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

Maharsh Soni ((016656770)
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

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

SJSU SAN JOSÉ 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

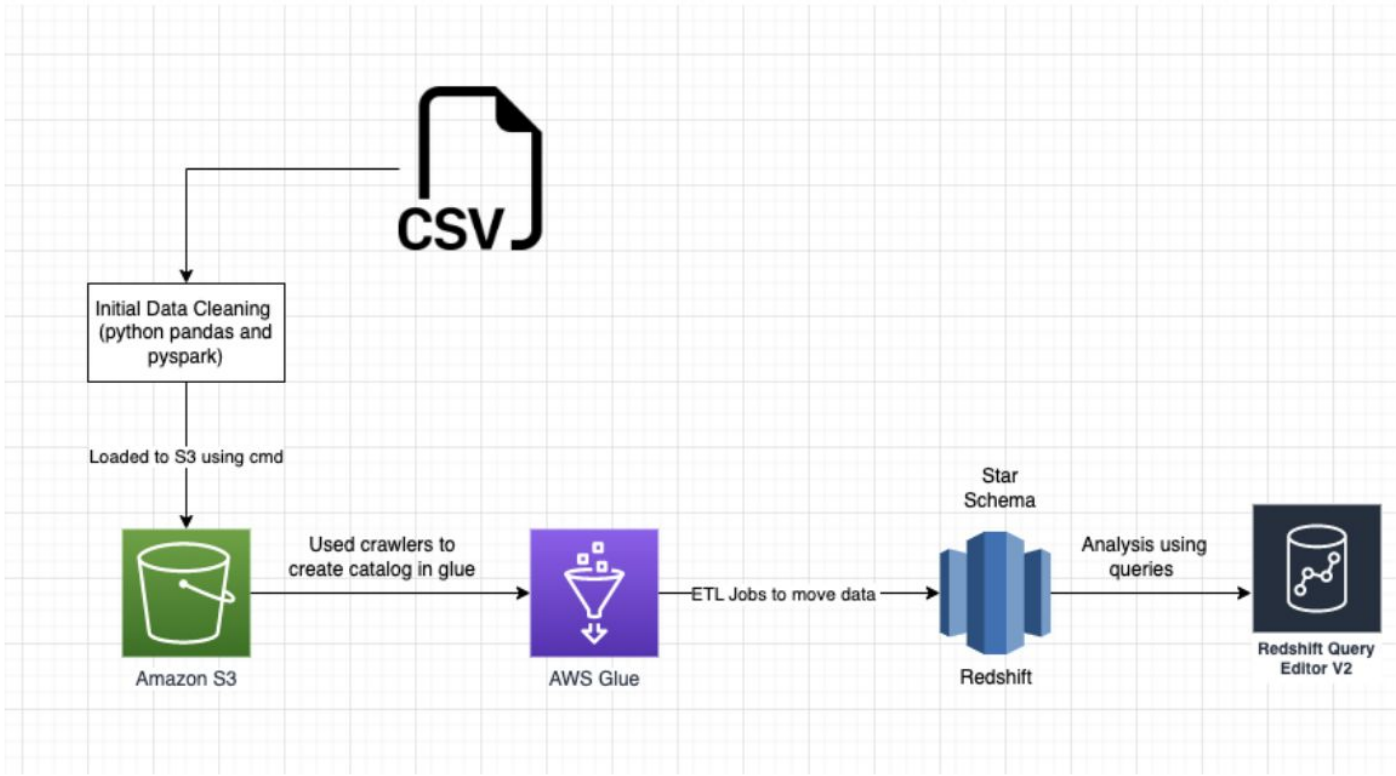
ETL Tools- PySpark, Glue

Storage- AWS S3

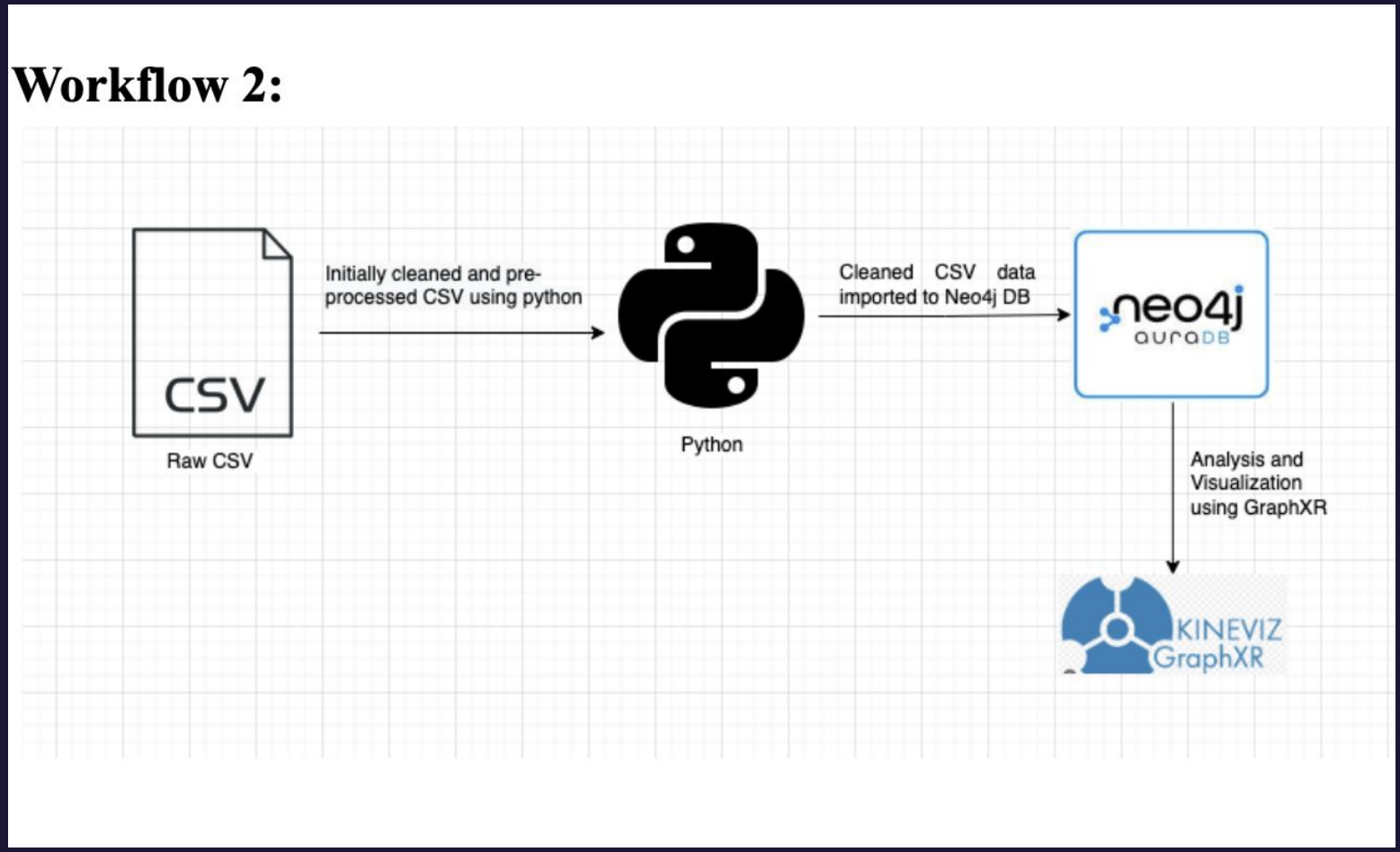
Programming- Python, SQL

Workflow1

Workflow 1:

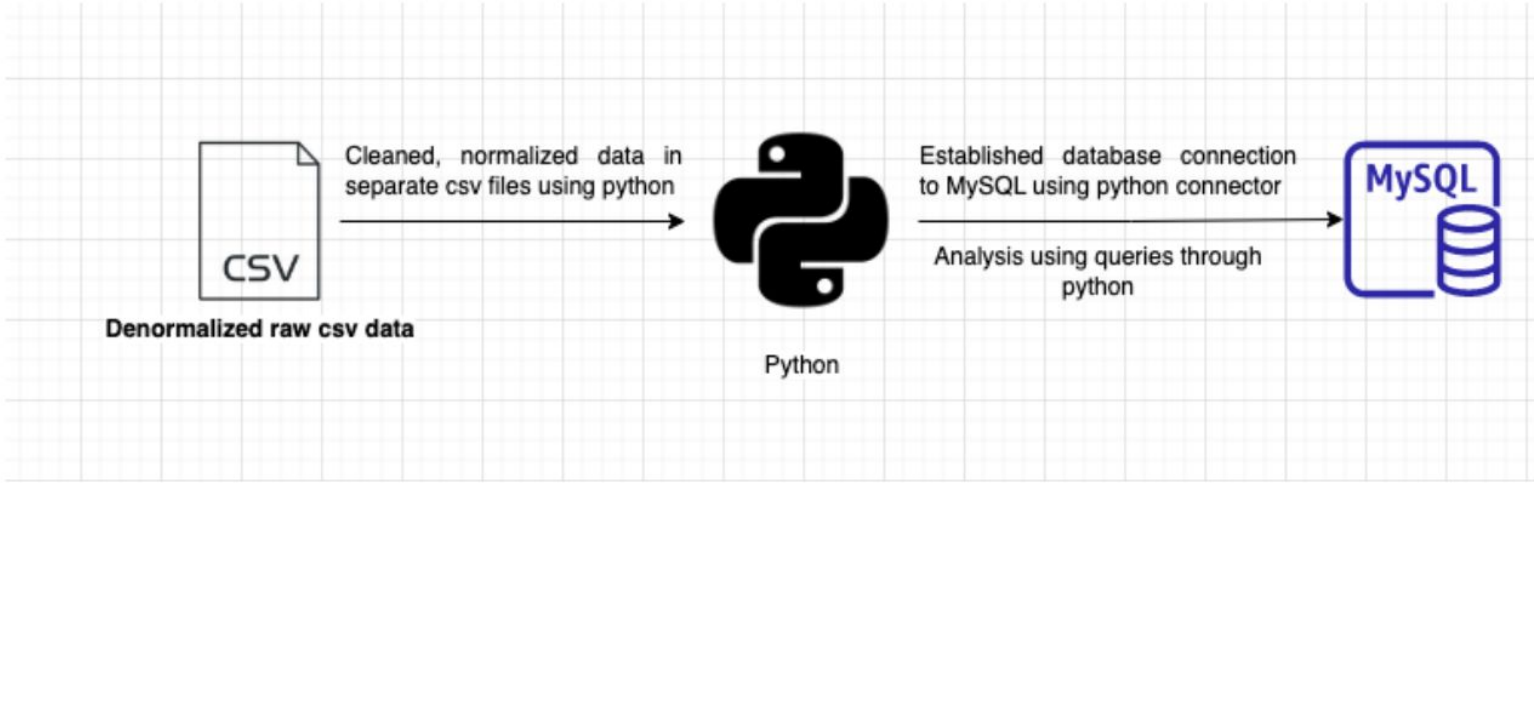


Workflow2


Workflow 2:
CSV
Raw CSV
Initially cleaned and pre-processed CSV using python
Python
Cleaned CSV data imported to Neo4j DB
neo4j
auraDB
Analysis and Visualization using GraphXR
KINEVIZ
GraphXR

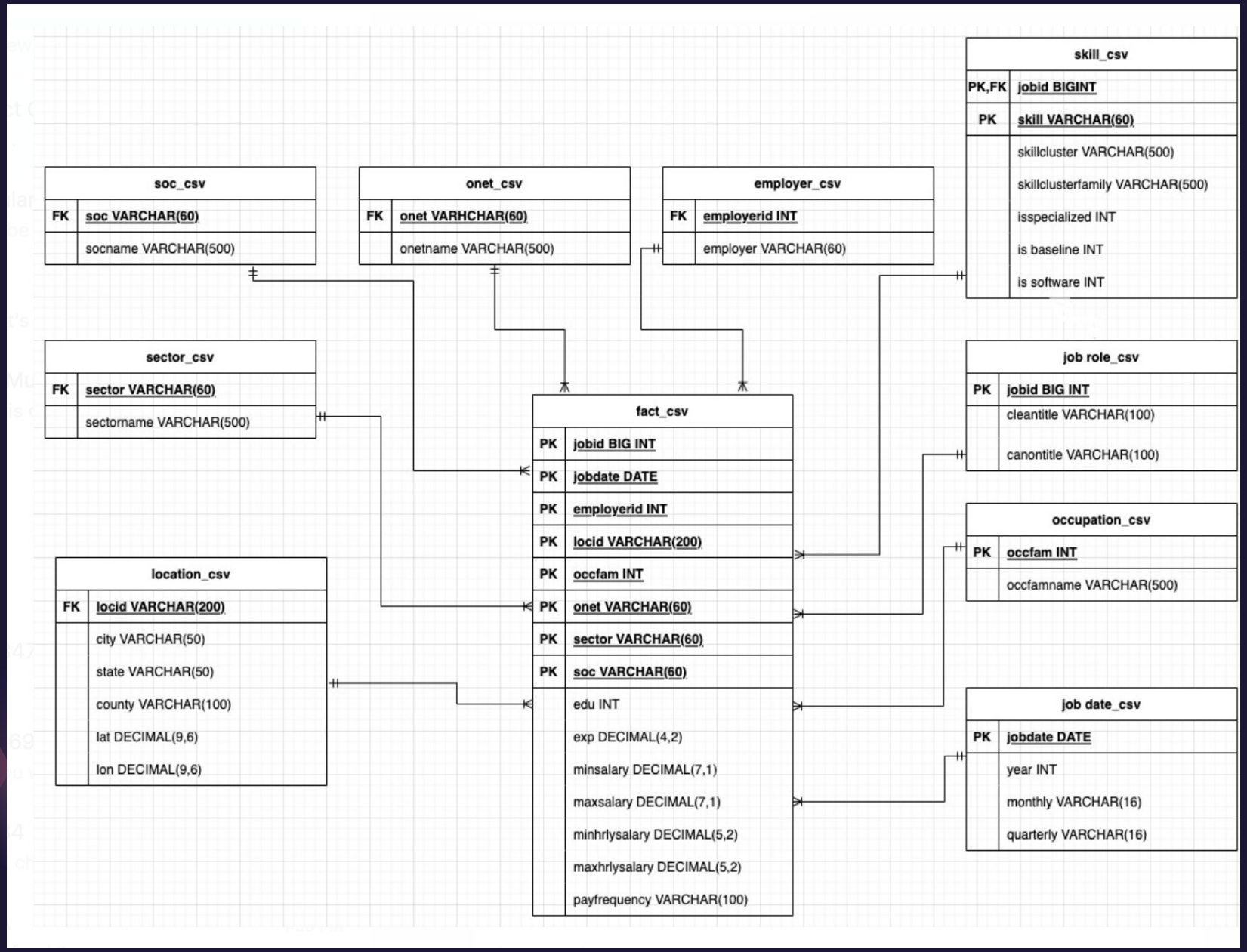
Workflow3

Workflow 3:



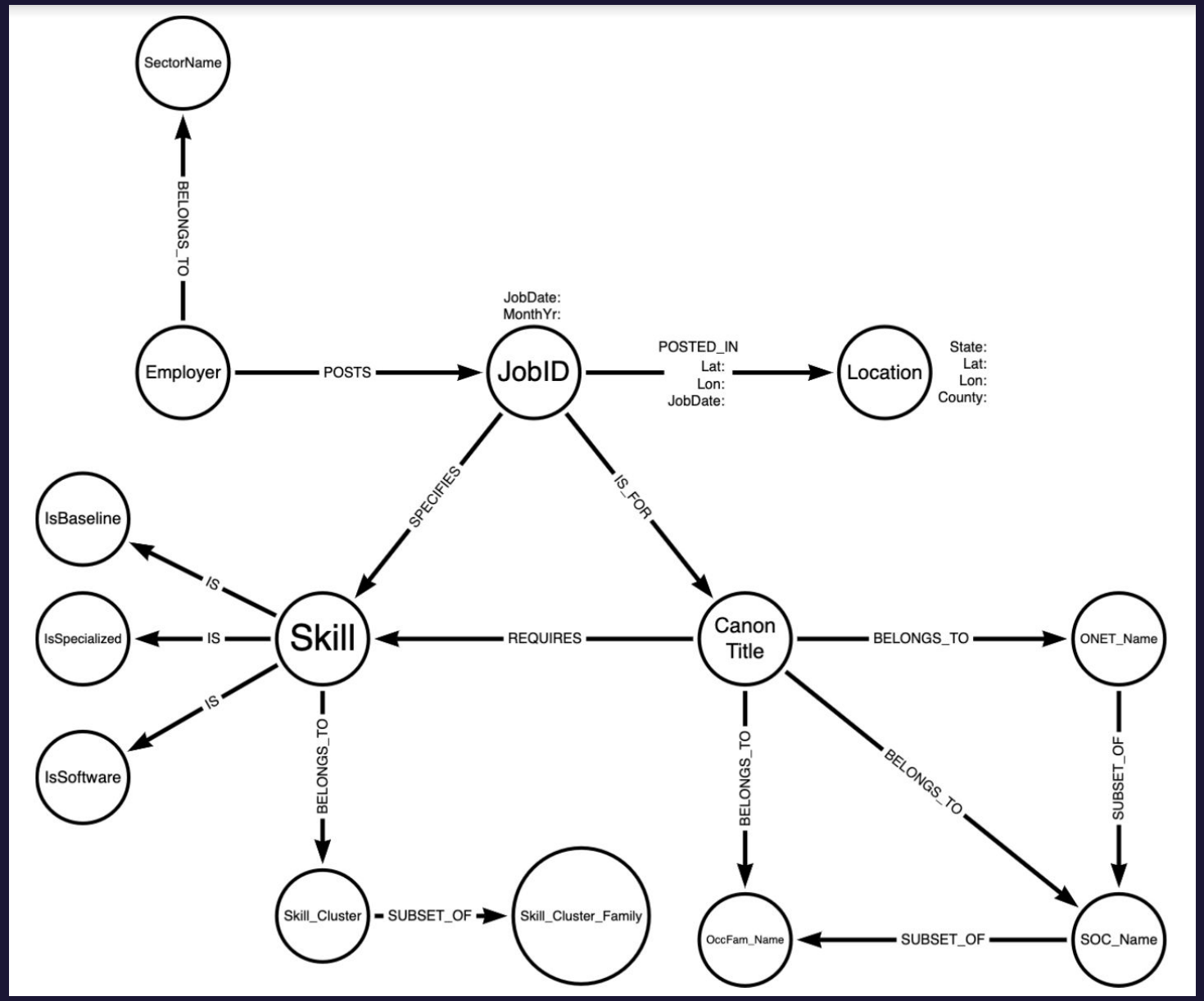
DATA MODEL-1

STAR SCHEMA



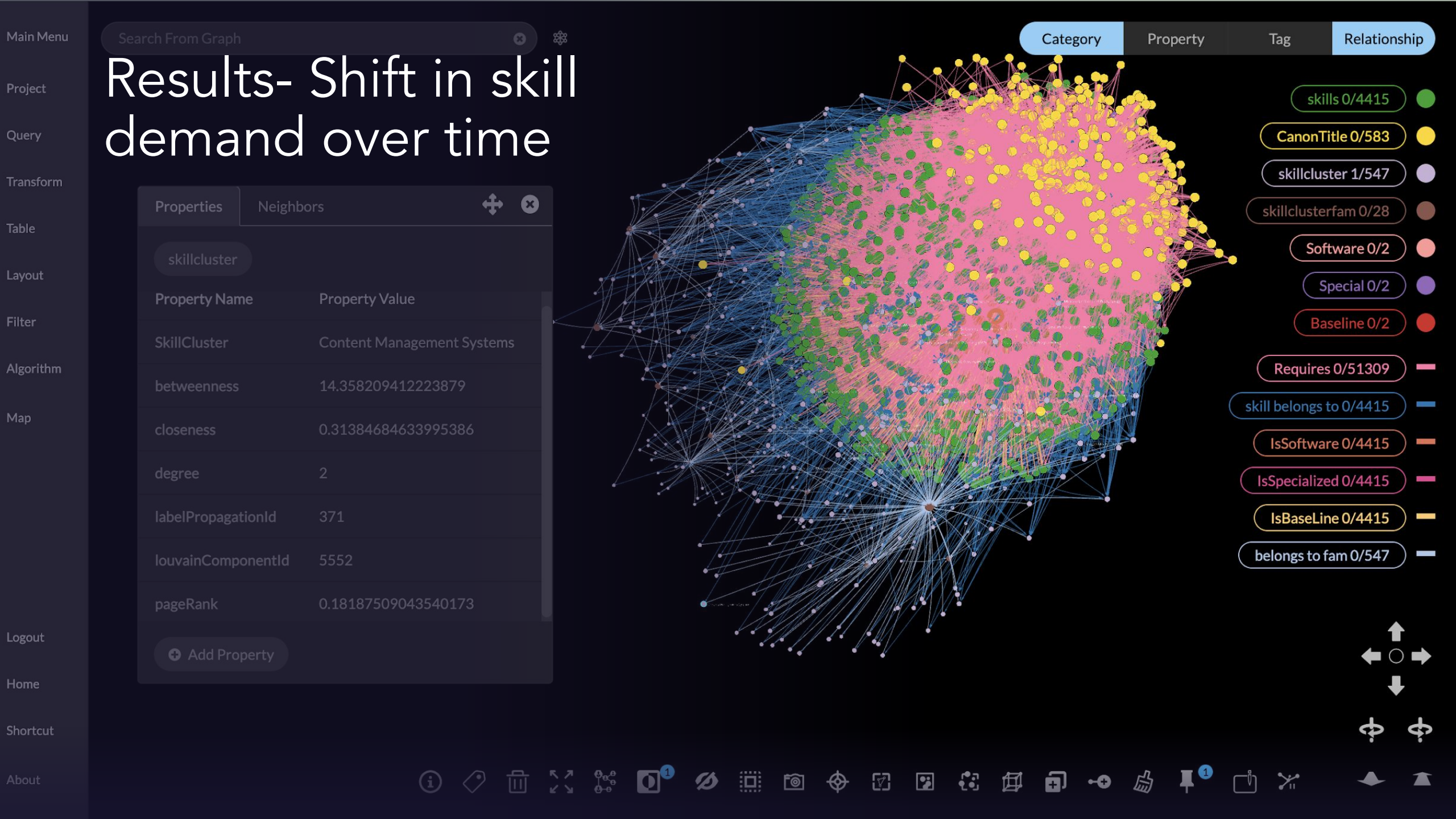
DATA MODEL-2

