Group 3 Final Project: Neptune Graph Database

SP 2022 DAV 6100: Information Architectures

Katz School of Science and Health

By

Atreish Ramlakhan
Mahlet Melese
Shichao Zhou
Yuxiao Shen

Group Members

Atreish Ramlakhan

M.S. Artificial Intelligence

Mahlet Melese

M.S. in Data Analytics & Visualization

Shichao Zhou

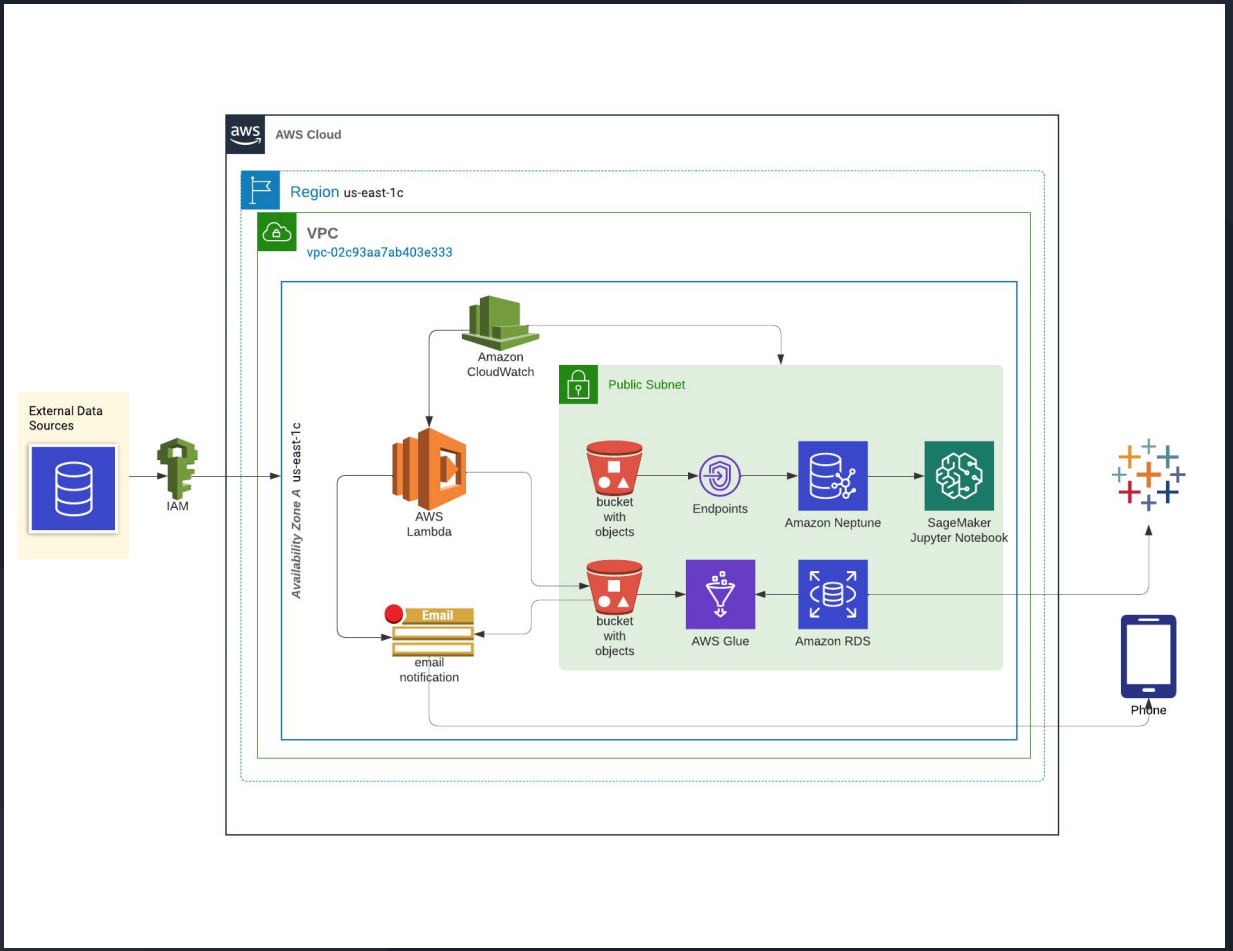
M.S. in Data Analytics & Visualization

Yuxiao (Henry) Shen

M.S. in Data Analytics & Visualization

Introduction

AWS Architecture



“Original” Data Structures

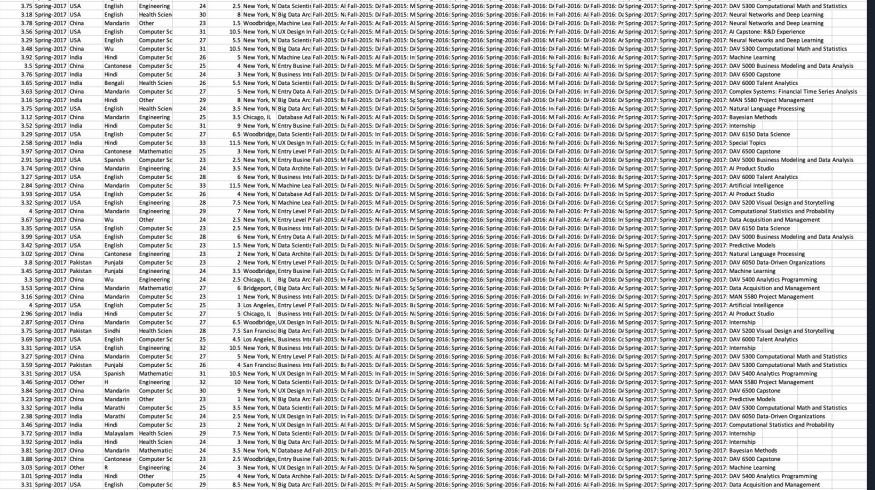
Student Focused Structure

Student ID	Course1	Course2	Course3	Course4	Course5	Course6	Course7	Course8	Course9	Course1
1168340921	Spring-2016: DAV 5200 Visual Design and Storytelling	Spring-2016: DAV 6100 Information Architectures	Spring-2016: DAV 5300 Computational Math and Statistics	Fall-2016: MAN 5580 Project Management	Fall-2016: DAV 6000 Talent Analytics	Fall-2016: Natural Language Processing	Spring-2017: DAV 5000 Business Modeling and Data Analysis	Spring-2017: DAV 6500 Capstone	Spring-2017: DAV 5100 Structured Data Management	Fall-201 Speci Topic
1191512830	Fall-2018: Predictive Models	Fall-2018: Neural Networks and Deep Learning	Fall-2018: Machine Learning	Spring-2019: Data Acquisition and Management	Spring-2019: Natural Language Processing	Spring-2019: AI Capstone: R&D Experience	Fall-2019: Numerical Methods	Fall-2019: Artificial Intelligence	Fall-2019: Computational Statistics and Probability	Spring-202 Bayesia Method
1232473375	Spring-2019: Data Visualization	Spring-2019: DAV 5400 Analytics Programming	Spring-2019: DAV 5200 Visual Design and Storytelling	Fall-2019: DAV 5000 Business Modeling and Data Analysis	Fall-2019: DAV 6000 Talent Analytics	Fall-2019: DAV 6200 Data Product Design	Spring-2020: DAV 6500 Capstone	Spring-2020: MAN 5580 Project Management	Spring-2020: DAV 5100 Structured Data Management	Fall-202 Speci Topic
1313020897	Fall-2016: Advanced Data Engineering	Fall-2016: DAV 5100 Structured Data Management	Fall-2016: DAV 5000 Business Modeling and Data Analysis	Spring-2017: DAV 5400 Analytics Programming	Spring-2017: DAV 5200 Visual Design and Storytelling	Spring-2017: DAV 6400 Internship	Fall-2017: DAV 6450 Independent Study	Fall-2017: Independent Study	Fall-2017: DAV 5300 Computational Math and Statistics	Spring-201 DAV 650 Capstor
1364243602	Spring-2020: MAN 5580 Project	Spring-2020: DAV 6000 Talent	Spring-2020: DAV 5200 Visual Design and	Fall-2020: DAV 6500 Capstone	Fall-2020: Internship	Fall-2020: DAV 5100 Structured Data	Spring-2021: DAV 5400 Analytics	Spring-2021: Independent Study	Spring-2021: Natural Language	Fall-202 DAV 500 Busines Modelir

