

Analyzing Jobs Data for understanding the Future of Work

DATA225 Group 2 Project

Bhavana Prasad Kote (016044899)
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

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

Motivation

- There is no information on the dynamic shifts in the job markets w.r.t:
 - Skill demand
 - Employee perception
- Prior literature on the future of work [Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals \(harvard.edu\)](#) calls for a new information system design that allows to study granular changes in the skill requirements with a better understanding of how heterogeneity in skill demands translates to firm production technology.



Technology Stack



Database- Neo4j Aura DB



Data Warehouse- Redshift



ETL Tools- PySpark, Glue



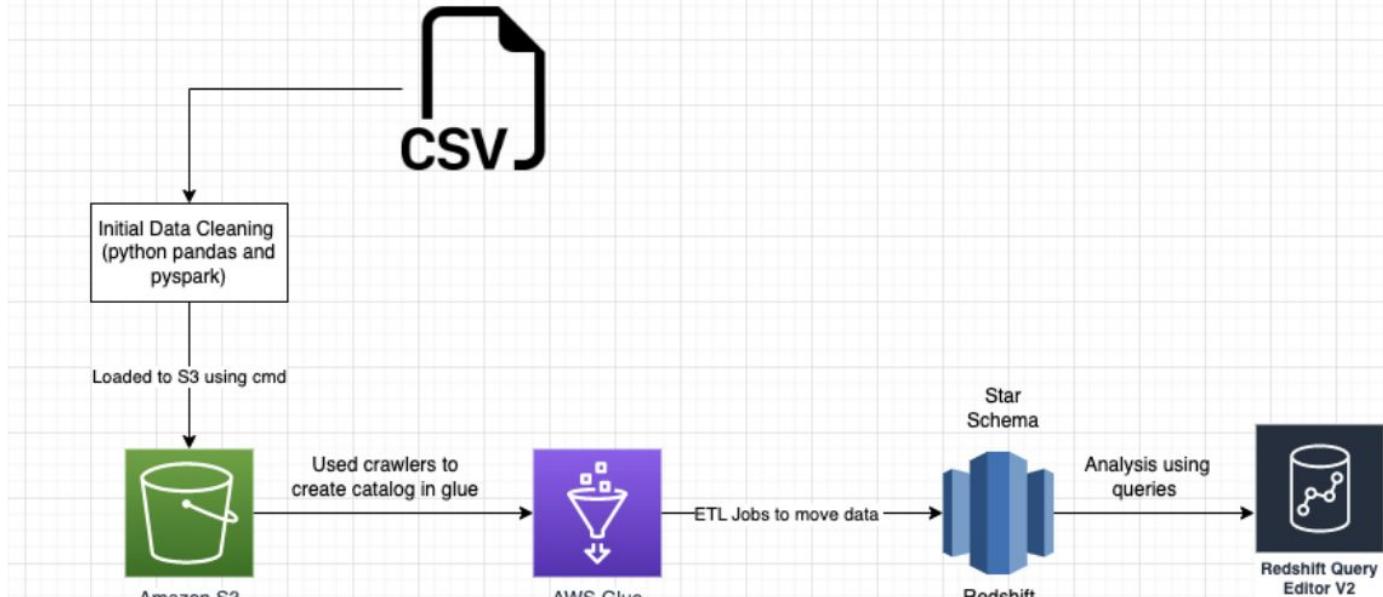
Storage- AWS S3



Programming- Python, SQL

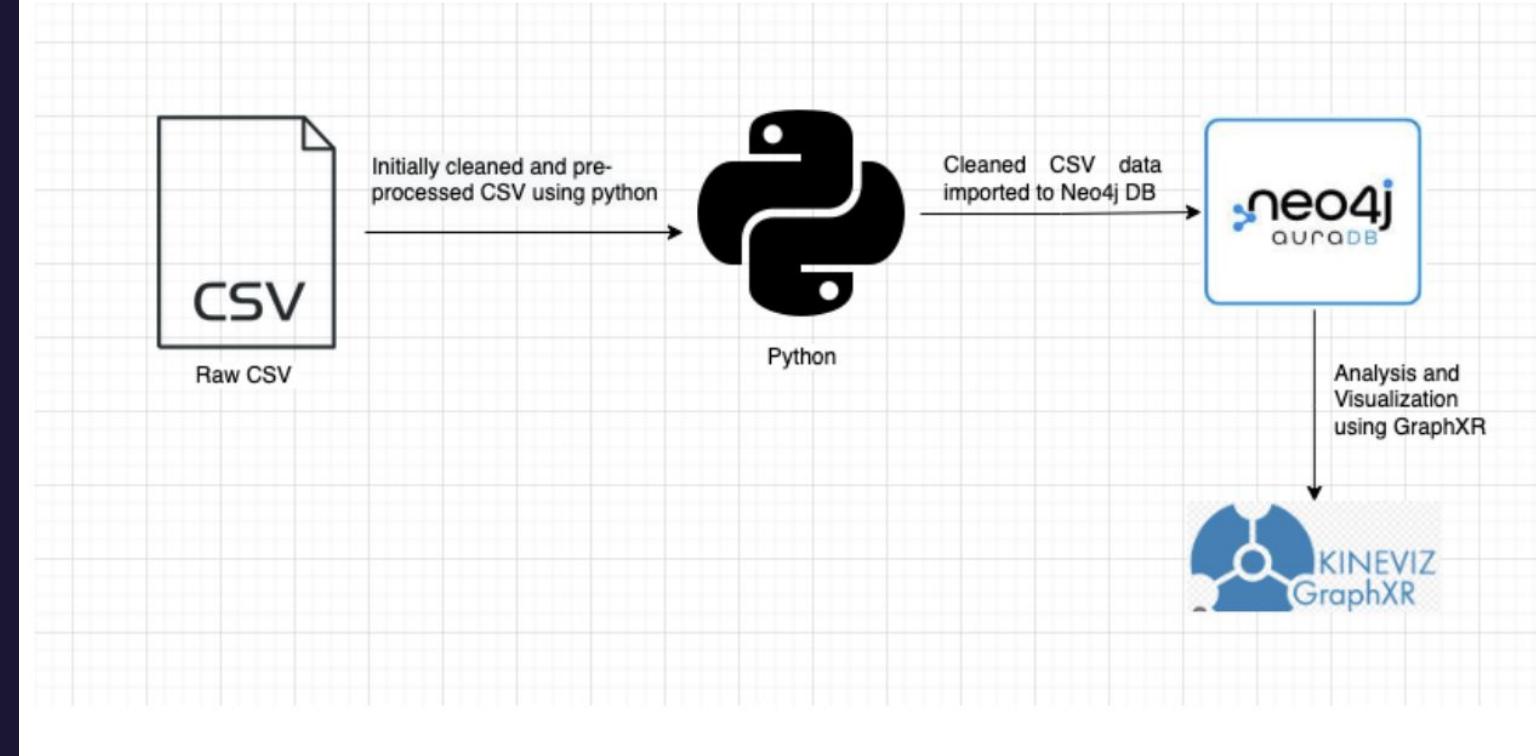
Workflow1

Workflow 1:



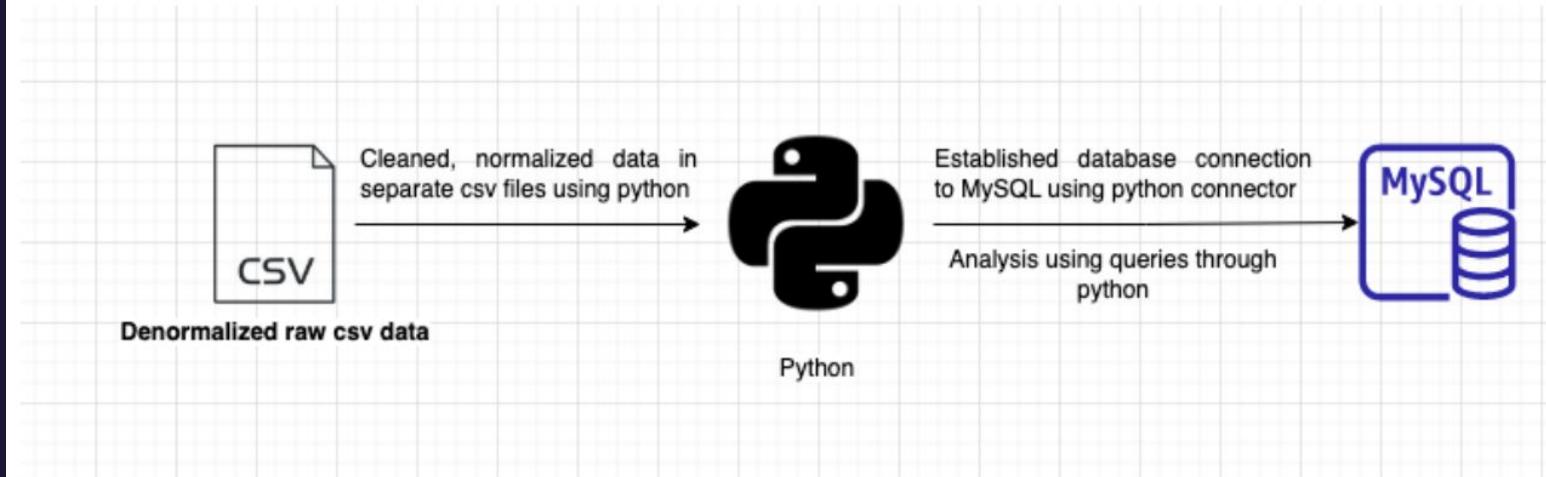
Workflow2

Workflow 2:



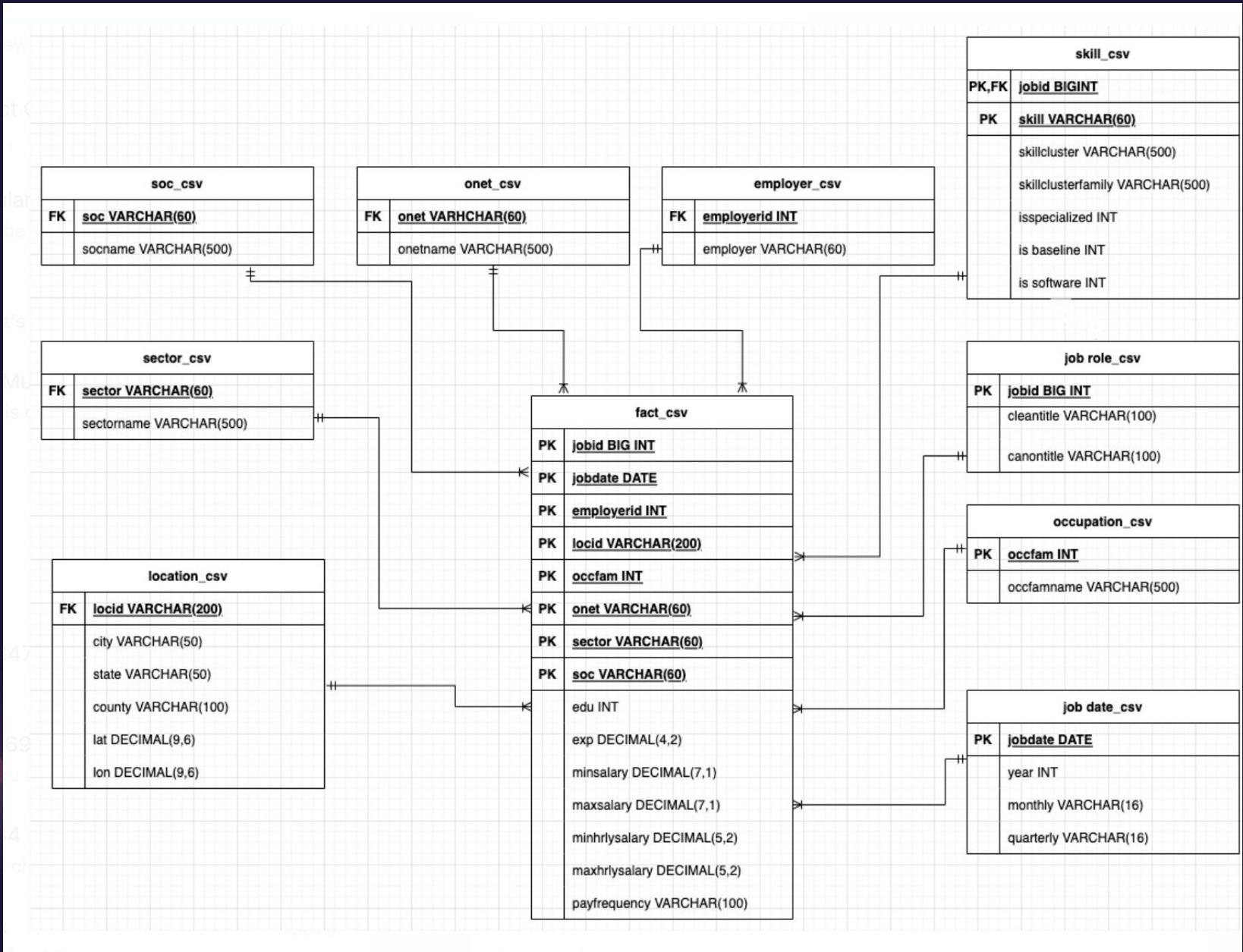
Workflow3

Workflow 3:



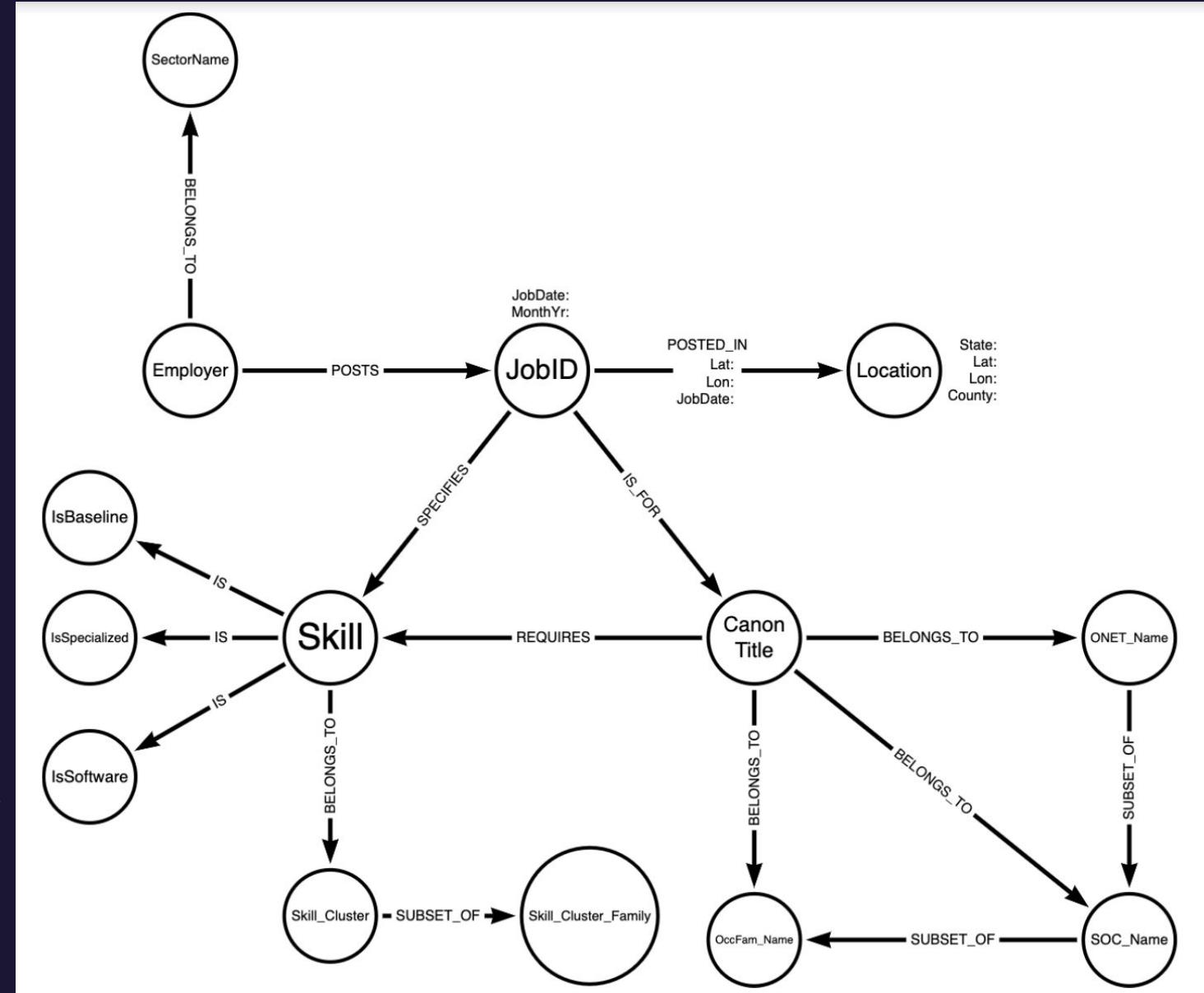
DATA MODEL-1

STAR SCHEMA



DATA MODEL-2

NEO4J DATA MODEL

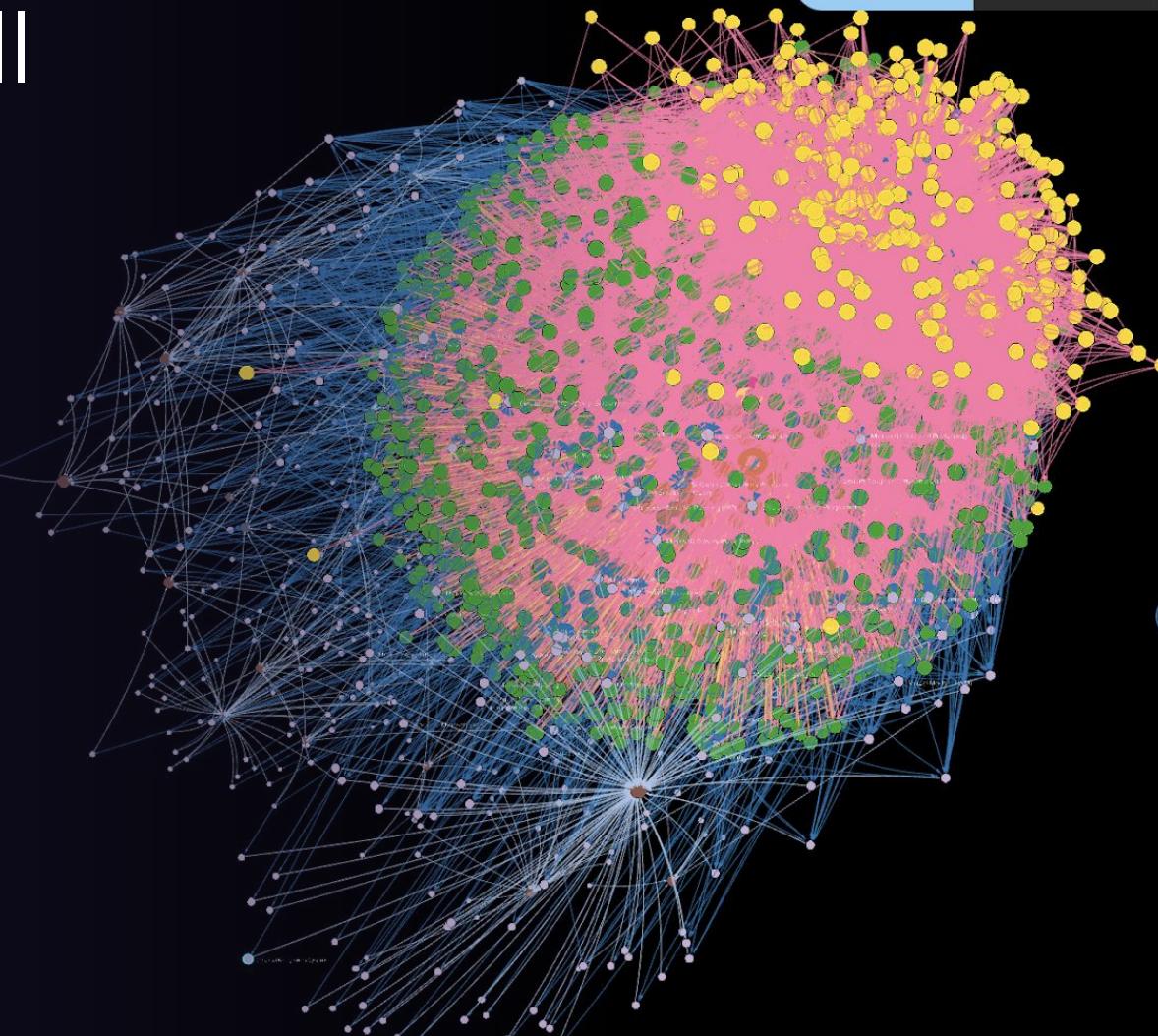


RESULTS FROM NEO4J

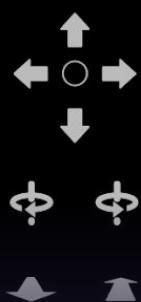


Results- Shift in skill demand over time

Properties	Neighbors
skillcluster	
Property Name	Property Value
SkillCluster	Content Management Systems
betweenness	14.358209412223879
closeness	0.31384684633995386
degree	2
labelPropagationId	371
louvainComponentId	5552
pageRank	0.18187509043540173
+ Add Property	



