup)
```

Out[10]:

```
['Triplebyte',
 'ClearOne Advantage',
 'Liberty Mutual Insurance',
 'Seen by Indeed',
 'Intuit',
 'Commonwealth Care Alliance, Inc.',
 'Tredence Inc.',
 'Mobile Insights',
 'Pathrise',
 'Conrelv Solutions Inc',
 'LOCKHEED MARTIN CORPORATION',
 'TISAA',
 'AETNA',
 'Buxton',
 'Foundation Medicine, Inc.',
 'Numero Data LLC',
 '1-800-Flowers',
 'Liberty Mutual Insurance',
 'Specific Diagnostics']
```

In [11]:

```python
def extract_location_from_result(soup):
    locations = []
    spans = soup.findAll("span", attrs={"class": "location"})
    for span in spans:
        locations.append(span.text)
    return(locations)
extract_location_from_result(soup)
```

Out[11]:

```
['New York, NY',
 'Remote',
 'Boston, MA',
 'Sunnyvale, CA 94089 (Lakewood area)',
 'Seattle, WA',
 'Wellesley, MA 02481',
 'Seattle, WA',
 'Boston, MA 02210 (South Boston area)',
 'Herndon, VA 20170',
 'New York, NY 10013 (Tribeca area)']
```

In [12]:

```python
def extract_salary_from_result(soup):
    salaries = []
    for div in soup.find_all(name="div", attrs={"class":"row"}):
        try:
            salaries.append(div.find("nobr").text)
        except:
            try:
                div_two = div.find(name="div", attrs={"class":"salarySnippet'
                salaries.append(div_two.text.strip())
            except:
                salaries.append("Not Posted")
    return(salaries)

extract_salary_from_result(soup)
```

Out[12]:

```
['$150,000 - $225,000 a year',
 '$70,000 - $80,000 a year',
 '$93,400 - $134,100 a year',
 'Not Posted',
 'Not Posted',
 'Not Posted',
 '$100,000 - $130,000 a year',
 '$85,000 - $115,000 a year',
 'Not Posted',
 'Not Posted',
 'Not Posted',
 'Not Posted',
 'Not Posted',
 'Not Posted',
 'Not Posted',
 '$70,000 - $85,000 a year',
 'Not Posted',
 '$89,700 - $148,800 a year',
 'Not Posted']
```

In [13]:

```
def extract_summary_from_result(soup):
    summaries = []
    spans = soup.findAll("div", attrs={"class": "summary"})
    for span in spans:
        summaries.append(span.text.strip())
    return(summaries)
extract_summary_from_result(soup)
```

Out[13]:

```
["You’ll report directly to Triplebytes' Head of Machine Learning and wi
ll work alongside a team of 6-8 machine learning engineers and data scie
ntists.",
 'We want to see a passion for machine-learning and research.\nBuild pre
dictive models and machine-learning algorithms.\nCombine models through
ensemble modeling.',
 'Demonstrated experience in deep learning, computer vision, natural lan
guage processing, and/or interpretable machine learning.',
 'With one application you can be considered for thousands of tech roles
from leading companies on Seen. Seen by Indeed is a free service that co
nnects you to…',
 'Intuit’s Innovation and Advanced Technology Group is hiring a Data Sci
entist to focus on Security and Anti-fraud.',
 'Use programming and mathematical tools to solve important problems.\nE
xperience with Python, git, SQL, healthcare data.',
 'Data analytics: 3 years (Preferred).\nLead and manage independently th
e onsite-offshore relation, at the same time adding value to the clien
t.',
 'Develop machine learning applications according to requirements.\nRun
machine learning tests and experiments.\nFamiliarity with machine learni
ng frameworks (like…',
 '0-3 years in data science.\nIn these positions you will be asked to ma
nipulate and utilize data in order to inform key business decisions and
model various…',
 'They will be creating models to use machine learning to identify that
customer, then using that info to do outreach to customer by contact and
marketing towards…',
 'Work on datasets with applied statistics and machine learning algorith
ms; Use exploratory data analysis techniques to identify meaningful rela
tionships,…',
 'Content Data Scientist This is true data scientist who will be working
on content efficiency modelling, taste personas, content acquisition, fo
recast, Avod…',
 'Demonstrates proficiency in several areas of data modeling, machine le
arning algorithms, statistical analysis, data engineering and data visua
lization.',
 'Content Data Scientist - This is true data scientist who will be worki
ng on content efficiency modelling, taste personas, content acquisition,
forecast, Avod…',
 'Your focus will be on clinical use cases, such as biomarker-based outc
omes analyses, examining correlates of genomics and clinical outcomes, c
linical utility of…',
 'Ability to break down and understand complex business problems, define
a solution and implement it using advanced quantitative methods.',
 'As Junior Data Scientist, you will be responsible for collecting, clea
ning, and extracting data from a variety of systems at 1-800-flowers wit
h intention to run…',
 'The position requires a Master’s degree, or foreign equivalent, in Sta
tistics, Mathematics, Economics, or another scientific field plus one
(1) year of…',
 'Used for bloodstream infection Specific’s solution provides results 2
days sooner than existing methods, saving patients suffering from drug-r
esistant infection…']
```

Automating the scrapping of Indeed

In [14]:

```
roles_list = ["data+scientist", "game+designer", "software+engineer"]

url = "https://www.indeed.com/jobs?q="

columns = ["job_title", "company_name", "location", "summary", "salary"]
df = pd.DataFrame(columns = columns)

for role in roles_list:
    first = 0
    while first <= 1000:
        if first == 0:
            req_url = url + role + "&l=United+States"
            first = 10
        else:
            req_url = url + role + "&l=United+States" + "&start=" + str(first
            first = first + 10
        soup = indeed_jobs_scrapper(req_url)
        for div in soup.find_all(name="div", attrs={"class":"row"}):

            num = (len(df) + 1)
            jobs_post = []

            #grabbing title
            for a in div.find_all(name="a", attrs={"data-tn-element":"jobTitl
                jobs_post.append(a["title"])

            #grabbing company
            company = div.find_all(name="span", attrs={"class":"company"})
            if len(company) > 0:
                for b in company:
                    jobs_post.append(b.text.strip())
            else:
                sec_try = div.find_all(name="span", attrs={"class": "result-l
                for span in sec_try:
                    jobs_post.append(span.text.strip())

            #grabbing location name
            spans = div.findAll("div", attrs={"class": "location"})
            if len(spans) == 0:
                jobs_post.append("Anywhere")
            else:
                jobs_post.append(spans[0].text)


            #grabbing summary text
            spans = div.findAll("div", attrs={"class": "summary"})
            for span in spans:
                jobs_post.append(span.text.strip())

            #grabbing salary
            try:
                jobs_post.append(div.find("nobr").text)
            except:
                try:
                    div_two = div.find(name="div", attrs={"class":"salarySnip
                    jobs_post.append(div_two.text.strip())
                except:
                    jobs_post.append("Not Posted")

            print(jobs_post)
            #appending list of job post info to dataframe at index num
            df.loc[num] = jobs_post
```

```
['Data Scientist', 'Triplebyte', 'Remote', "You'll report directly to Tr
iplebytes' Head of Machine Learning and will work alongside a team of 6-
8 machine learning engineers and data scientists.", '$150,000 - $225,000
 a year']
['Data Scientist, Medical Diagnostics', 'Specific Diagnostics', 'Mountai
n View, CA 94043', 'Used for bloodstream infection Specific's solution p
rovides results 2 days sooner than existing methods, saving patients suf
fering from drug-resistant infection…', 'Not Posted']
['Data Scientist', 'ClearOne Advantage', 'Baltimore, MD 21224 (Canton In
dustrial Area area)', 'We want to see a passion for machine-learning and
 research.\nBuild predictive models and machine-learning algorithms.\nCom
bine models through ensemble modeling.', '$70,000 - $80,000 a year']
['Analyst II, Data Science', 'Liberty Mutual Insurance', 'Boston, MA 021
01', 'The position requires a Master's degree, or foreign equivalent, in
 Statistics, Mathematics, Economics, or another scientific field plus one
 (1) year of…', '$89,700 - $148,800 a year']
['Data Scientist/Machine Learning Engineer', 'Mobile Insights', 'Anywher
e', 'Develop machine learning applications according to requirements.\nR
un machine learning tests and experiments.\nFamiliarity with machine lea
```

In [15]:

```
df.head()
```

Out[15]:

	job_title	company_name	location	summary	salary
1	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of...	150, 000−225,000 a year
2	Data Scientist, Medical Diagnostics	Specific Diagnostics	Mountain View, CA 94043	Used for bloodstream infection Specific's solu...	Not Posted
3	Data Scientist	ClearOne Advantage	Baltimore, MD 21224 (Canton Industrial Area area)	We want to see a passion for machine-learning ...	70, 000−80,000 a year
4	Analyst II, Data Science	Liberty Mutual Insurance	Boston, MA 02101	The position requires a Master's degree, or fo...	89, 700−148,800 a year
5	Data Scientist/Machine Learning Engineer	Mobile Insights	Anywhere	Develop machine learning applications accordin...	85, 000−115,000 a year

Creating an excel sheet of jobs

In [212]:
```
df.to_csv("./data/jobs.csv",encoding="utf-8",index=False)
```

In [16]:
```
unique_companies = set()
for i in df['company_name'].tolist():
    unique_companies.add(i)

unique_companies
```

Out[16]:
```
{'University of Pennsylvania Health System',
 'The Oakleaf Group',
 'LOGIXTech Solutions',
 '2U',
 'SelectMinds',
 'Ameriprise Financial',
 'BTMG USA',
 'Animus Studios',
 'Age of Learning',
 'Alt Shift USA',
 'Tuvli, LLC',
 'IT Synergy',
 'Open Clinica',
 'Parametric',
 'The Ash Group',
 'bellevue university',
 'Affirmed Networks Inc.',
 'Navitus Health Solutions / Lumicera Health Service...',
 'Quantum Mechanix',
```

In [167]:
```
# importing libraries required for downloading data
import tweepy
import twitter

# keys for accesing twitter api
consumerKey = 'AChFuchA4E4ywFLw02TY5vDHF'
consumerSecret = 'ZhsHMVkC8UnVb6xs1fI9Y1vubjFk58kptUWNIWoAbyi7F6LtGz'
ACCESS_TOKEN = '2483851159-GOBy7a31beVCmRvaAMcDF2M70AjReBJfCdVxGux'
ACCESS_SECRET = 'V5LERc12DKFcI0nNHPlrSGzs19Lq8Z6GJf8TXyWO2mn1m'

auth = tweepy.OAuthHandler(consumer_key=consumerKey, consumer_secret=consumer
#Connect to the Twitter API using the authentication
api = tweepy.API(auth)
```

In [169]:

```
results = []

try:
    #Get the first 5000 items based on the search query
    for company in unique_companies:
        search_q = '%'+ company
        for tweet in tweepy.Cursor(api.search, q=search_q, since='2019-04-04'
            results.append(tweet)
except tweepy.error.TweepError:
    raise

# Verify the number of items returned print
len(results)
```

Youtube Api set up

In [19]:

```
youtube_api_key = "AIzaSyCDXrUjT7cSSTZ7CkknhPi7DmHw6_Mj2aw"

from apiclient.discovery import build

youtube = build('youtube', 'v3', developerKey=youtube_api_key)
type(youtube)
```

Out[19]:

```
googleapiclient.discovery.Resource
```

Accessing video search api

In [20]:
```
# for company in unique_companies:
req = youtube.search().list(q="slalom build careers",part="snippet", type="vi
items = req.execute()['items']

items[0]
```

Out[20]:
```
{'kind': 'youtube#searchResult',
 'etag': '"nxOHAKTVB7baOKsQgTtJIyGxcs8/S9LlTo9MHP7aVTnazZ-2zhegWvc"',
 'id': {'kind': 'youtube#video', 'videoId': 'ON7h1AFAm3c'},
 'snippet': {'publishedAt': '2015-02-16T22:59:14.000Z',
  'channelId': 'UCfZs5rUpJk3KuISZkEBU1qg',
  'title': 'Slalom Boston',
  'description': "Slalom Boston's office is growing like gangbusters. Combi
ning the best local talent with an energetic market and innovative clients,
Slalom Boston is a great ...",
  'thumbnails': {'default': {'url': 'https://i.ytimg.com/vi/ON7h1AFAm3c/def
ault.jpg',
    'width': 120,
    'height': 90},
   'medium': {'url': 'https://i.ytimg.com/vi/ON7h1AFAm3c/mqdefault.jpg',
    'width': 320,
    'height': 180},
   'high': {'url': 'https://i.ytimg.com/vi/ON7h1AFAm3c/hqdefault.jpg',
    'width': 480,
    'height': 360}},
  'channelTitle': 'Slalom',
  'liveBroadcastContent': 'none'}}
```

Accessing video stats api

In [21]:
```
reqStats =  youtube.videos().list(part="statistics",id="ON7h1AFAm3c")
reqStats.execute()
```

Out[21]:
```
{'kind': 'youtube#videoListResponse',
 'etag': '"nxOHAKTVB7baOKsQgTtJIyGxcs8/BAqpPz3yxJF8uYw_uRNqkqk_Aog"',
 'pageInfo': {'totalResults': 1, 'resultsPerPage': 1},
 'items': [{'kind': 'youtube#video',
   'etag': '"nxOHAKTVB7baOKsQgTtJIyGxcs8/JG3D6WTKBNFnoEzqQ7hYWAnRCc0"',
   'id': 'ON7h1AFAm3c',
   'statistics': {'viewCount': '994',
    'likeCount': '6',
    'dislikeCount': '0',
    'favoriteCount': '0',
    'commentCount': '0'}}]}
```

Automating the youtube search process

In [210]:

```
# data frame for youtube videos
columns = ["videoId", "channelId", "title", "description","viewCount", "likeC
youtube_df = pd.DataFrame(columns=columns)

# dictionary to create a dataFrame for youtube channels data
channelsData = {"channelId": [], "channelTitle": []}
# list of unique channels
channels = []

for company in list(unique_companies)[0:20]:
    req = youtube.search().list(q= company +" careers",part="snippet", type='
    items = req.execute()['items']
    for item in items:
        # index to append data to dataFrame
        index = len(youtube_df)+1

        # list to keep all data regarding youtube video
        video_data = []
        video_data.append(item["id"]["videoId"])
        video_data.append(item["snippet"]["channelId"])

        # if channel not in list add it and in dict
        channel = item["snippet"]["channelId"]
        if channel not in channels:
            channels.append(channel)
            channelsData["channelId"].append(channel)
            channelsData["channelTitle"].append(item["snippet"]["channelTitle

        video_data.append(item["snippet"]["title"])
        video_data.append(item["snippet"]["description"])

        # req api for this video's statistics on youtube
        reqStats =  youtube.videos().list(part="statistics",id="ON7h1AFAm3c")
        video_stats = reqStats.execute()["items"][0]

        # add statistics data to list
        video_data.append(video_stats["statistics"]["viewCount"])
        video_data.append(video_stats["statistics"]["likeCount"])
        video_data.append(video_stats["statistics"]["dislikeCount"])
        video_data.append(video_stats["statistics"]["favoriteCount"])
        video_data.append(video_stats["statistics"]["commentCount"])
        video_data.append(company)

        youtube_df.loc[index] = video_data
```

Building channels Dataframe from collected data

In [34]:
```
channels_df = pd.DataFrame.from_dict(channelsData)
channels_df.head()
```

Out[34]:

	channelId	channelTitle
0	UCXubLFOt4iiX2_0tYQD8gRA	Penn State Health
1	UCUngw5TivNYk845EpJQ1Mfg	Penn Commercial Business/Technical School
2	UCb_JeH-0SzKbKOqSe0DZnmA	Pennsylvania College of Technology
3	UC36NIm8ikeZ4tDRx__BjJnA	Penn Medicine
4	UCSC8V1ez4zt3rviyPWzk9Sg	Cincinnati Children's

Creating youtube channels data in excel

In [213]:
```
channels_df.to_csv("./data/youtubeChannels.csv",encoding="utf-8",index=False)
```

Youtube dataframe

In [37]:
```
youtube_df.head()
```

Out[37]:

eoId	channelId	title	description	viewCc
biM8	UCXubLFOt4iiX2_0tYQD8gRA	Penn State Health - Careers	At Penn State Health, we work to provide the b...	
Jbx4	UCUngw5TivNYk845EpJQ1Mfg	Your Career in Healthcare Starts at Penn Comme...	Dr. John D. Six, M.D., Vice President of Medic...	
olsw	UCb_JeH-0SzKbKOqSe0DZnmA	Health Information Degrees at Penn College	https://www.pct.edu/academics/hs/healthIT Heal...	
Ib2w	UC36NIm8ikeZ4tDRx__BjJnA	Penn Medicine&#39;s Global Nurse Program	In response to worldwide nursing concerns, the...	
Kvw	UCSC8V1ez4zt3rviyPWzk9Sg	Immunology Graduate Program | Cincinnati Child...	The study of immunology is critical to our sur...	

In [214]:
```
youtube_df.to_csv("./data/youtubeVideos.csv",encoding="utf-8",index=False)
```

Glassdoor Reviews

In [63]:

```
def glassdoor_ratings_scrapper(url):
    headers = { 'accept': 'text/html,application/xhtml+xml,application/xml;q=
                'accept-encoding': 'gzip, deflate, sdch, br',
             'accept-language': 'en-GB,en-US;q=0.8,en;q=0.6',
              'referer': 'https://www.glassdoor.com/',
    'upgrade-insecure-requests': '1',
    'user-agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML,
    'Cache-Control': 'no-cache',
    'Connection': 'keep-alive'
    }

    location_headers = {
        'accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,imag
        'accept-encoding': 'gzip, deflate, sdch, br',
        'accept-language': 'en-GB,en-US;q=0.8,en;q=0.6',
        'referer': 'https://www.glassdoor.com/',
        'upgrade-insecure-requests': '1',
        'user-agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KH
        'Cache-Control': 'no-cache',
        'Connection': 'keep-alive'
        }
    r = requests.get(url, headers=headers)
    soup = BeautifulSoup(r.text, 'html.parser')
    return soup

soup = glassdoor_ratings_scrapper("https://www.glassdoor.com/Reviews/Google-R
```

In [64]:

```
soup
```

```
ry:32":{"type":"COUNTRY","id":32,"identString":"N,32","name":"Benin","co
ntainsEmployerHQ":false,"states":[{"type":"id","generated":false,"id":"S
tate:-32","typename":"State"}],"__typename":"Country"},"State:-32":{"typ
e":"STATE","id":-32,"identString":null,"name":null,"containsEmployerHQ":
false,"metros":[{"type":"id","generated":false,"id":"Metro:1254","typena
me":"Metro"}],"__typename":"State"},"Metro:1254":{"type":"METRO","id":12
54,"identString":"M,1254","name":"Porto-Novo, Benin Area","containsEmplo
yerHQ":false,"cities({\"onlyIfOther\":true})":null,"__typename":"Metr
o"},"Country:30":{"type":"COUNTRY","id":30,"identString":"N,30","nam
e":"Bolivia","containsEmployerHQ":false,"states":[{"type":"id","generate
d":false,"id":"State:3895","typename":"State"}],"__typename":"Countr
y"},"State:3895":{"type":"STATE","id":3895,"identString":"S,3895","nam
e":"Cochabamba","containsEmployerHQ":false,"metros":[{"type":"id","gener
ated":false,"id":"Metro:3049","typename":"Metro"}],"__typename":"Stat
e"},"Metro:3049":{"type":"METRO","id":3049,"identString":"M,3049","nam
e":"Cochabamba, Bolivia Area","containsEmployerHQ":false,"cities({\"only
IfOther\":true})":null,"__typename":"Metro"},"Country:36":{"type":"COUNT
RY","id":36,"identString":"N,36","name":"Brazil","containsEmployerHQ":fa
lse,"states":[{"type":"id","generated":false,"id":"State:3919","typenam
```

In [156]:

```
def scrape_reviews(soup, company):
    companyName = []
    reviewSummary = []
    reviewLink = []
    pros = []
    cons = []
    # iterating list of reviews in a page
    for li in soup.find_all(name="li", attrs={"class":"empReview"}):
        companyName.append(company)
        # header for summary
        h2 = li.find(name="h2", attrs={"class":"summary"})
        # link for individual review
        a = h2.find(name="a", attrs={"class": "reviewLink"})
        reviewSummary.append(a.text)
        reviewLink.append("https://www.glassdoor.com/" + a.get('href'))
        # div for pros and cons
        div = li.find(name="div", attrs={"class": "row"})
        p = div.find_all(name="p", attrs={"class": "mt-0"})
        pros.append(p[0].text)
        cons.append(p[1].text)
#         print(reviewLink)

    return reviewSummary,reviewLink,pros,cons,companyName

scrape_reviews(soup, "Google")
```

Out[156]:

```
(['"One of the best places to work."',
  '"Moving at the speed of light, burn out is inevitable"',
  '"Great balance between big-company security and fun, fast-moving proj
ects"',
  '"The best place I\'ve worked and also the most demanding."',
  '"Amazing culture"',
  '"Great"',
  '"Best in Class"',
  '"Cool"',
  '"N/A"',
  '"A machine"'],
 ['https://www.glassdoor.com//Reviews/Employee-Review-Google-RVW3286794
4.htm',
  'https://www.glassdoor.com//Reviews/Employee-Review-Google-RVW2757802.
htm',
  'https://www.glassdoor.com//Reviews/Employee-Review-Google-RVW4204034.
htm',
  'https://www.glassdoor.com//Reviews/Employee-Review-Google-RVW5873129.
htm',
```

In [157]:

```
# Slalom Consulting , Accenture, Snapchat, Twitter, Amazon Web Services, Appl
```

In [158]:

```
eviewsLinks = ["https://www.glassdoor.com/Reviews/Google-Reviews-E9079.htm",
               "https://www.glassdoor.com/Reviews/Slalom-Build-Reviews-E250458
               "https://www.glassdoor.com/Reviews/Accenture-Reviews-E4138.htm"
               "https://www.glassdoor.com/Reviews/Snap-Reviews-E671946.htm",
               "https://www.glassdoor.com/Reviews/Twitter-Reviews-E100569.htm"
               "https://www.glassdoor.com/Reviews/Amazon-Reviews-E6036.htm",
               "https://www.glassdoor.com/Reviews/Apple-Reviews-E1138.htm",
               "https://www.glassdoor.com/Reviews/Atlassian-Reviews-E115699.ht
               "https://www.glassdoor.com/Reviews/Bloomberg-L-P-Reviews-E3096.
               "https://www.glassdoor.com/Reviews/Boeing-Reviews-E102.htm",
               "https://www.glassdoor.com/Reviews/Bose-Reviews-E3098.htm"
              ]

ompaniesList = ["Google",  "Slalom Consulting" , "Accenture", "Snapchat", "Twi
```

Automating Review Scraping Process

In [163]:

```
# data frame for Glassdoor Reviews

def scapeAllReviews():
    columns = ["ReviewSummary", "link", "pros", "cons","companyId"]
    reviews_df = pd.DataFrame(columns=columns)
    i = 0
    for link in reviewsLinks:

        # calling functions for soup and scraping reviews
        soup = glassdoor_ratings_scrapper(link)
        reviewSummary,reviewLink,pros,cons,companyName = scrape_reviews(soup,
        i = i + 1