Course Focused Structure

Course Name	Student ID	First Name	Last Name	Katz School Major	GPA	Graduation Semester	Country of Origin	Languages	Undergraduate Major	Age at Graduation	Years of Experience	Location	Dec
Fall-2015: AI Capstone: R&D Experience	4784813239	Calahan	Martin	Artificial Intelligence	2.95	Spring-2017	USA	English	Computer Science	27	5.5	New York, NY	Ent
	4800891893	Johnson	Angela	Artificial Intelligence	3.51	Spring-2017	China	Mandarin	Mathematics	24	2.5	New York, NY	Ent
Fall-2015: AI Product Studio	8479614053	Pockrus	Clifford	Data Analytics and Visualization	4.00	Spring-2017	China	Mandarin	Other	29	8.0	New York, NY	B S
	9290159667	Potter	Eugene	Artificial Intelligence	3.63	Spring-2017	Pakistan	Sindhi	Mathematics	27	6.0	New York, NY	Ar
Fall-2015: Advanced Data Engineering	4991566949	Berry	Susan	Data Analytics and Visualization	3.55	Spring-2017	China	Mandarin	Computer Science	25	3.0	New York, NY	B Inte De
...	...	...	...	...	...	...	...	...	...	...	...	...	
Spring-2022: MAN 5580				Data Analytics		Spring-			Computer			Los	Da

Alumni



Salary.com

	A	B	C	D	E	F	G	H
1	Title	Location	Description	nTile10	nTile25	nTile50	nTile75	nTile90
2	Entry Data A	New York, N	The Entry Da	59417	66701	74701	85201	94761
3	Data Scientis	New York, N	Data Scientis	71385	79225	87835	94930	101390
4	Machine Lea	New York, N	Machine lear	120959	133726	147749	163993	178782
5	Business Inte	New York, N	Business Inte	84662	96747	110021	124339	137376
6	Entry Level P	New York, N	Web Applica	86240	96633	108048	123254	137098
7	UX Design In	New York, N	UI/Usability	70071	81053	93115	102323	110707
8	Data Archite	New York, N	Data Archite	56510	76862	99216	121634	142045
9	Big Data Arc	New York, N	A data archit	142697	160526	180108	209540	236337
10	Database Ad	New York, N	The Databas	63651	72300	81800	94800	106636
11	Entry Busine	New York, N	The Entry Bu	59876	67260	75370	84300	92430
12	Entry Data A	Chicago, IL	The Entry Da	52337	58801	65901	75101	83477
13	Data Scientis	Chicago, IL	Data Scientis	62952	69866	77459	83716	89413
14	Machine Lea	Chicago, IL	Machine lear	106670	117929	130295	144620	157662
15	Business Inte	Chicago, IL	Business Inte	74661	85318	97024	109651	121147
16	Entry Level P	Chicago, IL	Web Applica	76052	85218	95284	108694	120903
17	UX Design In	Chicago, IL	UI/Usability	61794	71478	82115	90236	97629
18	Big Data Arc	Chicago, IL	A data archit	125840	141563	158832	184787	208418
19	Database Ad	Chicago, IL	The Databas	56152	63800	72200	83700	94170
20	Entry Busine	Chicago, IL	The Entry Bu	52800	59310	66460	74340	81514
21	Entry Data A	Woodbridge,	The Entry Da	55373	62201	69701	79401	88232
22	Data Scientis	Woodbridge,	Data Scientis	66579	73890	81921	88538	94563
23	Machine Lea	Woodbridge,	Machine lear	112815	124722	137801	152951	166744
24	Business Inte	Woodbridge,	Business Inte	78962	90233	102613	115967	128126
25	Entry Level P	Woodbridge,	Web Applica	80433	90126	100773	114955	127867
26	UX Design In	Woodbridge,	UI/Usability	65353	75595	86845	95434	103253
27	Data Archite	Woodbridge,	Data Archite	52705	71686	92535	113444	132481
28	Big Data Arc	Woodbridge,	A data archit	133089	149717	167981	195432	220424
29	Database Ad	Woodbridge,	The Databas	59488	67500	76300	88500	99607
30	Entry Busine	Woodbridge,	The Entry Bu	55847	62730	70290	78620	86204
31	Entry Data A	San Francisc	The Entry Da	61744	69301	77601	88501	98425
32	Data Scientis	San Francisc	Data Scientis	74174	82320	91267	98639	105351
33	Machine Lea	San Francisc	Machine lear	125685	138951	153521	170400	185766
34	Business Inte	San Francisc	Business Inte	87970	100527	114319	129197	142743
35	Entry Level P	San Francisc	Web Applica	89609	100408	112269	128069	142454
36	UX Design In	San Francisc	UI/Usability	72809	84220	96753	106321	115032
37	Data Archite	San Francisc	Data Archite	58717	79865	103092	126386	147594
38	Big Data Arc	San Francisc	A data archit	148272	166797	187145	217727	245570
39	Database Ad	San Francisc	The Databas	66087	75100	85000	98600	110982
40	Entry Busine	San Francisc	The Entry Bu	62224	69890	78310	87590	96039
41	Entry Data A	Los Angeles,	The Entry Da	55873	62701	70201	80101	89114
42	Data Scientis	Los Angeles,	Data Scientis	67112	74482	82577	89247	95320
43	Machine Lea	Los Angeles,	Machine lear	113718	125720	138904	154175	168079
44	Business Inte	Los Angeles,	Business Inte	79594	90955	103434	116895	129151
45	Entry Level P	Los Angeles,	Web Applica	81077	90848	101579	115875	128890
46	UX Design In	Los Angeles,	UI/Usability	65876	76201	87540	96197	104079
47	Data Archite	Los Angeles,	Data Archite	53126	72260	93276	114352	133541
48	Big Data Arc	Los Angeles,	A data archit	134154	150916	169325	196996	222188
49	Database Ad	Los Angeles,	The Databas	59897	68000	76900	89200	100399
50	Entry Busine	Los Angeles,	The Entry Bu	56292	63230	70850	79250	86898
51	Entry Data A	Bridgeport, C	The Entry Da	56173	63001	70501	80401	89414
52	Data Scientis	Bridgeport, C	Data Scientis	67409	74812	82943	89643	95743
53	Machine Lea	Bridgeport, C	Machine lear	114223	126278	139520	154859	168825
54	Business Inte	Bridgeport, C	Business Inte	79947	91359	103893	117414	129725
55	Entry Level P	Bridgeport, C	Web Applica	81437	91251	102030	116389	129463
56	UX Design In	Bridgeport, C	UI/Usability	66169	76539	87929	96624	104541
57	Data Archite	Bridgeport, C	Data Archite	53362	72581	93690	114860	134134
58	Big Data Arc	Bridgeport, C	A data archit	134750	151585	170077	197870	223174
59	Database Ad	Bridgeport, C	The Databas	60106	68300	77300	89600	100799
60	Entry Busine	Bridgeport, C	The Entry Bu	56536	63510	71170	79600	87275

LinkedIn

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Full Name	Location	Most Recent	Job Title	Company Url										
2	Atreish Ram	New York, N	Katz School	Graduate Te	https://www.linkedin.com/company/16181365/										
3	Yuxiao (Henr	New York, N	The AAT Proj	Full Stack PH	https://www.linkedin.com/search/results/all/?keywords=The+AAT+Project+%28America%E2%80%99s+Amazing+Teens%2C+LLC%29										
4	Shichao Zhou	New York, N	S&P Global M	Data Analyst	https://www.linkedin.com/company/162892/										
5	Mahlet Mele	New York, New York, United States													
6															

Data Profile: Student Dummy Dataset

Dataset Summary

Source of Information	Randomly generated dataset
Number of Records	1319
Frequency of updates	Yearly
Data type and structure	
Number of columns	14
Granularity	Age

