EMPOWERING JOB SEEKERS IN THE MARKET WITH INFORMATION ABOVE AND BEYOND THE STATED JOB DESCRIPTION

Bhavana Prasad Kote (016044899)
Fall 2022 MS in Data Analytics
San Jose State University

Maharsh Soni ((016656770)
Fall 2022 MS in Data Analytics
San Jose State University

Rashmi Shree Veeraiah (016099395)
Fall 2022 MS in Data Analytics
San Jose State University

Sachin Kumar Srinivasa Murthy (016594773)
Fall 2022 MS in Data Analytics
San Jose State University

Abstract- The dynamic shifts in the skill requirements across all occupations are prevalent in the job market. Job seekers must have accurate information on the changing market needs that can help them upskill and unlearn specific tools to adapt and advance their career role at their preferred firm. Hence, we propose a new information systems approach that provides insights into the shift in skill requirements at varied granularities by analyzing thousands of job postings. The primary objective of our proposed system is to guide job seekers with information above and beyond the stated job description.

Keywords: *Job Markets, OLAP, SQL, Neo4J, GraphXR, TF-IDF, AWS S3, AWS Redshift, AWS Glue, Apache Spark, Data Warehousing, Star Schema, ETL, Graph Databases, Big Data*

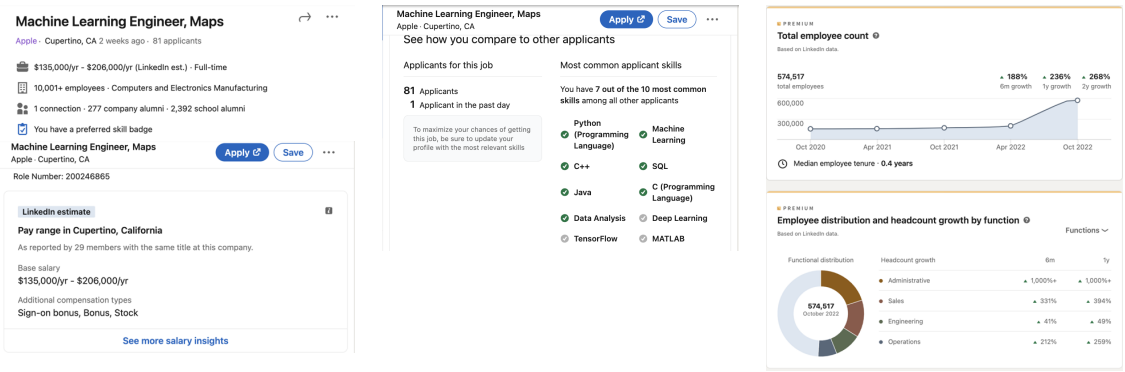
I. MOTIVATION

The dynamics of the labor market are changing rapidly in all dimensions. Firstly, the demand for skills in a defined occupation has changed significantly over the last few years and is predicted to evolve further due to technological advancements and the shift in firms' mindset towards sustainable development goals. Secondly, the driving factors for job seekers these days are not only total compensation but also other factors such as excellent work culture, diversity in the workplace, and the impact of their role and contributions on society. It is crucial to track the dynamic skill demand shifts in the job market to inform job seekers and help them stay relevant to the market's needs. Firms also need insights into the changing employee perspectives to offer opportunities for growth and support the employees in fulfilling their objectives.

We surveyed the three most popular job sites LinkedIn[1], Indeed[2], and Glassdoor[3]. We found that they own a wealth of data on job postings, salaries, and employee reviews. They perform analytics to provide terrific insights for their clients (employers and job-seekers) to inform them of the various trends in the job markets. All these three platforms have their secret-sauce/blue-ocean strategy of connecting the dots and providing insights.

LinkedIn stands out significantly by combining the power of the social network with job postings, offering a whole new dimension of insights. LinkedIn can provide insights that other player's in the market cannot easily offer, such as "education, skillset, location, and seniority of prior applications," "headcount of the firm by job functions," "time-series information on new hires," "the social connections of an applicant at a firm," and many more as provided in figure 1 (a). Indeed offers the "work happiness report," which is unique and provides incredible analytical insights into employee outlook at a firm. They also offer other analytical solutions, see Figure 1(b), such as "resume insights," "about the company," "employee reviews," and "salary insights by location and experience," which are quite similar to other competitors. Glassdoor offers similar analytical insights to its competitors, such as ratings, reviews, and salary statistics. However, Glassdoor's analytical career path insights are unlike its competitors' ones. By analyzing the career growth trajectories of the users on their platform, Glassdoor makes incredible suggestions to its end-users (job seekers) on how to get into a role and chase career growth in their profession.

Further, we also surveyed government websites, such as ONET[4] and the US Department of Labor[5], that focus on tracking the dynamics of the job markets to understand the nation's economic health and take necessary corrective actions. The data-backed insights offered by the five websites we discussed earlier have inspired us to **replicate the insights they offer at different granularities**, such as firm-level, industry-level, location-level, and various combinations of the granular classification combinations. Additionally, we found a lack of insights into the evolving demand in the job market for skills/ knowledge of technology tools and the shift in employee perceptions over time. Hence, we aim to develop a novel tool that tracks and informs on the changing skills demand and employee perceptions over time at various granular levels.



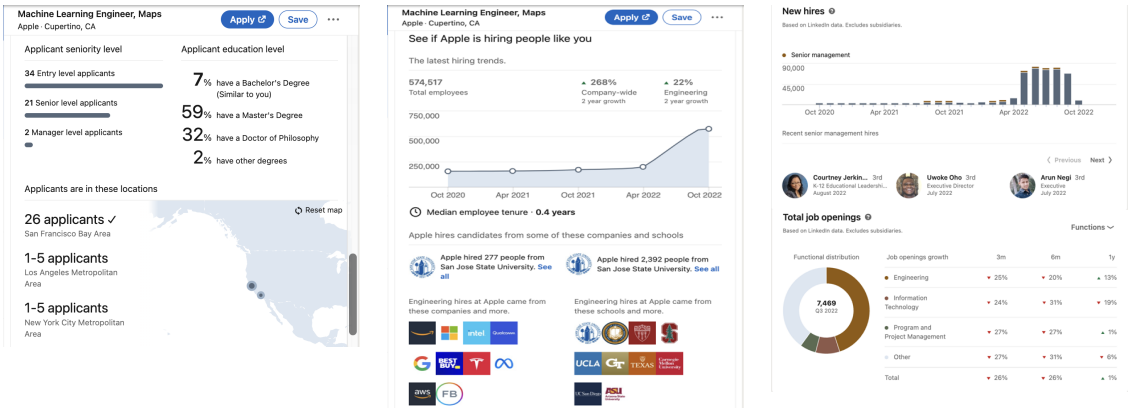


Figure 1(a)- Insights offered by LinkedIn

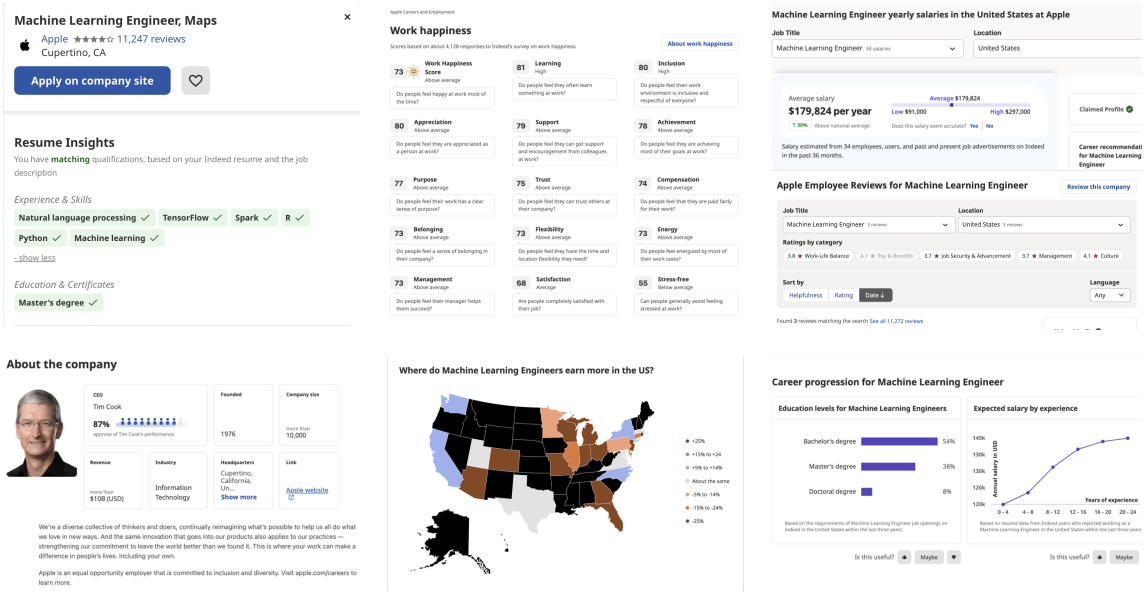


Figure 1(b)- Insights offered by Indeed

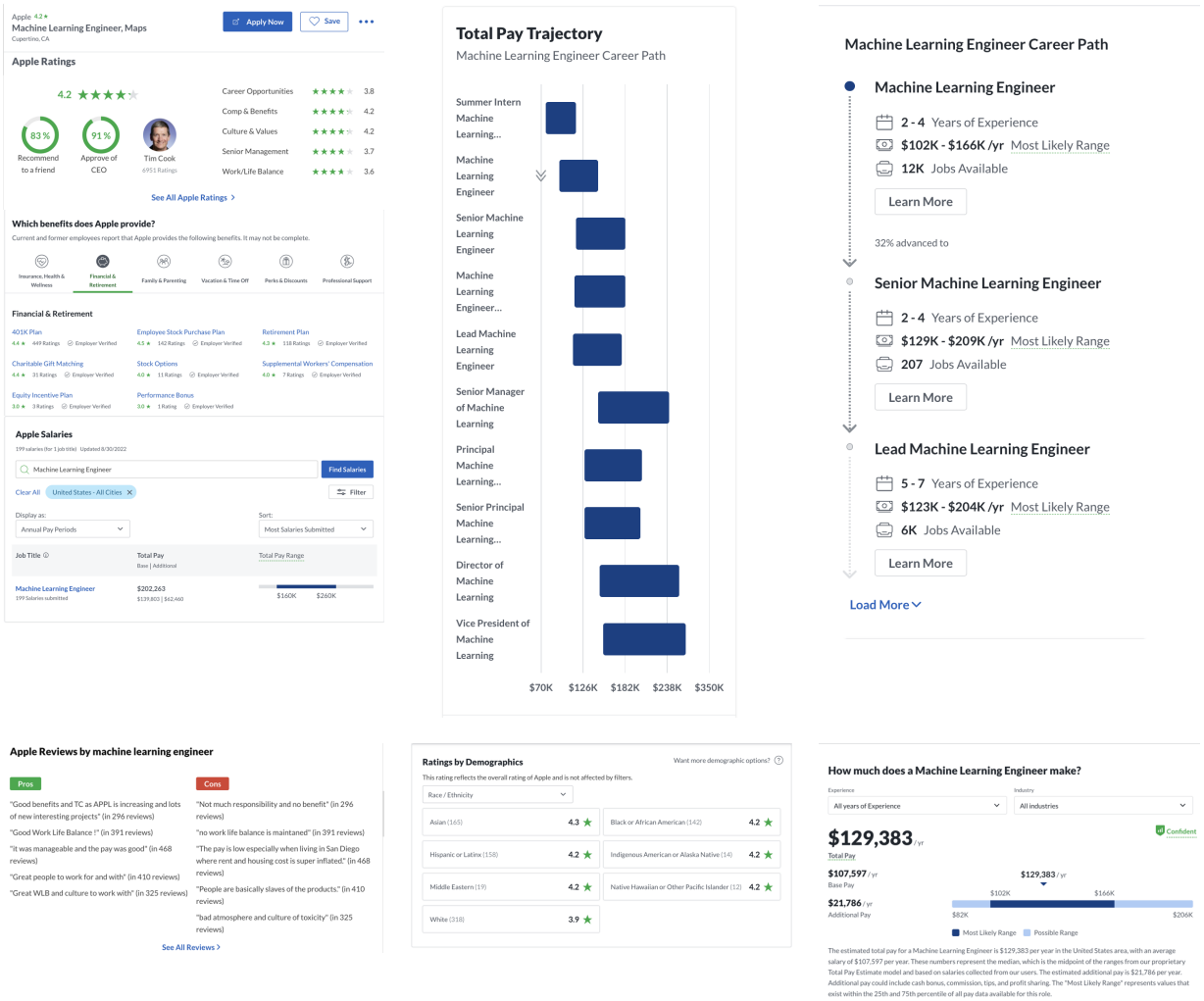


Figure 1(c)- Insights offered by Glassdoor

As job-seekers in the market, we believe it is vital to understand the increase and decrease in demand for specific technical skills. For example, many machine learning frameworks and new ML models are released often. It is hard to learn everything, so we need to understand the demand for knowledge on certain advancements in the domain besides the core knowledge to succeed as Data Scientists. Similarly, full stack developers skilled in the SOAP framework but having a hard time skilling up and learning the REST framework will be able to estimate the decline in demand for the SOAP framework quantitatively and understand the immediate need to skill up.

In sum, our motivation to develop a data-driven information system solution is driven by our personal needs as job seekers in the current market, and our in-depth research on the jobs sites, and labor markets.

II. LITERATURE SURVEY

A. DATABASE TECHNOLOGIES USED IN THE DOMAIN

All three of the chosen job advertisement platforms that we surveyed offer a democratic work culture that allows employees to write about their key technological achievements and the details of the technical solutions they use to develop complex solutions at scale. This has helped us to understand the backend technologies that power analytics at these companies. LinkedIn uses Apache Spark as its analytics engine to enable the processing of large-scale data, GraphQL [14] to query data APIs, Apache Pinot [15], super tables, Teradata, Hive SQL [16], Neo4j [17], and other solutions in their database for analytics pipeline. Indeed, LevelDB[18] and Imhotep Cluster [19] on AWS to power their analytical solutions. Glassdoor uses Cassandra and Apache Solr in their querying, searching, and designing of analytical solutions.

B. CURRENT ADVANCEMENTS IN THE DATABASE TECHNOLOGIES

Amazon Redshift Amazon Redshift is a columnar MPP data warehouse. Amazon Redshift integrates with AWS system and gives capabilities to ingest and perform ETL on semistructured data by using PartQL extension of SQL.[21]

Neo4j In a case study of recommendation systems, the paper [22] describes the research carried out utilizing Neo4j application which is an open-source graph database with strong support for connected or linked data applications. The declarative graph query language Cypher is used by Neo4j, which enables expressive and effective data querying and visualization in a property graph model.

C. PRIOR WORK ON USING NEO4J FOR MODELING JOB POSTINGS

Giabelli et al.(2021) proposed a novel recommender system named "skills2job," which provides job recommendations based on a user's skills. The system processes over 2.5 million online job vacancies posted in three European countries and evaluates the quality of several embeddings through an intrinsic evaluation. Moreover, the authors compute a measure of skill importance for each occupation in each country, the Revealed Comparative Advantage (RCA). The best vector model, one for each country, along with the RCA, feeds a graph database that serves as the keystone for the recommender system. Results are evaluated through a user study of 10 labor market experts, which shows high precision. The proposed approach can model and understand complex labor market phenomena and overcome existing limitations.

(Giabelli et al., 2020) presented a system, GraphLMI, that models labor market information collected from Online Job Vacancies (OJVs) as a graph database to support decision and policy-making activities. The authors designed and implemented the GraphLMI data model and processed over 5 million OJVs across Europe. The resulting graph database allows for the computation of occupation and skill similarities and relevance and for the enrichment of the European standard taxonomy of occupations and skills (ESCO). The authors demonstrated how the system can be queried to support policy and decision-making activities.

The paper "Skills2Graph: Processing million Job Ads to face the Job Skill Mismatch Problem" (Giabelli, Malandri, Mercorio, Mezzanzanica, & Seveso, 2021)\cite{ijcai2021p708} presents Skills2Graph, a tool designed to match job seekers with suitable job opportunities by analyzing 2.5 million Online Job Vacancies (OJVs) from the United Kingdom, France, and Germany. The tool uses a combination of co-occurrence statistics and distributional semantics to identify the most relevant jobs for the users' professional skills. The results of a user study with 10 labor market experts showed a high precision of the recommendations provided by Skills2Graph, with a high nDCG score of 0.985 and 0.984, indicating a strong correlation between the experts' scores and the rankings generated by Skills2Graph.