- skills 0/4415
- CanonTitle 0/583
- skillcluster 1/547
- skillclusterfam 0/28
- Software 0/2
- Special 0/2
- Baseline 0/2
- Requires 0/51309
- skill belongs to 0/4415
- IsSoftware 0/4415
- IsSpecialized 0/4415
- IsBaseLine 0/4415
- belongs to fam 0/547



Main Menu

Project

Query

Transform

Table

Layout

Filter

Algorithm

Map

Logout

Home

Shortcut

About

Results- Shift in skill demand over time

Properties	Neighbors
skillcluster	
Property Name	Property Value
SkillCluster	Instruction
betweenness	70.63576137864519
closeness	0.31202014618266694
degree	2
labelPropagationId	4581
louvainComponentId	4828
pageRank	0.19250045509253172
+ Add Property	



Results- Shift in skill demand over time

The image displays two side-by-side screenshots of a web-based graph visualization tool, likely GraphXR, showing the results of a skill demand analysis across different quarters.

Left Screenshot (2021Q1Data):

- Category Selection:** Shows a "Show selection only" checkbox and a list of categories: Baseline, CanonTitle, Software, Special, skillcluster (selected), skillclusterfam, and skills.
- Relationship Metrics:** A list of selected relationship metrics: SkillCluster (selected), betweenness, closeness, degree, and pageRank.
- Table View:** A table with columns: SkillCluster, betweenness, closeness, degree, and pageRank. One row is selected, showing the following values:

SkillCluster	betweenness	closeness	degree	pageRank
Signal Process	222.7116914	0.312399632	7	0.341257498

Right Screenshot (2019Q1Data):

- Category Selection:** Shows a "Show selection only" checkbox and a list of categories: Baseline, CanonTitle, ONET, Software, Special, skillcluster (selected), skillclusterfam, and skills.
- Relationship Metrics:** A list of selected relationship metrics: SkillCluster (selected), closeness, degree, louvainComponentId, and pageRank.
- Table View:** A table with columns: SkillCluster, closeness, degree, louvainCompo, and pageRank. One row is selected, showing the following values:

SkillCluster	closeness	degree	louvainCompo	pageRank
Signal Process	0.313328810	8	4532	0.447510404

Main Menu

Project

Query

Transform

Table

Layout

Filter

Algorithm

Map

Logout

Home

Shortcut

Search From Graph

Category

Relationship

Show selection only

Baseline

CanonTitle

ONET

Software

Special

skillcluster

skillclusterfam

skills

Skill closeness degree louvainComponentId pageRank Select All
(0/4025)

Add Row

Filter Rows

More Actions
(0/4025)

tableau

500

Previous 1 Next

<input type="checkbox"/> Skill	<input type="checkbox"/> closeness	<input type="checkbox"/> degree	<input type="checkbox"/> louvainComponentId	<input type="checkbox"/> pageRank
Tableau	0.4563420768954366	58	4554	0.15009762032289353

2019 demand for tableau

Relationship

skills 0/4025

cluster 1/525

onTitle 0/497

ONET 0/32

sterfam 0/28

Software 0/2

Special 0/2

ires 0/42320

ugs to 0/4025

ware 0/4025

ilized 0/4025

eLine 0/4025

Onet 0/753

to fam 0/525

Social Work

Physical

Employed Relat

Institutio

Analyst

Search From Graph Category Property Tag

Category Relationship Show selection only

Baseline CanonTitle Software Special skillcluster skillclusterfam skills

Skill x betweenness x closeness x degree x pageRank x

Select All (0/4415) Add Row Filter Rows More Actions (0/4415) tableau 500 | Previous 1 Next

<input type="checkbox"/>	Skill	<input type="checkbox"/>	betweenness	<input type="checkbox"/>	closeness	<input type="checkbox"/>	degree	<input type="checkbox"/>	pageRank	<input type="checkbox"/>	<input type="button" value="▲"/>
<input type="checkbox"/>	Tableau	<input type="checkbox"/>	5520.036408542301	<input type="checkbox"/>	0.456913499344692	<input type="checkbox"/>	75	<input type="checkbox"/>	0.15012557224042453	<input type="checkbox"/>	<input type="checkbox"/>

2020 demand for tableau

2020 demand for tableau



Main Menu

Project

Query

Transform

Table

Layout

Filter

Algorithm

Map

2021 Demand for tableau

Agricultural Research

Prev

Next

Export

Select All (551)

Enhanced Table

Interpretations and Translations



Category

Relationship

Category

Property

Tag

Relationship

Show selection only

Baseline

CanonTitle

Software

Special

skillcluster

skillclusterfam

skills

Skill

betweenness

closeness

degree

pageRank

Select All
(0/4276)