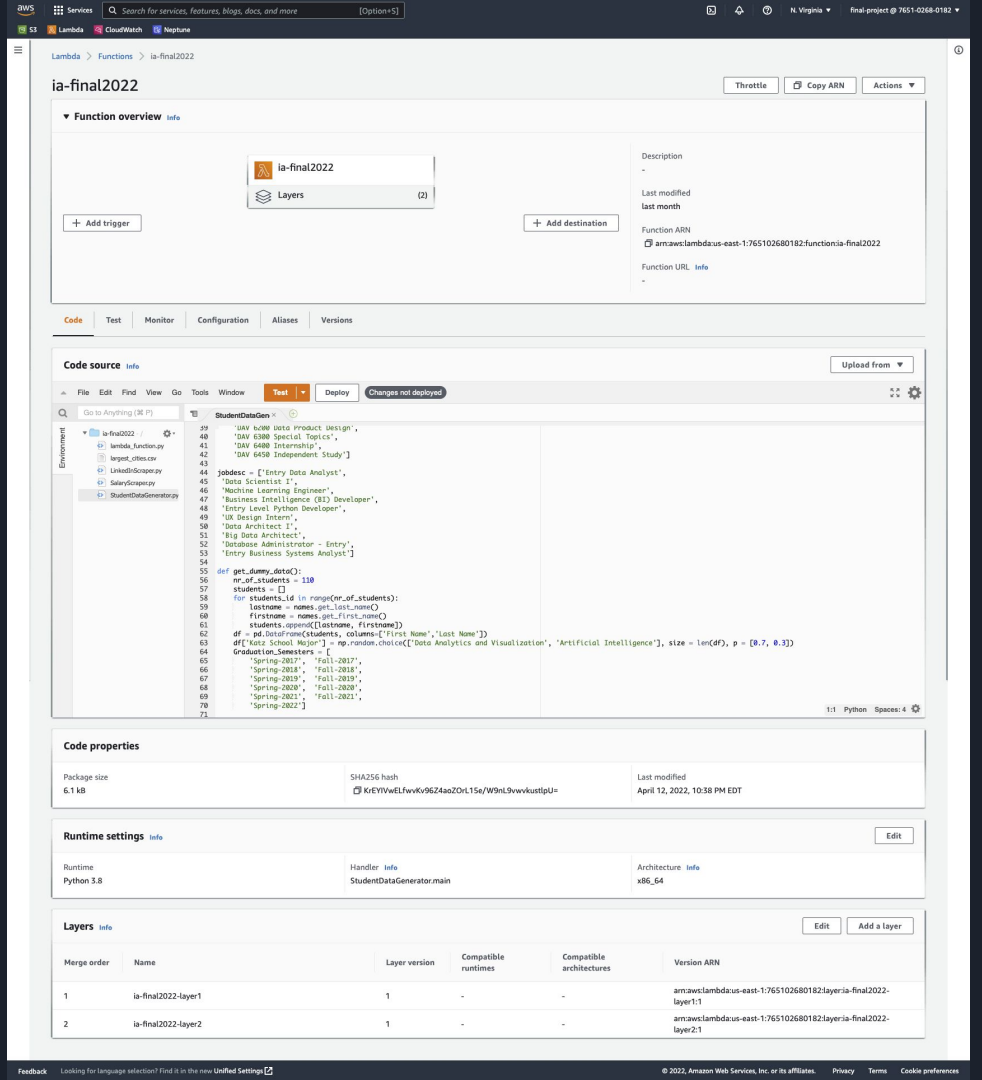
Lambda Functions and Layers

Email notification

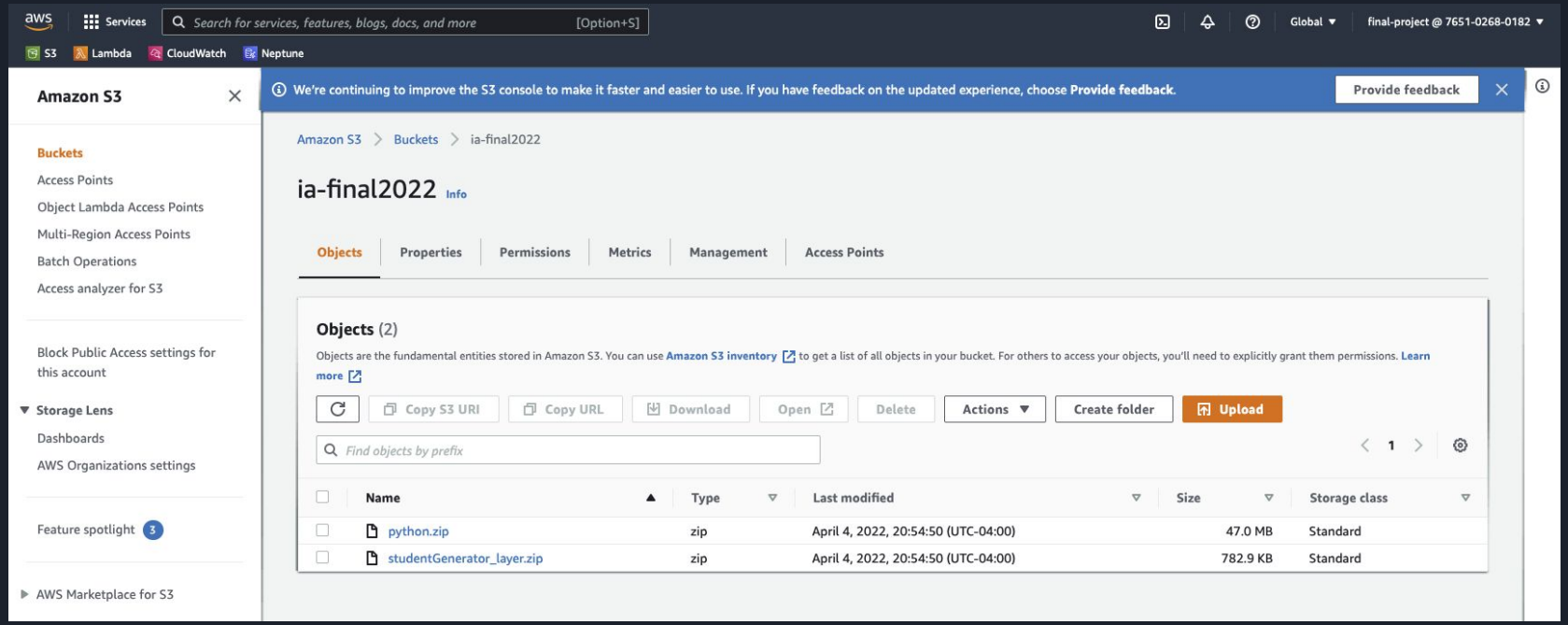
AWS Notifications May 2

to me

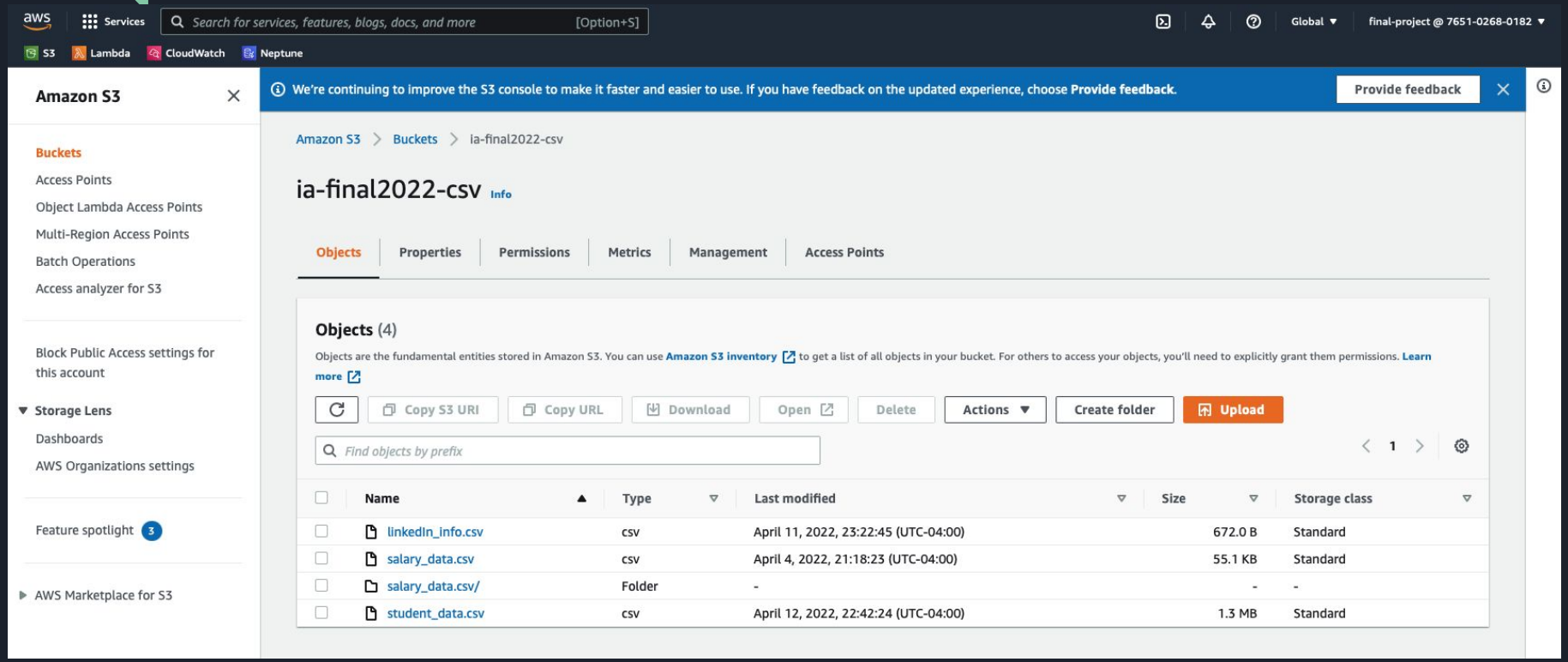
{"Records":[{"eventVersion":"2.1","eventSource":"aws:s3","awsRegion":"us-east-1","eventTime":"2022-05-02T21:42:54.961Z","eventName":"ObjectCreated:Put","userIdentity":{"principalId":"AWS:AROA3EI5MNB3PX3HEXU4Q:GlueJobRunnerSession"},"requestParameters":{"sourceIPAddress":"54.198.162.35"},"responseElements":{"x-amz-request-id":"JRPX0JZ3JA42YJNS","x-amz-id-2":"TYimg/zgfkaADCSkhRVbkriymPRGNIZ1rchQB8QJbxXY5G9uFXuDfLeNHBxcxpI5cPUqV3PCrNXcGvZLOLqLUgp2S/d4Wy2r"},"s3":{"s3SchemaVersion":"1.0","configurationId":"SendEmail","bucket":{"name":"ia-final2022-csv","ownerIdentity":{"principalId":"A2KZCXEDR9BHX9"},"arn":"arn:aws:s3:::ia-final2022-csv"},"object":{"key":"salary_data.csv/run-1651527756696-part-r-00000","size":56574,"eTag":"d3bdd90fdb7206c8dcd3ce402e3baa9e","sequencer":"006270505EDC728E9F"}}}]}