III. INNOVATION AND IMPACT OF THE PROJECT

The primary objective of our information system design is to offer an analytical solution that can quantify the growth or decline of demand for skills and knowledge of tools. There is plenty of research on the future of work [7] that cite the need for better-quantified solutions for tracking the demand for skills in the job markets. Prior work on analyzing the skills in order [6] provides information on the skill disruption indices by job title. However, due to the high granularity of skills, they have not been able to provide what skills are changing. So we believe that our proposed solution is **innovative** and can help someone looking for growth opportunities in their current line of work understand the market's needs. Secondly, the employment index is the driving factor of all economies, big or small. Understanding what the market is lacking can help policymakers design interventions for skilling up the workforce at the right time to ensure a healthy economy. In sum, if our proposed analytical solution is implemented into practice by job platforms, it can help job seekers on a massive scale with proper guidance and can aid in producing research on the domain of economic policy and the future of work.

Our project's other analytical solutions are the replicated analytical solutions offered in the existing market by Indeed, Glassdoor, and LinkedIn. We aim to provide details on the efficacy of our OLAP designs and the ETL tools that enabled us to develop our solutions. This can add value to future database systems as researchers refer to the use cases of the solutions that we've used in our project.

IV. SIGNIFICANCE TO REAL WORLD

Top Economic Policy researchers and consulting firms such as Mckinsey[8] and Deloitte[9] have produced numerous articles on the "Future of Work," all citing the significant shifts in skill demand due to various factors such as automation, technological distribution, and external shocks like covid-19, etc. However, job-portal platforms such as LinkedIn, Indeed, or Glassdoor offer no analytical solution to understand these shifts.

Further, the prior research on change in skill demand is only limited to the top-level skill clusters like technical skills, communication skills, etc. Therefore, we have enough evidence to believe that our solution is novel and offers granular-level insights into the increase/decrease in demand for skills and knowledge of specific tools. Additionally, our key can also provide insights into the shifts in employee perceptions tracking the changes in human-behaviors post external shocks such as COVID-19 and finding data-backed evidence on the reason for movements such as the great resignation, remote work and the dissatisfaction of employees towards going back to the office.

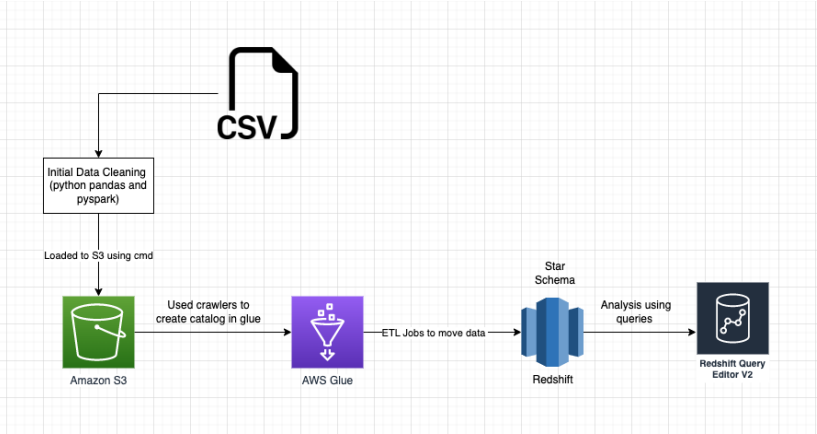
V. PROJECT WORKFLOW

Multiple database tools and technologies are used in our project. Based on different problem statements, we chose 3 workflows.

- We used workflow 1 & 3 to replicate the feasible analytical solutions offered by existing job portals in the market.
- We used workflow 2 to develop one of our visual solutions that tracks skill demand in the current job market.

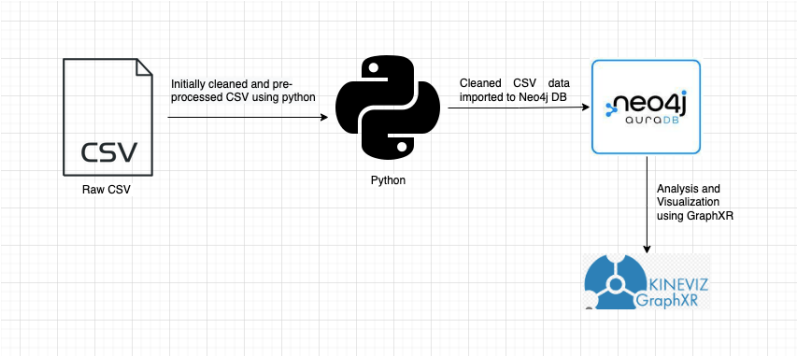
Workflow-1:

- Raw csv data was preprocessed and cleaned using Python's PySpark and Pandas libraries.
- Above data was then loaded to the Amazon S3 bucket (cloud storage) on AWS which is used as our storage. Data from S3 was crawled to glue data catalog using crawlers in AWS glue.
- We performed ETL of data using ETL jobs on glue and loaded the data to Amazon Redshift (Analytical tool) as a star schema with dimension tables and fact table.
- Using Redshift Query Editor V2, we performed our data analysis using various SQL queries.



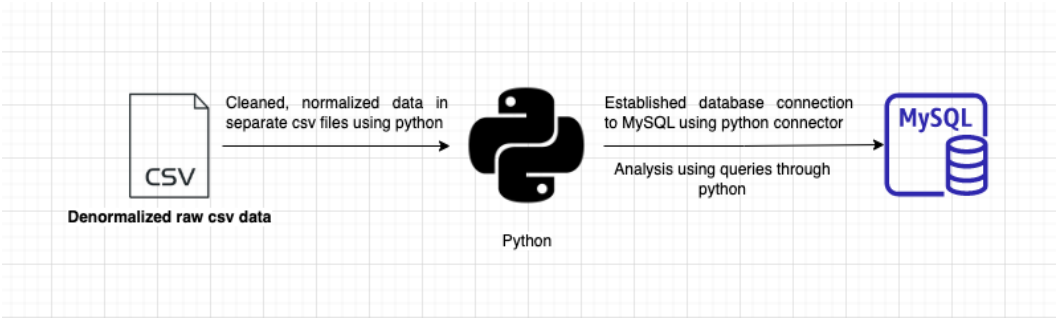
Workflow-2:

- The Multi-valued attribute, Skill and it's associated features such as Skill Cluster, Skill Cluster Family, IsSoftware, IsBaseline, and IsSpecialized were initially stored in a separate CSV file to adhere for First Normal Form.
- To this CSV file, we mapped other essential attributes as required by the Neo4J data model to show relationships using python pandas module.
- The new CSV file was imported into Neo4J AuraDB workspace, and the data model was constructed.
- Using the API Connection String of neo4J Data Model, we connect the Graph Database to GraphXR Platform to perform visual analytics.



Workflow-3:

- Normalized, cleaned data in CSV files are imported into MySQL server using MySQL workbench.
- Established connection to MySQL server using Python/MySQL connector from Python application.
- Analysis of our data was done using SQL queries in MySQL workbench and python code.



ETL

1. Using Python's PySpark module and pandas library

- Raw csv dataset was loaded to PySpark's RDD structure in python.
- Performed Data Transformations, Feature Engineering, Filtering rows
- Example of filtering out Internship jobs

```
print(df.groupBy('Internship').count().sort(desc('count')).show())
print(df.select(countDistinct("Internship")).show())
```

```
+----------+---------+
|Internship|    count|
+----------+---------+
|         0|113907545|
|         1|  1125092|
|      null|        5|
+----------+---------+

None
+-------------------------+
|count(DISTINCT Internship)|
+-------------------------+
|                        2|
+-------------------------+

None
```

```
d1= d1.filter(d1.Internship=='0')
print(d1.count(), len(d1.columns))
```

```
113870020 53
```

```
print(d1.groupBy('Internship').count().sort(desc('count')).show())
```

```
+----------+---------+
|Internship|    count|
+----------+---------+
|         0|113870020|
+----------+---------+

None
```

- Divided The Dataset into multiple tables(10 tables - 9 Dimensions & 1 fact table) and stored them to separate CSV files.

```
# dim_onet
dfon= df[['ONET', 'ONETName']]
dfon.drop_duplicates(inplace=True)
dfon.reset_index(inplace=True, drop=True)
dfon.shape
```

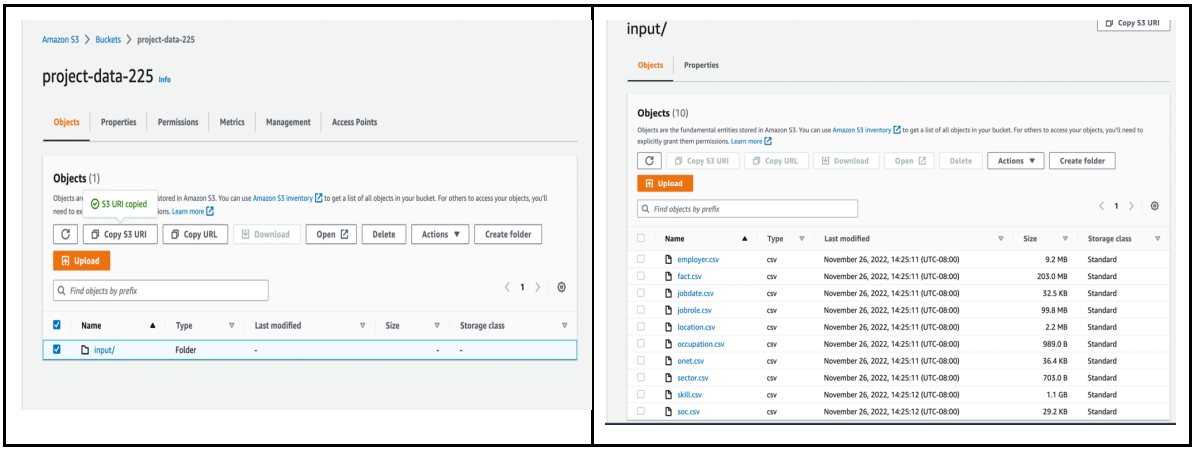
```
/var/folders/q4/8qvknrhs5d9dq4q2zxqgqzxr0000gn/T/ipykernel_89091/21853
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas
  dfon.drop_duplicates(inplace=True)
(837, 2)
```

```
# dim_emp
dfe= df[['EmployerID', 'Employer']]
dfe.drop_duplicates(inplace=True)
dfe.reset_index(inplace=True, drop=True)
dfe.shape
```

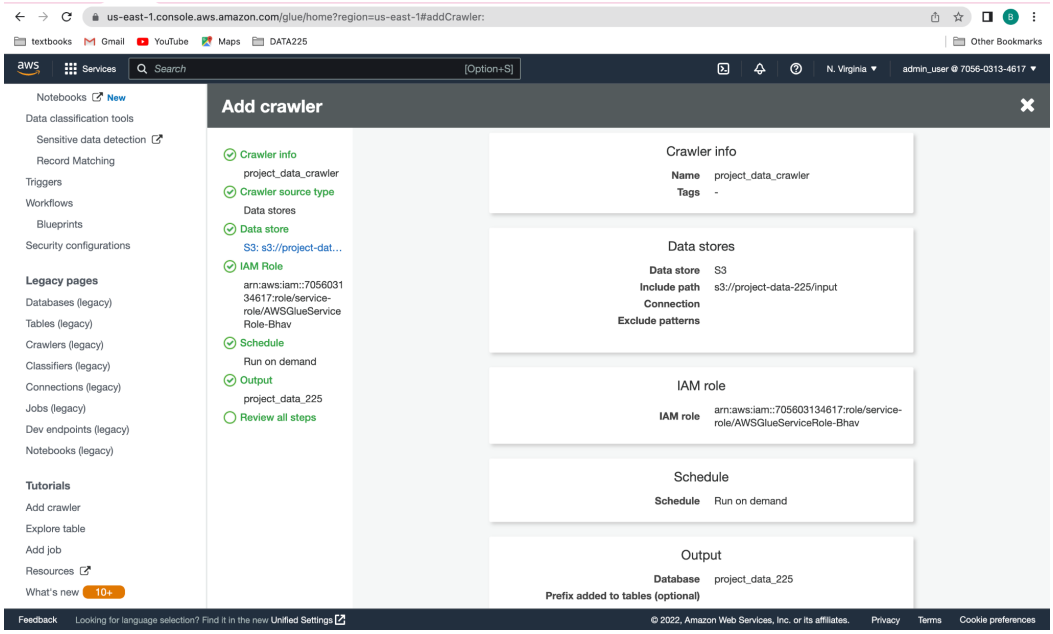
```
/var/folders/q4/8qvknrhs5d9dq4q2zxqgqzxr0000gn/T/ipykernel_89091/12007
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas
  dfe.drop_duplicates(inplace=True)
(313703, 2)
```

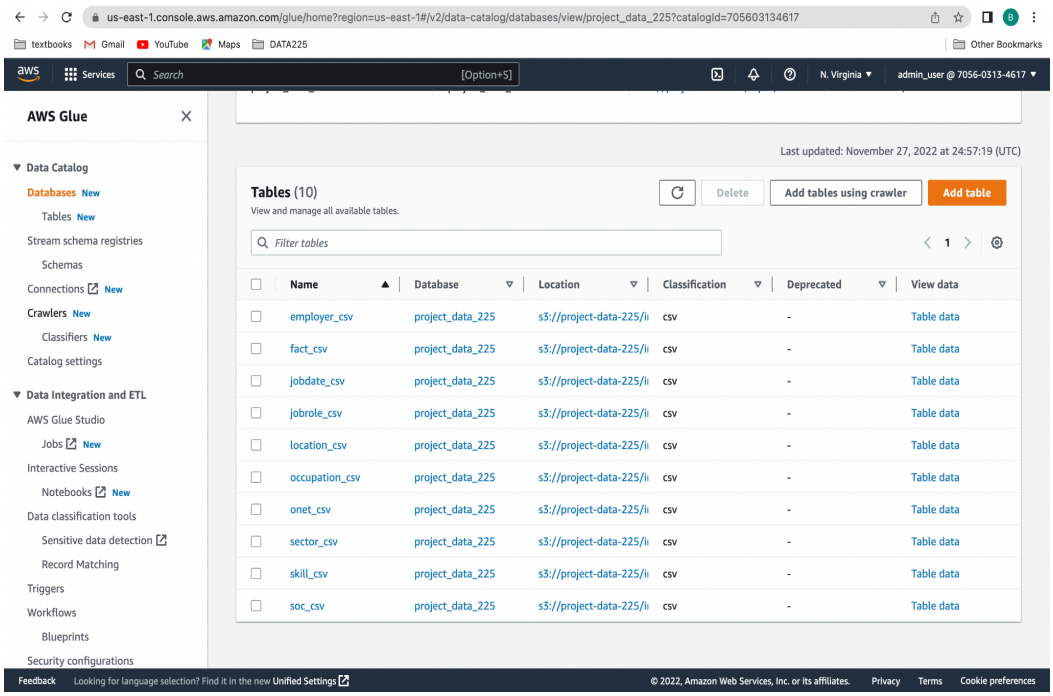
2. Using AWS

- Preprocessed csv files were loaded into S3 bucket which is a data lake on AWS using command line interface.