Add Row

Filter Rows

More Actions
(0/4276)

tableau

500

Previous 1 Next



Skill



betweenness



closeness



degree



pageRank



Tableau

6843.7613918629195

0.4552064181848571

73

0.15018583154779772

skills 0/4276



CanonTitle 0/587



cluster 551/551



clusterfam 0/28



Software 0/2



Special 0/2



ires 0/49175



ngs to 0/4276



ware 0/4276



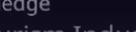
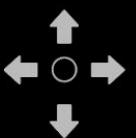
lized 0/4276



eLine 0/4276



to fam 0/551



Sample Results from Redshift Queries



Salaries by firm, and role across states



```
Run Limit 100 Explain Isolated session redshift-cluster-1 jobs_data_225
4
5 -- Pay Range Estimate for a Job by a firm in multiple locations. Top 10
6 -- Ex: SE at Amazon in (Bay, Seattle, Georgia,...) in separate rows
7 select count(*) RecordsPerGroup, r.canontitle, e.employer, l.state, avg(f.minsalary) as avg_minsal,
8 avg(f.maxsalary) as avg_maxsal
9 from fact_csv f inner join jobrole_csv r ON f.jobid = r.jobid
10 inner join employer_csv e on f.employerid = e.employerid
11 inner join location_csv l on f.locid = l.locid
12 where e.employer='Google Inc.' and r.canontitle= 'Software Development Engineer'
13 group by r.canontitle, e.employer, l.state
14 order by avg_maxsal desc;
15
```

Result 1 (6)

recordspergroup	canontitle	employer	state	avg_minsal	avg_maxsal
13	Software Development E...	Google Inc.	California	122906.53846153847	170153.84615384
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

Elapsed time: 79 ms Total rows: 6

Pay Range Estimate for a role

```
23 --- Pay Range Estimate for a Role.
24 select count(*) RecordsPerGroup, r.canontitle, avg(f.minsalary) as avg_minsal, avg(f.maxsalary) as avg_maxsal
25 from fact_csv f inner join jobrole_csv r ON f.jobid = r.jobid
26 where r.canontitle= 'Software Development Engineer'
27 group by r.canontitle;
28
29
```

Result 1 (1)

Export

Chart

recordspergroup	canontitle	avg_minsal	avg_maxsal
4395	Software Development Engineer	95680.09831626849	126959.43089874857

Elapsed time: 14 ms Total rows: 1

Experience and Education influencing payscale

```
30 -- Experience and Education Required for a role (Like) for a given pay range.
31 select count(*) RecordsPerGroup, r.canontitle, f.edu, max(f.maxhrlysalary) as max_maxsal,
32 CASE WHEN f.exp<=4 THEN '0-4'
33 WHEN f.exp>4 AND f.exp<=8 THEN '4-8'
34 WHEN f.exp>8 AND f.exp<=12 THEN '8-12'
35 WHEN f.exp>12 THEN '12+'
36 END AS expCat
37 from fact_csv f inner join jobrole_csv r ON f.jobid = r.jobid
38 where r.canontitle='Information Technology Specialist'
39 group by r.canontitle, f.edu, expCat
40 order by edu, expcat, max_maxsal;
41
```

Result 1 (19)

Export Chart ×

recordspergroup	canontitle	edu	max_maxsal	expcat
3	Information Technology S...	16	81.25	12+
156	Information Technology S...	16	82.93	4-8
35	Information Technology S...	16	83	8-12
351	Information Technology S...	18	78.94	0-4
2	Information Technology S...	18	75.87	4-8
106	Information Technology S...	21	80.05	0-4

Elapsed time: 20 ms Total rows: 19

Career Trajectory

```
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

Top 10 required skills for Data Analyst