S3 Bucket
Python packages & ‘names’ generator package

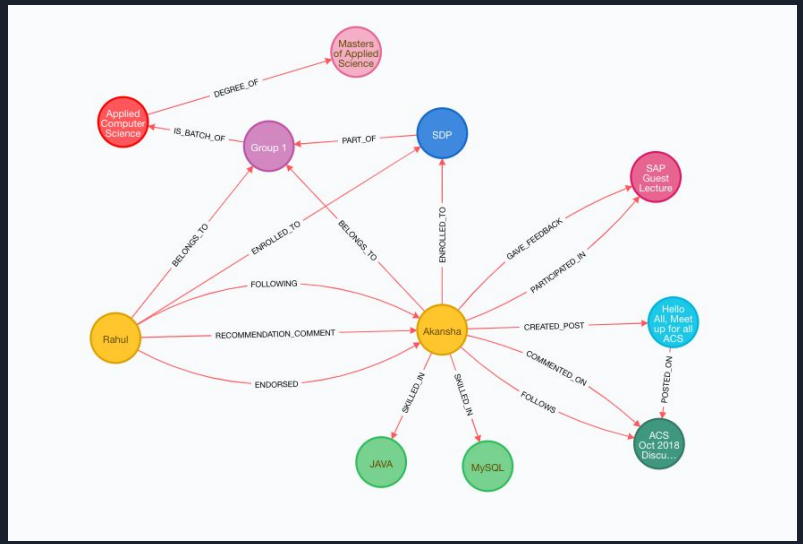


S3 Bucket
Dummy, LinkedIn, Salary.com



Graph Databases(Neptune)

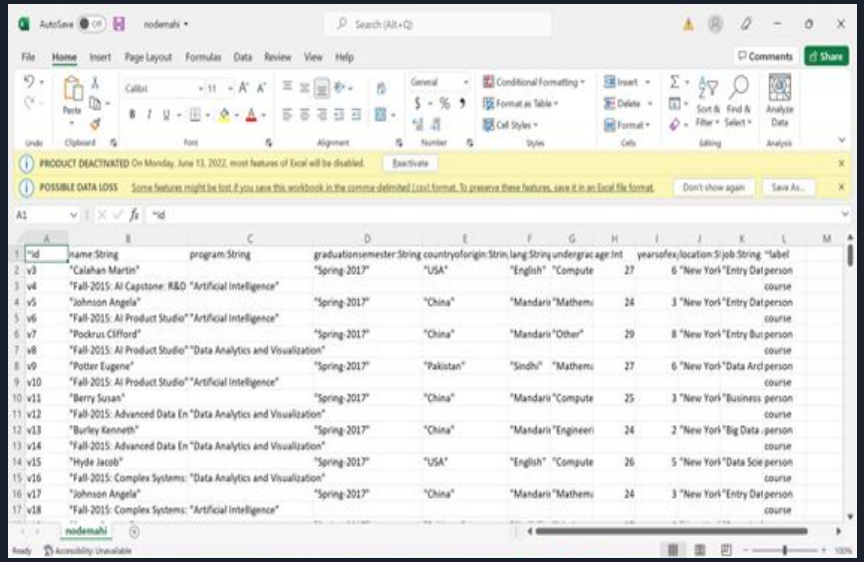
- **Used to create relationship between data.**
- **Store data as Vertices of graph and the relationships as edge.**



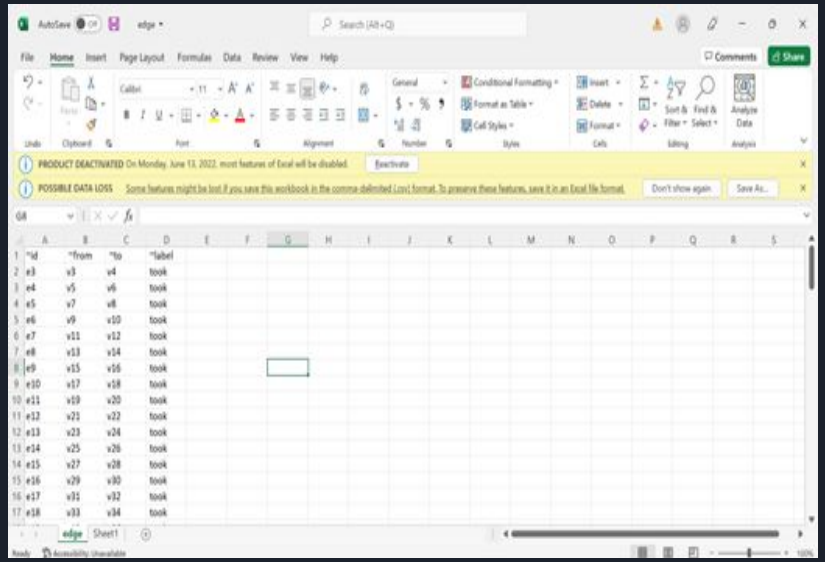
Graph Database

Several nodes exist in the graph database which allow for edges or relationships across students. For example, Rahul is a student who is following Akansha. Both Rahul and Akansha below to Group 1 as part of the Applied Computer Science Class, which is part of the Master of Applied Science Program.

Neptune



~id	name:String	program:String	graduationsemester:String	countryoforigin:Strin	lang:String	undergrac	age:Int	yearsofex	location:S	job:String	~label
v3	"Calahan Martin"		"Spring-2017"	"USA"	"English"	"Compute	27	6	"New York"	"Entry Dat	person
v4	"Fall-2015: AI Capstone: R&D	"Artificial Intelligence"									course
v5	"Johnson Angela"		"Spring-2017"	"China"	"Mandarin	"Mathem	24	3	"New York"	"Entry Dat	person
v6	"Fall-2015: AI Product Studio"	"Artificial Intelligence"									course
v7	"Pockrus Clifford"		"Spring-2017"	"China"	"Mandarin	"Other"	29	8	"New York"	"Entry Bui	person
v8	"Fall-2015: AI Product Studio"	"Data Analytics and Visualization"									course
v9	"Potter Eugene"		"Spring-2017"	"Pakistan"	"Sindhi"	"Mathem	27	6	"New York"	"Data Arcl	person
v10	"Fall-2015: AI Product Studio"	"Artificial Intelligence"									course
v11	"Berry Susan"		"Spring-2017"	"China"	"Mandarin	"Compute	25	3	"New York"	"Business	person
v12	"Fall-2015: Advanced Data En	"Data Analytics and Visualization"									course
v13	"Burley Kenneth"		"Spring-2017"	"China"	"Mandarin	"Engineeri	24	2	"New York"	"Big Data	person
v14	"Fall-2015: Advanced Data En	"Data Analytics and Visualization"									course
v15	"Hyde Jacob"		"Spring-2017"	"USA"	"English"	"Compute	26	5	"New York"	"Data Scie	person
v16	"Fall-2015: Complex Systems:	"Data Analytics and Visualization"									course
v17	"Johnson Angela"		"Spring-2017"	"China"	"Mandarin	"Mathem	24	3	"New York"	"Entry Dat	person
v18	"Fall-2015: Complex Systems:	"Artificial Intelligence"									course