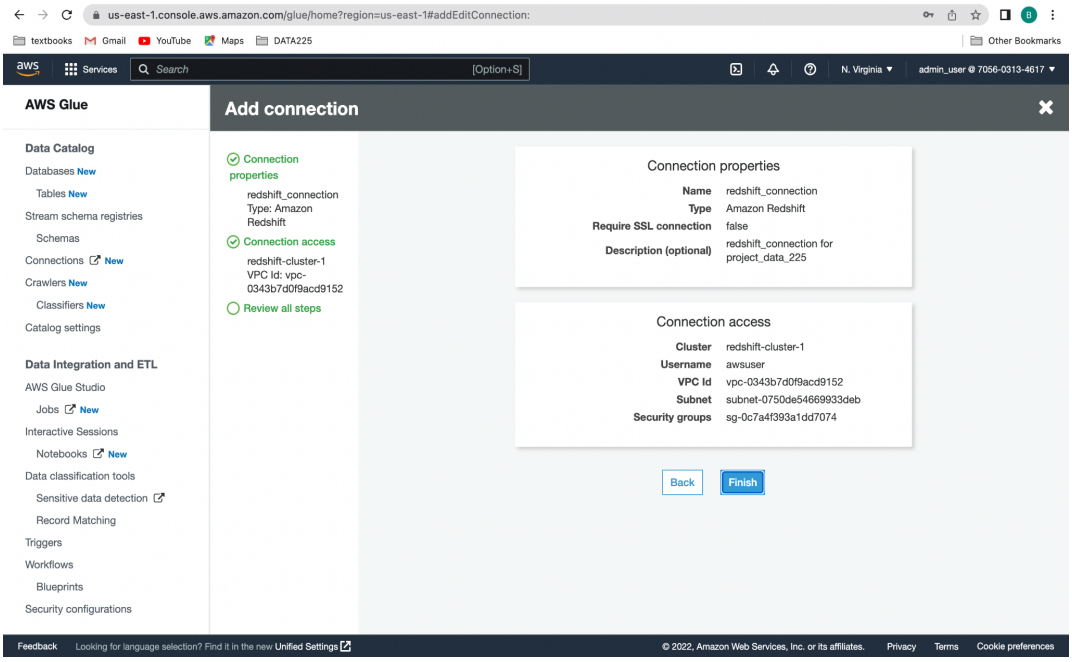


- Using crawler, data was crawled to a data catalog on AWS glue which is an ETL tool.

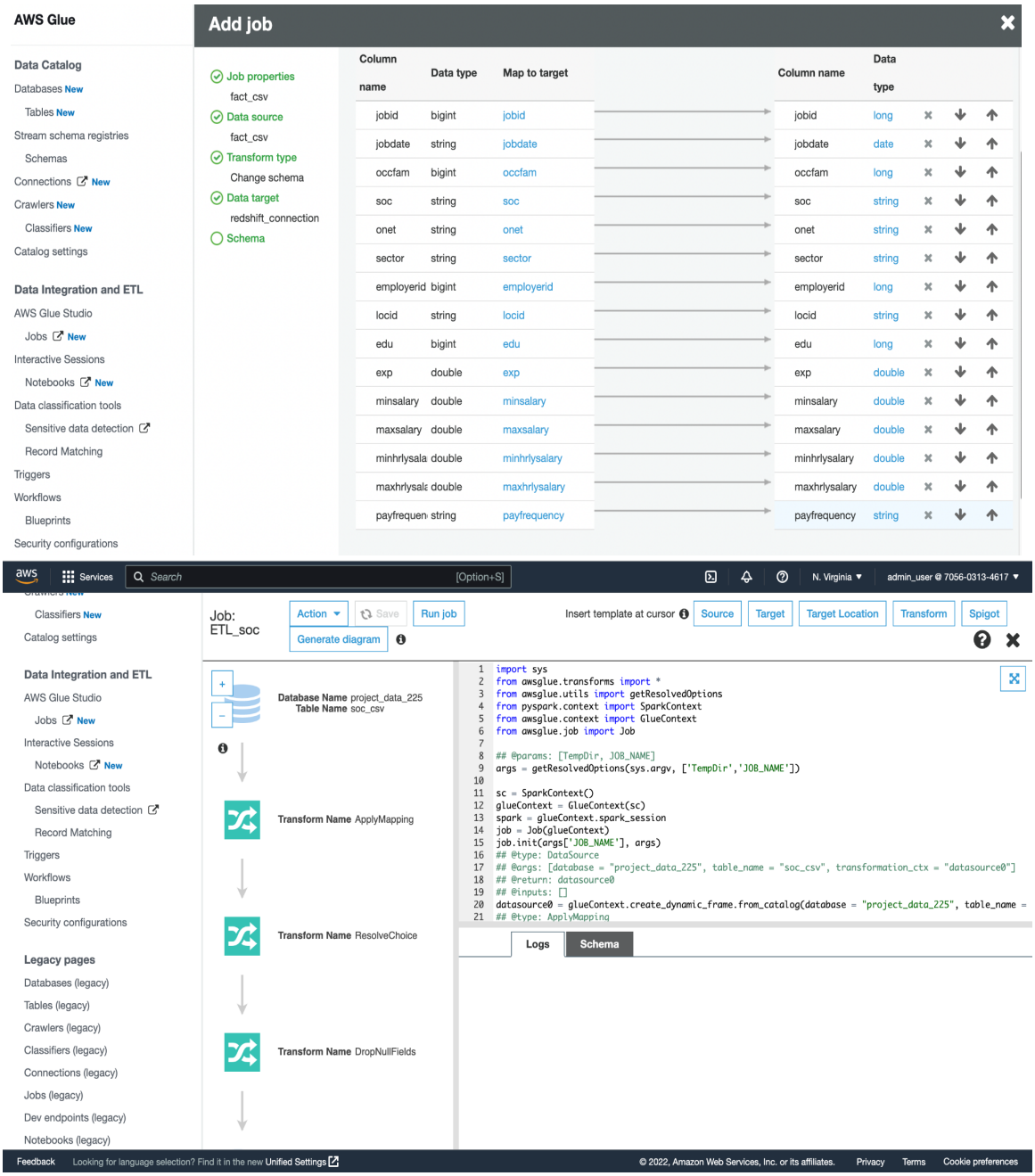




- Then, we created a cluster on Amazon redshift and the database. Created and tested connection to Redshift on glue.

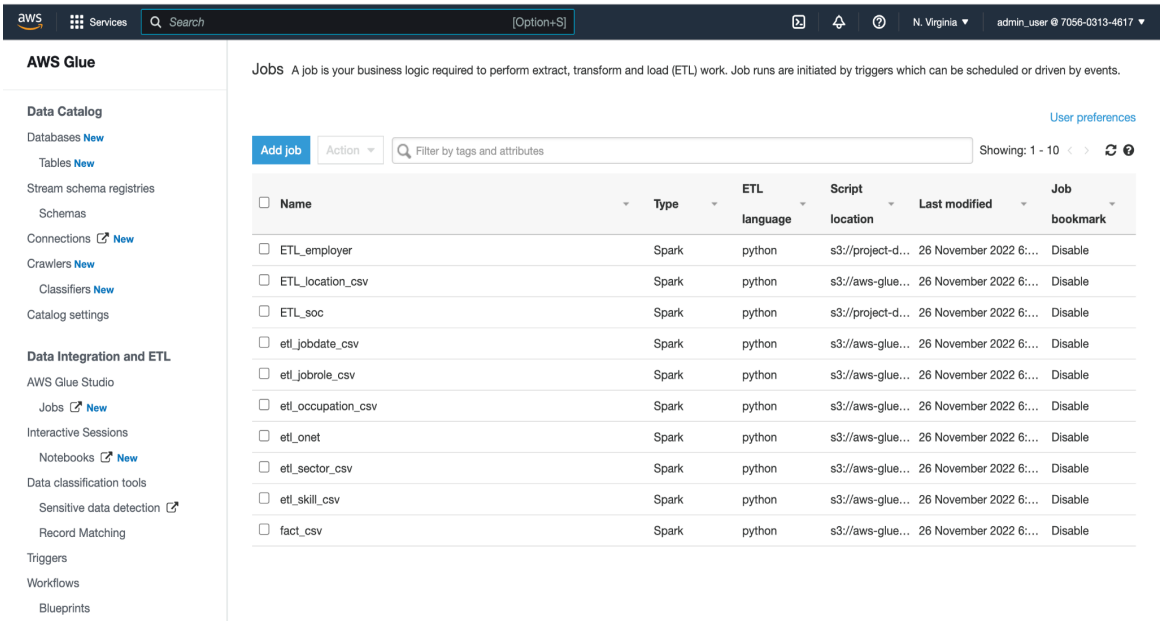


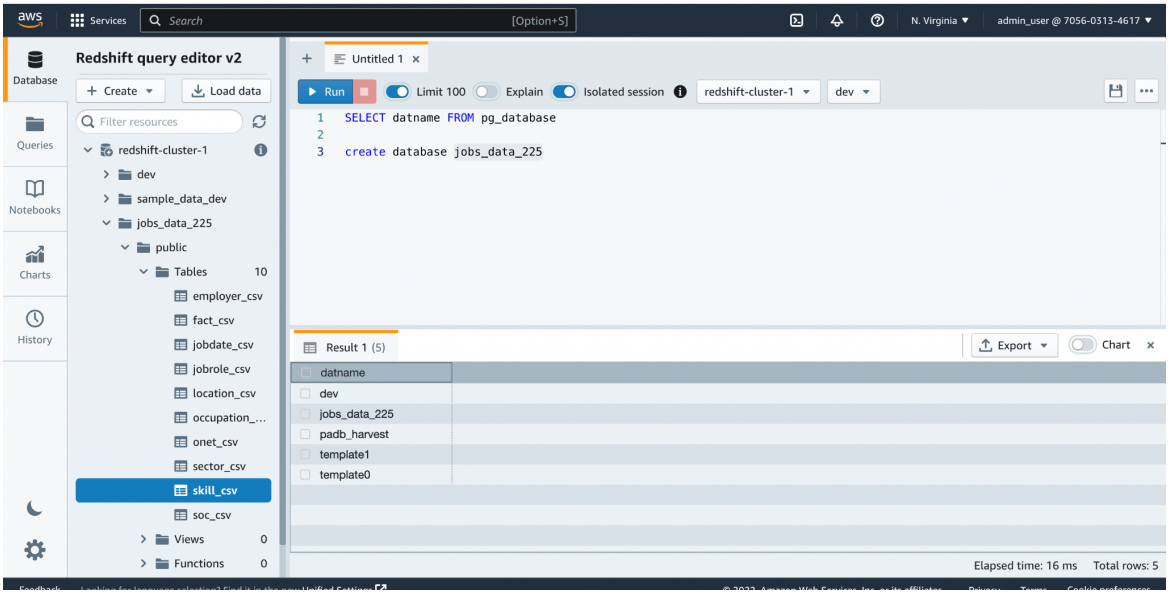
- At this stage, we performed data type transformations, extraction of month year, quarter year and year from job_date field, removal of duplicates etc using ETL jobs in AWS glue.



- The data was then loaded into Redshift from glue using the glue job created above.

- The above process was followed for all the other csv files using multiple ETL jobs





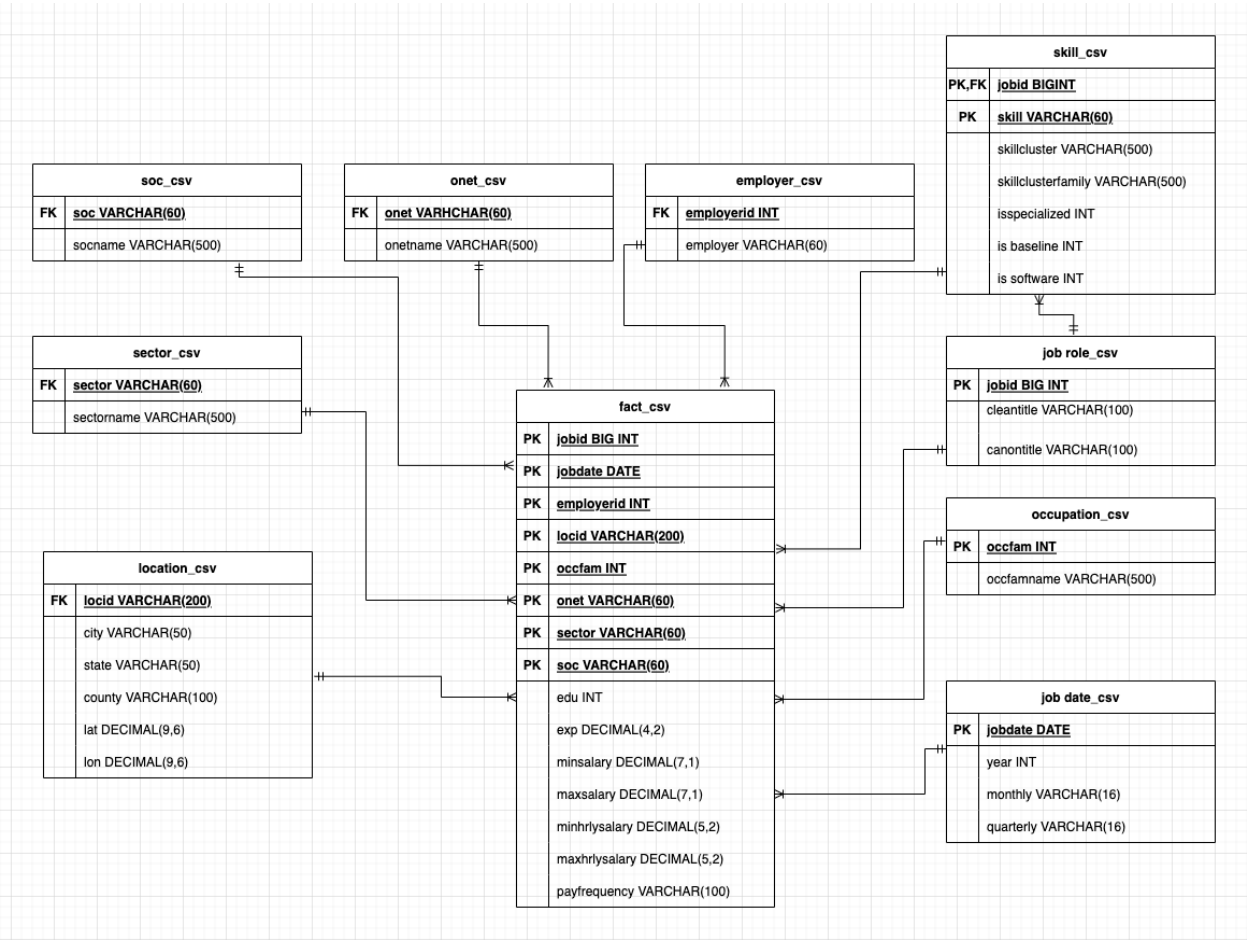
- Above screenshot shows all the tables loaded into Amazon Redshift Data Warehouse.

VI. DATA MODELING

Data Warehouse:

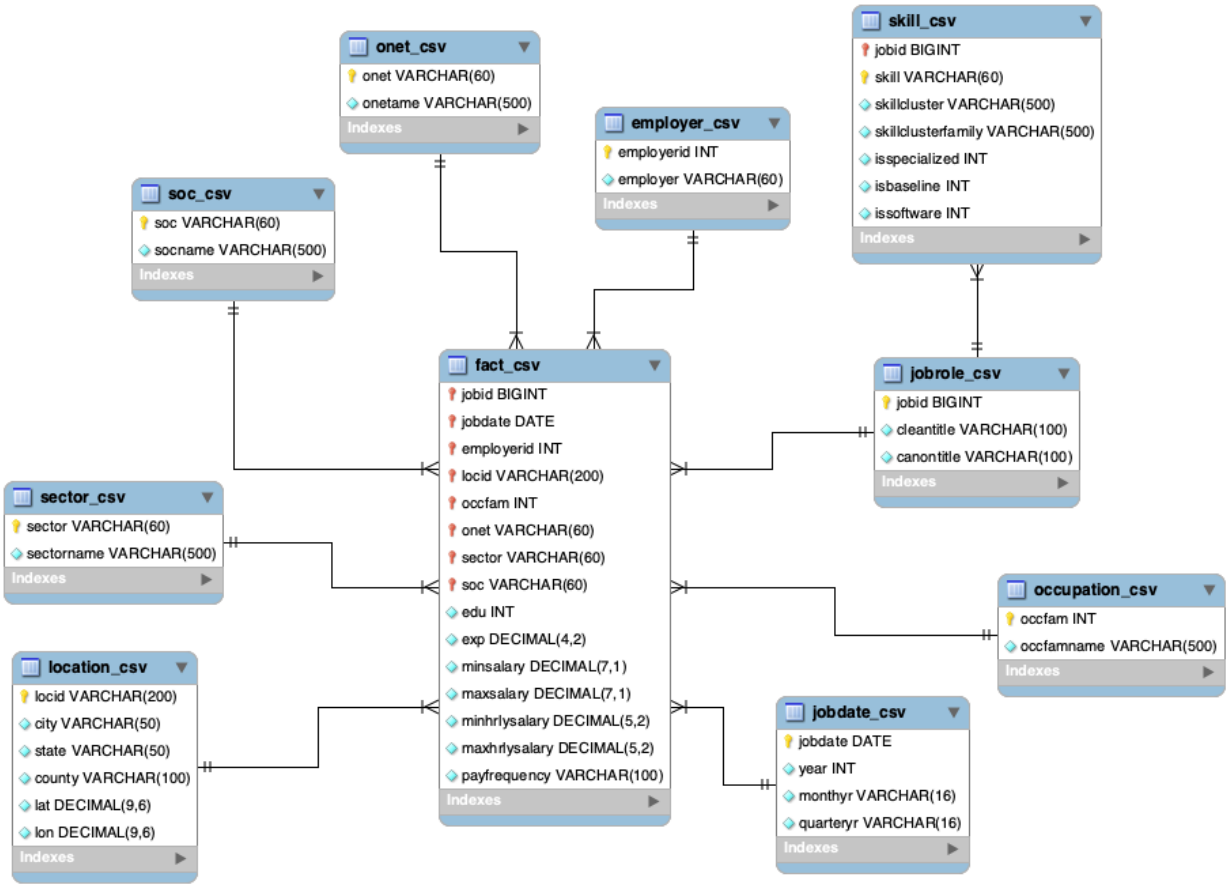
Below data model is designed based on the star schema for the data warehouse. Our model for DW consists of 9 dimensions which pertains to each entity type that describes the objects in fact table and will mostly be static and one fact table consisting of all the quantitative measures and foreign keys from all the dimension tables as its composite primary key.