        # creating a dict from recieved lists
        reviewDict = {}
        reviewDict["ReviewSummary"] = reviewSummary
        reviewDict["link"] = reviewLink
        reviewDict["pros"] = pros
        reviewDict["cons"] = cons
        reviewDict["companyId"] = companyName

        # create a dataframe of reviews for particular company using dict abo
        companyReview_df = pd.DataFrame.from_dict(reviewDict)
        # ignore index and append to reviewDf for all companies
        if i == 1:
            reviews_df = pd.DataFrame.from_dict(reviewDict)
        else:
            reviews_df = reviews_df.append(companyReview_df, ignore_index = T

    return reviews_df

reviews = scapeAllReviews()
```

```
                                       ReviewSummary  \
0                               "Amazing experience."
1                                             "Solid"
2                                          "Software"
3             "They really care about their employees"
4                                              "Nice"
5                                "Great Company to grow"
6  "I believe this is what the kids would call a ...
7                          "Best Place I've Ever Worked"
8    "Great company to start your consulting career"
9                                 "Really Great Company"

                                                link  \
0  https://www.glassdoor.com//Reviews/Employee-Re... (https://www.glassd
oor.com//Reviews/Employee-Re...)
1  https://www.glassdoor.com//Reviews/Employee-Re... (https://www.glassd
oor.com//Reviews/Employee-Re...)
2  https://www.glassdoor.com//Reviews/Employee-Re... (https://www.glassd
oor.com//Reviews/Employee-Re...)
3  https://www.glassdoor.com//Reviews/Employee-Re... (https://www.glassd
```

In [165]:
```
reviews.tail()
```

Out[165]:

	ReviewSummary	link	pros	cons	co
105	"Great Work Environment, Great People"	https://www.glassdoor.com//Reviews/Employee-Re...	Smart, Interesting, Innovative people to work ...	Office is still set up in cubicles which makes...	
106	"Once great, now solid"	https://www.glassdoor.com//Reviews/Employee-Re...	-Good (not amazing) benefits\r\n-decent corpor...	-Very little support for retail stores\r\n-mic...	
107	"Micromanaged from the First Day"	https://www.glassdoor.com//Reviews/Employee-Re...	Smart colleagues; competent engineers; nice ca...	Heavy in corporate bureaucracy; feels like wor...	
108	"Bose: Company in Decline"	https://www.glassdoor.com//Reviews/Employee-Re...	None whatsoever to speak of.	Closing all retail stores\r\nHit-or-miss manag...	
109	"I enjoyed my time at Bose"	https://www.glassdoor.com//Reviews/Employee-Re...	I had many good co-workers. We built great so...	upper management did not always know what was ...	

Entering Glassdoor Reviews in excel

In [215]:
```
reviews.to_csv("./data/glassdoorReviews.csv",encoding="utf-8",index=False)
```

Scraping all Ratings

In [196]:

```
def scrape_ratings(soup, company):
    overall = []
    recommended = []
    companyName = [company]

    span = soup.find(name="div", attrs={"class": "v2__EIReviewsRatingsStyles
    overall.append(span.text)

    span = soup.find(name="tspan", attrs={"class": "donut__DonutStyle__donut
    recommended.append(span.text)

    return overall,recommended,companyName

soup = glassdoor_ratings_scrapper("https://www.glassdoor.com/Reviews/Google-
scrape_ratings(soup, "Google")
```

Out[196]:

```
(['4.4'], ['89'], ['Google'])
```

In [200]:

```
def scapeAllRatings():

    columns = ["rating", "recommended", "companyId"]
    ratings_df = pd.DataFrame(columns=columns)

    i = 0
    for link in reviewsLinks:

        # calling functions for soup and scraping reviews
        soup = glassdoor_ratings_scrapper(link)
        overall,recommended,companyName = scrape_ratings(soup, CompaniesList
        i = i + 1

        # creating a dict from recieved lists
        reviewDict = {}
        reviewDict["ratings"] = overall
        reviewDict["recommended"] = recommended
        reviewDict["companyId"] = companyName

        # create a dataframe of reviews for particular company using dict ab
        companyRating = pd.DataFrame.from_dict(reviewDict)
        # ignore index and append to reviewDf for all companies
        if i == 1:
            ratings_df = pd.DataFrame.from_dict(reviewDict)
        else:
            ratings_df = ratings_df.append(companyRating, ignore_index = Tru

    return ratings_df
```

In [206]:

```
ratings = scapeAllRatings()
```

In [205]:
```
ratings.head()
```

Out[205]:

	ratings	recommended	companyId
0	4.4	89	Google
1	4.5	92	Slalom Consulting
2	3.8	77	Accenture
3	3.4	65	Snapchat
4	4.0	82	Twitter

Entering company rating data into Excel

In [216]:
```
ratings.to_csv("./data/glassdoorRatings.csv",encoding="utf-8",index=False)
```

ERD

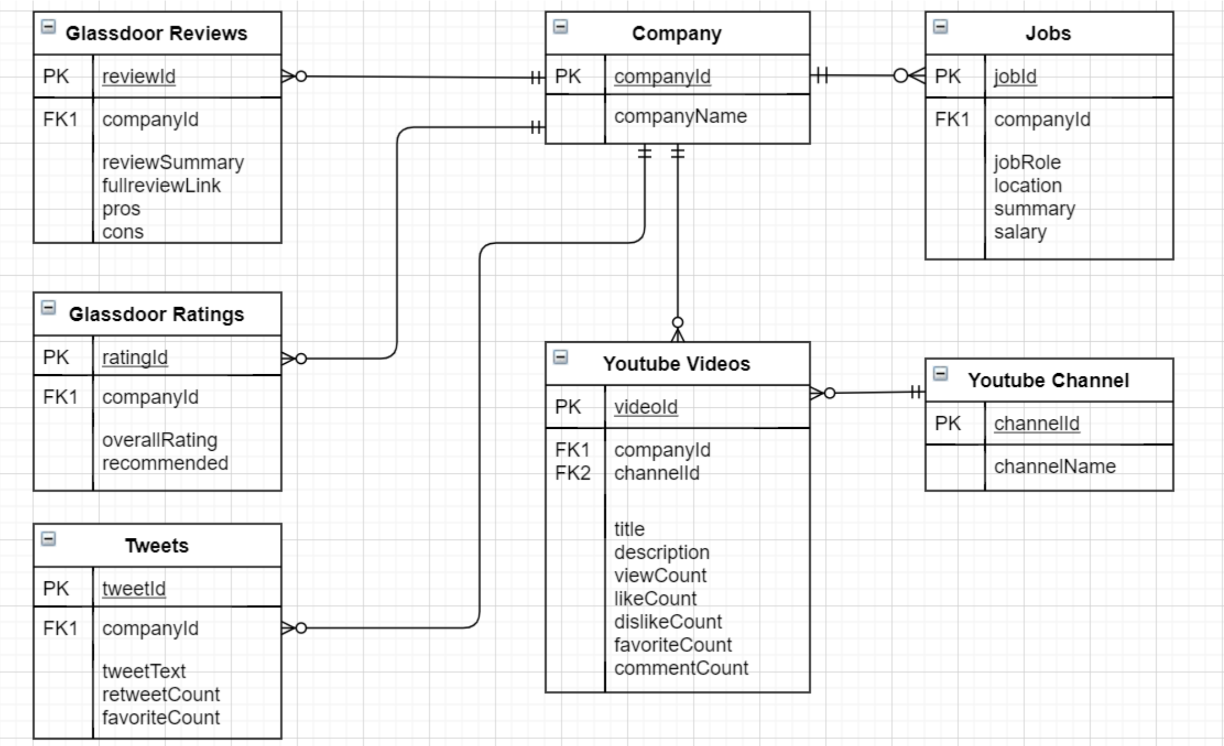


Table Description

TABLES

• COMPANY • JOBS • TWEETS • YOUTUBE_VIDEOS • YOUTUBE_CHANNELS • GLASSDOOR_RATINGS • GLASSDOOR_REVIEWS

1) COMPANY_TBL (company table):

- This table contains the following attributes:
- COMPANY_ID(PK): As we collected the data of the 300 finance company's we created a unique id for each of the company. This column is the primary key of the table and each row can be uniquely identified using this primary key.
- COMPANY_NAME: This attribute or the column contains all the company names which we collected for the finance department.

2) JOBS_TBL:

- This table has the following attributes:
- JOB_ID:
- When we check for any company website for the jobs we can see that there will be a job id uniquely defined for each and every different type of the job they post. This JOB_ID refers to the same thing.
- JOB_Role:
- Job position refers to the what type of position he is applying for. For example, he may apply to software Engineering, Data scientist etc.
- COMPANY_ID:
- This is the foreign key in this table which refers to the primary key in the COMPANY_TBL.
- Foreign keys are used to provide the perfect link between the tables.
- JOB_LOCATION: Job location refers to on which location this job is available.
- JOB_SALARY: salary range of job
- SUMMARY: summary provided by company about the job rol

3) YOUTUBE_CHANNEL:

- This table has the following attributes and the primary key in this table is channel_id
- CHANNEL_ID: Each channel in the YouTube is given the unique id. This unique id is called the CHANNEL_ID.
- CHANNEL_TITLE: Channel_title is the title of the YouTube channel

4) YOUTUBE_VIDEO_DATA:

- This table has the following attributes and the primary key in this table is the VIDEO_ID and the foreign key in this table are COMPANY_ID which references COMPANY_TBL and the CHANNEL_ID references the YOUTUBE_DATA table.
- VIDEO_ID: VIDEO_ID represents the unique id given to each video posted in the YouTube.
- COMPANY_ID: This is the foreign key which references the COMPANY_ID in the COMPANY_TBL.
- VIEW_COUNT: VIEW_COUNT represents the number of views for that table.
- COMMENT_COUNT: COMMENT_COUNT represents the number of counts for that video.
- LIKE_COUNT: LIKE_COUNT represents the number of likes for that video.
- DISLIKE_COUNT: DISLIKE_COUNT represents the number of dislikes for that video.
- FAVORITE_COUNT: FAVORITE_COUNT represents the number of favorites for that video.
- CHANNEL_ID: This is the unique id given to each category in the category table. This is the foreign key in the table.

5) GLASSDOOORS_DOOR_RATINGS

- This table has the following attributes and the foreign keys in this table
- COMPANY_ID: This is the foreign key which references the COMPANY_ID in the COMPANY_TBL.
- RATING_OVERALL: It represents the overall Glassdoor rating of that particular company.
- RECOMMENDED: It represents how much people recommend that company

6) GLASSDOOR_REVIEWS:

- This table has the following attributes:
- COMPANY_ID: This is the foreign key which references the COMPANY_ID in the COMPANY_TBL.
- REVIEW_TITLE: Title given to each review.
- PROS: Pros about the company.
- CONS: Cons about the company
- REVIEW_ID: It represents the unique id given to the each and every review.

1) NORMALIZATION:

- After the tables are created then the next step is data normalization. Normalization is used to reduce the data redundancy. We can’t eliminate the data redundancy completely, but we can reduce the redundancy by dividing the repeating columns in the particular table into a new table and generate a unique Id to that table. Now instead of repeating of all the columns we will give this unique id to the table and it acts a s link between them.

1) 1ST NORMALIZATION FORM:

- A table in 1NF should be atomic and have non repeating rows and columns.
- Our tables are in 1NF as they satisfy each requirement of first Normalization form.

2) 2ND NORMALIZATION FORM:

- There should not be any 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.

3) 3RD NORMALIZATION FORM:

- A table is said to be in 3NF if no non primary attribute in the table should be dependent on other nonprimary attribute in the table.
- Our tables are in 3NF as they satisfy every requirement of third Normalization form.

Merging excel tables with common column

In [218]:

```
# companies with id
df_q = pd.read_excel('Unique.xlsx')

# ratings with company name
df_1 = pd.read_csv('./data/jobs.csv')
df_new = pd.merge(df_1, df_q, left_on='company_name', right_on='company_name'
# df_new.to_excel('./data/final/jobs.xlsx')

# ratings with company name
df_1 = pd.read_csv('./data/glassdoorRatings.csv')
df_new = pd.merge(df_1, df_q, left_on='company_name', right_on='company_name'
# df_new.to_excel('./data/final/glassdoor_ratings.xlsx')


# reviews with company name
df_1 = pd.read_csv('./data/glassdoorReviews.csv')
df_new = pd.merge(df_1, df_q, left_on='company_name', right_on='company_name'
# df_new.to_excel('./data/final/glassdoor_reviews.xlsx')

# youtube videos with company name
df_1 = pd.read_csv('./data/youtubeVideos.csv')
df_new = pd.merge(df_1, df_q, left_on='company_name', right_on='company_name'
# df_new.to_excel('./data/final/youtube_videos.xlsx')
```

All final excel tables with primary and foriegn keys

In [224]:

```
df = pd.read_excel("./data/final/companies.xlsx")
df.head()
```

Out[224]:

	company_id	company_name
0	1	Global Science & Technology, Inc.
1	2	SpiralTech Superior Dental Implants
2	3	Tonk Tonk Games, Inc
3	4	Arrayo
4	5	ERNIESYS

In [225]:

```
df = pd.read_excel("./data/final/job_postings.xlsx")
df.head()
```

Out[225]:

	jobposting_id	job_title	company_name	location	summary	salary	company_i
0	j1	Data Scientist	Triplebyte	Remote	You’ll report directly to Triplebytes' Head of...	150, 000−225,000 a year	47
1	j2	Data Scientist	Triplebyte	Remote	You’ll report directly to Triplebytes' Head of...	150, 000−225,000 a year	47
2	j3	Data Scientist	Triplebyte	Remote	You’ll report directly to Triplebytes' Head of...	150, 000−225,000 a year	47
3	j4	Data Scientist	Triplebyte	Remote	You’ll report directly to Triplebytes' Head of...	150, 000−225,000 a year	47
4	j5	Data Scientist	Triplebyte	Remote	You’ll report directly to Triplebytes' Head of...	150, 000−225,000 a year	47

In [226]:

```
df = pd.read_excel("./data/final/glassdoor_reviews.xlsx")
df.head()
```

Out[226]:

	reviewid	ReviewSummary	link	pros	cons
0	rev1	"One of the best places to work."	https://www.glassdoor.com//Reviews/Employee-Re...	Amazing place the work. Great culture, great p...	Very difficult to get promoted.
1	rev2	"Moving at the speed of light, burn out is ine...	https://www.glassdoor.com//Reviews/Employee-Re...	1) Food, food, food. 15+ cafes on main campus...	1) Work/life balance. What balance? All thos...
2	rev3	"Great balance between big-company security an...	https://www.glassdoor.com//Reviews/Employee-Re...	* If you're a software engineer, you're among ...	* It *is* becoming larger, and with it comes g...
3	rev4	"The best place I've worked and also the most ...	https://www.glassdoor.com//Reviews/Employee-Re...	You can't find a more well-regarded company th...	I live in SF so the commute can take between 1...
4	rev5	"Amazing culture"	https://www.glassdoor.com//Reviews/Employee-Re...	very caring about the individual, great benefi...	very smart people, hence a very intense work e...

In [227]:

```
df = pd.read_excel("./data/final/glassdoor_ratings.xlsx")
df.head()
```

Out[227]:

	ratingid	ratings	recommended	company_id
0	r1	4.4	89	586
1	r2	4.5	92	895
2	r3	3.8	77	757
3	r4	3.4	65	203
4	r5	4.0	82	516

In [228]:

```
df = pd.read_excel("./data/final/youtube_videos.xlsx")
df.head()
```

Out[228]:

	videoId	channelId	title	descri
0	h5yJ-_bbiM8	UCXubLFOt4iiX2_0tYQD8gRA	Penn State Health - Careers	At Penn State Health, we work to pr th
1	Pabq5ElJbx4	UCUngw5TivNYk845EpJQ1Mfg	Your Career in Healthcare Starts at Penn Comme...	Dr. John D. Six, M.D., Vice Preside Me
2	pL-Ra2hoIsw	UCb_JeH-0SzKbKOqSe0DZnmA	Health Information Degrees at Penn College	https://www.pct.edu/academics/hs/hea H
3	zouPcloHb2w	UC36NIm8ikeZ4tDRx__BjJnA	Penn Medicine&#39;s Global Nurse Program	In response to worldwide nu concerns,
4	NlpK-1b5Kvw	UCSC8V1ez4zt3rviyPWzk9Sg	Immunology Graduate Program | Cincinnati Child...	The study of immunology is critical t

In [232]:

```
df = pd.read_csv("./data/final/youtube_channels.csv")
df.head()
```

Out[232]:

	channelId	channelTitle
0	UCXubLFOt4iiX2_0tYQD8gRA	Penn State Health
1	UCUngw5TivNYk845EpJQ1Mfg	Penn Commercial Business/Technical School
2	UCb_JeH-0SzKbKOqSe0DZnmA	Pennsylvania College of Technology
3	UC36NIm8ikeZ4tDRx__BjJnA	Penn Medicine
4	UCSC8V1ez4zt3rviyPWzk9Sg	Cincinnati Children's

Now that we have all the final tables with primary keys and foriegn keys in excel.

We created schema in workbench and exported all the excel files

below are the screenshots for our sql workbench

CREATE TABLE `jobs` . `company` (`COMPANY_ID` INT NOT NULL, `COMPANY_NAME` VARCHAR(45) NOT NULL, PRIMARY KEY (`COMPANY_ID`));

CREATE TABLE `jobs` . `jobpostings` (`jobposting_id` INT NOT NULL, `job_title` VARCHAR(45) NULL, `company_name` VARCHAR(45) NULL, `location` VARCHAR(45) NULL, `summary` VARCHAR(45) NULL, `salary` INT NULL, `company_id` INT NULL, PRIMARY KEY (`jobposting_id`));

CREATE TABLE `jobs` . `youtubevideos` (`videoId` INT NOT NULL, `channelId` VARCHAR(4500) NULL, `title` VARCHAR(4500) NULL, `description` VARCHAR(4500) NULL, `viewCount` INT NULL, `likeCount` INT NULL, `dislikeCount` INT NULL, `favoriteCount` INT NULL, `commentCount` INT NULL, `company_name` VARCHAR(450) NULL, `company_id` INT NULL, PRIMARY KEY (`videoId`));

CREATE TABLE `jobs` . `youtubechannels` (`channelId` VARCHAR(450) NOT NULL, `channelTitle` VARCHAR(450) NULL, PRIMARY KEY (`channelId`));

CREATE TABLE `jobs` . `glassdoorratings` (`ratingid` INT NOT NULL, `ratings` FLOAT NULL, `recommended` INT NULL, `company_id` INT NULL, PRIMARY KEY (`ratingid`));

CREATE TABLE `jobs` . `glassdoorreviews` (`reviewid` VARCHAR(45) NOT NULL, `ReviewSummary` VARCHAR(450) NULL, `link` VARCHAR(450) NULL, `pros` VARCHAR(4500) NULL, `cons` VARCHAR(4500) NULL, `company_id` INT NULL, PRIMARY KEY (`reviewid`));

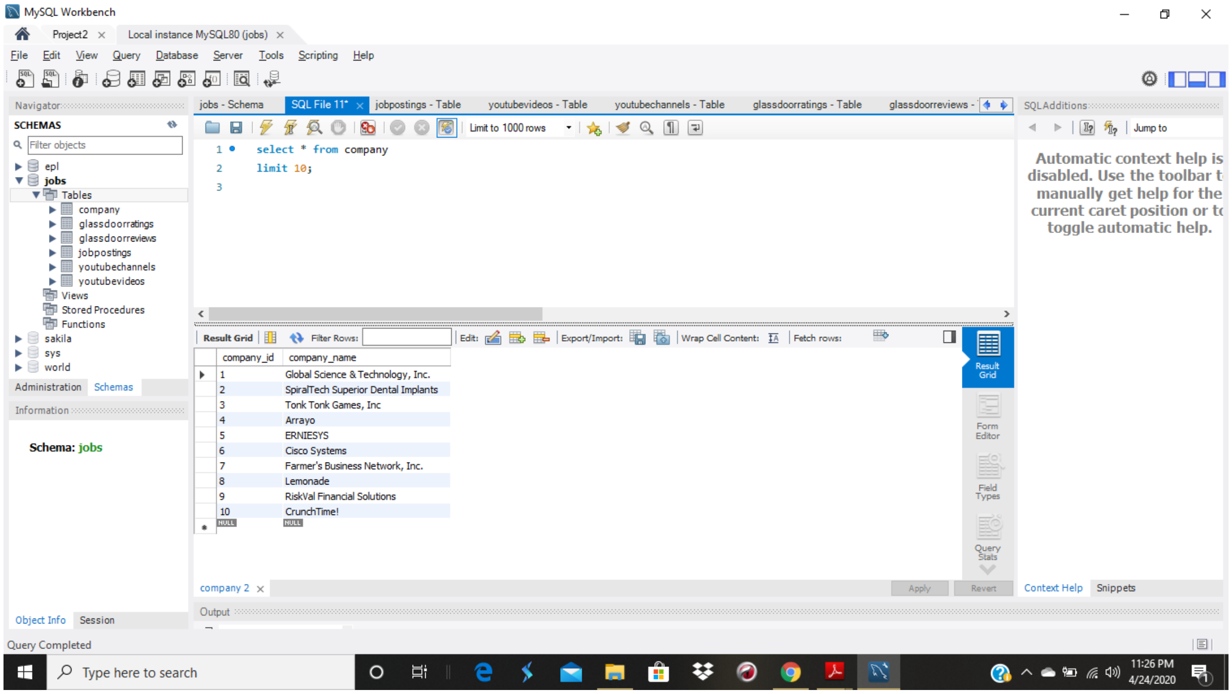
ALTER TABLE `jobs` . `jobpostings` ADD INDEX `company_id_idx` (`company_id` ASC) VISIBLE;

ALTER TABLE `jobs` . `jobpostings` ADD CONSTRAINT `company_id` FOREIGN KEY (`company_id`) REFERENCES `jobs` . `company` (`company_id`) ON DELETE NO ACTION ON UPDATE NO ACTION;

ALTER TABLE `jobs` . `youtubevideos` ADD INDEX `channelId_idx` (`channelId` ASC) VISIBLE, ADD INDEX `company_id_idx` (`company_id` ASC) VISIBLE; ; ALTER TABLE `jobs` . `youtubevideos` ADD CONSTRAINT `company_id` FOREIGN KEY (`company_id`) REFERENCES `jobs` . `company` (`company_id`) ON DELETE NO ACTION ON UPDATE NO ACTION, ADD CONSTRAINT `channelId` FOREIGN KEY (`channelId`) REFERENCES `jobs` . `youtubechannels` (`channelId`) ON DELETE NO ACTION ON UPDATE NO ACTION;

ALTER TABLE `jobs` . `glassdoorratings` CHANGE COLUMN `company_id` `ratings_company_id` INT(11) NULL DEFAULT NULL , ADD INDEX `ratings_company_id` (`ratings_company_id` ASC) VISIBLE; ; ALTER TABLE `jobs` . `glassdoorratings` ADD CONSTRAINT `ratings_company_id` FOREIGN KEY (`ratings_company_id`) REFERENCES `jobs` . `company` (`company_id`) ON DELETE NO ACTION ON UPDATE NO ACTION;

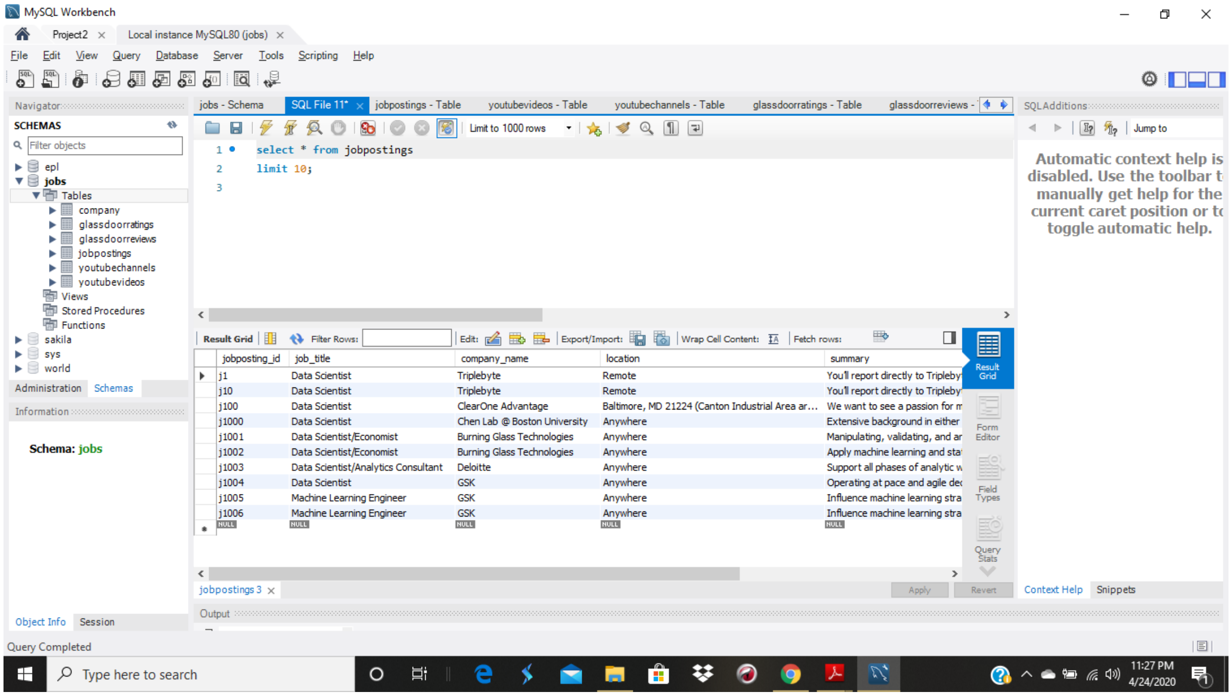
ALTER TABLE `jobs` . `glassdoorreviews` CHANGE COLUMN `company_id` `reviews_company_id` INT(11) NULL DEFAULT NULL , ADD INDEX `reviews_company_id` (`reviews_company_id` ASC) VISIBLE; ; ALTER TABLE `jobs` . `glassdoorreviews` ADD CONSTRAINT `reviews_company_id` FOREIGN KEY (`reviews_company_id`) REFERENCES `jobs` . `company` (`company_id`) ON DELETE NO ACTION ON UPDATE NO ACTION;