```
1 ---top 10 most asked skills for a given job title(Data Analyst)
2 select top 10 sk.skill, r.canontitle, count(sk.skill) as cnt
3 from fact_csv f inner join skill_csv sk
4 on f.jobid = sk.jobid
5 inner join jobrole_csv r
6 on f.jobid = r.jobid
7 inner join employer_csv e
8 on e.employerid = f.employerid
9 where r.canontitle in ('Data Analyst')
10 group by r.canontitle, sk.skill
11 order by cnt desc;
```

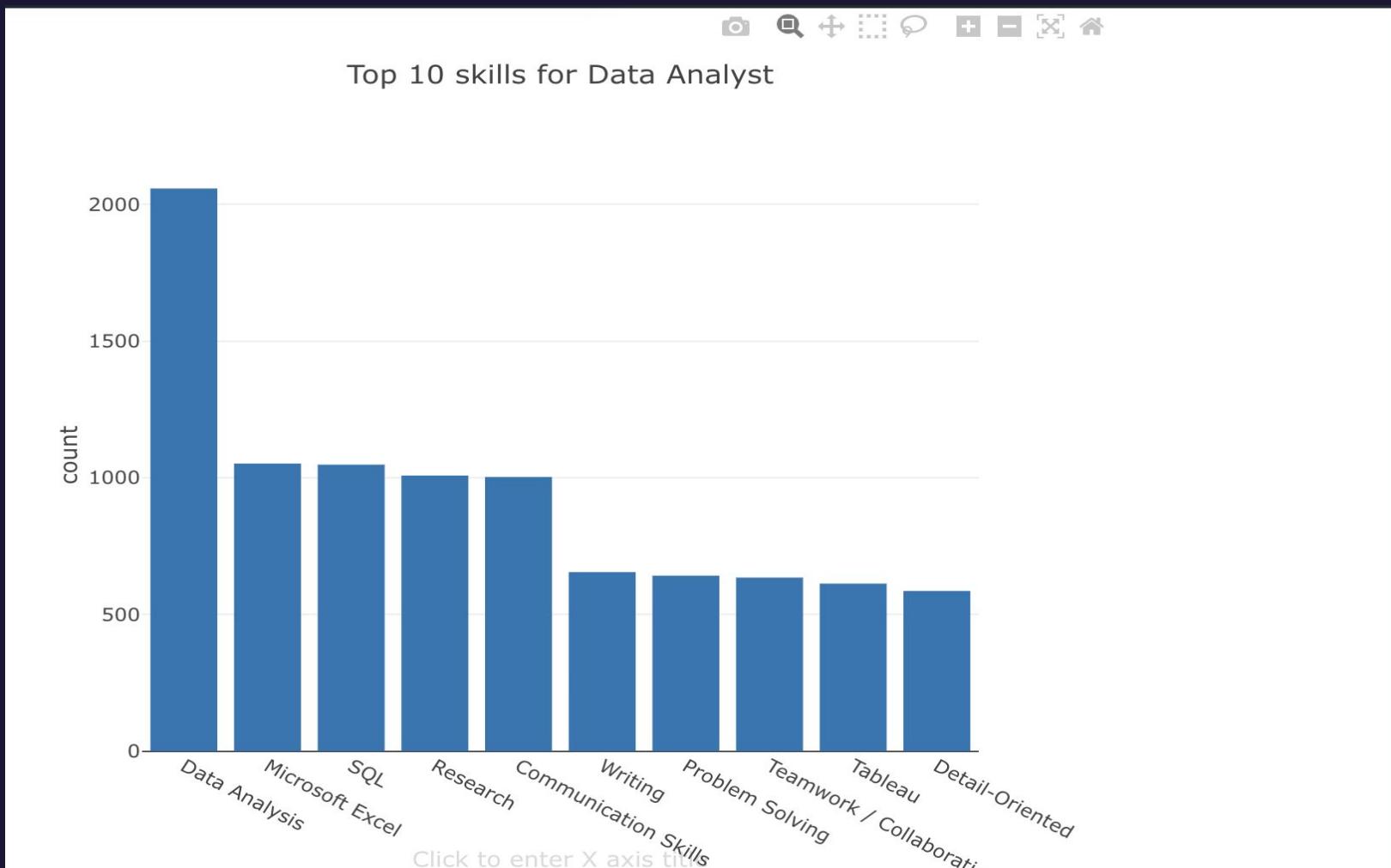
Result 1 (10)

Export Chart X

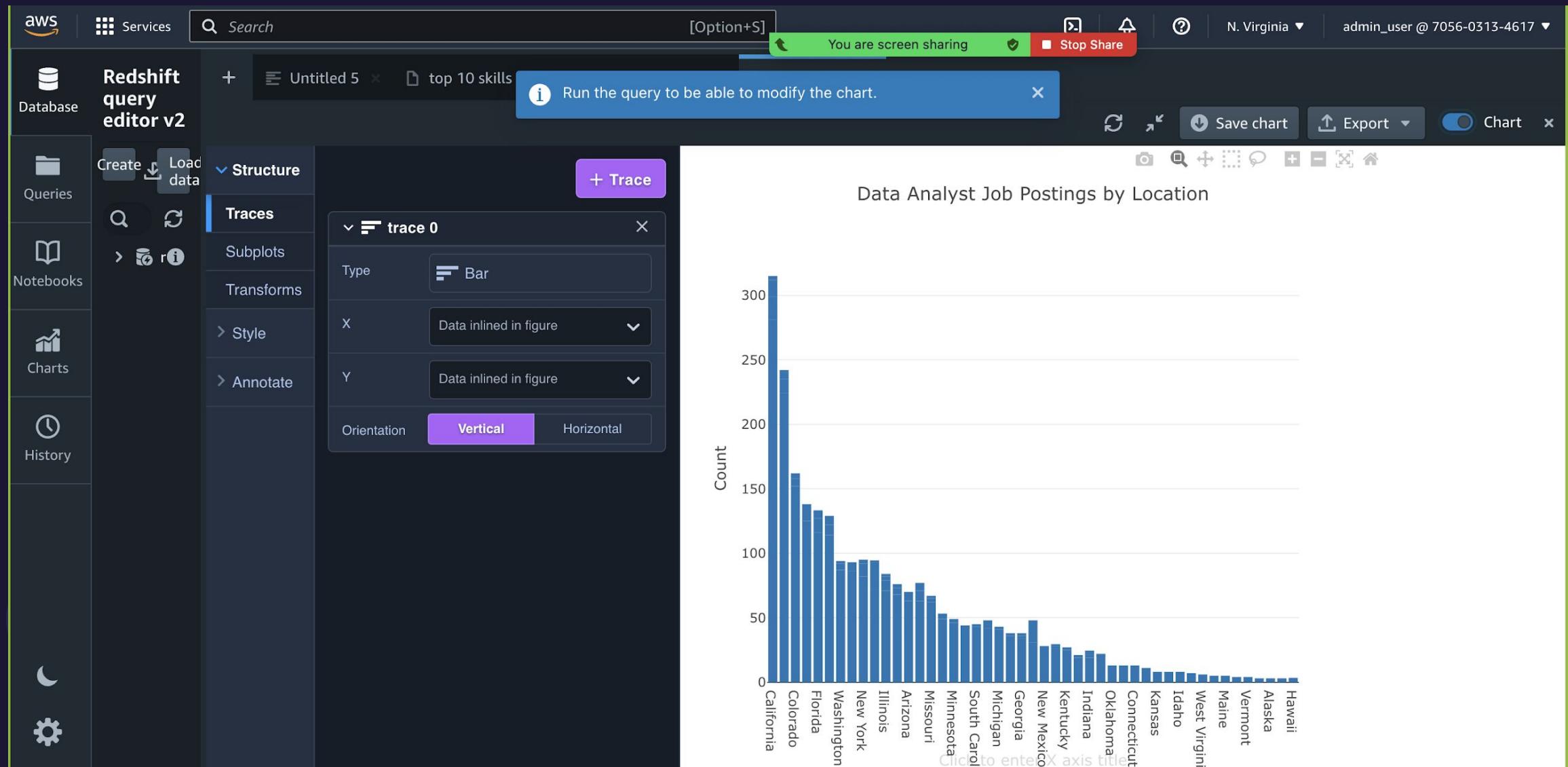
canontitle	skill	cnt
Data Analyst	Data Analysis	2058
Data Analyst	Microsoft Excel	1052
Data Analyst	SQL	1048
Data Analyst	Research	1008
Data Analyst	Communication Skills	1003
Data Analyst	Writing	655
Data Analyst	Problem Solving	642
Data Analyst	Teamwork / Collabora...	635
Data Analyst	Tableau	613

Elapsed time: 18 ms Total rows: 10

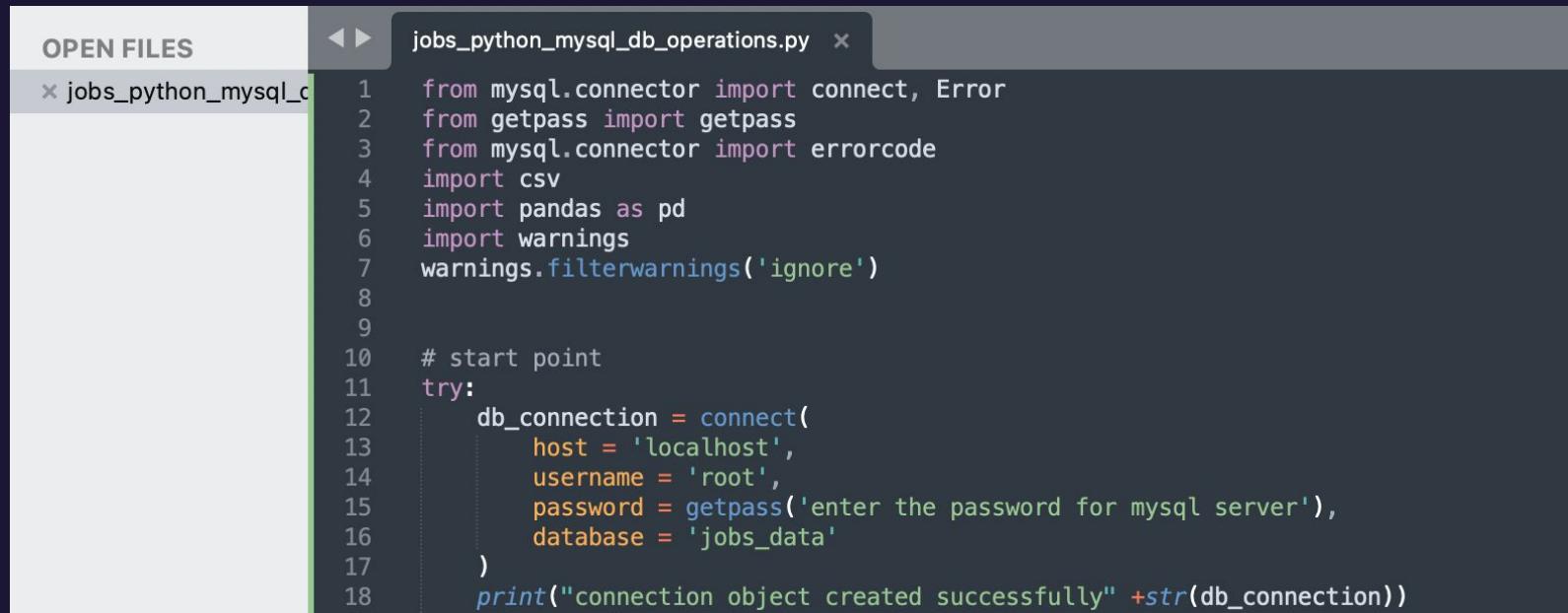
Top 10 required skills for Data Analyst



Job Postings for Data Analyst across states



Python SQL Connectivity



The image shows a screenshot of a code editor with a dark background. On the left, there's a sidebar titled "OPEN FILES" containing two entries: "jobs_python_mysql_db_operations.py" (which is currently selected) and "jobs_python_mysql_c". The main area displays the following Python script:

```
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))

```



Python SQL Connectivity

```
    '
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)
71
```

```
-----*****-----
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 join 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
```

Key Learnings



We learned how to use PySpark to deal with datasets of size 100GB+ with just an 8 GB RAM Machine.



We understand how the Spark Backend works, and processes data in RDDs



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.



We learned how to use AWS Solution Stack to build end-to-end Data Engineering/ Data Warehousing Solutions.



We were able to understand the query design behind the scenes that give us results on job sites.

Technical Difficulties

Due to the Neo4J Aura DB's limitations with the free tier, we were unable to use all the data in our Neo4J Data Model. We struggled a bit to come up with an ideal way of handling this, and this took incredible amount of time.

The Learning curve with the AWS Solution Stack was challenging, we had very little time to explore the tools to its fullest capabilities.

The Technical difficulty with reducing the size of the dataset from 100 million jobs to 1 million Jobs was demanding. Learning pyspark ahead of time helped us tackle this difficulty. This ETL script is the biggest piece of code file in our project.

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.

Conclusion



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 the employee perceptions to track the health of 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.

Future Work



Thank you!

Any questions?