Initial model design for Data warehouse:



STAR SCHEMA DIAGRAM

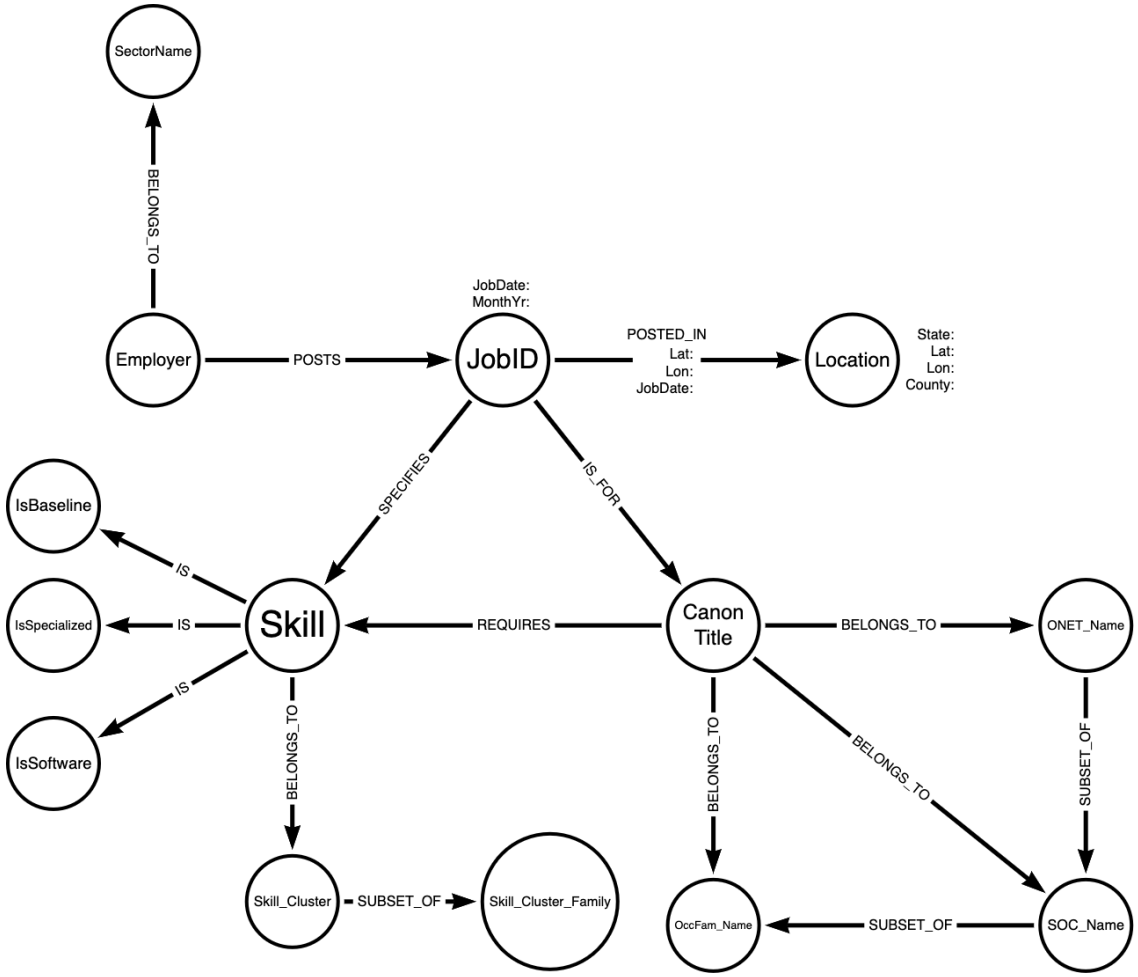
Reverse engineered data model:



STAR SCHEMA DIAGRAM

Neo4j Data Model:

The below Neo4j data model represents 14 nodes and the corresponding edges represent the relationship among them. Nodes and edges also have properties such as jobDate, monthyr etc



VII. ANALYTICS & DATA DRIVEN DECISIONS

1. Analysis using RDBMS:

- We first established a connection between MySQL server and python application using the python connector/MySQL.

jobs_python_mysql_db_operations.py

```
from mysql.connector import connect, Error
from getpass import getpass
from mysql.connector import errorcode
import csv
import pandas as pd
import warnings
warnings.filterwarnings('ignore')


# start point
try:
    db_connection = connect(
        host = 'localhost',
        username = 'root',
        password = getpass('enter the password for mysql server'),
        database = 'jobs_data'
    )
    print("connection object created successfully" +str(db_connection))
```

- Created the database ‘jobs_data’, all the tables and imported data into the tables from csv files on MySQL Workbench.
- Once all the data was imported into tables, we performed some analysis using MySQL through Python.

Query 1: Number of Job Openings by Occupation for specified (window) timed intervals By company with job_postings > 50

```
    cursor = db_connection.cursor()

    #Job Openings by Occupation for specified (window) timed intervals By company with job_postings > 50
    join1 = ("select d.year, o.occfamname, e.employer, count(*) as job_postings "
    "from fact_csv f inner join jobdate_csv d "
    "on f.jobdate = d.jobdate "
    "inner join occupation_csv o "
    "on f.occfam = o.occfam "
    "inner join employer_csv e "
    "on f.employerid = e.employerid "
    "where o.occfamname = \'Computer and Mathematical Occupations\' "
    "group by o.occfamname, e.employer, d.year "
    "having d.year between \'2019\' and \'2021\' "
    "and job_postings > 50 "
    "order by year, occfamname;")

    print("----------**********----------")
    print('executing query: '+join1)
    cursor.execute(join1)
    records = cursor.fetchall()
    print('output of the query')
    for row in records:
        print(row)
    print('------ with column names ------')
    df3 = pd.read_sql(join1, db_connection)
    print(df3)

    cursor.close()
    db_connection.close()



except Error as e:
    print("exception occurred ")
    print(e)


finally:
    if db_connection is not None:
        db_connection.close()
        print("Database connection closed")
```

Output:

```
bhavana@Bhavanas-MacBook-Pro python % python3 jobs_python_mysql_db_operations.py
enter the password for mysql server
connection object created successfully<mysql.connector.connection_cext.CMySQLConnection object at 0x103dcb2b0>
----------**********------------
executing query: select d.year, o.occfamname, e.employer, count(*) as job_postings from fact_csv f inner join jobdate_csv d on f.jobdate = d.jobdate inner join occupation_csv o on f.occfam = o.occfam inne
r join employer_csv e on f.employerid = e.employerid where o.occfamname = 'Computer and Mathematical Occupations' group by o.occfamname, e.employer, d.year having d.year between '2019' and '2021' and job_
postings > 50 order by year, occfamname;
output of the query
     year                          occfamname                          employer  job_postings
0    2019  Computer and Mathematical Occupations                    Charles Schwab           133
1    2019  Computer and Mathematical Occupations                   Department Army           405
2    2019  Computer and Mathematical Occupations                          US Navy           237
3    2019  Computer and Mathematical Occupations  US Department of Homeland Security            71
4    2019  Computer and Mathematical Occupations     Federal Aviation Administration            51
..    ...                                    ...                               ...           ...
129  2021  Computer and Mathematical Occupations                        Accenture          1078
130  2021  Computer and Mathematical Occupations                Change Healthcare           166
131  2021  Computer and Mathematical Occupations                   Sanford Health            95
132  2021  Computer and Mathematical Occupations       AECOM Technology Corporation            67
133  2021  Computer and Mathematical Occupations                     Edward Jones            73

[134 rows x 4 columns]
Database connection closed
```

With this query, we analysed the number of job postings for Computer and mathematical occupations posted by employers in 2019, 2020 and 2021. There are significant no of employers with more than 50 job openings for different job roles.

Query 2: Job Openings by Occupation for specified (window) timed intervals: By Location

```
48      #Job Openings by Occupation for specified (window) timed intervals: By Location
49
50      join2 = ("select d.year, o.occfamname, l.state, count(*) as job_postings "
51      "from fact_csv f inner join jobdate_csv d "
52      "on f.jobdate = d.jobdate "
53      "inner join occupation_csv o "
54      "on f.occfam = o.occfam "
55      "inner join location_csv l "
56      "on f.locid = l.locid "
57      "group by o.occfamname, l.state, d.year "
58      "having d.year between '2019' and '2021' "
59      "order by year, occfamname;")
60
61      print("----------**********-------------")
62      print('executing query: '+join2)
63      cursor.execute(join2)
64      records = cursor.fetchall()
65      print('output of the query')
66      #for row in records:
67      #    print(row)
68      print('------ with column names ------')
69      df4 = pd.read_sql(join2, db_connection)
70      print(df4)
```

output:

```
----------**********------------
executing query: select d.year, o.occfamname, l.state, count(*) as job_postings from fact_csv f inner join jobdate_csv d on f.jobdate = d.jobdate inner join occupation_csv o on f.occfam = o.occfam inner j
oin location_csv l on f.locid = l.locid group by o.occfamname, l.state, d.year having d.year between '2019' and '2021' order by year, occfamname;
output of the query
------ with column names ------
      year                                  occfamname          state  job_postings
0     2019       Architecture and Engineering Occupations        Alabama           246
1     2019       Architecture and Engineering Occupations           Ohio           373
2     2019       Architecture and Engineering Occupations       Oklahoma           205
3     2019       Architecture and Engineering Occupations         Alaska           110
4     2019       Architecture and Engineering Occupations  North Carolina           621
...    ...                                           ...            ...           ...
3396  2021  Transportation and Material Moving Occupations      Louisiana           839
3397  2021  Transportation and Material Moving Occupations      Wisconsin          1107
3398  2021  Transportation and Material Moving Occupations        Indiana          1881
3399  2021  Transportation and Material Moving Occupations           Iowa           817
3400  2021  Transportation and Material Moving Occupations        Wyoming           288

[3401 rows x 4 columns]
Database connection closed
```

Query 3: Number of job postings for ‘Data Analyst’ job role in each state in the United States in descending order

```
#states with no of job postings for DA in descending order

join3 = ("select l.state, r.canontitle, count(*) as job_postings "
"from fact_csv f inner join location_csv l "
"on f.locid = l.locid "
"inner join jobrole_csv r "
"on r.jobid = f.jobid "
"group by l.state, r.canontitle "
"having r.canontitle = \'Data Analyst\' "
"order by job_postings desc;")

print("----------**********----------")
print('executing query: '+join3)
cursor.execute(join3)
records = cursor.fetchall()
print('output of the query')
#for row in records:
#   print(row)
print('------ with column names ------')
df4 = pd.read_sql(join3, db_connection)
print(df4)
```

Output:

```
output of the query
------ with column names ------
                   state    canontitle  job_postings
0             California  Data Analyst           251
1                  Texas  Data Analyst           210
2               Colorado  Data Analyst           140
3               Virginia  Data Analyst           116
4           Pennsylvania  Data Analyst           105
5                Florida  Data Analyst           103
6             Washington  Data Analyst            83
7   District of Columbia  Data Analyst            81
8               New York  Data Analyst            75
9                 Oregon  Data Analyst            71
10              Illinois  Data Analyst            69
11              Maryland  Data Analyst            63
12               Arizona  Data Analyst            60
13        North Carolina  Data Analyst            59
14              Missouri  Data Analyst            55
15         Massachusetts  Data Analyst            45
16             Minnesota  Data Analyst            45
17        South Carolina  Data Analyst            41
18            New Jersey  Data Analyst            39
19                  Ohio  Data Analyst            37
20             Tennessee  Data Analyst            36
```

We can see that the highest number of job postings for Data Analyst role are posted for California location with 251 openings. This is followed by Texas with 210 postings. Job Opportunities are higher in states where most of the companies headquarters are located.

Query 4: Find out the no of employers (companies) offering the job postings for software related positions

```
#no of employers offering job postings in software industry

join4 = ("select r.canontitle, count(distinct e.employer) as no_of_employers "
"from jobrole_csv r inner join fact_csv f "
"on r.jobid = f.jobid "
"inner join employer_csv e "
"on f.employerid = e.employerid "
"inner join occupation_csv o "
"on f.occfam = o.occfam "
"where o.occfamname = \'Computer and Mathematical Occupations\' "
"group by r.canontitle "
"order by no_of_employers desc;")

print("----------**********----------")
print('executing query: '+join4)
cursor.execute(join4)
records = cursor.fetchall()
print('output of the query')
#for row in records:
#   print(row)
print('------ with column names ------')
df4 = pd.read_sql(join4, db_connection)
print(df4)
```

output:

```
output of the query
------ with column names ------
                            canontitle  no_of_employers
0         Software Development Engineer            1323
1                         Data Analyst            1035
2                 Systems Administrator             980
3     Information Technology Specialist             774
4                       Systems Analyst             724
...                                 ...             ...
1176      Wordpress Developer/Designer               1
1177        Wordpress Plugin Developer               1
1178            Workstation Technician               1
1179                      Writer/Editor              1
1180                   Xamarin Developer             1

[1181 rows x 2 columns]
Database connection closed
```

The above analysis shows that large number of companies (1323 distinct employers) offer Software Development Engineer positions, which is followed by Data Analyst, System Administrator and Information Technology Specialist.