NEO4J DATA MODEL

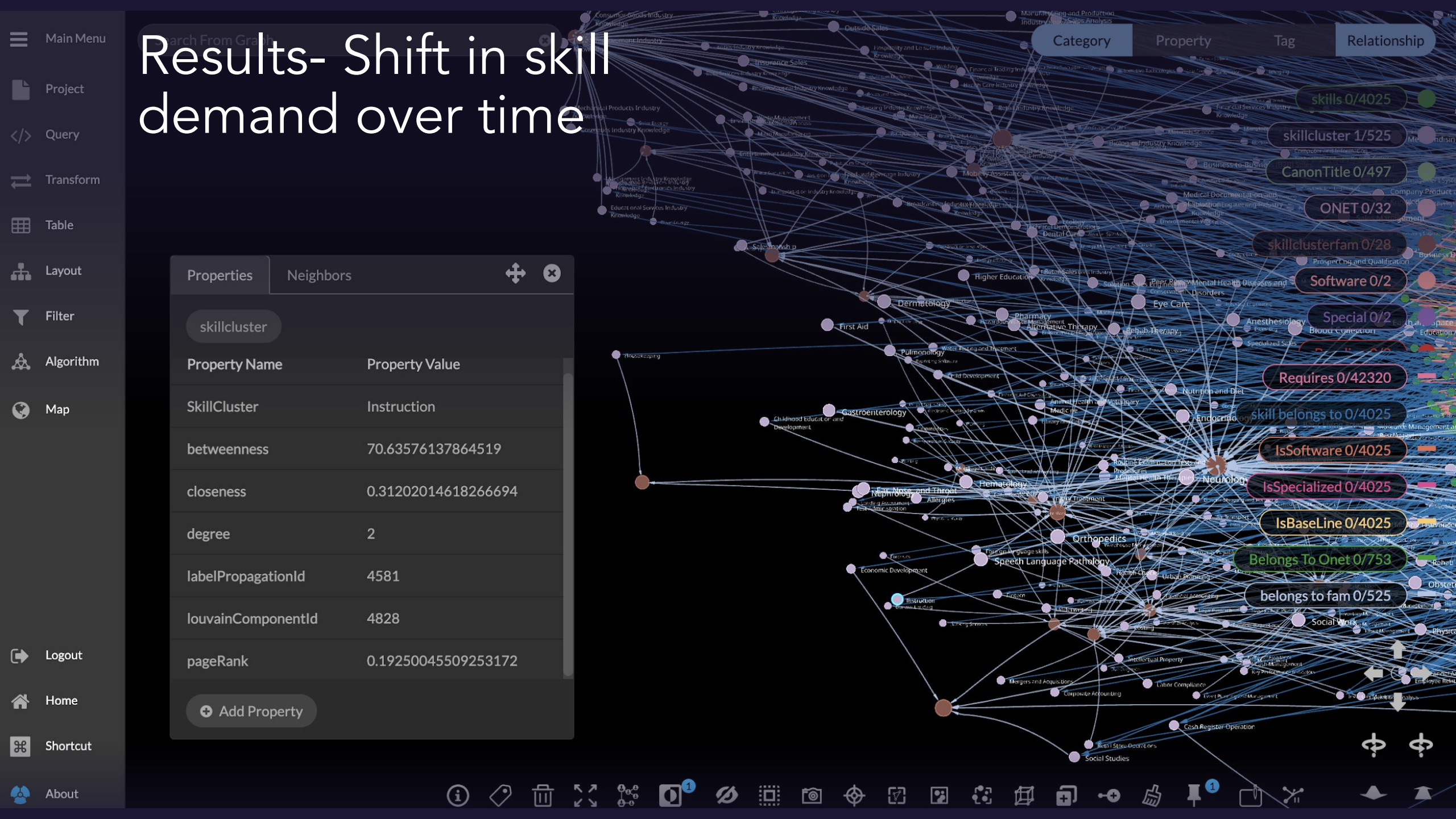


RESULTS FROM NEO4J

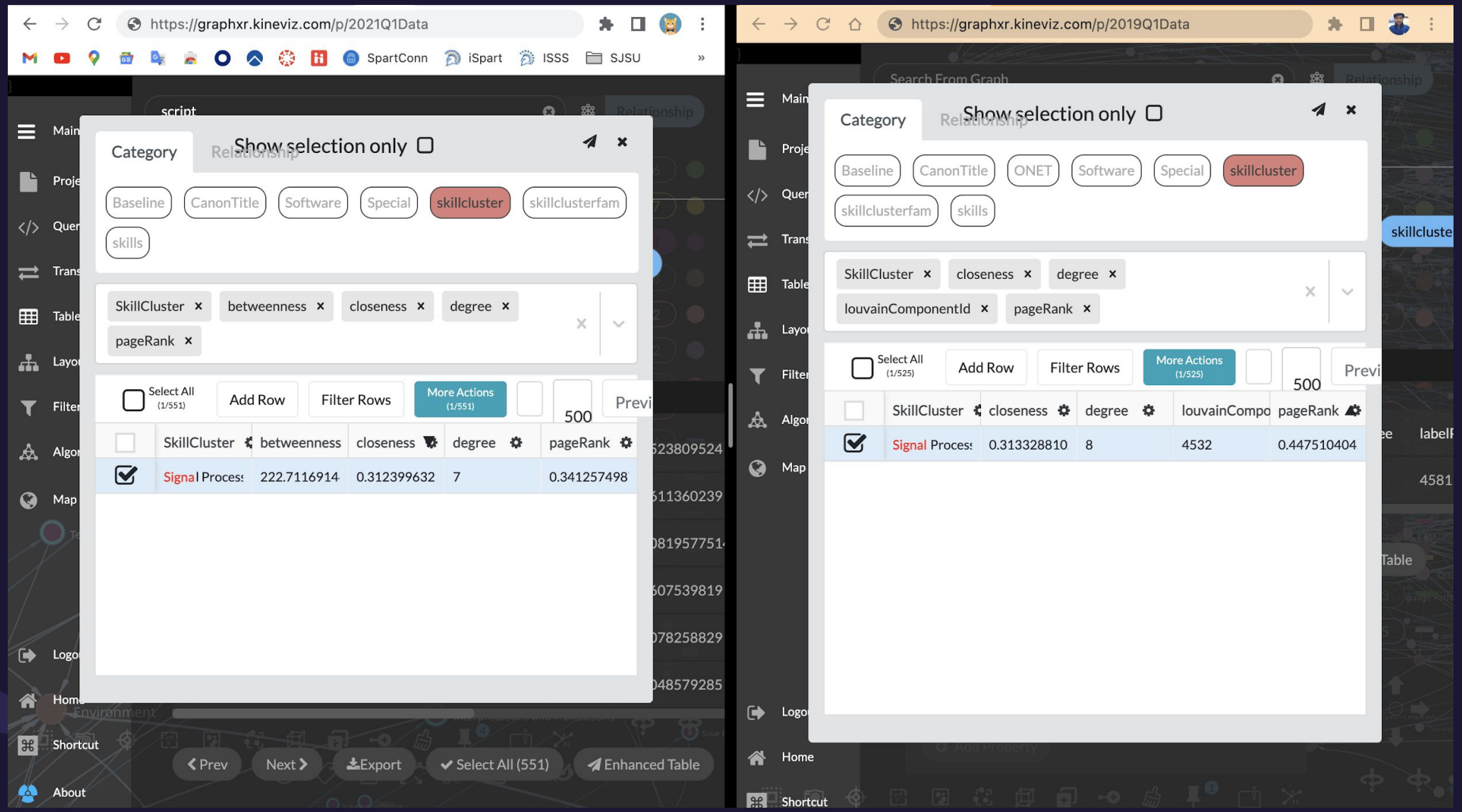