```
select * from company
limit 10;
```

company_id	company_name
1	Global Science & Technology, Inc.
2	SpiralTech Superior Dental Implants
3	Tonk Tonk Games, Inc
4	Arrayo
5	ERNIESYS
6	Cisco Systems
7	Farmer's Business Network, Inc.
8	Lemonade
9	RiskVal Financial Solutions
10	CrunchTime!
NULL	NULL

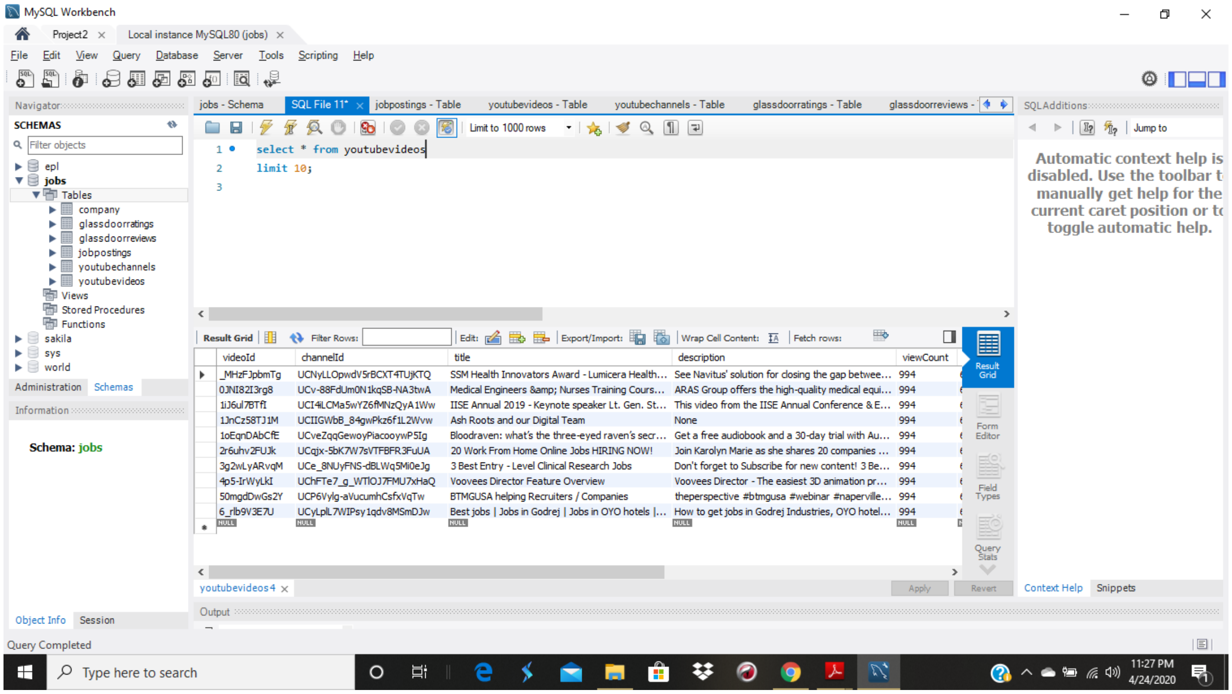
Query Completed



```
select * from jobpostings
limit 10;
```

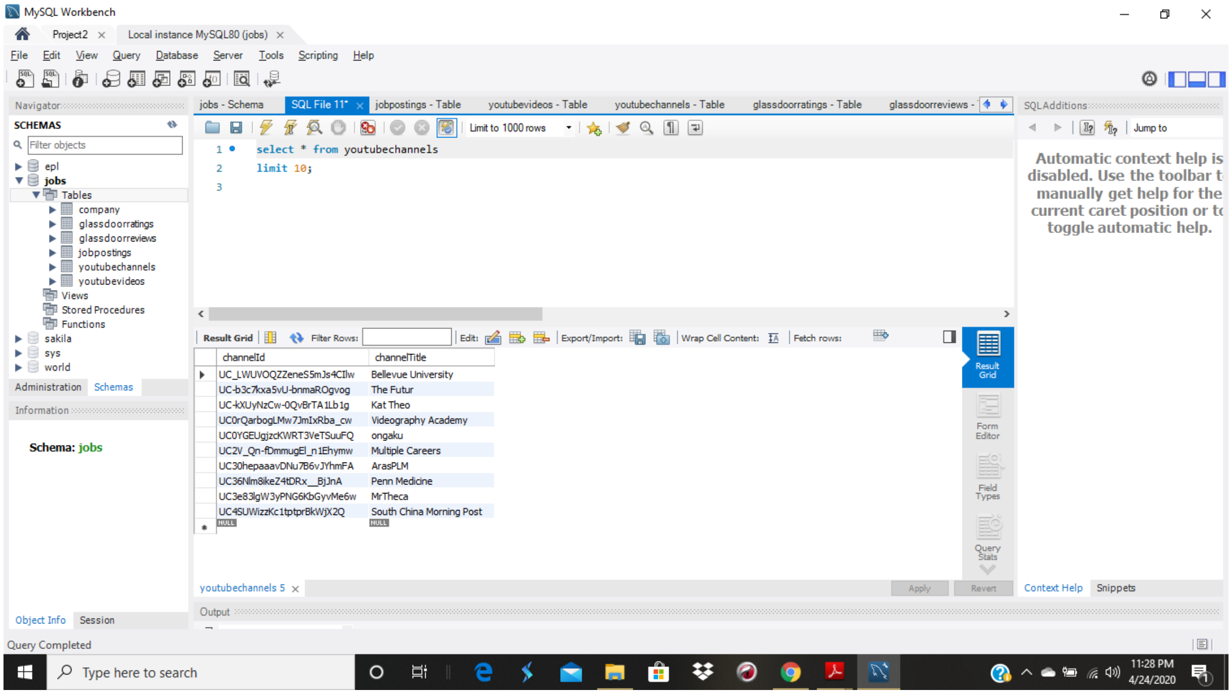
jobposting_id	job_title	company_name	location	summary
j1	Data Scientist	Triplebyte	Remote	You'll report directly to Tripleby
j10	Data Scientist	Triplebyte	Remote	You'll report directly to Tripleby
j100	Data Scientist	ClearOne Advantage	Baltimore, MD 21224 (Canton Industrial Area ar...	We want to see a passion for m
j1000	Data Scientist	Chen Lab @ Boston University	Anywhere	Extensive background in either
j1001	Data Scientist/Economist	Burning Glass Technologies	Anywhere	Manipulating, validating, and ar
j1002	Data Scientist/Economist	Burning Glass Technologies	Anywhere	Apply machine learning and sta
j1003	Data Scientist/Analytics Consultant	Deloitte	Anywhere	Support all phases of analytic w
j1004	Data Scientist	GSK	Anywhere	Operating at pace and agile de
j1005	Machine Learning Engineer	GSK	Anywhere	Influence machine learning stra
j1006	Machine Learning Engineer	GSK	Anywhere	Influence machine learning stra
NULL	NULL	NULL	NULL	NULL

Query Completed



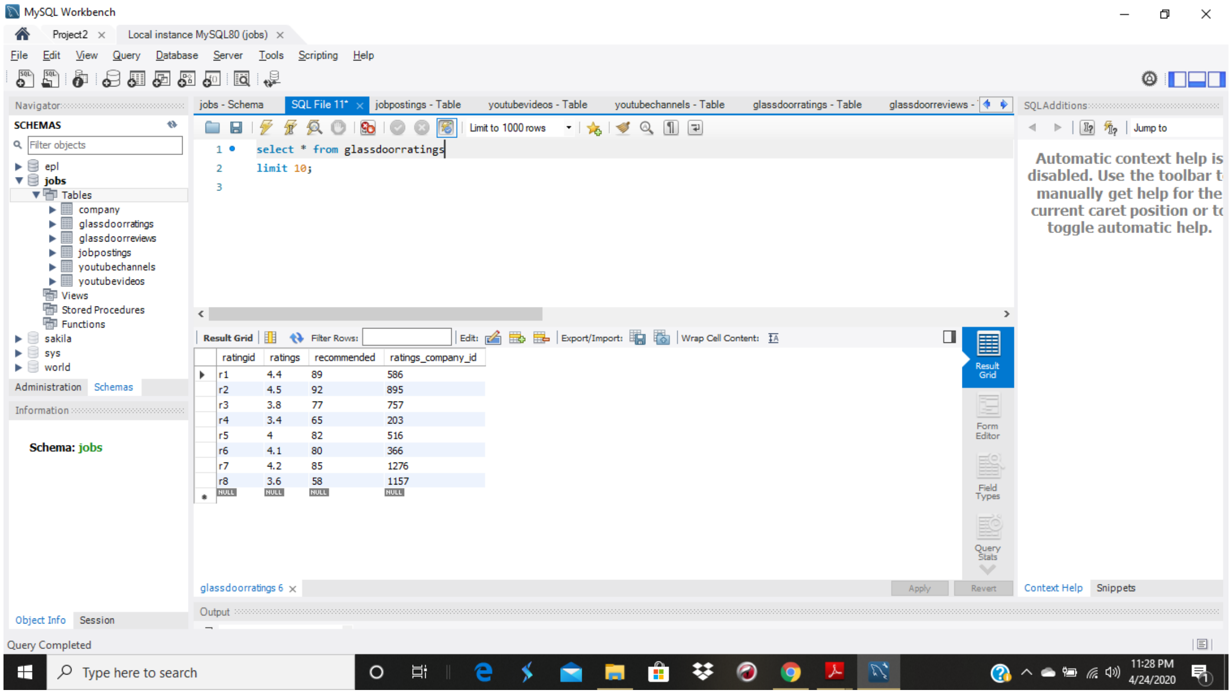
```
select * from youtubevideos
limit 10;
```

	videoId	channelId	title	description	viewCount
▶	_MHzFJpbmTg	UCNyLLOpwdV5rBCXT4TUjKTQ	SSM Health Innovators Award - Lumicera Health…	See Navitus' solution for closing the gap betwee…	994
	0JNI82I3rg8	UCv-88FdUm0N1kqSB-NA3twA	Medical Engineers &amp; Nurses Training Cours…	ARAS Group offers the high-quality medical equi…	994
	1iJ6ul7BTfI	UCI4iLCMa5wYZ6fMNzQyA1Ww	IISE Annual 2019 - Keynote speaker Lt. Gen. St…	This video from the IISE Annual Conference & E…	994
	1JnCz58TJ1M	UCIIGWbB_84gwPkz6f1L2Wvw	Ash Roots and our Digital Team	None	994
	1oEqnDAbCfE	UCveZqqGewoyPiacooywP5Ig	Bloodraven: what's the three-eyed raven's secr…	Get a free audiobook and a 30-day trial with Au…	994
	2r6uhv2FUJk	UCqjx-5bK7W7sVTFBFR3FuUA	20 Work From Home Online Jobs HIRING NOW!	Join Karolyn Marie as she shares 20 companies …	994
	3g2wLyARvqM	UCe_8NUyFNS-dBLWq5Mi0eJg	3 Best Entry - Level Clinical Research Jobs	Don't forget to Subscribe for new content! 3 Be…	994
	4p5-IrWyLkI	UChFTe7_g_WTlOJ7FMU7xHaQ	Voovees Director Feature Overview	Voovees Director - The easiest 3D animation pr…	994
	50mgdDwGs2Y	UCP6Vylg-aVucumhCsfxVqTw	BTMGUSA helping Recruiters / Companies	theperspective #btmgusa #webinar #naperville…	994
	6_rlb9V3E7U	UCyLplL7WIPsy1qdv8MSmDJw	Best jobs | Jobs in Godrej | Jobs in OYO hotels |…	How to get jobs in Godrej Industries, OYO hotel…	994
*	NULL	NULL	NULL	NULL	NULL