Query 5: Trend analysis of number of job postings by year in each occupation

```
#job postings by year and occupation

join5 = ("select d.year, o.occfamname, count(*) as job_postings "
"from fact_csv f inner join jobdate_csv d "
"on f.jobdate = d.jobdate "
"inner join occupation_csv o "
"on f.occfam = o.occfam "
"group by d.year, o.occfamname "
"order by occfamname, d.year;")

print("----------**********----------")
print('executing query: '+join5)
cursor.execute(join5)
records = cursor.fetchall()
print('output of the query')
#for row in records:
#   print(row)
print('------ with column names ------')
df4 = pd.read_sql(join5, db_connection)
print(df4)
```

output:

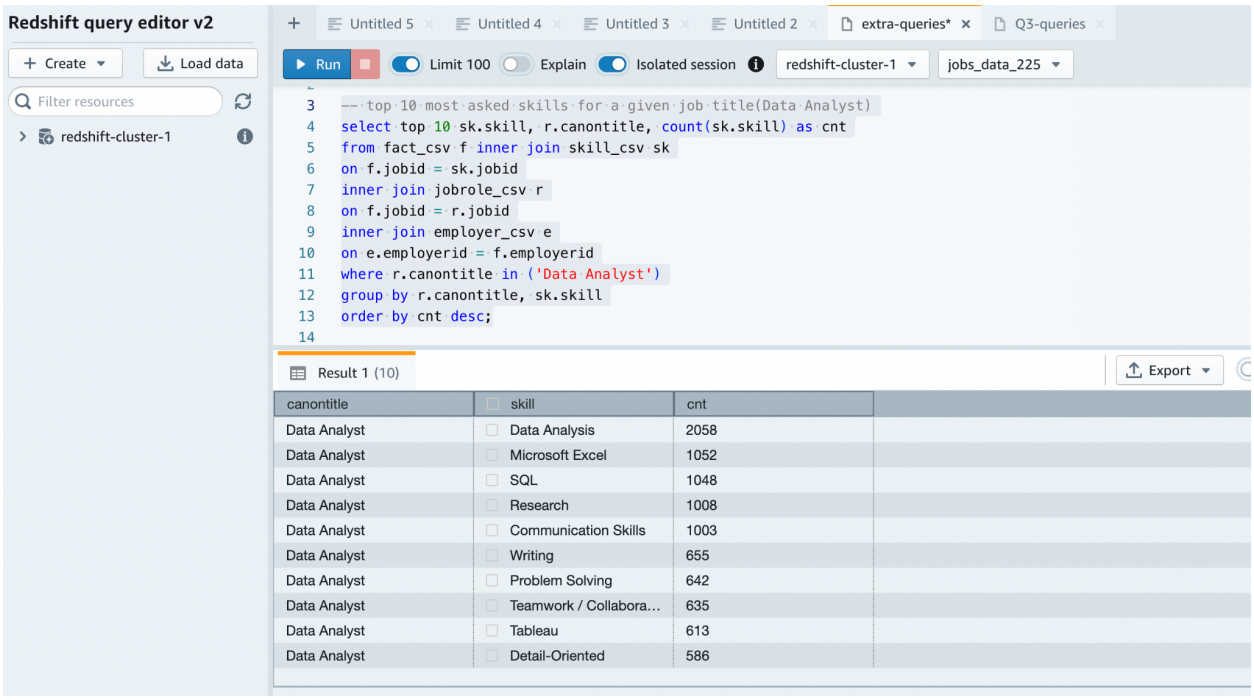
```
output of the query
------ with column names ------
    year                                          occfamname  job_postings
0   2019          Architecture and Engineering Occupations         14495
1   2020          Architecture and Engineering Occupations         10637
2   2021          Architecture and Engineering Occupations         14769
3   2022          Architecture and Engineering Occupations            12
4   2019  Arts, Design, Entertainment, Sports, and Media...          4917
..   ...                                                ...           ...
82  2022                    Sales and Related Occupations            86
83  2019     Transportation and Material Moving Occupations         26627
84  2020     Transportation and Material Moving Occupations         44904
85  2021     Transportation and Material Moving Occupations         68217
86  2022     Transportation and Material Moving Occupations            35

[87 rows x 3 columns]
```

With this analysis, We can clearly understand the trend in the number of job postings in each year. With the number of startups increasing and companies growing in size, there are more job opportunities opening every year compared to the previous year.
(Note: numbers in 2022 are less since our data covers only half of the first quarter of 2022)

2. Using Amazon Redshift Query Editor V2

Below are some of the interesting analysis performed using the the analytical tool query editor v2 provided by Amazon Redshift which is a data warehouse

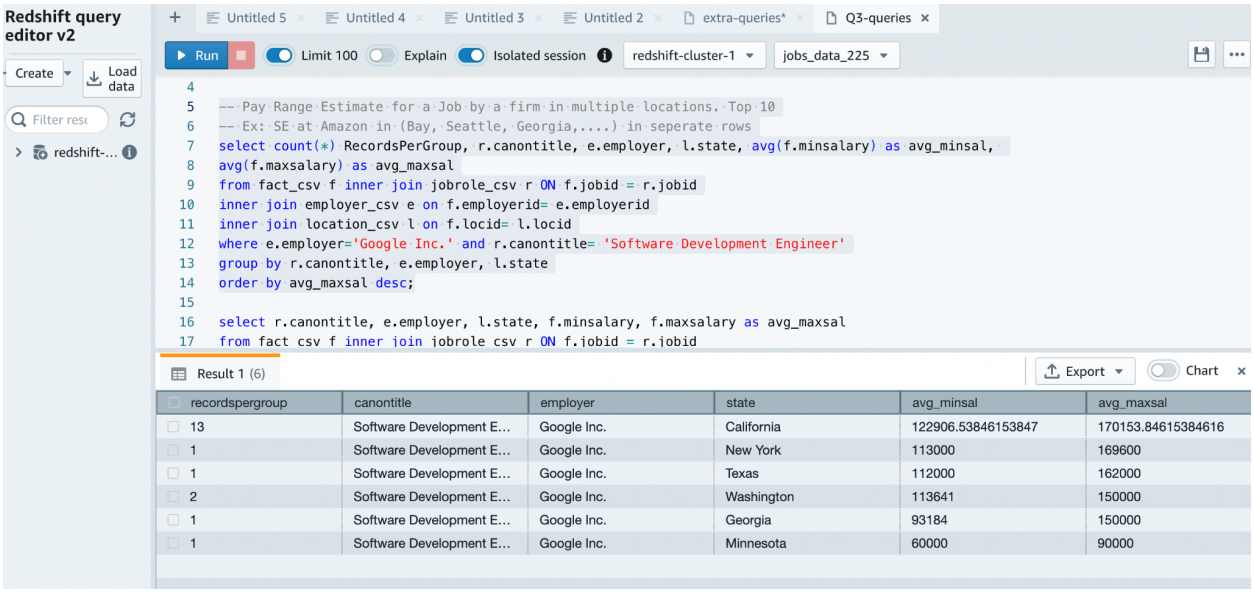
- Which are the top 10 skills required or necessary (basic requirements) for a Data Analyst job



canontitle	skill	cnt
Data Analyst	Data Analysis	2058
Data Analyst	Microsoft Excel	1052
Data Analyst	SQL	1048
Data Analyst	Research	1008
Data Analyst	Communication Skills	1003
Data Analyst	Writing	655
Data Analyst	Problem Solving	642
Data Analyst	Teamwork / Collabora...	635
Data Analyst	Tableau	613
Data Analyst	Detail-Oriented	586

Data Analysis, Microsoft Excel, SQL and communication skills are the most important skills for a data Analyst in the industry.

- Find the pay range estimate for a given job by a firm/company in multiple locations
  Example: Pay range for a Software Development Engineer at Google or Amazon in different locations



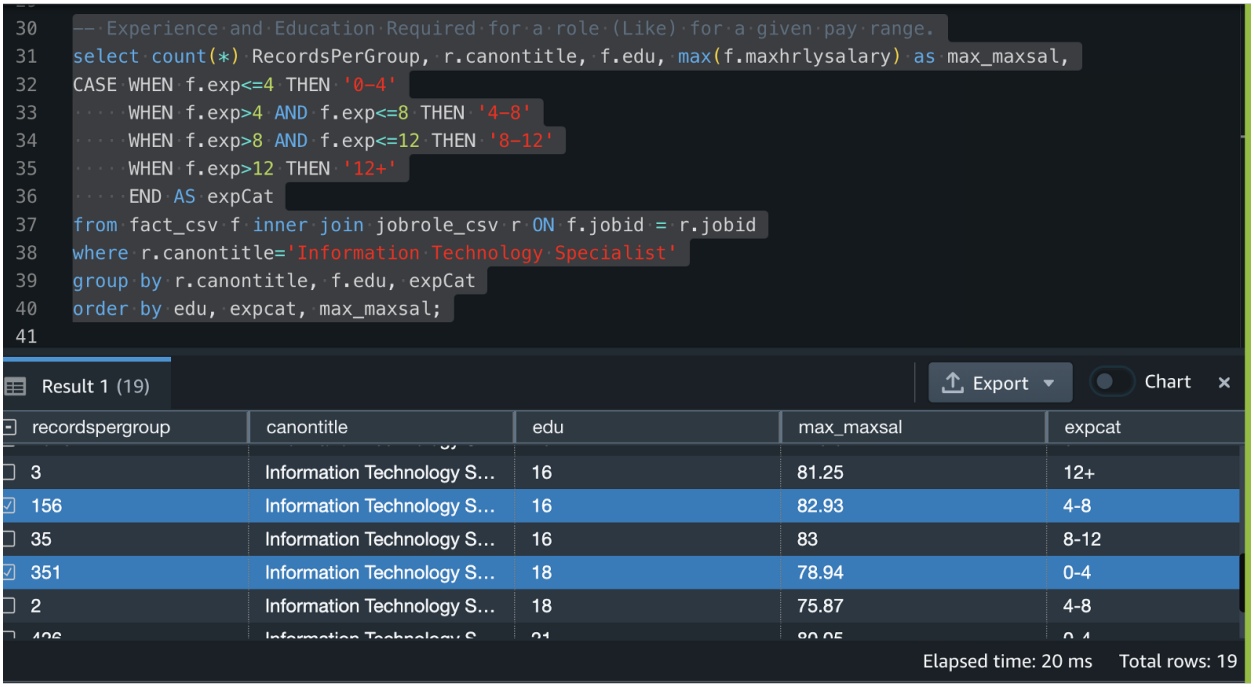
recordspergroup	canontitle	employer	state	avg_minsal	avg_maxsal
13	Software Development E...	Google Inc.	California	122906.53846153847	170153.84615384616
1	Software Development E...	Google Inc.	New York	113000	169600
1	Software Development E...	Google Inc.	Texas	112000	162000
2	Software Development E...	Google Inc.	Washington	113641	150000
1	Software Development E...	Google Inc.	Georgia	93184	150000
1	Software Development E...	Google Inc.	Minnesota	60000	90000

It is very important for job seekers to learn the difference in pay by location when deciding on their career in a location. Hence, we execute the above query to inform the job seekers by inputting their role, and firm name to understand the average max and min salaries across locations.

- Pay range estimate for a job role across all locations irrespective of the company and location.
  As a software engineering student, I would like to know the pay range for a software development engineer in the industry.

The Above result is a replication of Glassdoor's Salary Estimate for a Job Title irrespective of location, and firm filters.

- List down the education and experience required for a job role for a particular pay range



In the above result we see that those with bachelor's degree require twice the amount of experience to earn commensurate salaries as that of their master's degree holder colleagues.

- Find the career trajectory for a domain. Provide the average salary, education and experience
  In the below query, we have taken an example of 'clinical research'.

```
41
42  -- Career Trajectory for a domain. (ML, Clinical Research, Store Manager)
43  select r.canontitle, avg(f.edu) as edu, round(avg(f.exp)) as exp, round(avg(f.maxsalary)) as maxsal
44  from fact_csv f inner join jobrole_csv r ON f.jobid = r.jobid
45  where r.canontitle like '%Clinical Research%'
46  group by r.canontitle
47  order by maxsal, exp, edu;
48
```

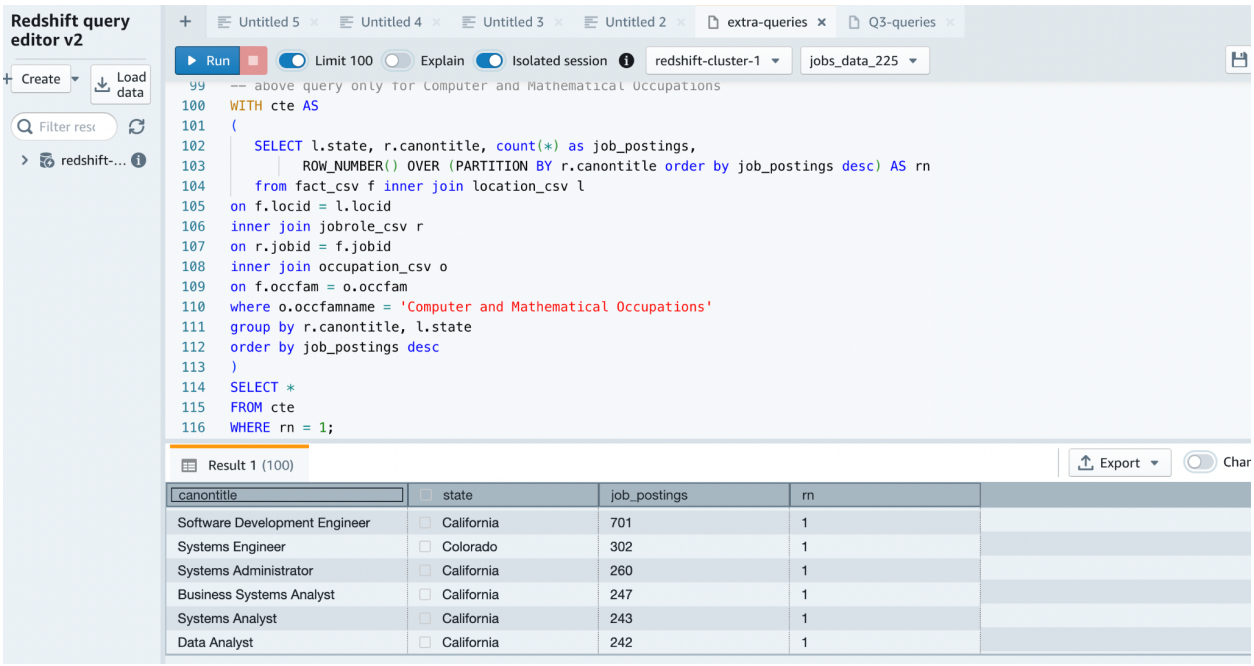
Result 1 (5) | Export | Chart

canontitle	edu	exp	maxsal
Clinical Research Assistant	7	2	47483
Clinical Research Coordinator	13	3	62685
Clinical Research Specialist	14	5	72250
Clinical Research Associate	13	3	82621
Clinical Research Manager	15	5	91881

Elapsed time: 32 ms Total rows: 5

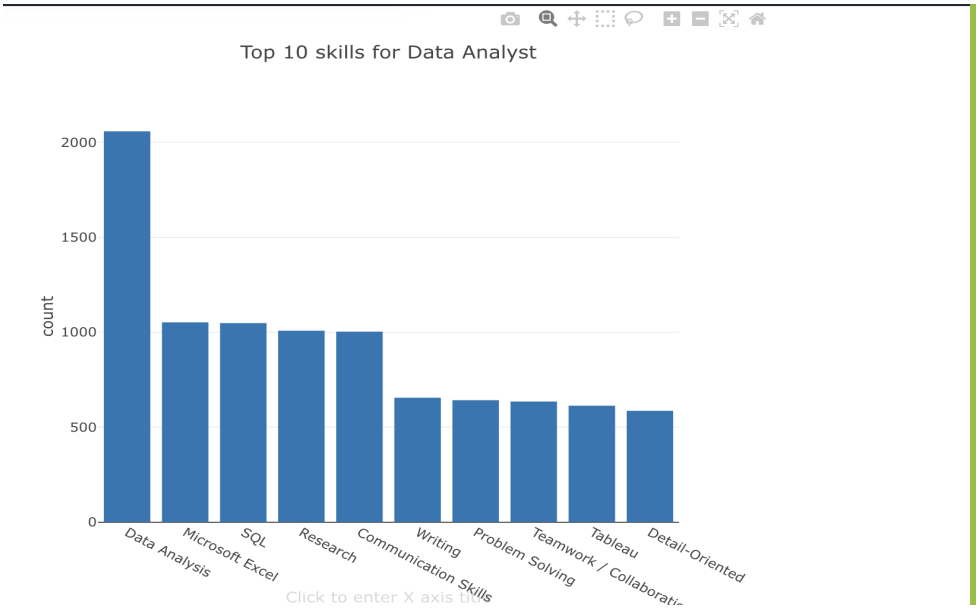
In the above result, we show the career path for someone in Clinical Research, and inform them how they should grow in their chosen occupation and what maximum salary they can expect at each stage of their career.