Results- Shift in skill demand over time



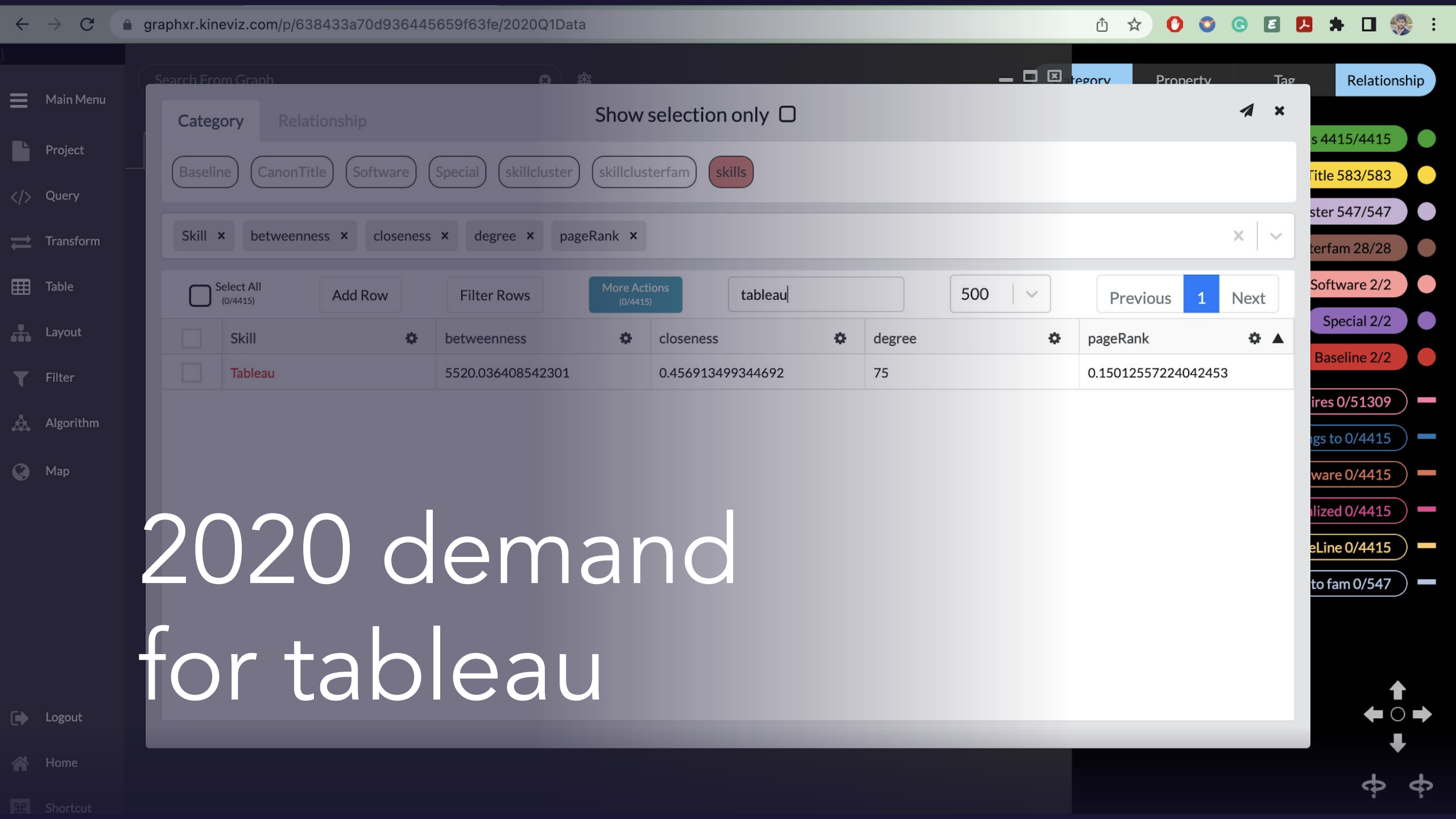
Results- Shift in skill demand over time