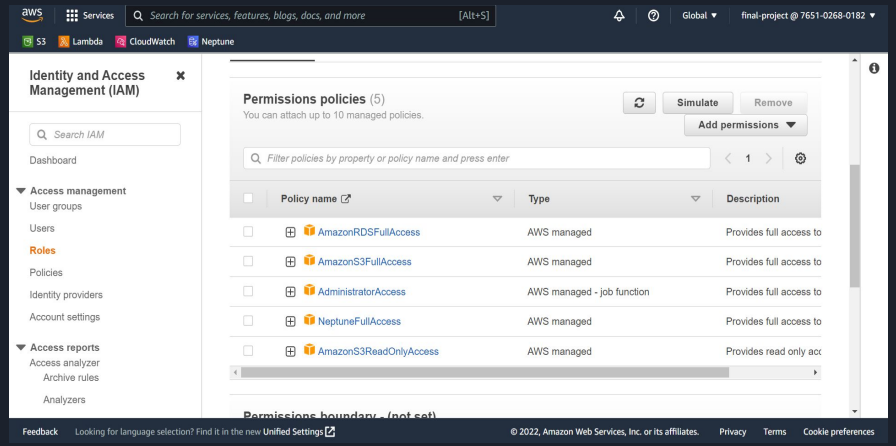


~id	~from	~to	~label
e3	v3	v4	took
e4	v5	v6	took
e5	v7	v8	took
e6	v9	v10	took
e7	v11	v12	took
e8	v13	v14	took
e9	v15	v16	took
e10	v17	v18	took
e11	v19	v20	took
e12	v21	v22	took
e13	v23	v24	took
e14	v25	v26	took
e15	v27	v28	took
e16	v29	v30	took
e17	v31	v32	took
e18	v33	v34	took

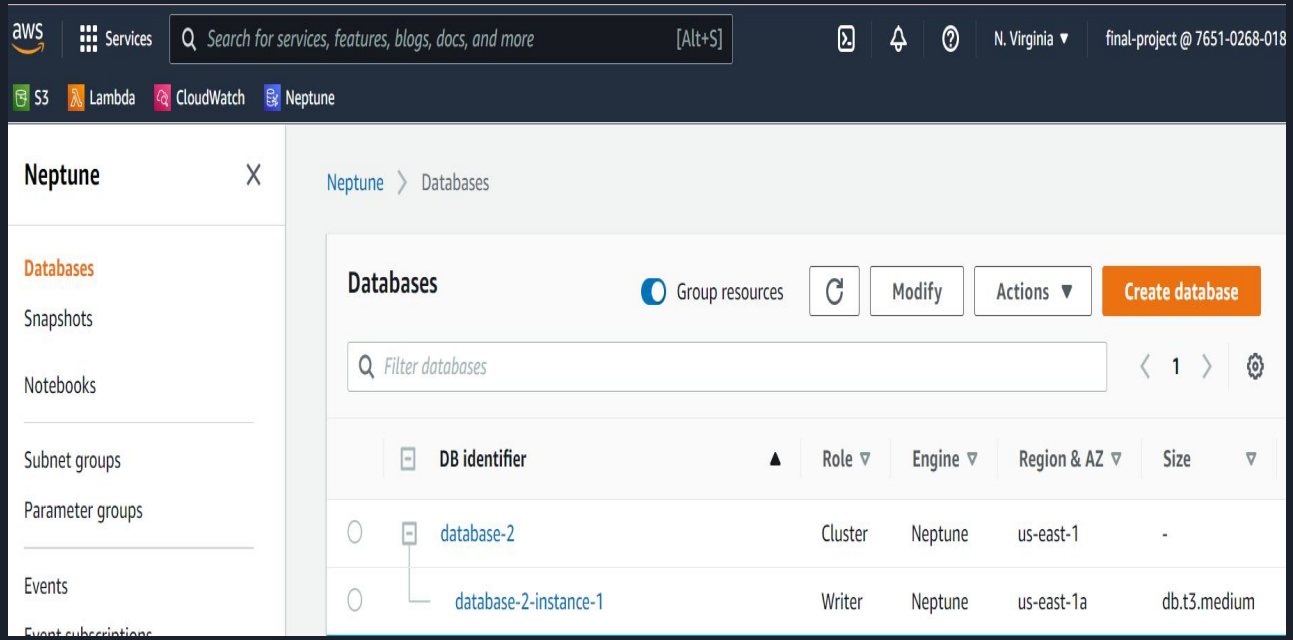
Create Neptune cluster

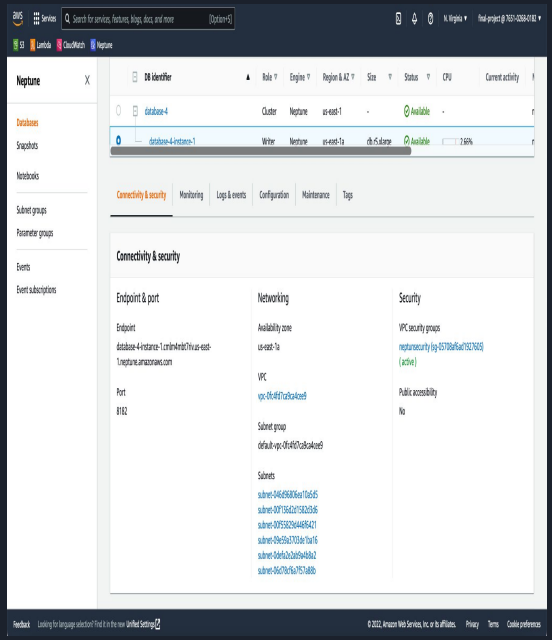
Creating new database cluster

- Setup VPC where the database cluster is located.
- Setup the security group
- Add IAM role