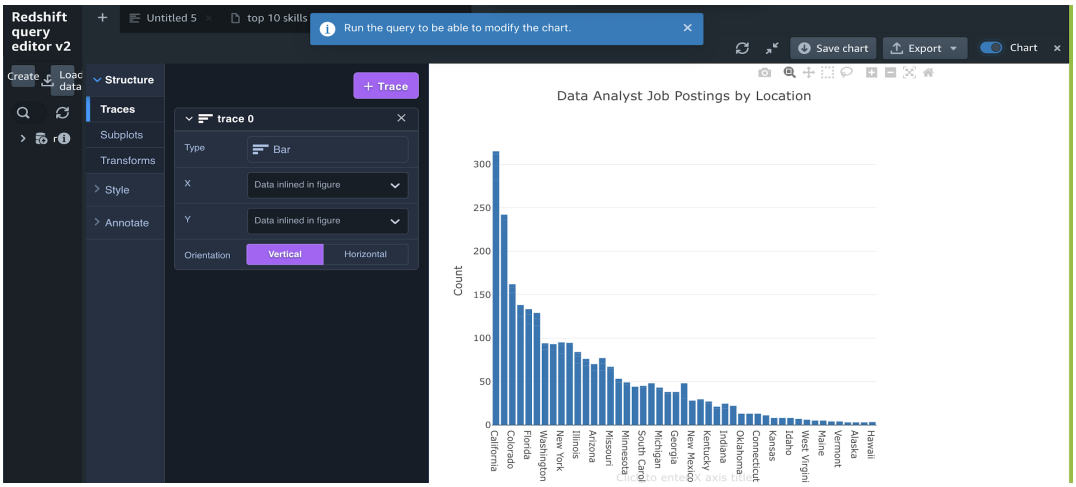
- Find out the state with highest job postings for each job title



there are more job postings in California for most of the software job titles in 'Computer and Mathematical Occupations' occupation

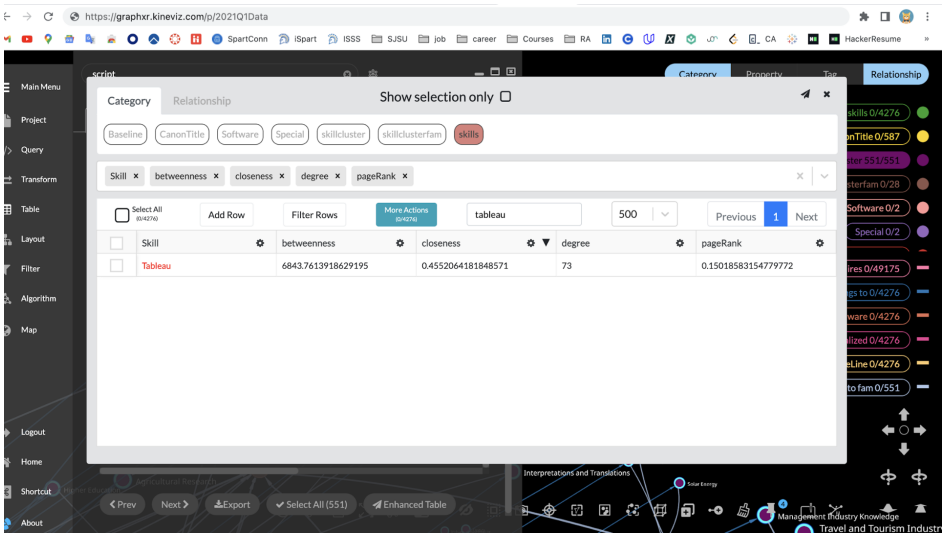
Some of the visual analysis using Amazon Redshift Query Editor Charts for above mentioned queries.



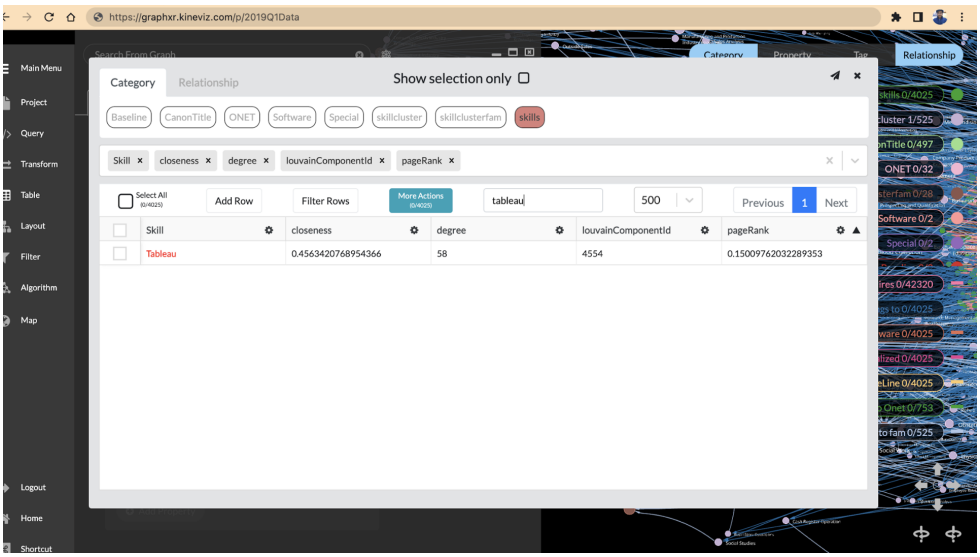


3. Using Neo4J+GraphXR

a. Shift in Skill Demand over time (Observing for Tableau)



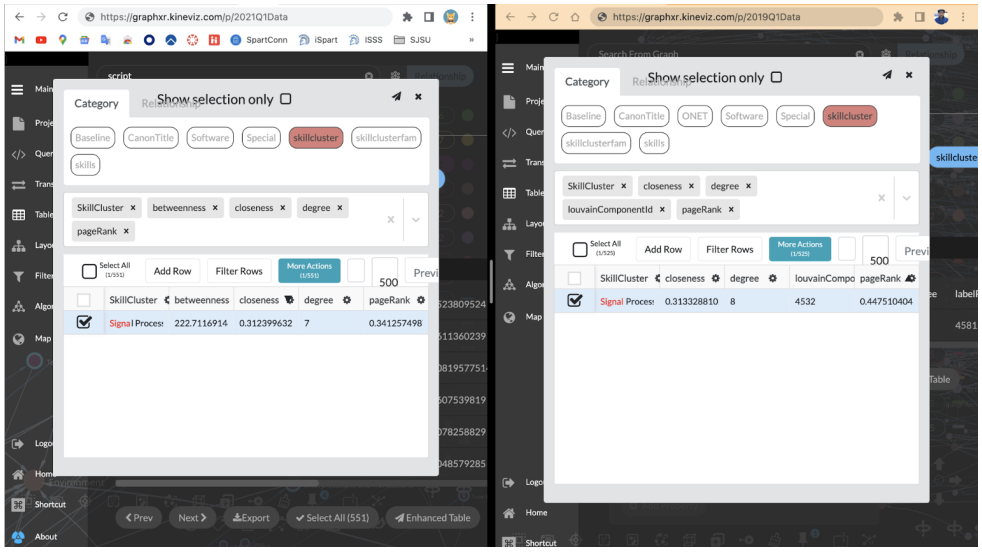
2021 Tableau Skill Demand



2019 Tableau Skill Demand

From the 2 results above we see that the skill demand for Tableau has grown over time. Quantitatively, the degree associated with Tableau in 2019 was **58,** and in 2021 it is **73.** This is very helpful to quantify the growth of demand for the skill Tableau.

b. Shift in demand for a skill cluster over time (Observing for Signal Processing)



2021 on LHS, 2019 on RHS

In the above result we can see that signal processing is a skill cluster that is declining over the years. The PageRank Score of Signal Processing was .44 in 2019, and it dropped to .34 in 2021.

VIII. TECHNICAL DIFFICULTY

Difficulties with data

- Initially we had accumulated a large amount of data, around 100 Million Job postings. We were unaware of the Free tier limitations and were planning to use this entire dataset in our analysis.
- After reducing the size of our dataset to fit AWS S3's free tier constraints, we then faced issues with the free tier limitations of Neo4J Aura DB. This was another overhead which obstructed us from deigning the final solution. We discuss the solution steps in detail in the Difficulties with Neo4J Section.
- Modeling the Data into Star Schema was difficult as there is no validation tool to check if the schema is correct or wrong.

Difficulties with AWS

- We faced errors while running AWS glue job to load data to Redshift, resolved by checking permissions, security groups and modified the inbound rules of one security group to open all ingress ports.
- Encountered errors due to Access Denied for S3 bucket exception while running glue job to load data to Redshift. We fixed the issue by creating a new IAM role for glue and attached it to the job after which the job ran successfully.
- While creating connections on glue from s3 to glue to redshift, we came across endpoint issues. Created a VPC endpoint to resolve the issues.

Difficulties with Neo4j + GraphXR

- Neo4j Aura DB comes with a constraint in free tier that is:
    - One account can have only one free aura db cluster
    - The free cluster cannot have more than 200k nodes and 400k edges.
- Our dataset across all 3 years had 1 million jobs, and 10 million skills hence we could not work with the free tier.

  Solution:
    - We filter out and consider only jobs belonging to occupation family = "computer and mathematicians"
    - We create 3 subsets of this dataset by considering datapoints of only first quarter of each year.
    - We create 3 separate accounts and create 3 databases with a predefined model file.

- Using GraphXR meant having extreme patience as it is prone to lagging, and page dying due to the scale of connections in our datasets.

IX. KEY LEARNINGS

1. We learned how to use PySpark to deal with datasets of size 100GB+ with just an 8 GB RAM Machine.
2. We understand how the Spark Backend works, and processes data in RDDs
3. We learned the concepts of Graph Databases so much so that we were able to apply it and extend the work of renowned Harvard and Yale Data Scientists/ Applied Economists [22].
4. We learned how to use AWS Solution Stack to build end-to-end Data Engineering/ Data Warehousing Solutions.
5. We were able to understand the query design in behind the scenes of job platforms that give us results on the front end.

X. CONCLUSION AND FUTURE WORK

We have successfully built a new information system design that can inform the shift in demand for skills, and skill Clusters overtime. Our solution is built using limited datapoints due to the free tier limitations. However, our solution can be extended with the entire length of the data set and it would still work flawlessly. We have also built queries that mimic the behind the scenes functionalities of the job sites to retrieve useful information. We used a Data Warehousing solution to build this query functionality.

We believe that our information system design can be extended with the employee reviews data to inform on the shift in employee perceptions over time by industry, sector, employer, location, etc… It is very important to understand employee perceptions to track the health of the economy and an industry over time. With funding/ access to unlimited use of Neo4j Aura DB, we can use our full dataset of 100 Million Job Postings to show the true shift in demand for skills over time. We also would like to propose the implementation of this solution to better inform the workforce on what skills to focus on to land better job opportunities. Right now, the only source of learning which tools to learn is from stack overflow developer survey. However, Job Postings are more accurate than the survey, and job portals must bring in a change and incorporate more informative solutions like the one we proposed in this project.

XII. REFERENCES

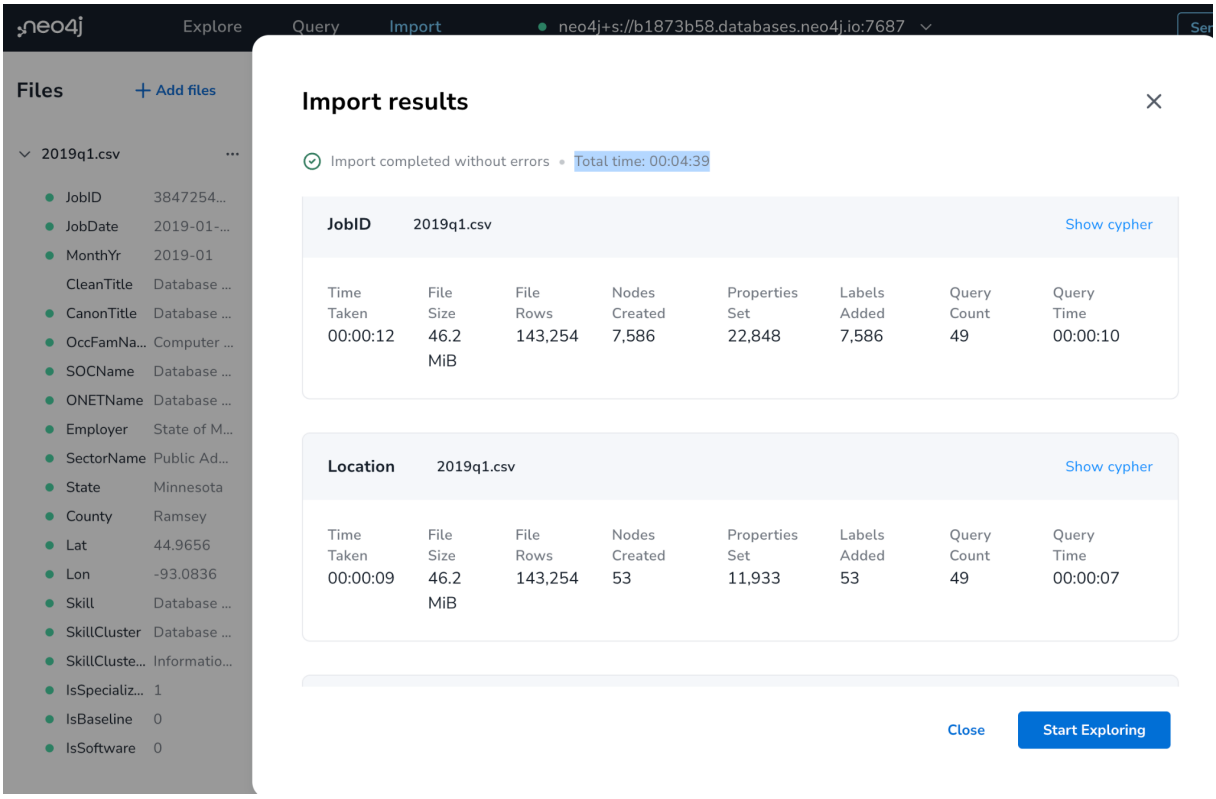
1. “LinkedIn Job Search” LinkedIn, https://www.linkedin.com/jobs/
2. Job Search | Indeed. https://www.indeed.com/
3. Glassdoor Job Search | Find the Job That Fits Your Life https://www.glassdoor.com/
4. O*NET OnLine, National Center for O*NET Development, www.onetonline.org/
5. U.S. Bureau of Labor Statistics, U.S. Bureau of Labor Statistics, https://www.bls.gov/
6. Sigelman, Matt. “Shifting Skills, Moving Targets, and Remaking the Workforce.” The Burning Glass Institute, The Burning Glass Institute, https://www.burningglassinstitute.org/research/shifting-skills-moving-targets-and-remaking-the-workforce
7. Howell, David R., and Susan S. Wieler. “Skill-Biased Demand Shifts and the Wage Collapse in the United States: A Critical Perspective.” Eastern Economic Journal, vol. 24, no. 3, 1998, pp. 343–66. JSTOR http://www.jstor.org/stable/40325878
8. Bughin, Jacques, et al. “Skill Shift: Automation and the Future of the Workforce. McKinsey & Company, 23 Jan. 2021, www.mckinsey.com/featured-insights/future-of-work/skill-shift-automation-and-the-future-of-the-workforce
9. Moueddene, Karim, et al. "Expected skills needs for the future of work." UK: Deloitte (2019): 1-20.
10. Benoit Dageville et. al 2016. The Snowflake Elastic Data Warehouse. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16). Association for Computing Machinery, New York, NY, USA, 215–226. https://doi.org/10.1145/2882903.2903741
11. Behm, Alexander. "Photon: A High-Performance Query Engine for the Lakehouse." CIDR. www. cidrdb. org. http://cidrdb. org/cidr2022/papers/a100-behm. pdf (2022)
12. Peyton, T., & Zigarmi, D. (2021). Employee perceptions of their work environment, work passion, and work intentions: A replication study using three samples. BRQ Business Research Quarterly, 0(0). https://doi.org/10.1177/23409444211002210
13. “LinkedIn's Graphql Journey for Integrations and Partnerships: How We Accelerated Development by 90%.” LinkedIn's GraphQL Journey for Integrations and Partnerships | LinkedIn Engineering, https://engineering.linkedin.com/blog/2022/linkedin-s-graphql-journey-for-integrations-and-partnerships.
14. “LinkedIn's Graphql Journey for Integrations and Partnerships: How We Accelerated Development by 90%.” LinkedIn's GraphQL Journey for Integrations and Partnerships | LinkedIn Engineering, https://engineering.linkedin.com/blog/2022/linkedin-s-graphql-journey-for-integrations-and-partnerships.
15. “From Daily Dashboards to Enterprise Grade Data Pipelines.” LinkedIn Engineering, https://engineering.linkedin.com/blog/2021/from-daily-dashboards-to-enterprise-grade-data-pipelines.
16. “Evolving Linkedin's Analytics Tech Stack.” LinkedIn Engineering, https://engineering.linkedin.com/blog/2021/evolving-linkedin-s-analytics-tech-stack.