```
select * from youtubechannels
limit 10;
```

	channelId	channelTitle
▶	UC_LWUVOQZZeneS5mJs4CIlw	Bellevue University
	UC-b3c7kxa5vU-bnmaROgvog	The Futur
	UC-kXUyNzCw-0QvBrTA1Lb1g	Kat Theo
	UC0rQarbogLMw7JmIxRba_cw	Videography Academy
	UC0YGEUgjzcKWRT3VeTSuuFQ	ongaku
	UC2V_Qn-fDmmugEl_n1Ehymw	Multiple Careers
	UC30hepaaavDNu7B6vJYhmFA	ArasPLM
	UC36Nlm8ikeZ4tDRx__BjJnA	Penn Medicine
	UC3e83lgW3yPNG6KbGyvMe6w	MrTheca
	UC4SUWizzKc1tptprBkWjX2Q	South China Morning Post
*	NULL	NULL



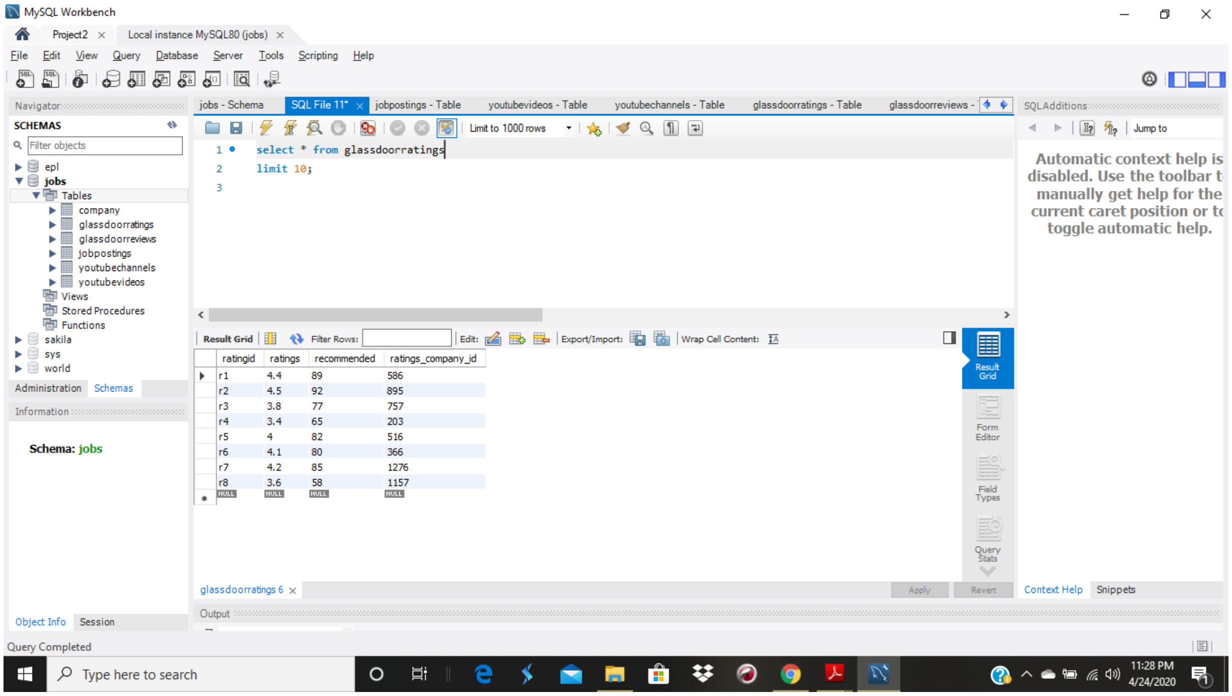
```
select * from glassdoorratings
limit 10;
```

ratingid	ratings	recommended	ratings_company_id
r1	4.4	89	586
r2	4.5	92	895
r3	3.8	77	757
r4	3.4	65	203
r5	4	82	516
r6	4.1	80	366
r7	4.2	85	1276
r8	3.6	58	1157
NULL	NULL	NULL	NULL

Schema: jobs

Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or tc toggle automatic help.

Query Completed



```
select * from glassdoorratings
limit 10;
```

ratingid	ratings	recommended	ratings_company_id
r1	4.4	89	586
r2	4.5	92	895
r3	3.8	77	757
r4	3.4	65	203
r5	4	82	516
r6	4.1	80	366
r7	4.2	85	1276
r8	3.6	58	1157
NULL	NULL	NULL	NULL

Schema: jobs

Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or tc toggle automatic help.

Query Completed

UseCases

select all from companies

SELECT * FROM COMPANIES

select software jobs from job postings

select * from jobs where job_role like '%software%'

select datascience jobs from job postings

select * from jobs where job_role like '%data%'

Get a list of all the companies with remote jobs

select * from jobs where location like '%remote%'

Get the id of the companies with the most YouTube video views

select companyId,videoTitle,views from youtubeVideos order by views desc limit 10

Select company with most job postings

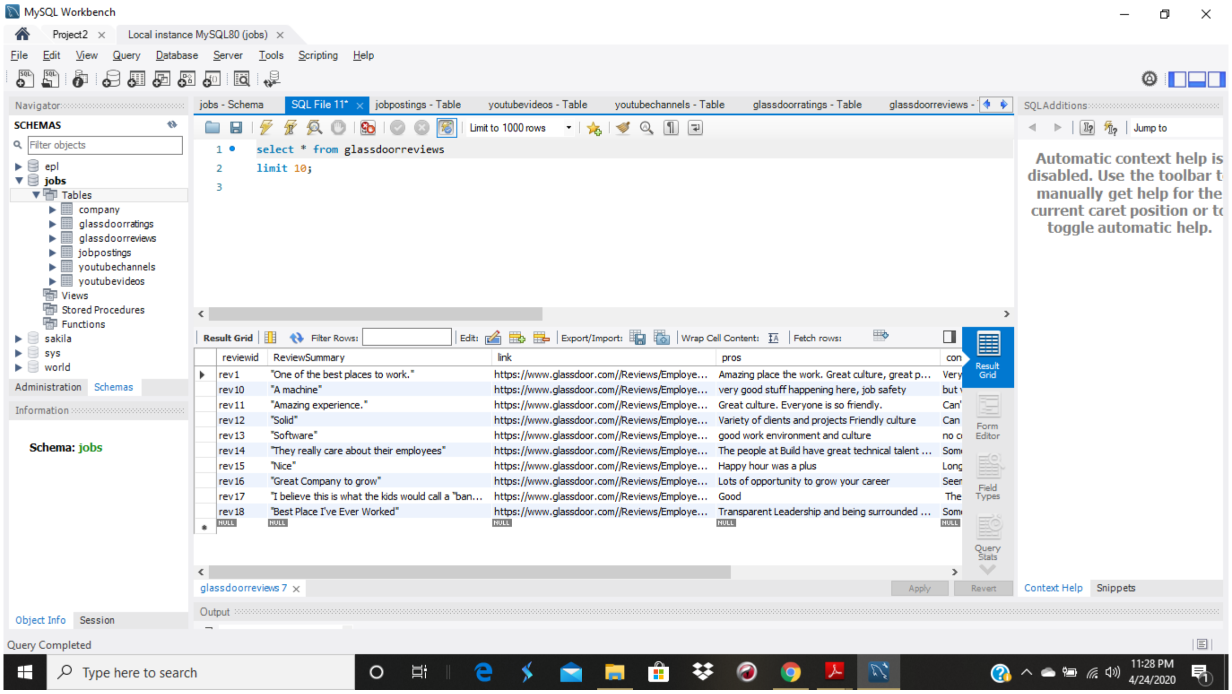
SELECT c.companyName, COUNT(j.jobId) AS jobs FROM JOBS_TBL j JOIN company c ON c.companyId = J.job_id GROUP BY c.companyName

Get the name of the video with more likes

select companyId,videoTitle,likes from youtubeVideos order by likes desc limit 10

Function: Get the role name given the JOB_ID

CREATE FUNCTION GET_ROLE_NAME (JOB_ID_IN IN VARCHAR2 , JOB_POSITION_OUT OUT VARCHAR2) RETURN VARCHAR2 AS BEGIN SELECT JOB_POSITION INTO JOB_POSITION_OUT FROM JOBS_TBL WHERE JOB_ID = JOB_ID_IN; RETURN JOB_POSITION_OUT; END GET_POSITION_NAME;



Schema: jobs