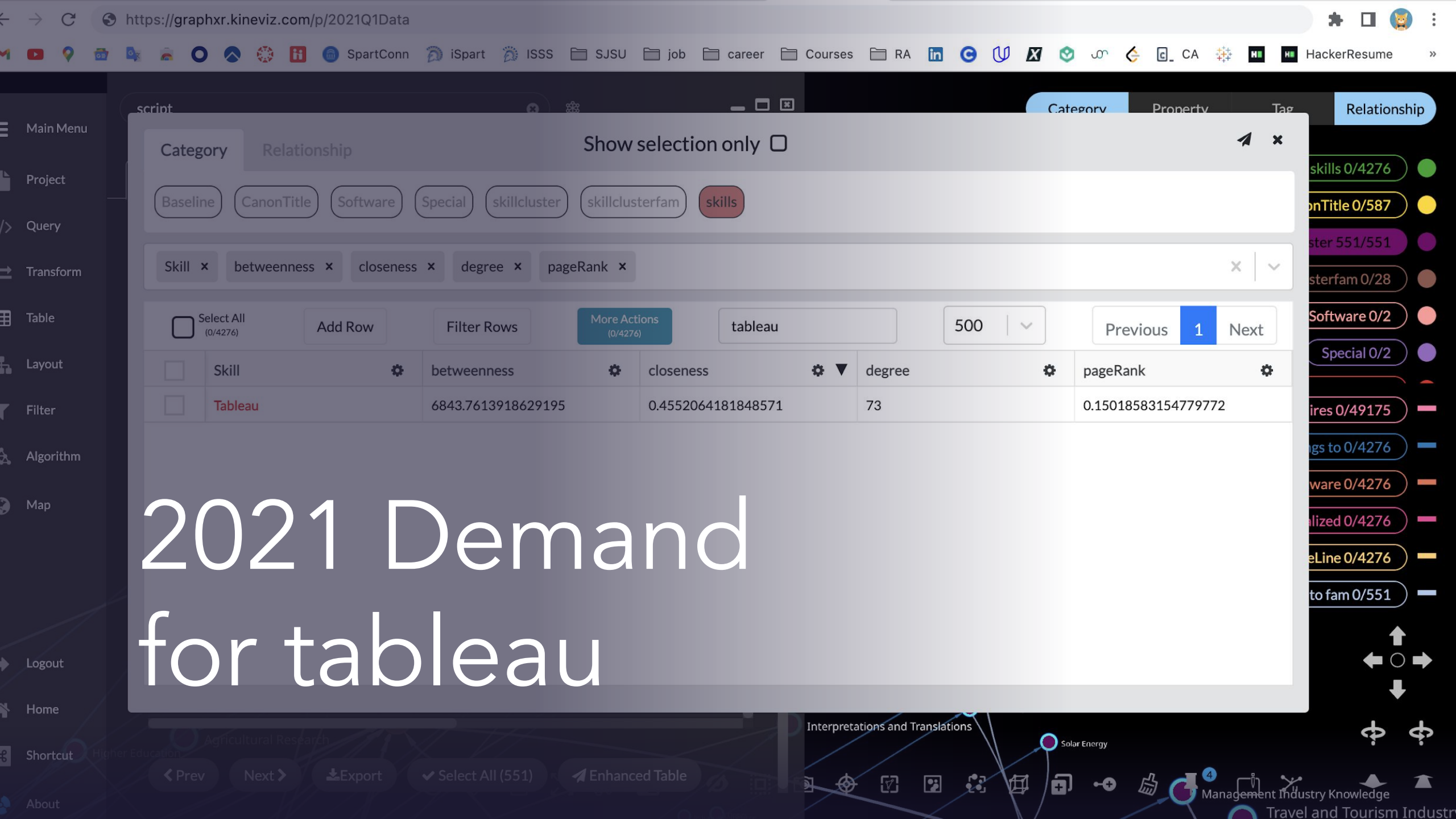


Results- Shift in skill demand over time



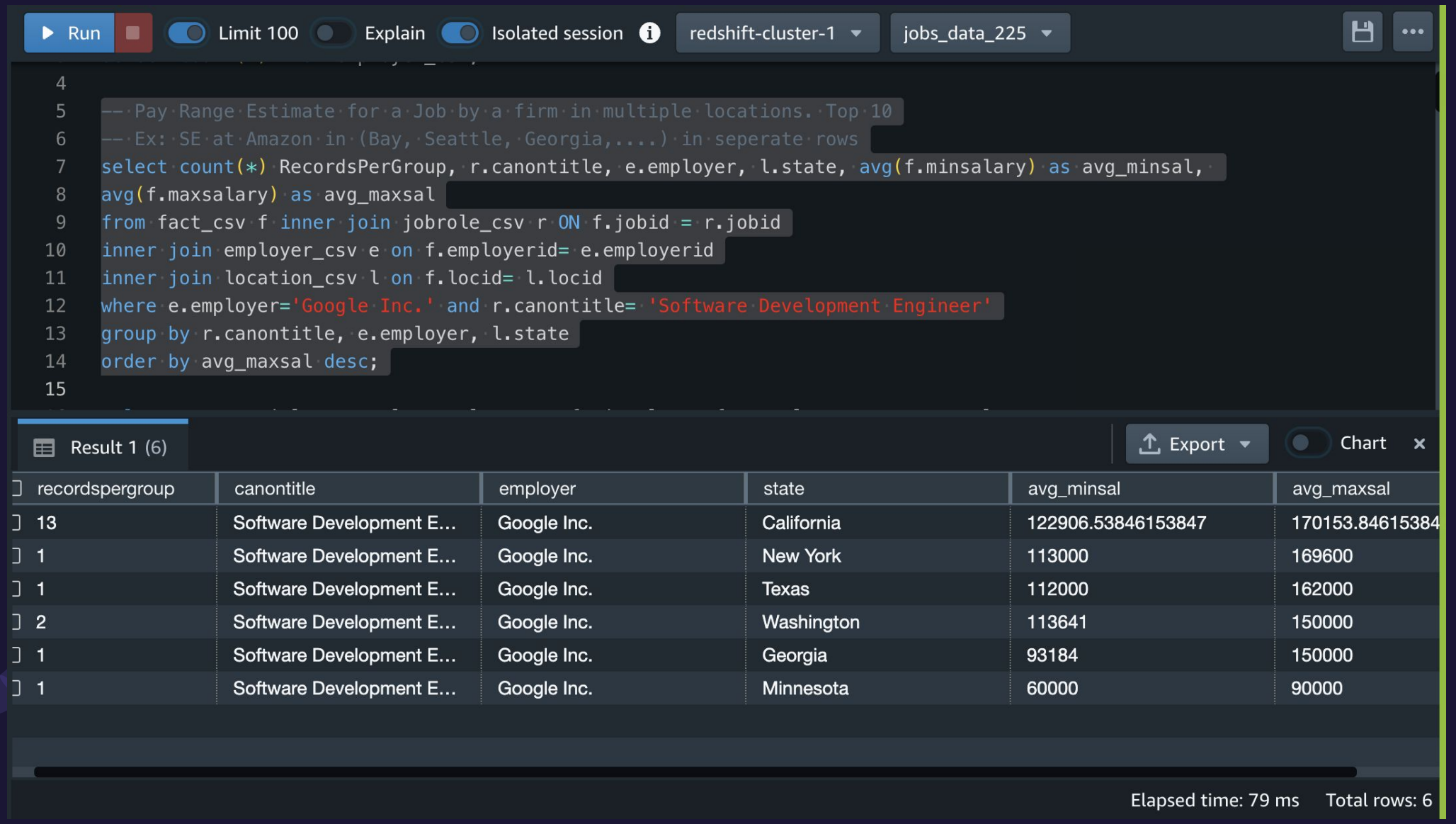

https://graphxr.kineviz.com/p/2019Q1Data
Main Menu
Project
Query
Transform
Table
Layout
Filter
Algorithm
Map
Logout
Home
Shortcut
Search From Graph
Category
Property
Tag
Relationship
Category
Relationship
Show selection only
Baseline
CanonTitle
ONET
Software
Special
skillcluster
skillclusterfam
skills
Skill ×
closeness ×
degree ×
louvainComponentId ×
pageRank ×
Select All (0/4025)
Add Row
Filter Rows
More Actions (0/4025)
tableau
500
Previous
1
Next
Skill
closeness
degree
louvainComponentId
pageRank
Tableau
0.4563420768954366
58
4554
0.15009762032289353
skills 0/4025
cluster 1/525
onTitle 0/497
ONET 0/32
sterfam 0/28
Software 0/2
Special 0/2
ires 0/42320
gs to 0/4025
ware 0/4025
lized 0/4025
eLine 0/4025
Onet 0/753
to fam 0/525
2019 demand
for tableau