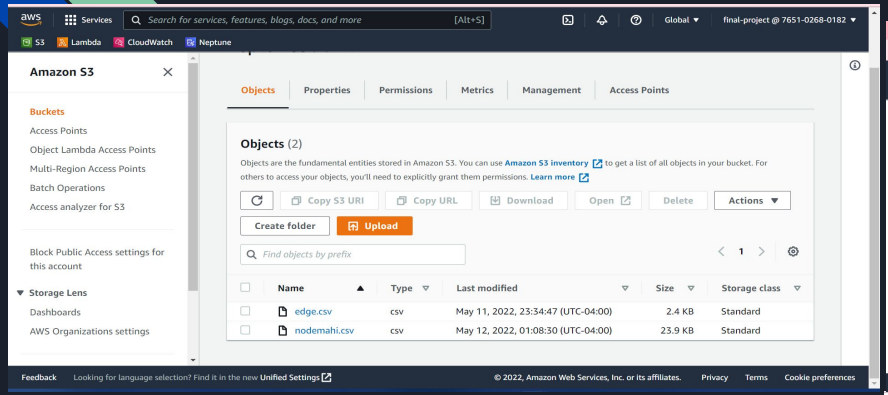


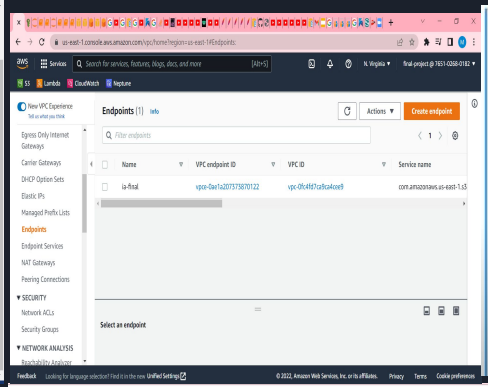
Neptune Database

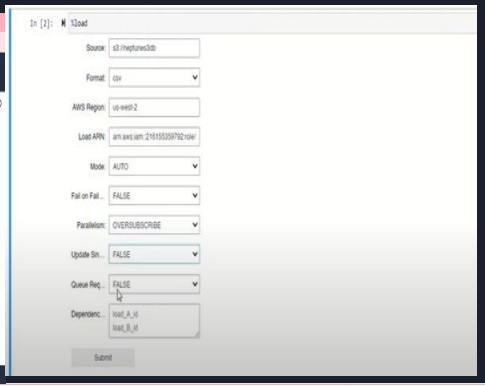


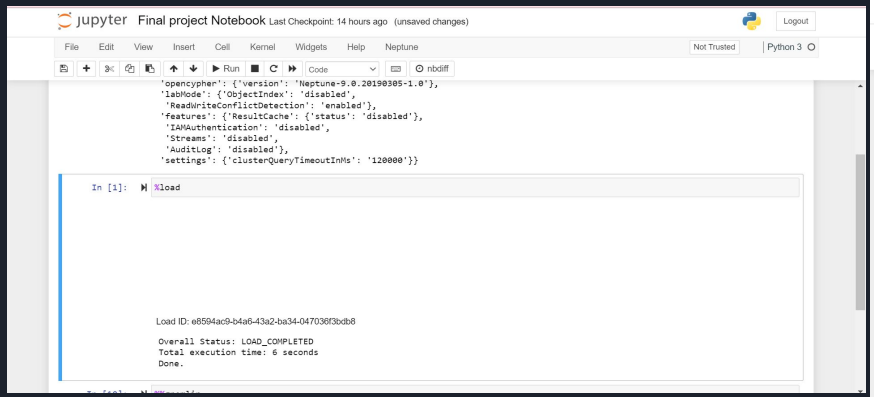


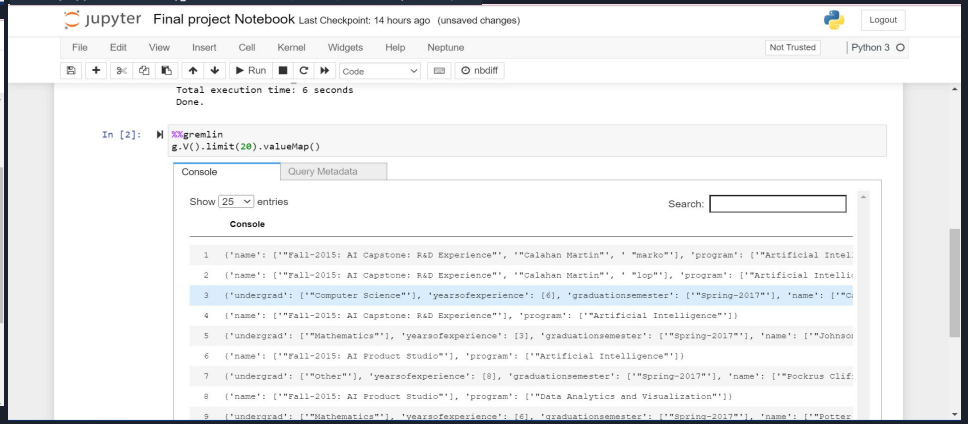
Load data from S3 to Neptune





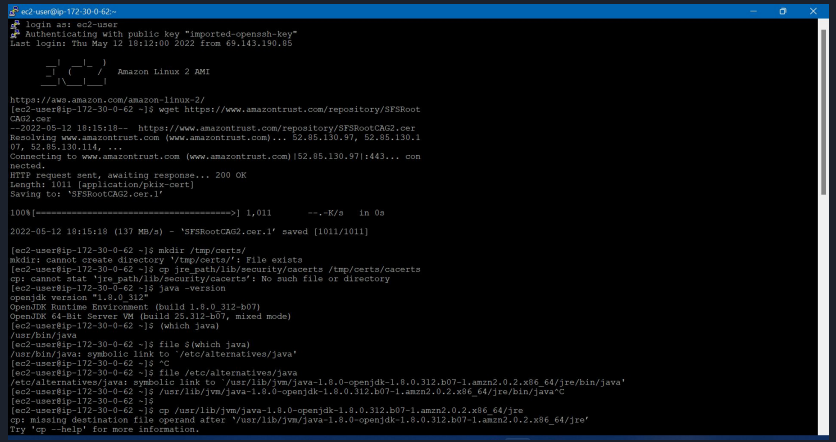


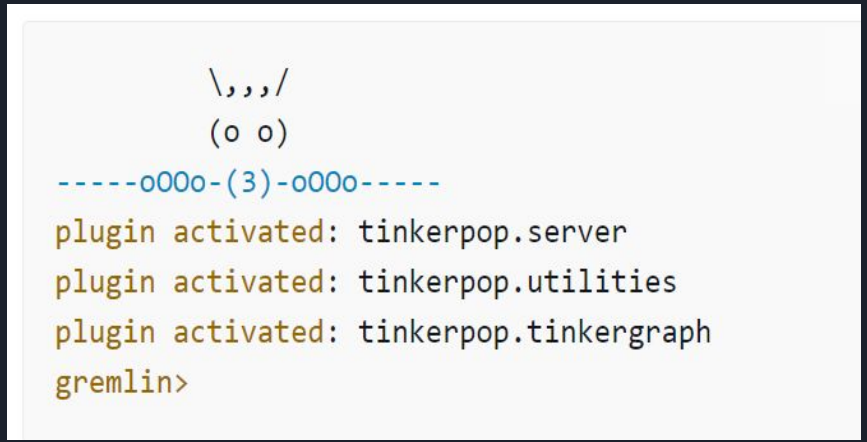




Install the Gremlin Console and Connect to Neptune

- ❏ Create EC2 instance with key pair
- ❏ Install Java 8 on EC2 instance
- ❏ wgethttps://archive.apache.org/dist/tinkerpop/3.5.2/apache-tinkerpop-gremlin-console-3.5.2-bin.zip