```
select * from glassdoorreviews
limit 10;
```

reviewid	ReviewSummary	link	pros	con
rev1	"One of the best places to work."	https://www.glassdoor.com//Reviews/Employe...	Amazing place the work. Great culture, great p...	Very
rev10	"A machine"	https://www.glassdoor.com//Reviews/Employe...	very good stuff happening here, job safety	but
rev11	"Amazing experience."	https://www.glassdoor.com//Reviews/Employe...	Great culture. Everyone is so friendly.	Can'
rev12	"Solid"	https://www.glassdoor.com//Reviews/Employe...	Variety of clients and projects Friendly culture	Can
rev13	"Software"	https://www.glassdoor.com//Reviews/Employe...	good work environment and culture	no c
rev14	"They really care about their employees"	https://www.glassdoor.com//Reviews/Employe...	The people at Build have great technical talent ...	Som
rev15	"Nice"	https://www.glassdoor.com//Reviews/Employe...	Happy hour was a plus	Long
rev16	"Great Company to grow"	https://www.glassdoor.com//Reviews/Employe...	Lots of opportunity to grow your career	Seer
rev17	"I believe this is what the kids would call a "ban...	https://www.glassdoor.com//Reviews/Employe...	Good	The
rev18	"Best Place I've Ever Worked"	https://www.glassdoor.com//Reviews/Employe...	Transparent Leadership and being surrounded ...	Som
NULL	NULL	NULL	NULL	NULL

Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or t toggle automatic help.

Query Completed

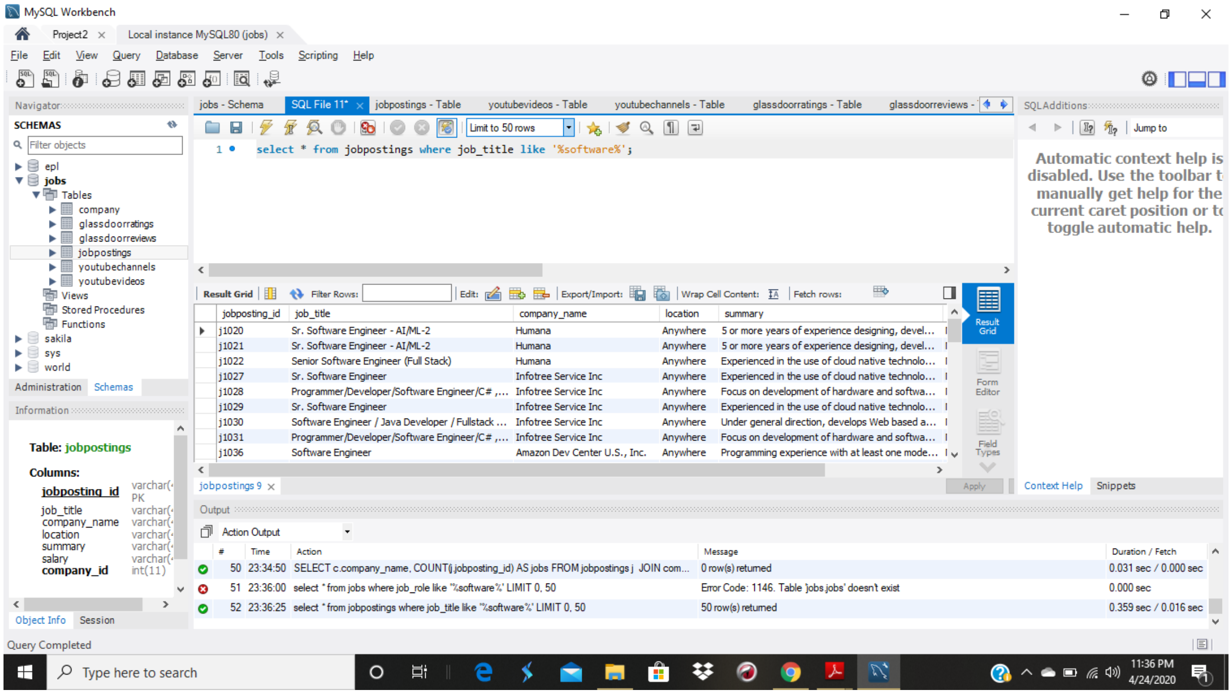


Table: jobpostings

Columns:

jobposting_id	varchar(PK
job_title	varchar(
company_name	varchar(
location	varchar(
summary	varchar(
salary	varchar(
company_id	int(11)

```
1 ● select * from jobpostings where job_title like '%software%';
```

jobposting_id	job_title	company_name	location	summary
j1020	Sr. Software Engineer - AI/ML-2	Humana	Anywhere	5 or more years of experience designing, devel...
j1021	Sr. Software Engineer - AI/ML-2	Humana	Anywhere	5 or more years of experience designing, devel...
j1022	Senior Software Engineer (Full Stack)	Humana	Anywhere	Experienced in the use of cloud native technolo...
j1027	Sr. Software Engineer	Infotree Service Inc	Anywhere	Experienced in the use of cloud native technolo...
j1028	Programmer/Developer/Software Engineer/C# ,...	Infotree Service Inc	Anywhere	Focus on development of hardware and softwa...
j1029	Sr. Software Engineer	Infotree Service Inc	Anywhere	Experienced in the use of cloud native technolo...
j1030	Software Engineer / Java Developer / Fullstack ...	Infotree Service Inc	Anywhere	Under general direction, develops Web based a...
j1031	Programmer/Developer/Software Engineer/C# ,...	Infotree Service Inc	Anywhere	Focus on development of hardware and softwa...
j1036	Software Engineer	Amazon Dev Center U.S., Inc.	Anywhere	Programming experience with at least one mode...

jobpostings 9

#	Time	Action	Message	Duration / Fetch
50	23:34:50	SELECT c.company_name, COUNT(j.jobposting_id) AS jobs FROM jobpostings j JOIN com...	0 row(s) returned	0.031 sec / 0.000 sec
51	23:36:00	select * from jobs where job_role like '%software%' LIMIT 0, 50	Error Code: 1146. Table 'jobs.jobs' doesn't exist	0.000 sec
52	23:36:25	select * from jobpostings where job_title like '%software%' LIMIT 0, 50	50 row(s) returned	0.359 sec / 0.016 sec

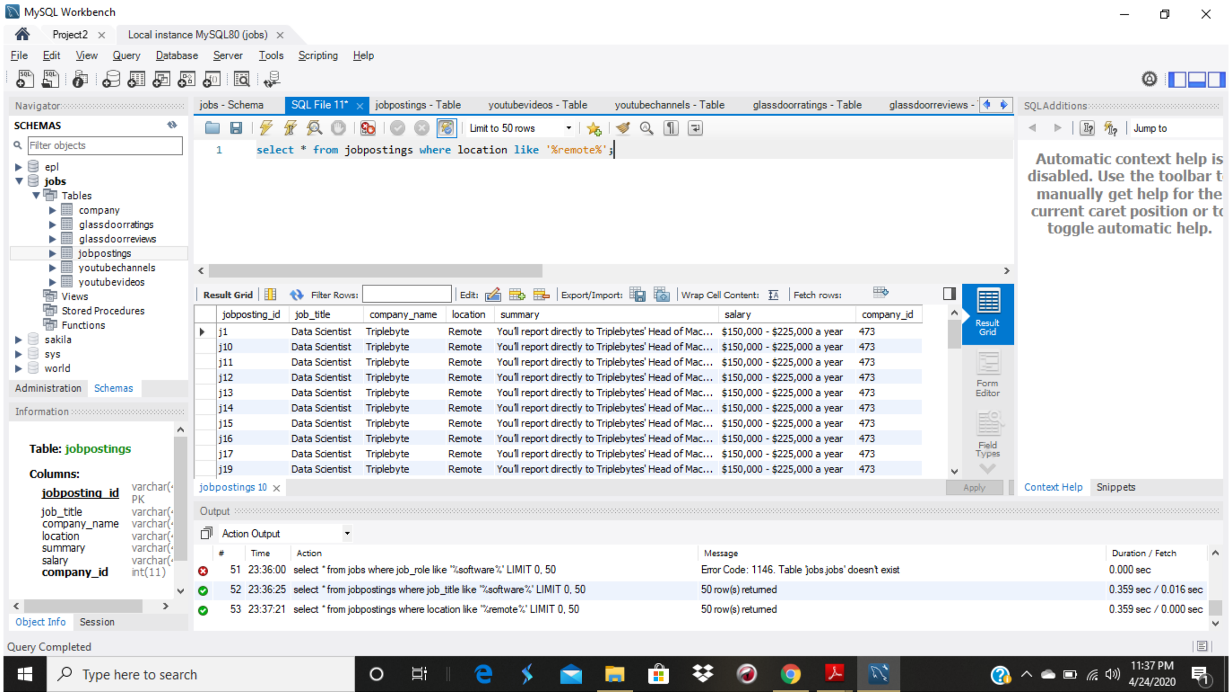


Table: jobpostings

Columns:

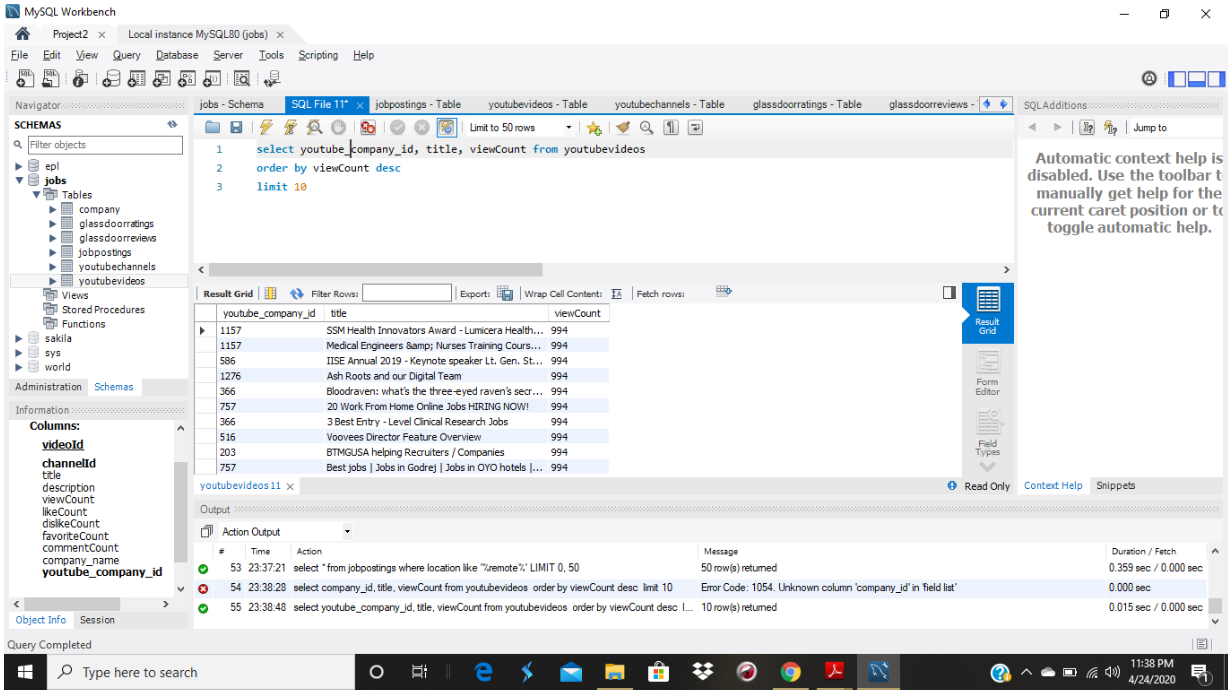
jobposting_id	varchar(PK
job_title	varchar(
company_name	varchar(
location	varchar(
summary	varchar(
salary	varchar(
company_id	int(11)