graphxr.kineviz.com/p/638433a70d936445659f63fe/2020Q1Data
Main Menu
Project
Query
Transform
Table
Layout
Filter
Algorithm
Map
Logout
Home
Shortcut
Search From Graph
Property
Tag
Relationship
Category
Relationship
Show selection only
Baseline
CanonTitle
Software
Special
skillcluster
skillclusterfam
skills
Skill ×
betweenness ×
closeness ×
degree ×
pageRank ×
Select All (0/4415)
Add Row
Filter Rows
More Actions (0/4415)
tableau
500
Previous
1
Next
Skill
betweenness
closeness
degree
pageRank
Tableau
5520.036408542301
0.456913499344692
75
0.15012557224042453
s 4415/4415
Title 583/583
ster 547/547
terfam 28/28
Software 2/2
Special 2/2
Baseline 2/2
ires 0/51309
gs to 0/4415
ware 0/4415
lized 0/4415
eLine 0/4415
to fam 0/547
2020 demand
for tableau


https://graphxr.kineviz.com/p/2021Q1Data
SpartConn
iSpart
ISSS
SJSU
job
career
Courses
RA
CA
HackerResume
Main Menu
Project
Query
Transform
Table
Layout
Filter
Algorithm
Map
Logout
Home
Shortcut
About
Category
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
2021 Demand for tableau
Property
Tag
skills 0/4276
onTitle 0/587
ster 551/551
sterfam 0/28
Software 0/2
Special 0/2
ires 0/49175
gs to 0/4276
ware 0/4276
lized 0/4276
eLine 0/4276
to fam 0/551
Interpretations and Translations
Solar Energy
Management Industry Knowledge
Travel and Tourism Industr
Prev
Next
Export
Select All (551)
Enhanced Table