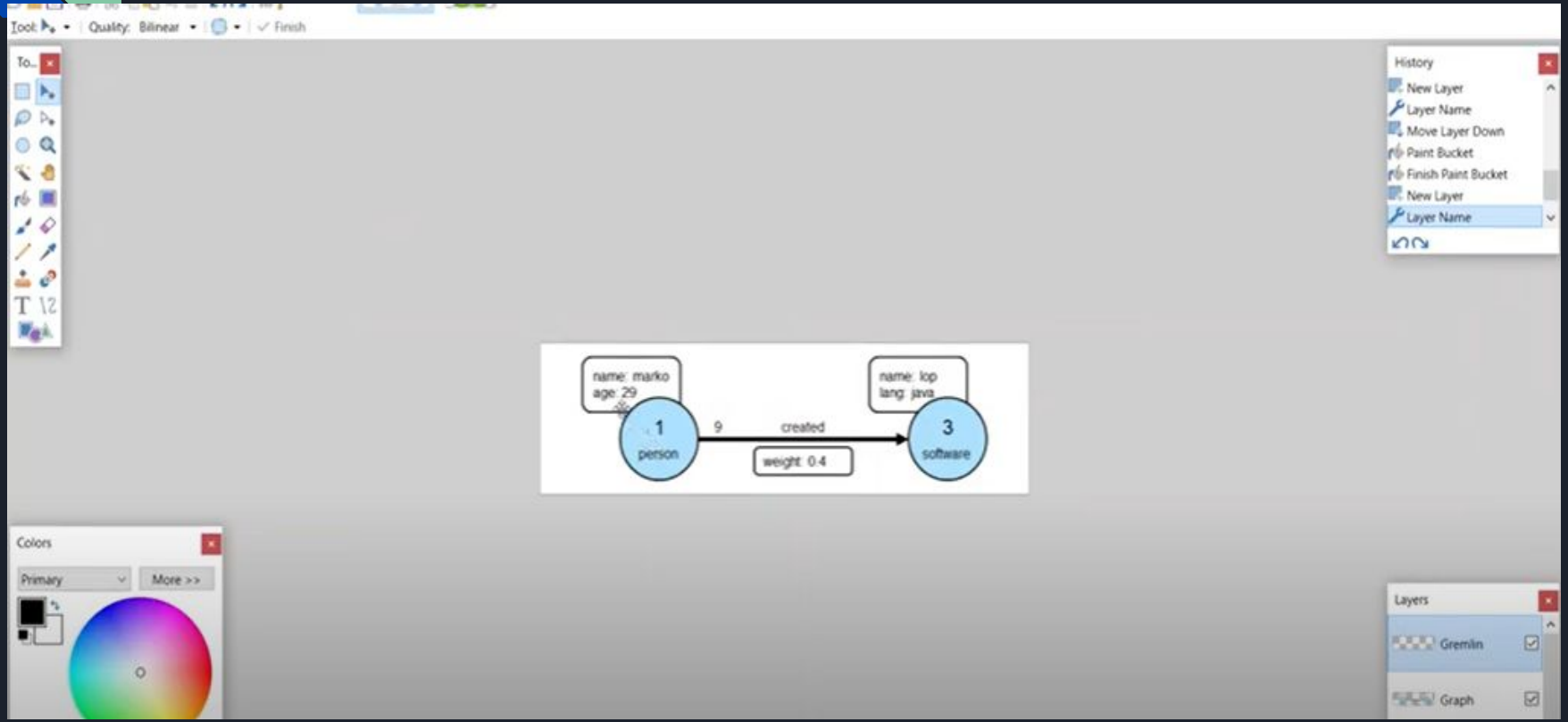


Create Vertex and Edge for Gremlin

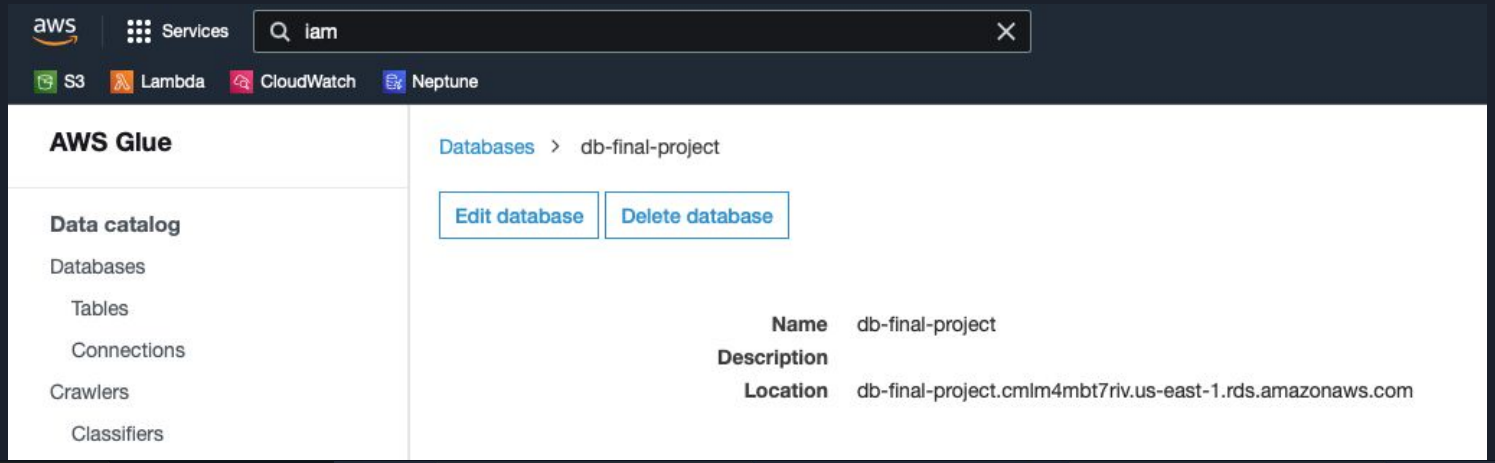
File Edit Format View Help

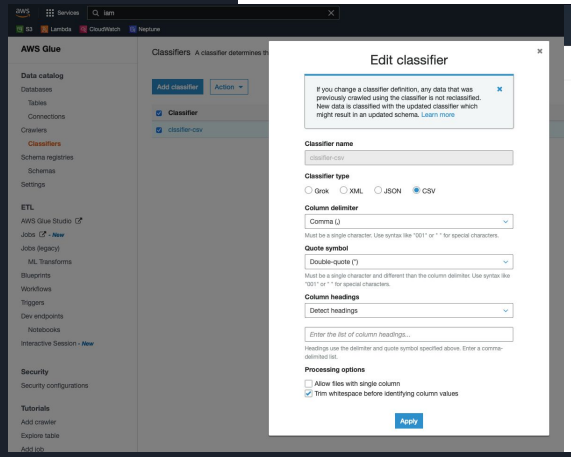
```
v1 = g.addV('person')\
      .property(id, 1)\
      .property('name', 'marko')\
      .property('age', 29)\
      .next()

// Create the "software" vertex and assign it to "v2"

v2 = g.addV('software')\
      .property(id, 3)\
      .property('name', 'lop')\
      .property('lang', 'java')\
      .next()

// Create the "created" edge from "v1" to "v2"

g.AddE('created')\
 .from(v1)\
 .to(v2)\
 .property(id, 9)\
```

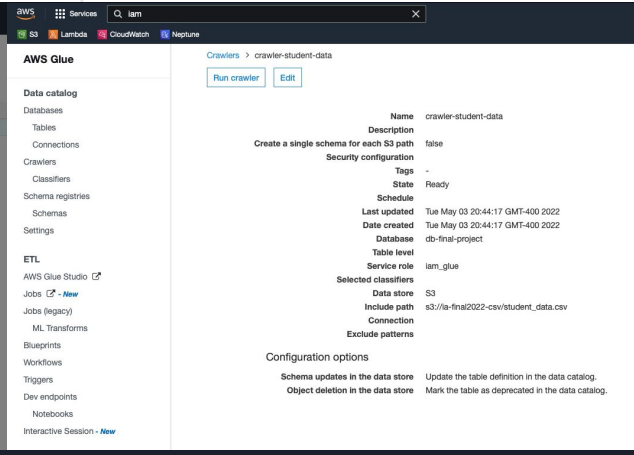
Con’t…

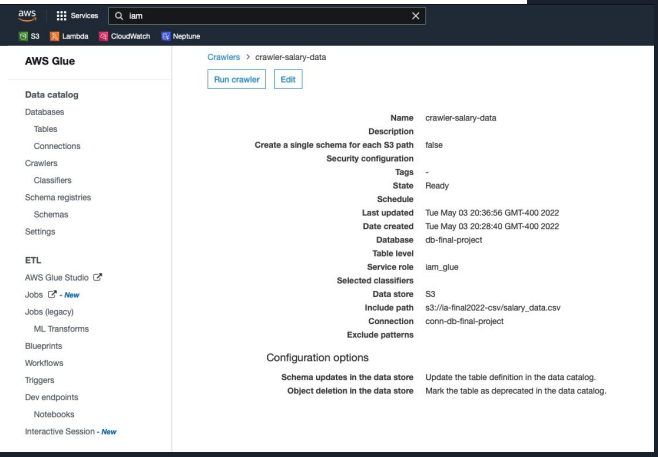

Tool:
Quality: Bilinear
Finish
History
New Layer
Layer Name
Move Layer Down
Paint Bucket
Finish Paint Bucket
New Layer
Layer Name
name: marko
age: 29
1
person
9
created
weight: 0.4
name: lop
lang: java
3
software
Colors
Primary
More >>
Layers
Gremlin
Graph

Glue

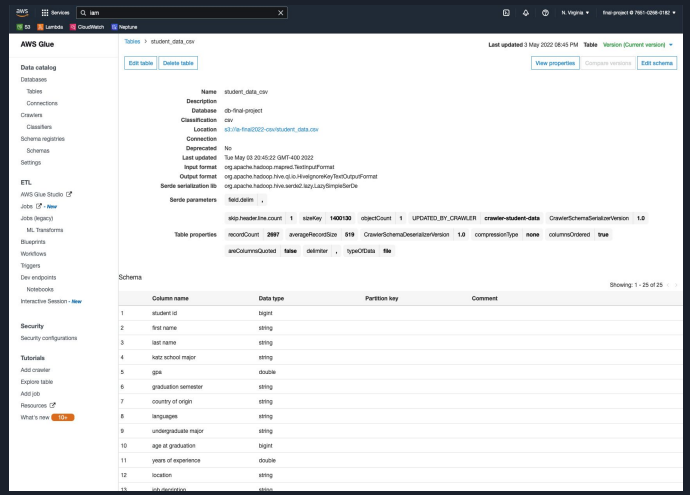

aws
Services
iam
S3
Lambda
CloudWatch
Neptune
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Databases > db-final-project
Edit database
Delete database
Name db-final-project
Description
Location db-final-project.cmlm4mbt7riv.us-east-1.rds.amazonaws.com

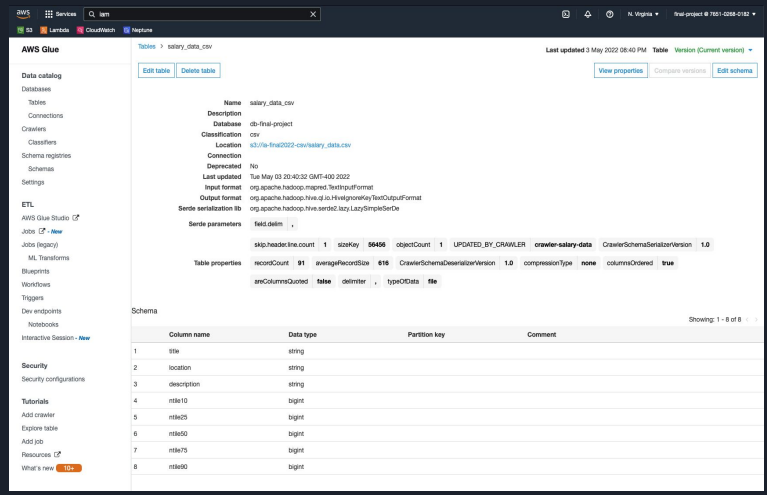

Edit classifier
Classifier name
Classifier type
Grok
XML
JSON
CSV
Column delimiter
Comma (,)
Quote symbol
Double-quote (")
Column headings
Detect headings
Processing options
Allow files with single column
Trim whitespace before identifying column values
Apply


Crawlers > crawler-student-data
Run crawler
Edit
Name crawler-student-data
Description
Create a single schema for each S3 path false
Security configuration
Tags -
State Ready
Schedule
Last updated Tue May 03 20:44:17 GMT-400 2022
Date created Tue May 03 20:44:17 GMT-400 2022
Database db-final-project
Table level
Service role iam_glue
Selected classifiers
Data store S3
Include path s3://ia-final2022-csv/student_data.csv
Connection
Exclude patterns
Configuration options
Schema updates in the data store Update the table definition in the data catalog.
Object deletion in the data store Mark the table as deprecated in the data catalog.


Crawlers > crawler-salary-data
Run crawler
Edit
Name crawler-salary-data
Description
Create a single schema for each S3 path false
Security configuration
Tags -
State Ready
Schedule
Last updated Tue May 03 20:36:56 GMT-400 2022
Date created Tue May 03 20:28:40 GMT-400 2022
Database db-final-project
Table level
Service role iam_glue
Selected classifiers
Data store S3
Include path s3://ia-final2022-csv/salary_data.csv
Connection conn-db-final-project
Exclude patterns
Configuration options
Schema updates in the data store Update the table definition in the data catalog.
Object deletion in the data store Mark the table as deprecated in the data catalog.