```
1 select * from jobpostings where location like '%remote%';
```

jobposting_id	job_title	company_name	location	summary	salary	company_id
j1	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j10	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j11	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j12	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j13	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j14	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j15	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j16	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j17	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473
j19	Data Scientist	Triplebyte	Remote	You'll report directly to Triplebytes' Head of Mac...	$150,000 - $225,000 a year	473

jobpostings 10

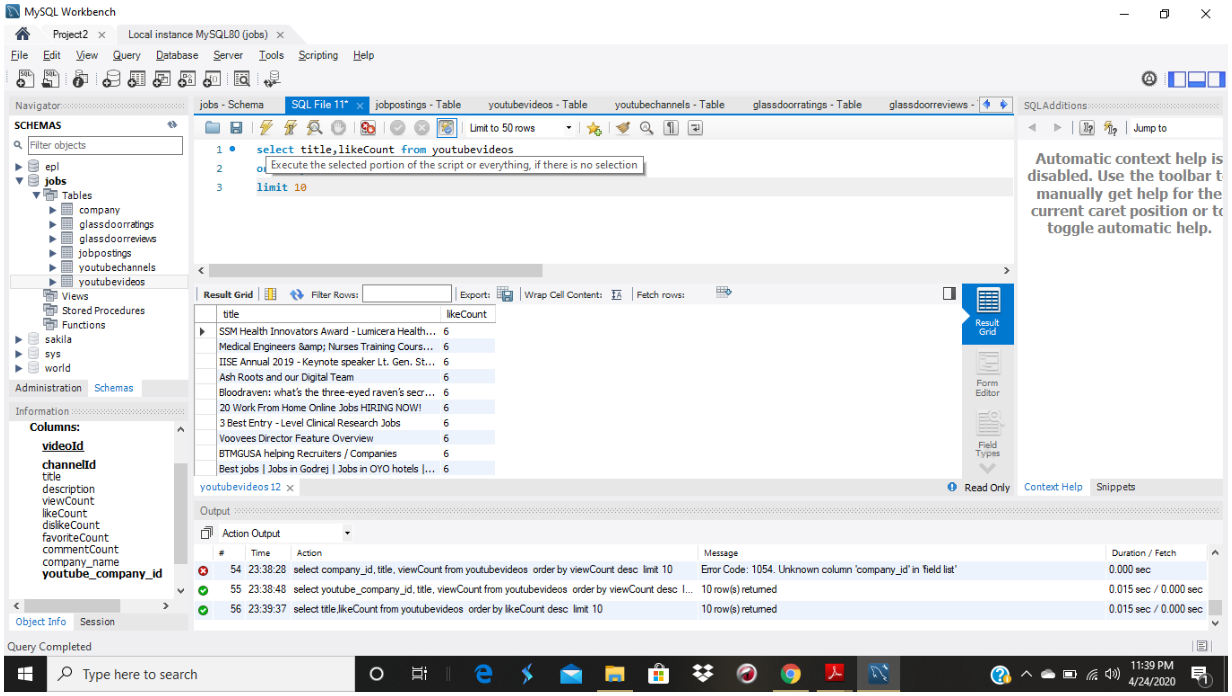
#	Time	Action	Message	Duration / Fetch
51	23:36:00	select * from jobs where job_role like '%software%' LIMIT 0, 50	Error Code: 1146. Table 'jobs.jobs' doesn't exist	0.000 sec
52	23:36:25	select * from jobpostings where job_title like '%software%' LIMIT 0, 50	50 row(s) returned	0.359 sec / 0.016 sec
53	23:37:21	select * from jobpostings where location like '%remote%' LIMIT 0, 50	50 row(s) returned	0.359 sec / 0.000 sec



```sql
select youtube_company_id, title, viewCount from youtubevideos
order by viewCount desc
limit 10
```

	youtube_company_id	title	viewCount
▶	1157	SSM Health Innovators Award - Lumicera Health...	994
	1157	Medical Engineers &amp; Nurses Training Cours...	994
	586	IISE Annual 2019 - Keynote speaker Lt. Gen. St...	994
	1276	Ash Roots and our Digital Team	994
	366	Bloodraven: what’s the three-eyed raven’s secr...	994
	757	20 Work From Home Online Jobs HIRING NOW!	994
	366	3 Best Entry - Level Clinical Research Jobs	994
	516	Voovees Director Feature Overview	994
	203	BTMGUSA helping Recruiters / Companies	994
	757	Best jobs | Jobs in Godrej | Jobs in OYO hotels |...	994

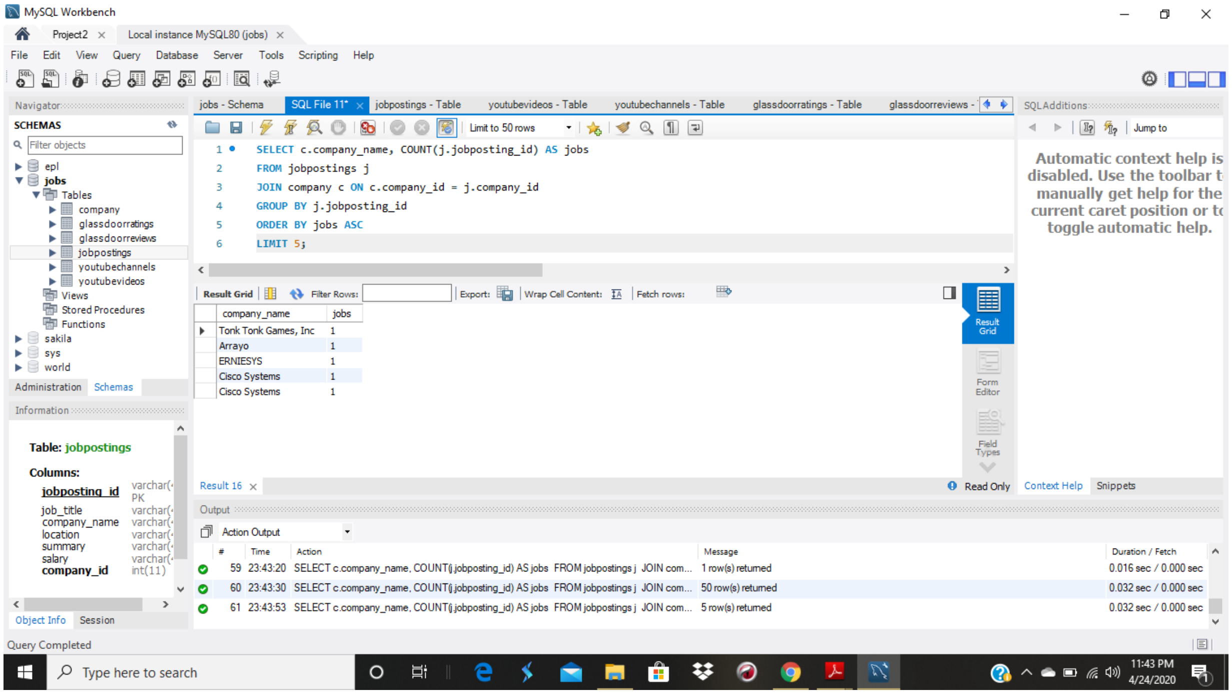
	#	Time	Action	Message	Duration / Fetch
✓	53	23:37:21	select * from jobpostings where location like '%remote%' LIMIT 0, 50	50 row(s) returned	0.359 sec / 0.000 sec
✗	54	23:38:28	select company_id, title, viewCount from youtubevideos order by viewCount desc limit 10	Error Code: 1054. Unknown column 'company_id' in 'field list'	0.000 sec
✓	55	23:38:48	select youtube_company_id, title, viewCount from youtubevideos order by viewCount desc l...	10 row(s) returned	0.015 sec / 0.000 sec



```sql
select title,likeCount from youtubevideos
o
limit 10
```

	title	likeCount
▶	SSM Health Innovators Award - Lumicera Health...	6
	Medical Engineers &amp; Nurses Training Cours...	6
	IISE Annual 2019 - Keynote speaker Lt. Gen. St...	6
	Ash Roots and our Digital Team	6
	Bloodraven: what’s the three-eyed raven’s secr...	6
	20 Work From Home Online Jobs HIRING NOW!	6
	3 Best Entry - Level Clinical Research Jobs	6
	Voovees Director Feature Overview	6
	BTMGUSA helping Recruiters / Companies	6
	Best jobs | Jobs in Godrej | Jobs in OYO hotels |...	6

	#	Time	Action	Message	Duration / Fetch
✗	54	23:38:28	select company_id, title, viewCount from youtubevideos order by viewCount desc limit 10	Error Code: 1054. Unknown column 'company_id' in 'field list'	0.000 sec
✓	55	23:38:48	select youtube_company_id, title, viewCount from youtubevideos order by viewCount desc l...	10 row(s) returned	0.015 sec / 0.000 sec
✓	56	23:39:37	select title,likeCount from youtubevideos order by likeCount desc limit 10	10 row(s) returned	0.015 sec / 0.000 sec



```
SELECT c.company_name, COUNT(j.jobposting_id) AS jobs
FROM jobpostings j
JOIN company c ON c.company_id = j.company_id
GROUP BY j.jobposting_id
ORDER BY jobs ASC
LIMIT 5;
```

company_name	jobs
Tonk Tonk Games, Inc	1
Arrayo	1
ERNIESYS	1
Cisco Systems	1
Cisco Systems	1

```
#   Time      Action                                                                            Message              Duration / Fetch
59  23:43:20  SELECT c.company_name, COUNT(j.jobposting_id) AS jobs  FROM jobpostings j  JOIN com...  1 row(s) returned    0.016 sec / 0.000 sec
60  23:43:30  SELECT c.company_name, COUNT(j.jobposting_id) AS jobs  FROM jobpostings j  JOIN com...  50 row(s) returned   0.032 sec / 0.000 sec
61  23:43:53  SELECT c.company_name, COUNT(j.jobposting_id) AS jobs  FROM jobpostings j  JOIN com...  5 row(s) returned    0.032 sec / 0.000 sec
```

Query Completed

AUDIT VALIDITY/ACCURACY

We say data is accurate only when it is neat and with no junk values. By using various commands like drop, del and lambda functions, all the unwanted junk values were deleted from the above rows and columns which gives valid and accuarate data report.

AUDIT COMPLETNESS

In real world, when a list of teams stats, player stats, player information, team information from a particular Player or Team or season is requested, a list of it will be displayed or presented, similarly when we compare it with above data too, we get proper real time data showing correct information for all the Matches played by teams/players. This can be extended for multiple seasons like which team is popular in that season.

AUDIT CONSISTENCY/UNIFORMITY

The datasets which have been used in this assignment show a uniform relationship between each of the dataset since they are linked to each other by a common attribute.

CONCLUSION

Primary focus of this assignment is to learn how to get the data from different sources, cleaning of data, checking null values present in the data, data munging and to reformat the data to fit a conceptual database model.

Later Created a SQL database of jobs so that job seekers and search for jobs mostly software and data jobs during the covid time

References

- https://medium.com/@msalmon00/web-scraping-job-postings-from-indeed-96bd588dcb4b (https://medium.com/@msalmon00/web-scraping-job-postings-from-indeed-96bd588dcb4b)
- https://developers.google.com/youtube/v3 (https://developers.google.com/youtube/v3)
- https://pbpython.com/pandas-list-dict.html (https://pbpython.com/pandas-list-dict.html)
- https://www.geeksforgeeks.org/python-pandas-dataframe-append/ (https://www.geeksforgeeks.org/python-pandas-dataframe-append/)

CONTRIBUTION

Your contribution towards project. How much code did you write and how much you took from other site or some other source.

I contributed By Own: 30%
Teammate contribution: 60%
Provided by the template : 10%

LICENSE

```
In [ ]:
```