Sample Results from Redshift Queries

Salaries by firm, and role across states



Pay Range Estimate for a role

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

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

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

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

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

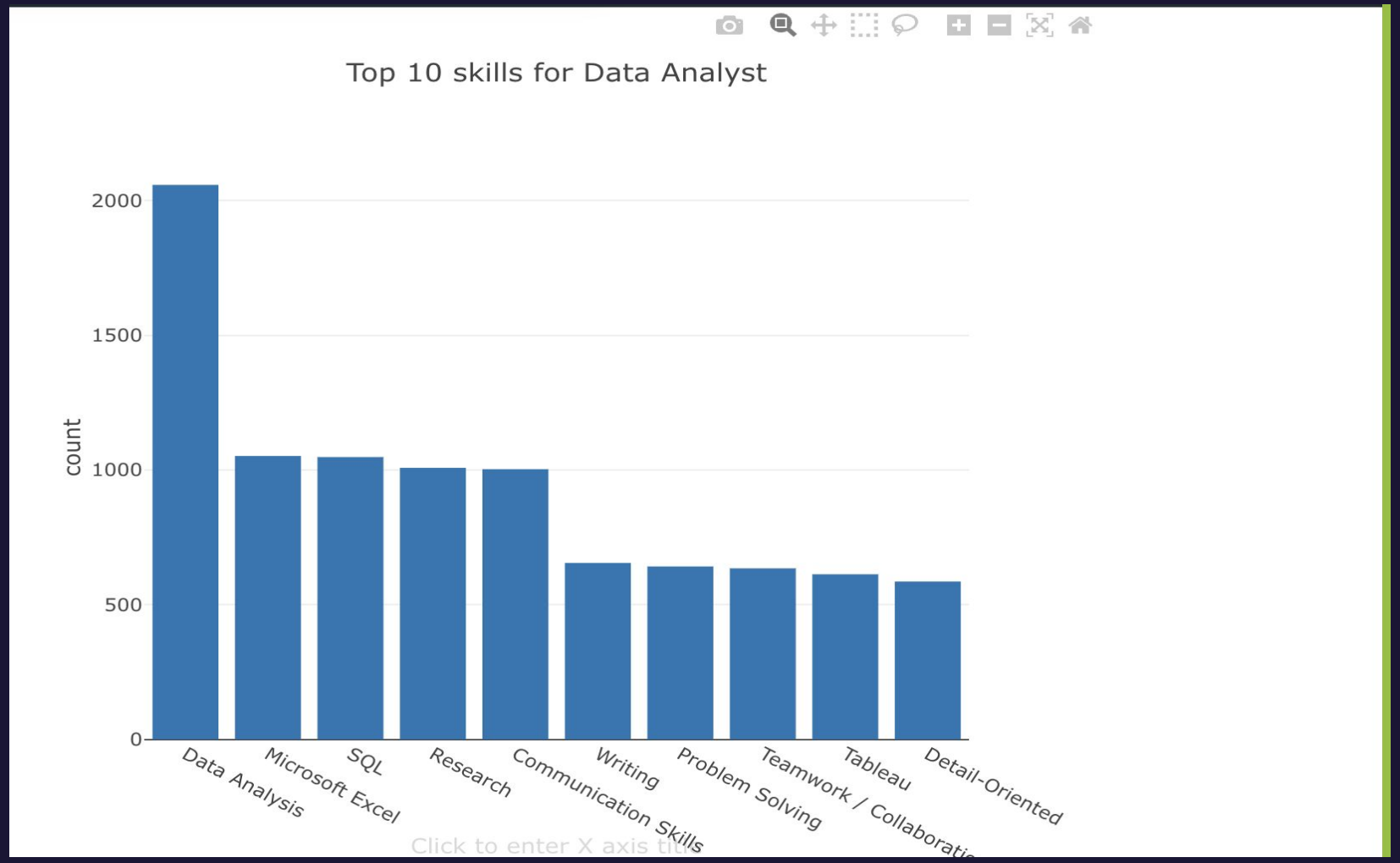
```sql
-- top 10 most asked skills for a given job title(Data Analyst)
select top 10 sk.skill, r.canontitle, count(sk.skill) as cnt
from fact_csv f inner join skill_csv sk
on f.jobid = sk.jobid
inner join jobrole_csv r
on f.jobid = r.jobid
inner join employer_csv e
on e.employerid = f.employerid
where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

Result 1 (10) | Export | Chart

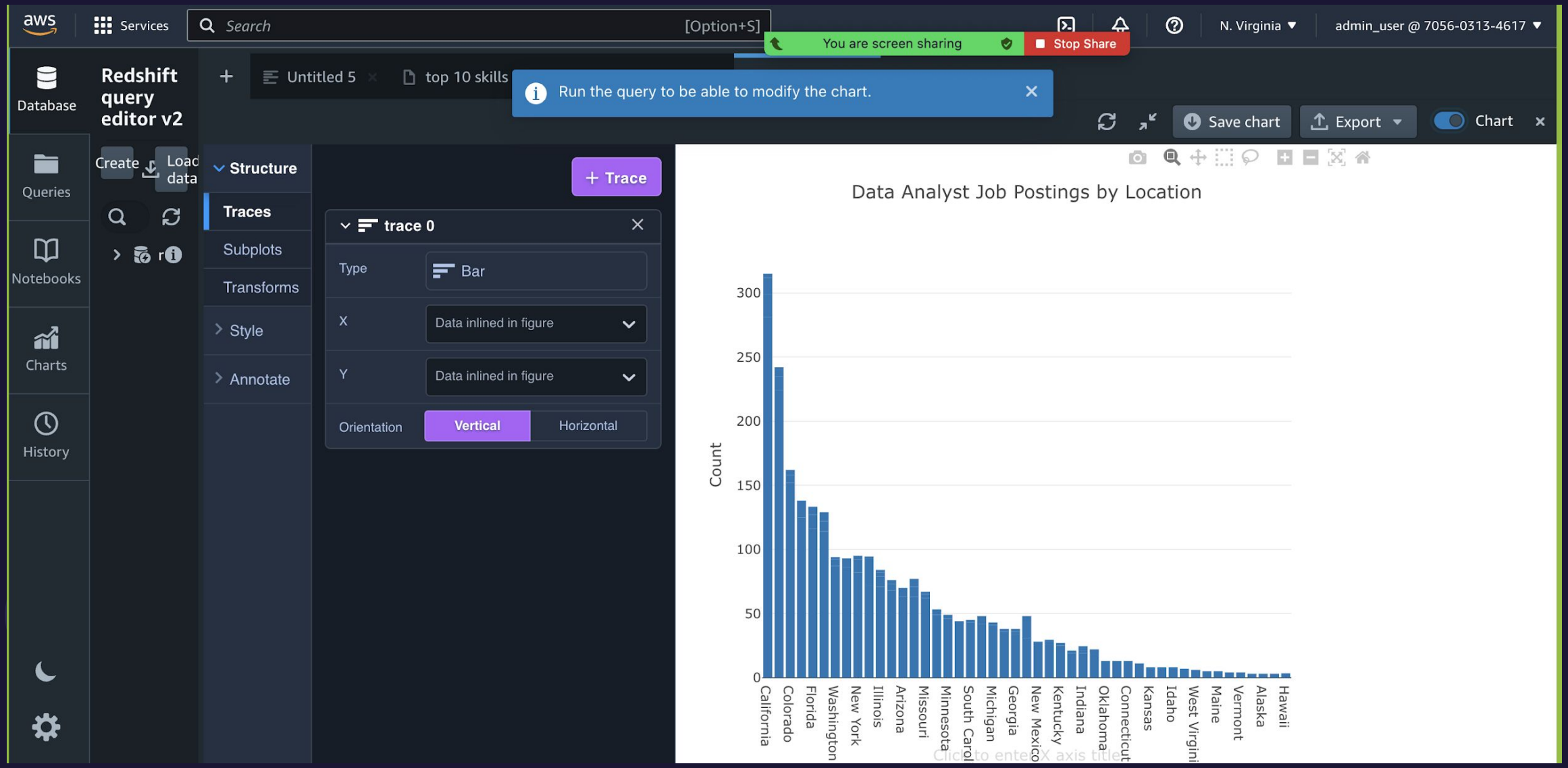
canontitle	skill	cnt
Data Analyst	Data Analysis	2058
Data Analyst	Microsoft Excel	1052
Data Analyst	SQL	1048
Data Analyst	Research	1008
Data Analyst	Communication Skills	1003
Data Analyst	Writing	655
Data Analyst	Problem Solving	642
Data Analyst	Teamwork / Collabora...	635
Data Analyst	Tableau	613

Elapsed time: 18 ms Total rows: 10

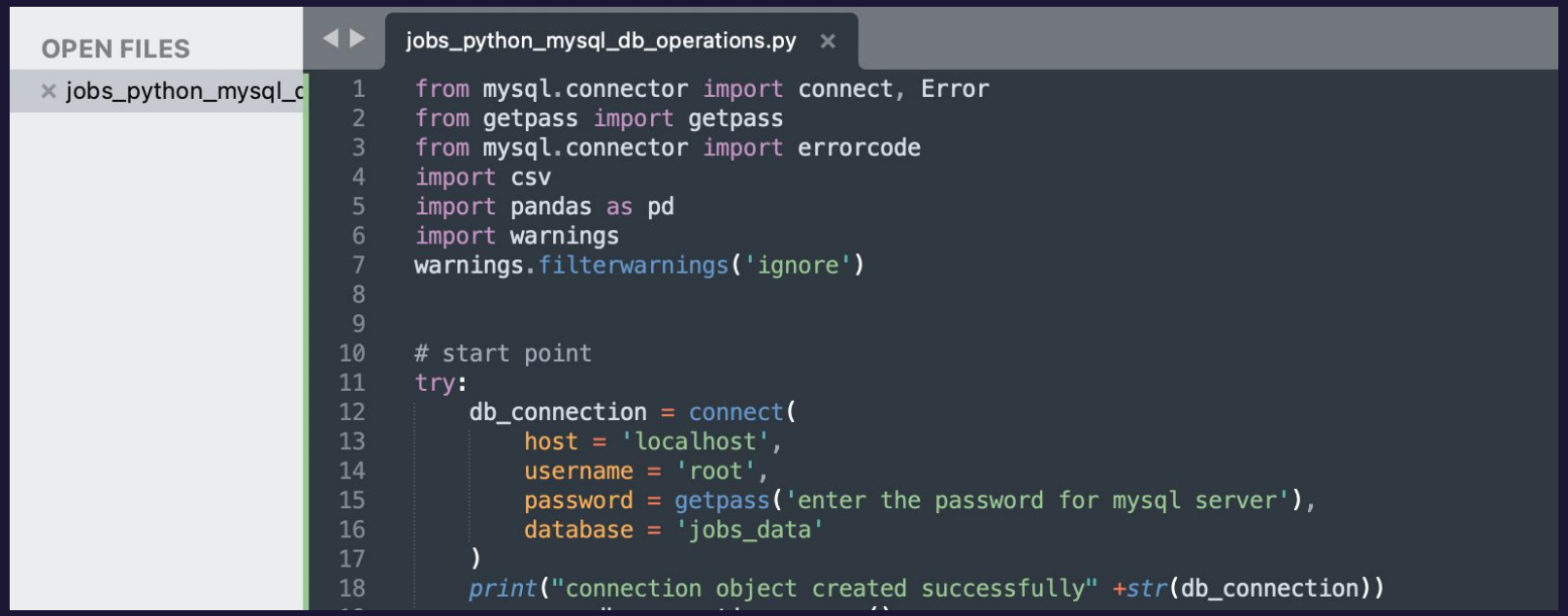
Top 10 required skills for Data Analyst



Job Postings for Data Analyst across states



Python SQL Connectivity



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)
```

```
----------**********-------------
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
```

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.

Conclusion

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 reterive 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 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 prosposed in this project.

Future Work

Thank you!

Any questions?