Glue


AWS Glue
Tables > student_data_csv
Last updated 3 May 2022 08:45 PM Table Version (Current version)
Edit table
Delete table
View properties
Compare versions
Edit schema
Name student_data_csv
Description
Database db-final-project
Classification csv
Location s3://a-final2022-csv/student_data.csv
Connection
Deprecated No
Last updated Tue May 03 20:45:22 GMT-400 2022
Input format org.apache.hadoop.mapred.TextInputFormat
Output format org.apache.hadoop.hive.ql.io.HivelgnoreKeyTextOutputFormat
Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters field.delim ,
skip.header.line.count 1 sizeKey 1400130 objectCount 1 UPDATED_BY_CRAWLER crawler-student-data CrawlerSchemaSerializerVersion 1.0
Table properties recordCount 2697 averageRecordSize 519 CrawlerSchemaDeserializerVersion 1.0 compressionType none columnsOrdered true
areColumnsQuoted false delimiter , typeOfData file
Schema
Showing: 1 - 25 of 25
Column name
Data type
Partition key
Comment
1 student id bigint
2 first name string
3 last name string
4 katz school major string
5 gpa double
6 graduation semester string
7 country of origin string
8 languages string
9 undergraduate major string
10 age at graduation bigint
11 years of experience double
12 location string


AWS Glue
Tables > salary_data_csv
Last updated 3 May 2022 08:40 PM Table Version (Current version)
Edit table
Delete table
View properties
Compare versions
Edit schema
Name salary_data_csv
Description
Database db-final-project
Classification csv
Location s3://a-final2022-csv/salary_data.csv
Connection
Deprecated No
Last updated Tue May 03 20:40:32 GMT-400 2022
Input format org.apache.hadoop.mapred.TextInputFormat
Output format org.apache.hadoop.hive.ql.io.HivelgnoreKeyTextOutputFormat
Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters field.delim ,
skip.header.line.count 1 sizeKey 56456 objectCount 1 UPDATED_BY_CRAWLER crawler-salary-data CrawlerSchemaSerializerVersion 1.0
Table properties recordCount 91 averageRecordSize 616 CrawlerSchemaDeserializerVersion 1.0 compressionType none columnsOrdered true
areColumnsQuoted false delimiter , typeOfData file
Schema
Showing: 1 - 8 of 8
Column name
Data type
Partition key
Comment
1 title string
2 location string
3 description string
4 title10 bigint
5 title25 bigint
6 title50 bigint
7 title75 bigint
8 title90 bigint

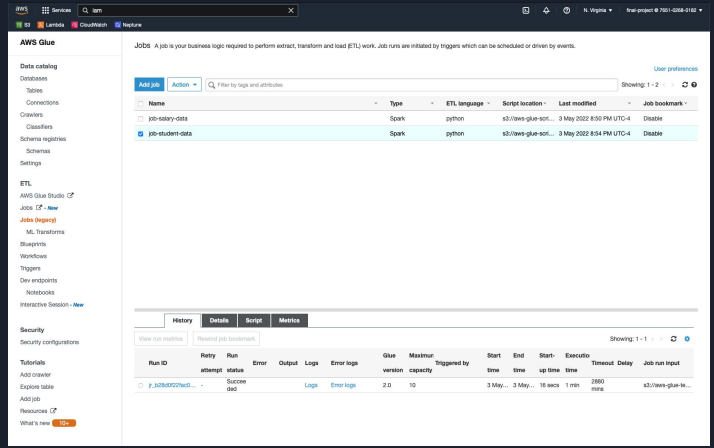
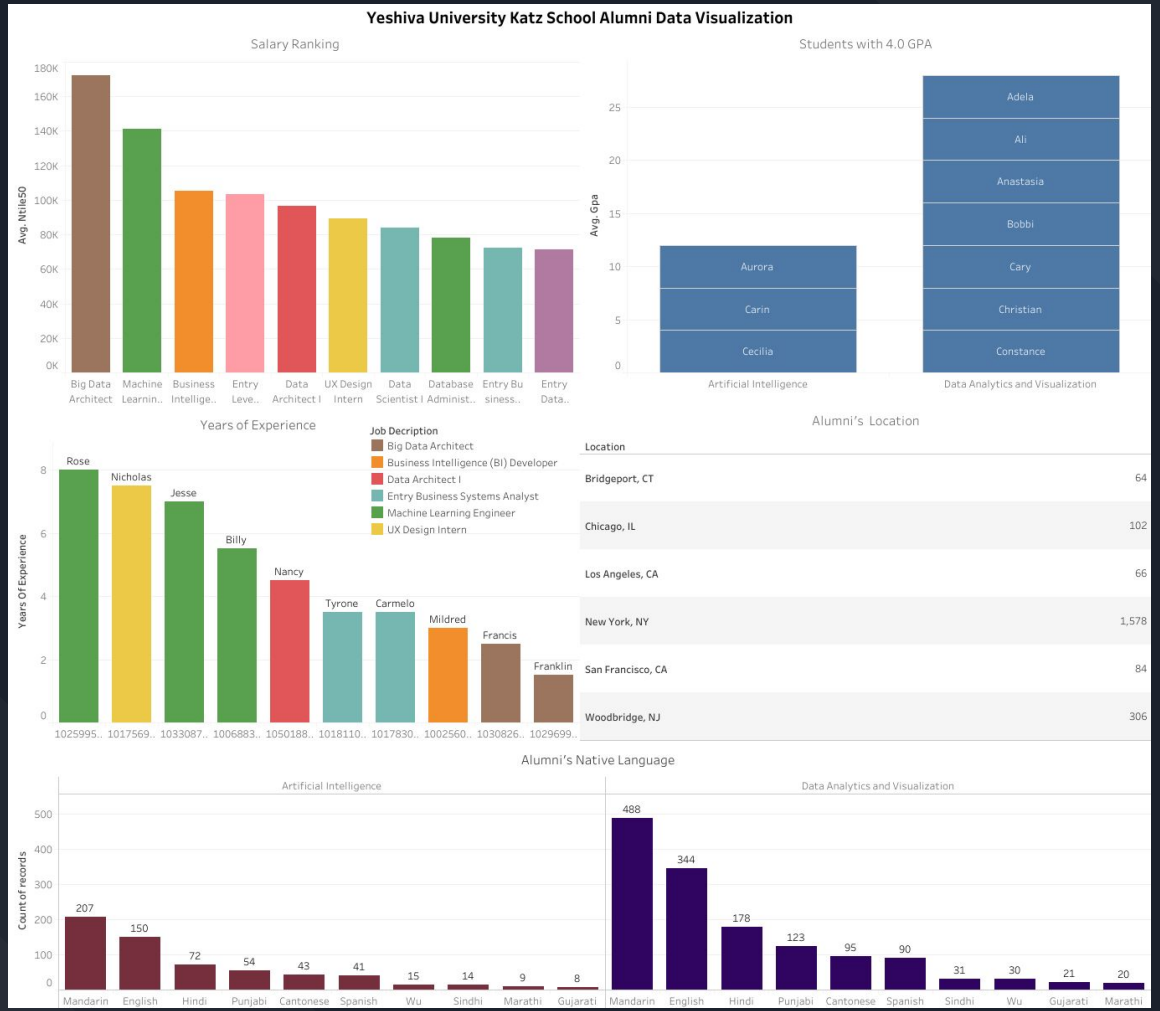

AWS Glue
Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.
Add job
Action
Name
Type
ETL language
Script location
Last modified
Job bookmark
job-salary-data Spark python s3://aws-glue-scri... 3 May 2022 8:50 PM UTC-4 Disable
job-student-data Spark python s3://aws-glue-scri... 3 May 2022 8:54 PM UTC-4 Disable
History
Details
Script
Metrics
Run ID
Retry attempt
Run status
Error
Output
Logs
Error logs
Glue version
Maximum capacity
Triggered by
Start time
End time
Start-up time
Execution time
Timeout
Delay
Job run input

Tableau Dashboard



Challenges and Solutions

Challenges

Lambda function execution

- Timeout Limit
- Package Size Limit
- Incompatible Dependencies

Neptune graph database

- Node and edge selection
- The connection between s3 and neptune database and notebook.
- Dataset loading into JupyterLab notebook
- CA Certificate failure

Solutions

Lambda function execution

- Extend the timeout to 15 mins
- Clean the script code
- Run time setting from 3.9 to 3.8

Neptune graph database

- Identify entities and relationships
- Create endpoint,add “all traffic” to inbound and outbound and connect to relevant VPC
- Convert dataset format to Gremlin csv
- Create the connection with gremlin.bat console.

Thanks for listening!

Any comments/questions?