17. Scully, Julie. “Serving over 1 Billion Documents per Day with Docstore V2.” Indeed Engineering Blog, 9 Feb. 2021, https://engineering.indeedblog.com/blog/2013/10/serving-over-1-billion-documents-per-day-with-docstore-v2/.
18. Engineering, Indeed. “Open Source Interactive Data Analytics with Imhotep.” Indeed Engineering Blog, 9 Feb. 2021, https://engineering.indeedblog.com/blog/2014/10/open-source-interactive-data-analytics-with-imhotep/.
19. He, Rolland. “Building a Nearline System for Scale and Performance: Part II.” Medium, Glassdoor Engineering Blog, 22 Feb. 2021, https://medium.com/glassdoor-engineering/building-a-nearline-system-for-scale-and-performance-part-ii-9e01bf51b23d.
20. Pandis, Ippokratis. "The evolution of Amazon redshift." Proceedings of the VLDB Endowment 14.12 (2021): 3162-3174.
21. Sen, S., Mehta, A., Ganguli, R. et al. Recommendation of Influenced Products Using Association Rule Mining: Neo4j as a Case Study. SN COMPUT. SCI. 2, 74 (2021). https://doi.org/10.1007/s42979-021-00460-8
22. Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals David Deming and Lisa B. Kahn, Journal of Labor Economics 2018 36:S1, S337-S369
23. A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso, Skills2job: A recommender system that encodes job offer embeddings on graph databases, Applied Soft Computing 101 (2021) 107049. doi:https://doi.org/10.1016/j.asoc.2020.107049.
24. A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, Graphlmi: A data driven system for exploring labor market information through graph databases, Multimedia Tools and Applications 81 (3) (2020) 3061–3090. doi:10.1007/s11042-020-09115-x.
25. A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso, Skills2graph: Processing million job ads to face the job skill mismatch problem, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4984–4987, demo Track. doi:10.24963/ijcai.2021/708. URL https://doi.org/10.24963/ijcai.2021/708

