Sources of Data

I used two main sources for data collection, which gave the cleanest data with more information - LinkedIn and Coursera

I used a request session to download an html page with job and course information, and scraped it using beautiful soup.

Files -

LinkedScraper.ipynb - downloads data on jobs and their skills required from LinkedIn and saves to CSV file.
CourseraScraper.ipynb - downloads data on courses from coursera and saves to CSV file.
DataExtraction.ipynb - picks up data from both csv files and cleans data, removes null values, converts to lower case, executes lemmatization, and removes stop words. This data is then used to extract skills through WordCloud data visualization. These skills are then used to look up relevant learnings resources from coursera.
This file also creates the database using create and insert queries.

Data Cleaning

- Adjusted the table for rows with missing information in any columns.
- Cleaned row entries with garbage value and filtered non English text data.
- Maintained and verified that the data for all tables in every instance is consistent.
- Added multiple rows for multi valued entries.
- Removed partial dependencies by adding new tables with primary key/foreign key relationships.
- Data in each table is stored in its most reduced form (atomic).

Use Cases

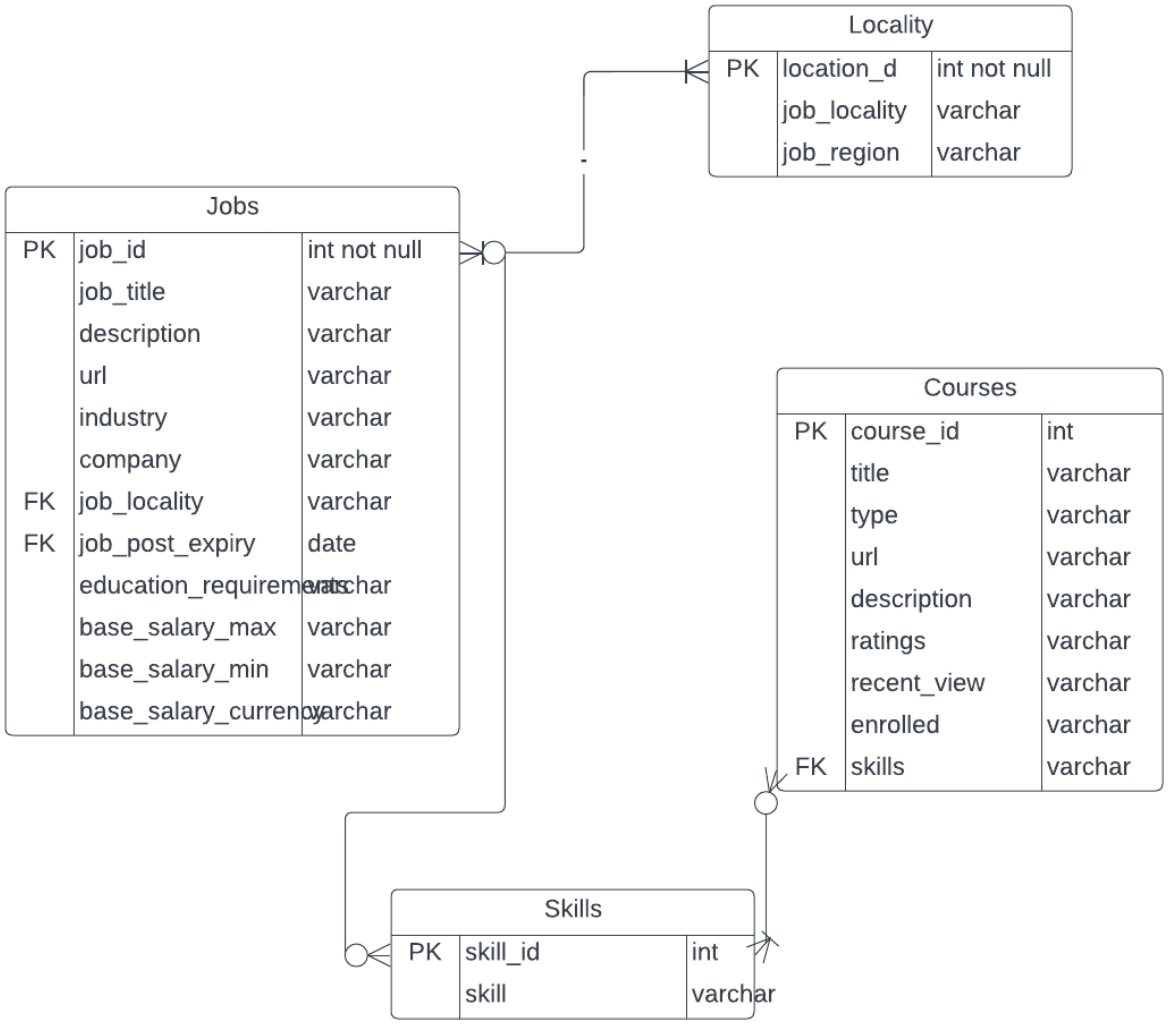
1. Find skills required for a Data Analyst Job Position
2. Find courses for my job position
3. Find jobs for set of skills in user's profile
4. How many paid courses are available on AI which provide a certificate?
5. Which is the most trending course on Natural Language Processing?

Queries -

```
query1 = "SELECT * FROM skills \
        INNER JOIN jobs ON jobs.skills=skills.skill_id \
```

```
            where jobs.job_title = 'Data Analyst' "
query2 = "SELECT * FROM courses \
             INNER JOIN skills ON skills.skill_id=courses.skills \
             INNER JOIN jobs ON skills.skill_id=jobs.skills \
            where jobs.job_title = 'Data Analyst' "
query3 = "SELECT * FROM jobs \
             INNER JOIN jobs ON jobs.skills=skills.skill_id \
            where skills.skill = 'ai' "
query4 = "SELECT * FROM courses \
             INNER JOIN jobs ON courses.skills=skills.skill_id \
            where skills.skill = 'ai' "
query5 = "SELECT * FROM courses \
             INNER JOIN jobs ON courses.skills=skills.skill_id \
            where skills.skill = 'ai' ORDER BY ratings DESC LIMIT 1"
```

Entity Relationship Diagram


Locality
PK
location_d
int not null
job_locality
varchar
job_region
varchar
Jobs
PK
job_id
int not null
job_title
varchar
description
varchar
url
varchar
industry
varchar
company
varchar
FK
job_locality
varchar
FK
job_post_expiry
date
education_requirements
varchar
base_salary_max
varchar
base_salary_min
varchar
base_salary_currency
varchar
Courses
PK
course_id
int
title
varchar
type
varchar
url
varchar
description
varchar
ratings
varchar
recent_view
varchar
enrolled
varchar
FK
skills
varchar
Skills
PK
skill_id
int
skill
varchar