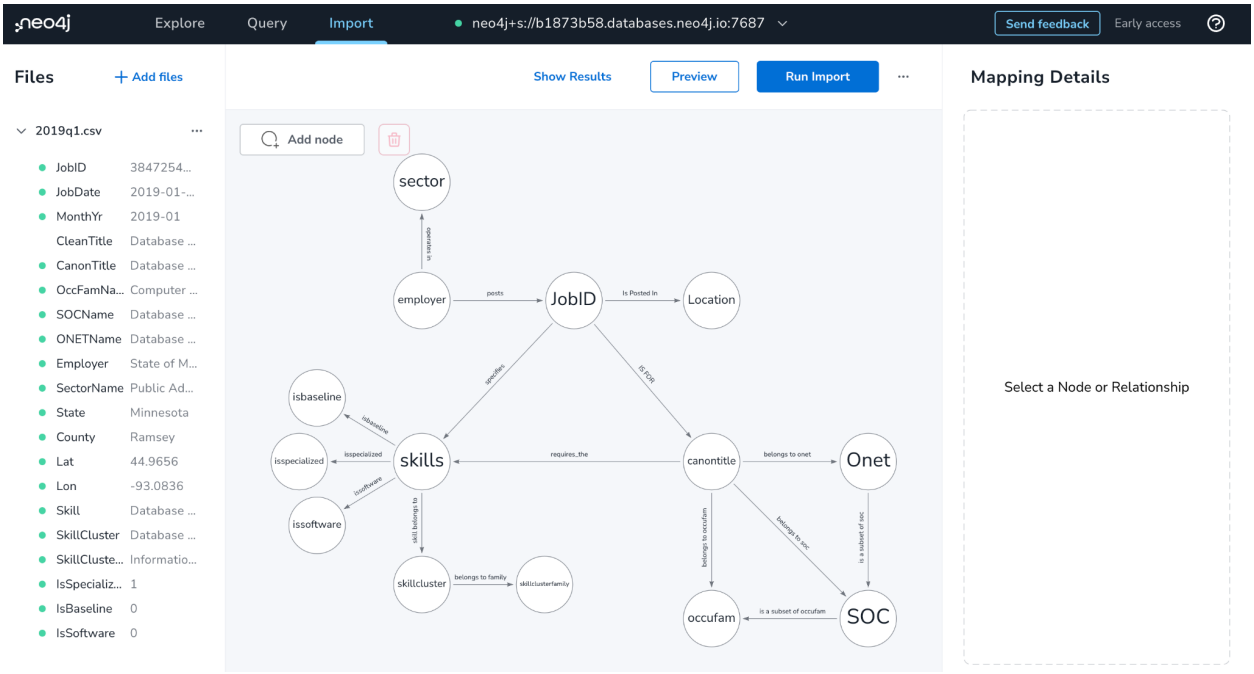
XIII. APPENDIX

1. Screenshots

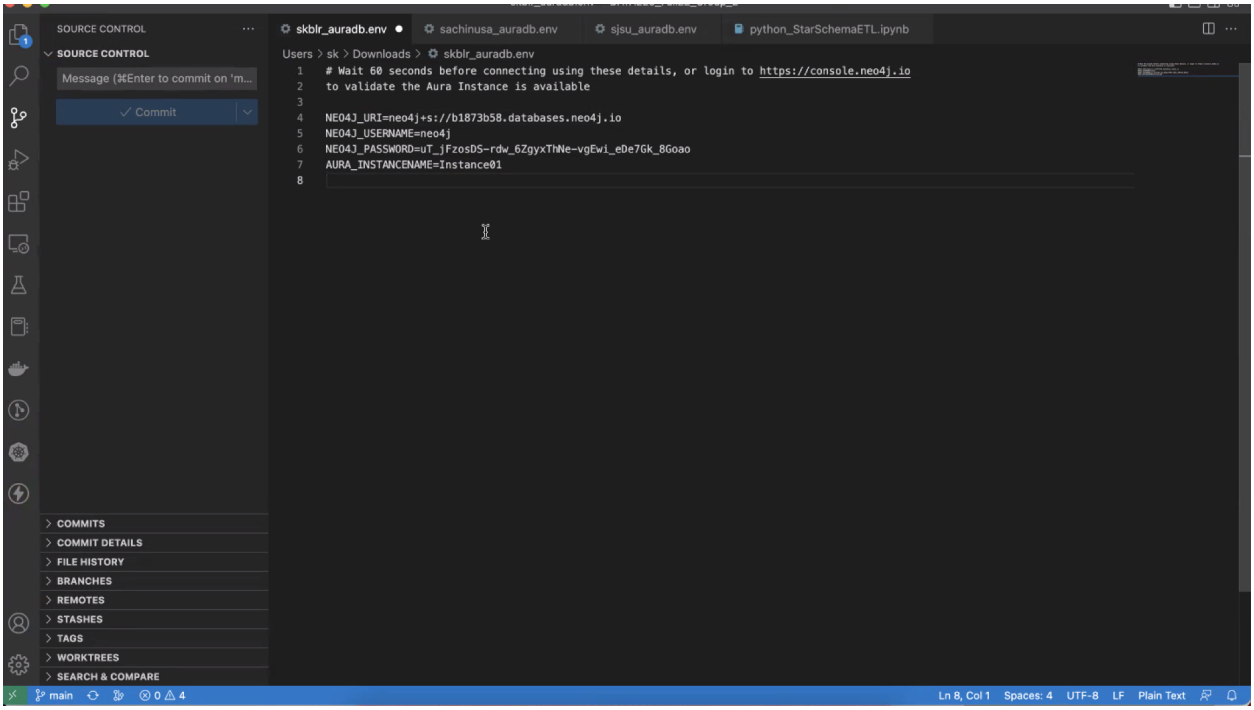
Importing Data into Neo4J Graph DB



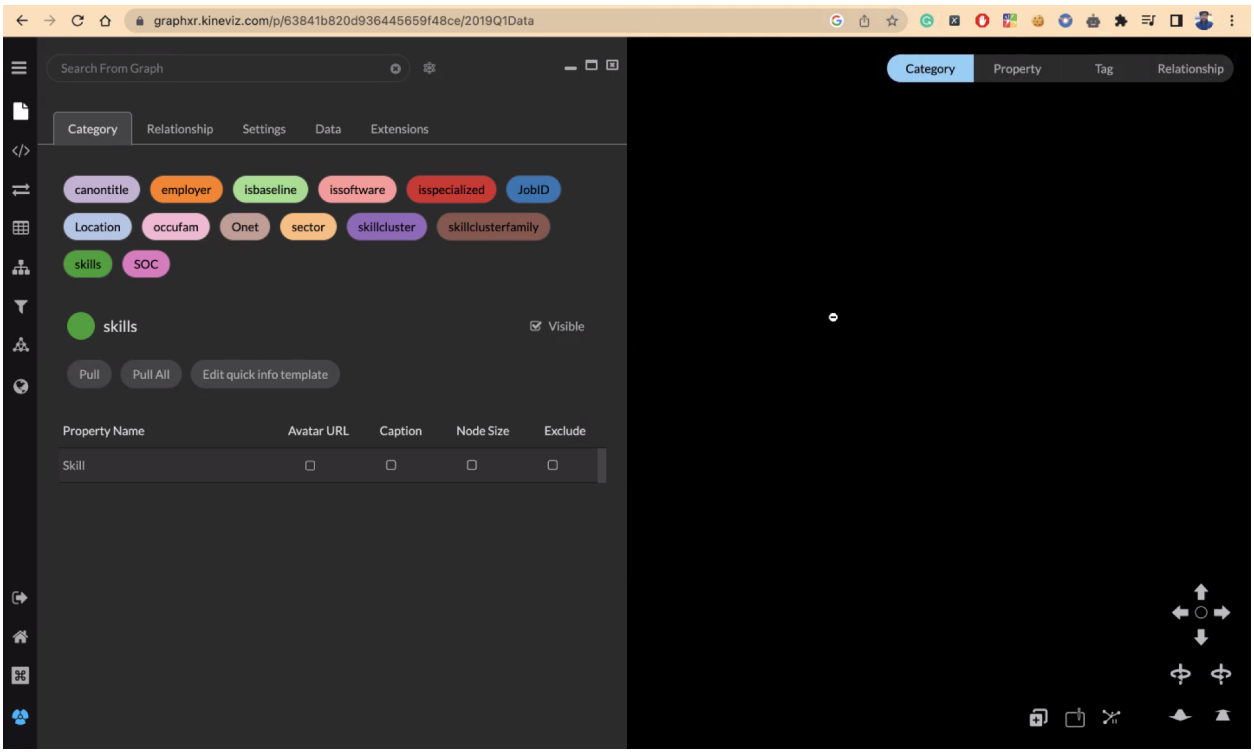
Modeling Data in Neo4J



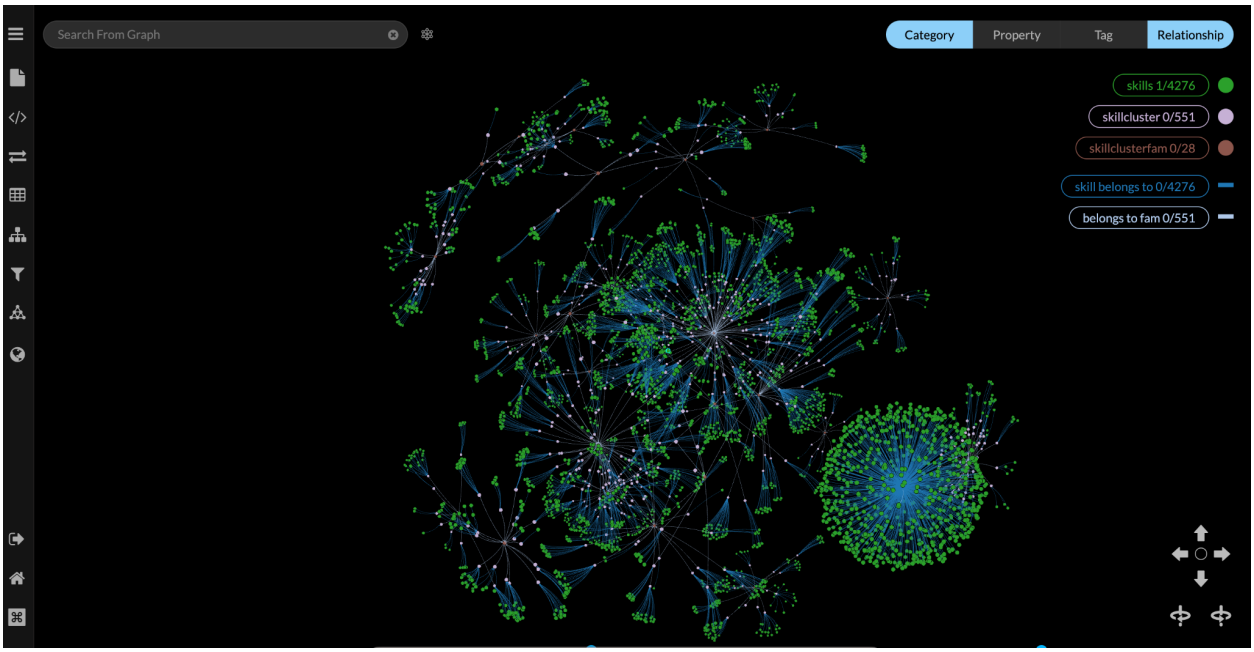
API Connection String to Neo4J Data



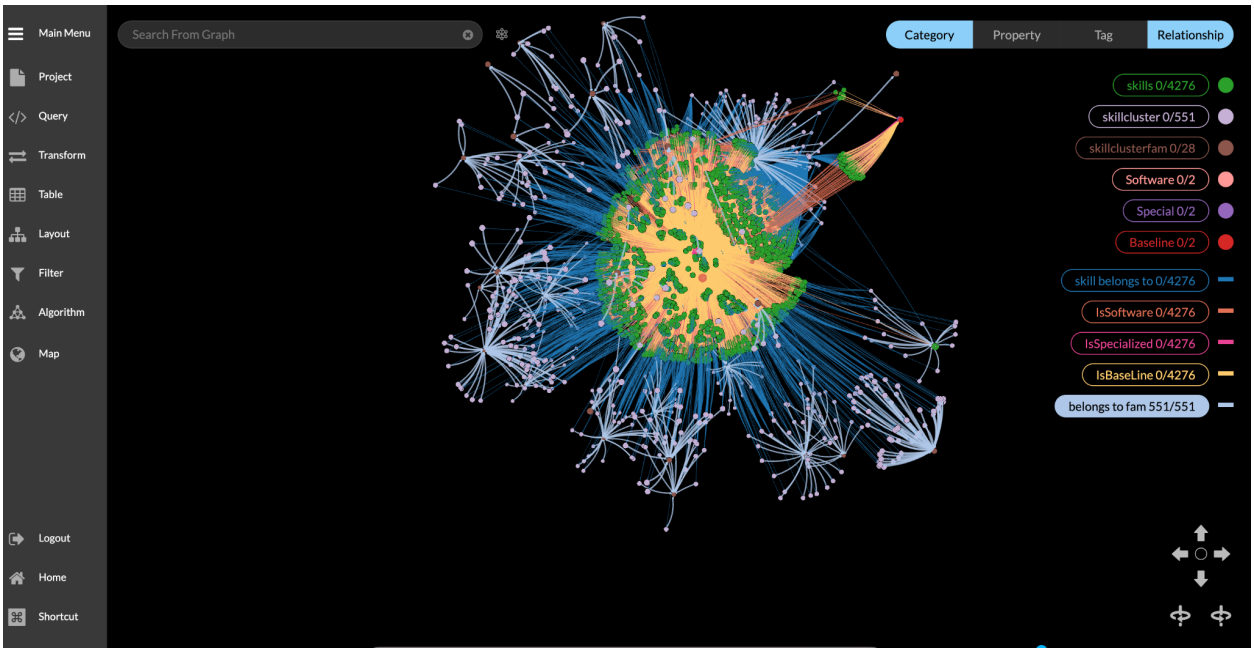
GraphXR Interface



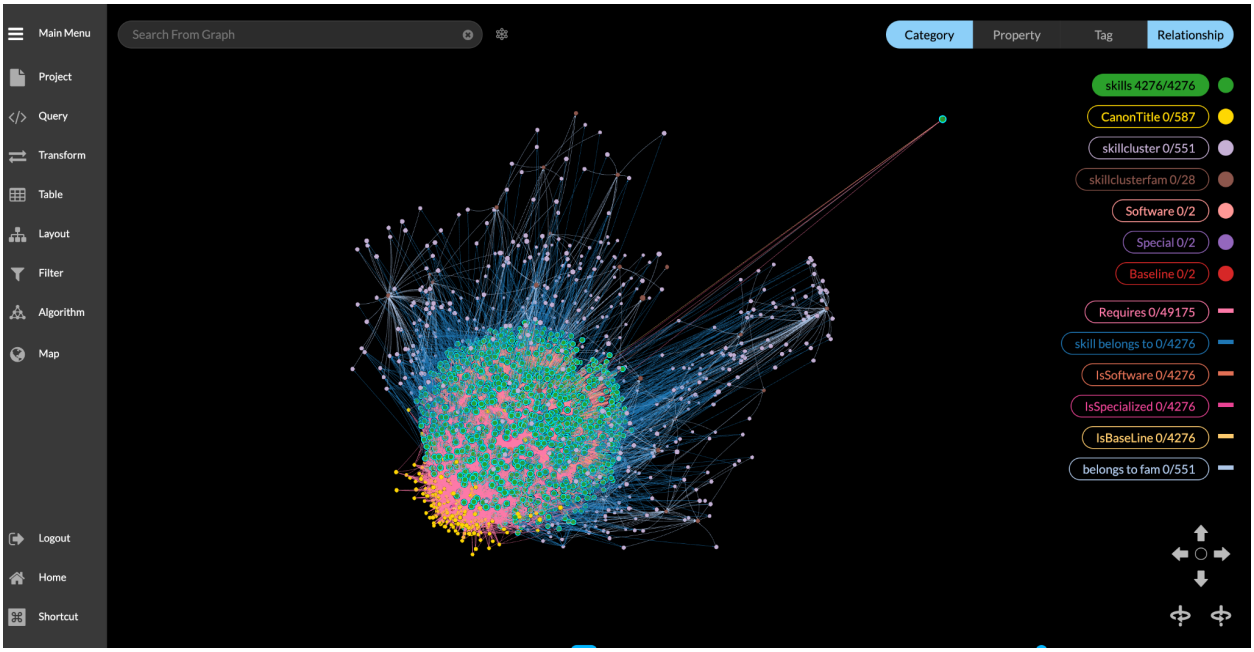
GraphXR Plot Skills and Skill Clusters



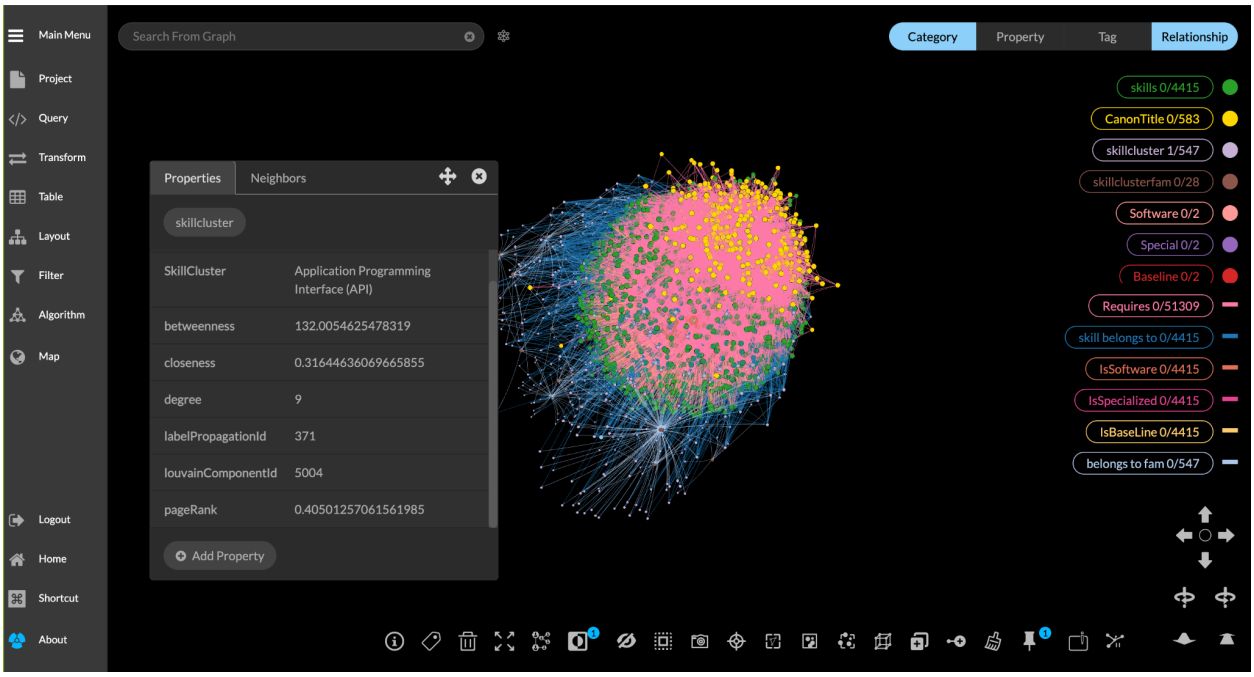
GraphXR Plot Skills, and all its associated features



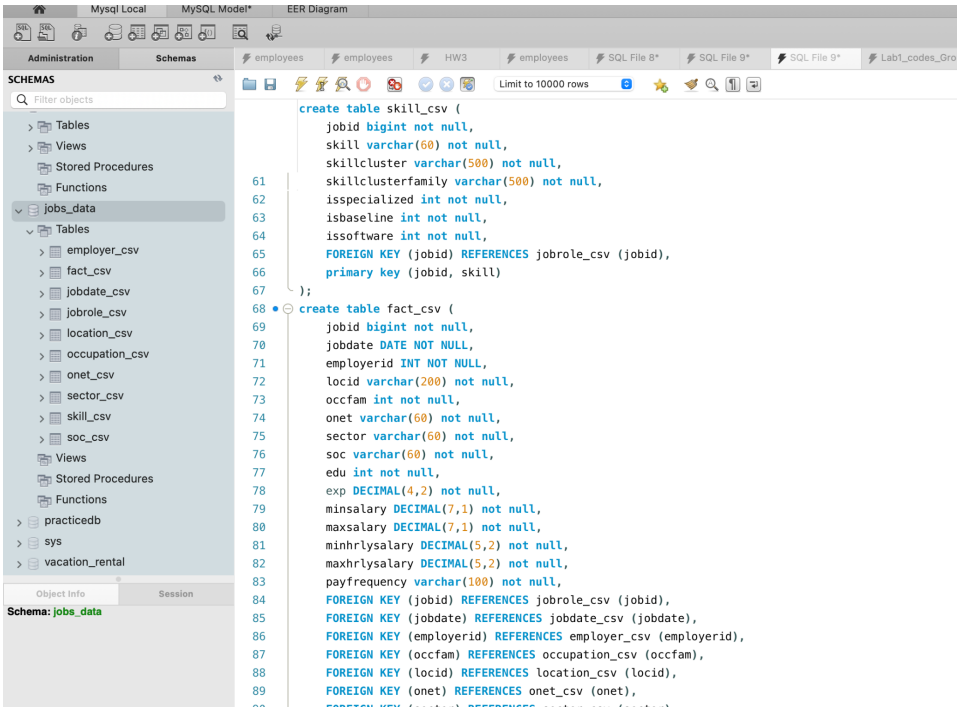
GraphXR Plot Skills, and all its associated features, and Job Title.



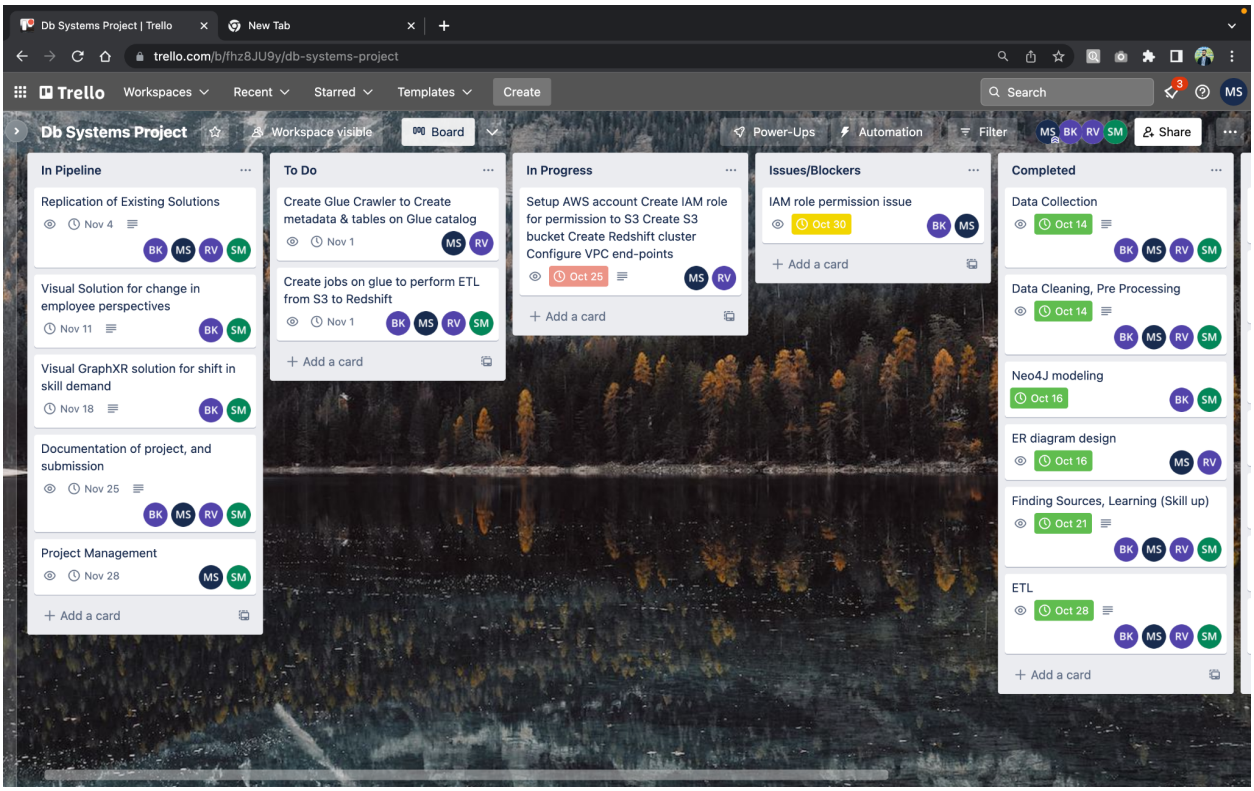
GraphXR Plot Skills, and all its associated features, and Job Title. Showing Centrality Scores for a random Skill Cluster.

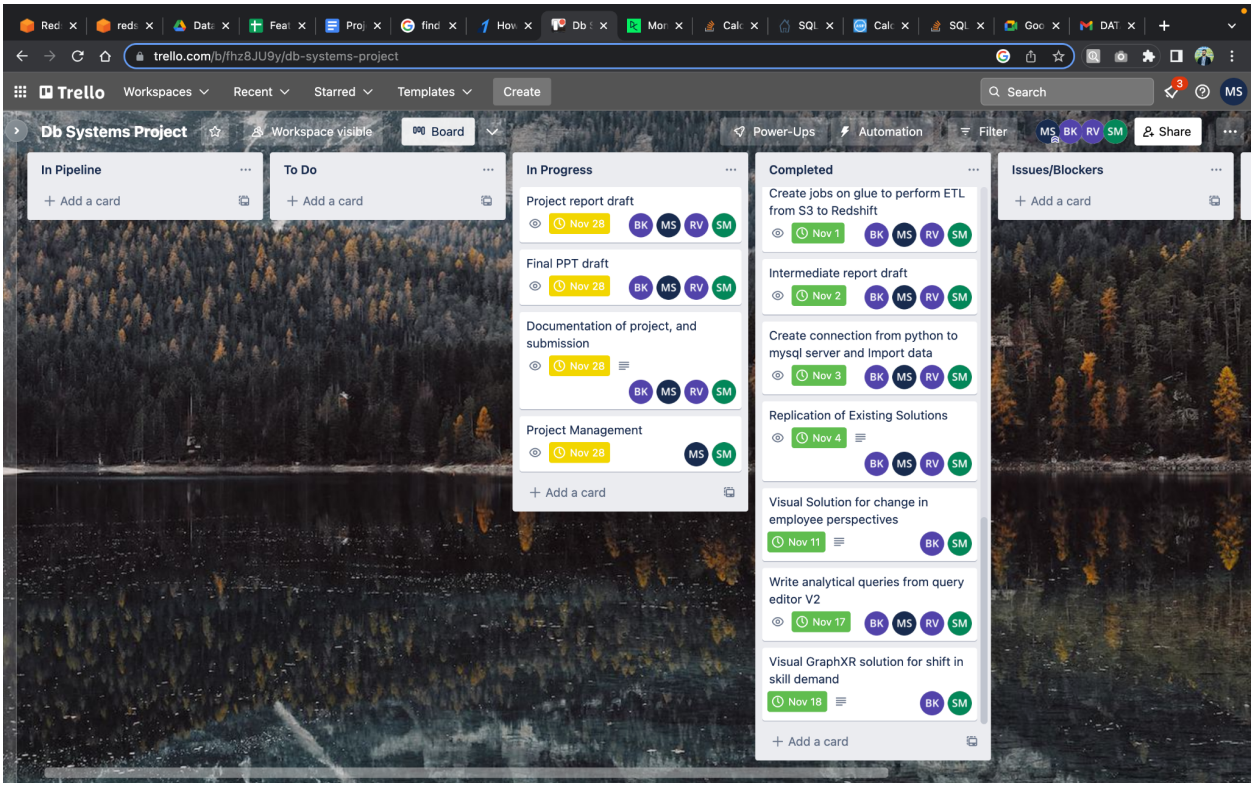


Creating tables in MySQL workbench



Used Trello to manage project task tracking


Db Systems Project | Trello
New Tab
trello.com/b/fhz8JU9y/db-systems-project
Trello
Workspaces
Recent
Starred
Templates
Create
Search
Db Systems Project
Workspace visible
Board
Power-Ups
Automation
Filter
Share
In Pipeline
Replication of Existing Solutions
Nov 4
Visual Solution for change in employee perspectives
Nov 11
Visual GraphXR solution for shift in skill demand
Nov 18
Documentation of project, and submission
Nov 25
Project Management
Nov 28
Add a card
To Do
Create Glue Crawler to Create metadata & tables on Glue catalog
Nov 1
Create jobs on glue to perform ETL from S3 to Redshift
Nov 1
Add a card
In Progress
Setup AWS account Create IAM role for permission to S3 Create S3 bucket Create Redshift cluster Configure VPC end-points
Oct 25
Add a card
Issues/Blockers
IAM role permission issue
Oct 30
Add a card
Completed
Data Collection
Oct 14
Data Cleaning, Pre Processing
Oct 14
Neo4J modeling
Oct 16
ER diagram design
Oct 16
Finding Sources, Learning (Skill up)
Oct 21
ETL
Oct 28
Add a card


trello.com/b/fhz8JU9y/db-systems-project
Trello
Workspaces
Recent
Starred
Templates
Create
Search
Db Systems Project
Workspace visible
Board
Power-Ups
Automation
Filter
Share
In Pipeline
Add a card
To Do
Add a card
In Progress
Project report draft
Nov 28
Final PPT draft
Nov 28
Documentation of project, and submission
Nov 28
Project Management
Nov 28
Add a card
Completed
Create jobs on glue to perform ETL from S3 to Redshift
Nov 1
Intermediate report draft
Nov 2
Create connection from python to mysql server and Import data
Nov 3
Replication of Existing Solutions
Nov 4
Visual Solution for change in employee perspectives
Nov 11
Write analytical queries from query editor V2
Nov 17
Visual GraphXR solution for shift in skill demand
Nov 18
Add a card
Issues/Blockers
Add a card