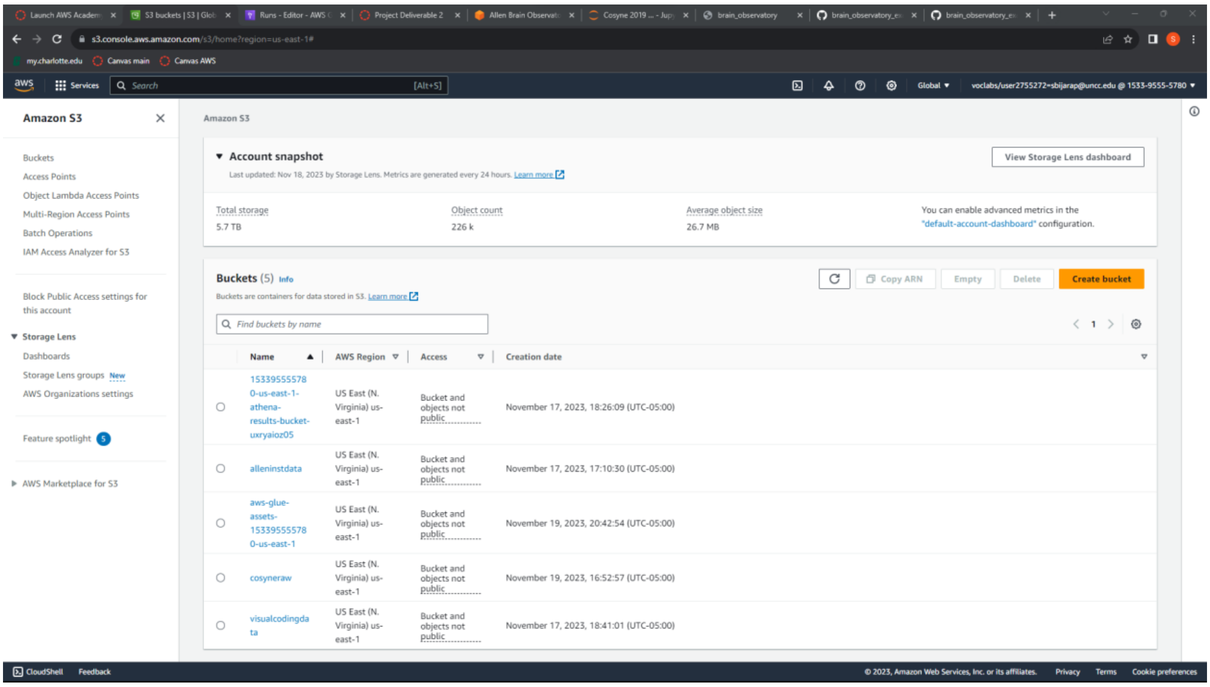
Project Deliverable 2

Team 10:
Sanjana
Tarun
Swathi
Archana
Shashank

Set Up AWS Environment - Account Setup: Team Access to AWS Learner Lab and Github, create IAM Roles

https://github.com/RoshnaSanjanaKommareddy/Cloud_Project_10#cloud_project_10.
The whole team had access to AWS Learner Lab and IAM roles were created.

S3 Data Storage:
Create an S3 bucket to store the dataset(s)
Upload data to the S3 Bucket



Account snapshot

Last updated: Nov 18, 2023 by Storage Lens. Metrics are generated every 24 hours. Learn more

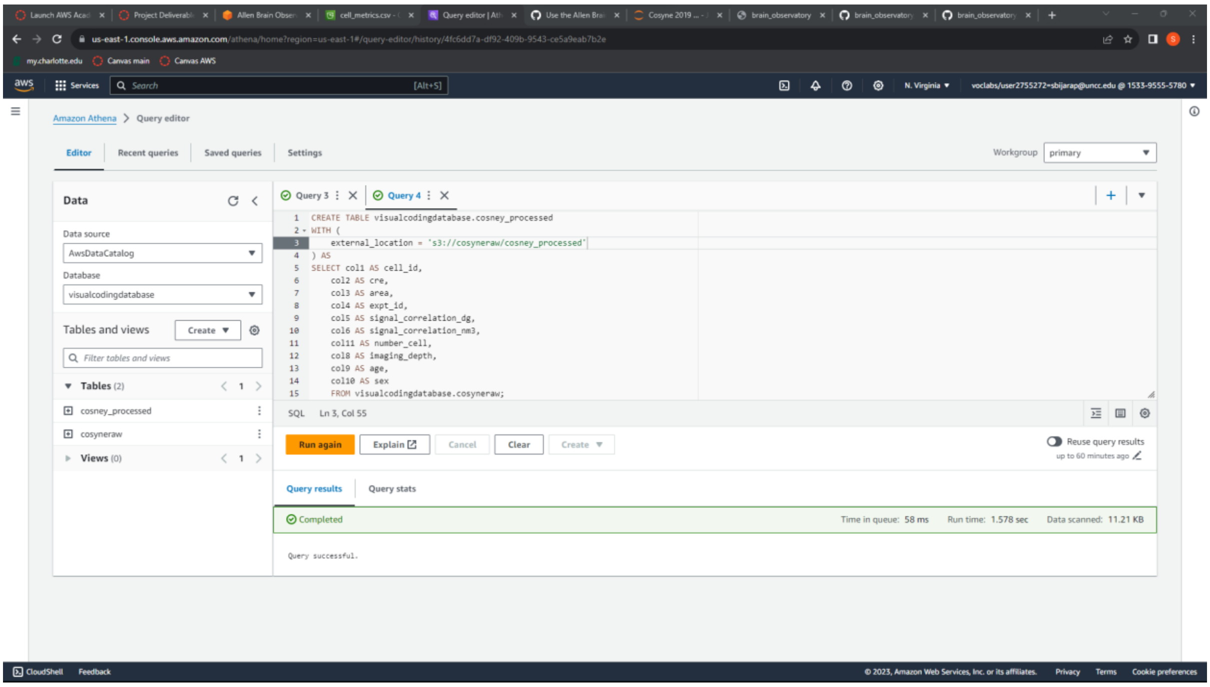
Total storage	Object count	Average object size	
5.7 TB	226 k	26.7 MB	You can enable advanced metrics in the "default-account-dashboard" configuration.

Buckets (5) Info

Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	153395555780-us-east-1-athena-results-bucket-uxryaioz05	US East (N. Virginia) us-east-1	Bucket and objects not public	November 17, 2023, 18:26:09 (UTC-05:00)
○	alleninstdata	US East (N. Virginia) us-east-1	Bucket and objects not public	November 17, 2023, 17:10:30 (UTC-05:00)
○	aws-glue-assets-153395555780-us-east-1	US East (N. Virginia) us-east-1	Bucket and objects not public	November 19, 2023, 20:42:54 (UTC-05:00)
○	cosyneraw	US East (N. Virginia) us-east-1	Bucket and objects not public	November 19, 2023, 16:52:57 (UTC-05:00)
○	visualcodingdata	US East (N. Virginia) us-east-1	Bucket and objects not public	November 17, 2023, 18:41:01 (UTC-05:00)

Data Exploration for Insight and Pre-processing
Use Amazon Athena to query the transformed data.
Use SQL queries for meaningful insights from the dataset for data exploration
Create visualizations using Amazon QuickSight.

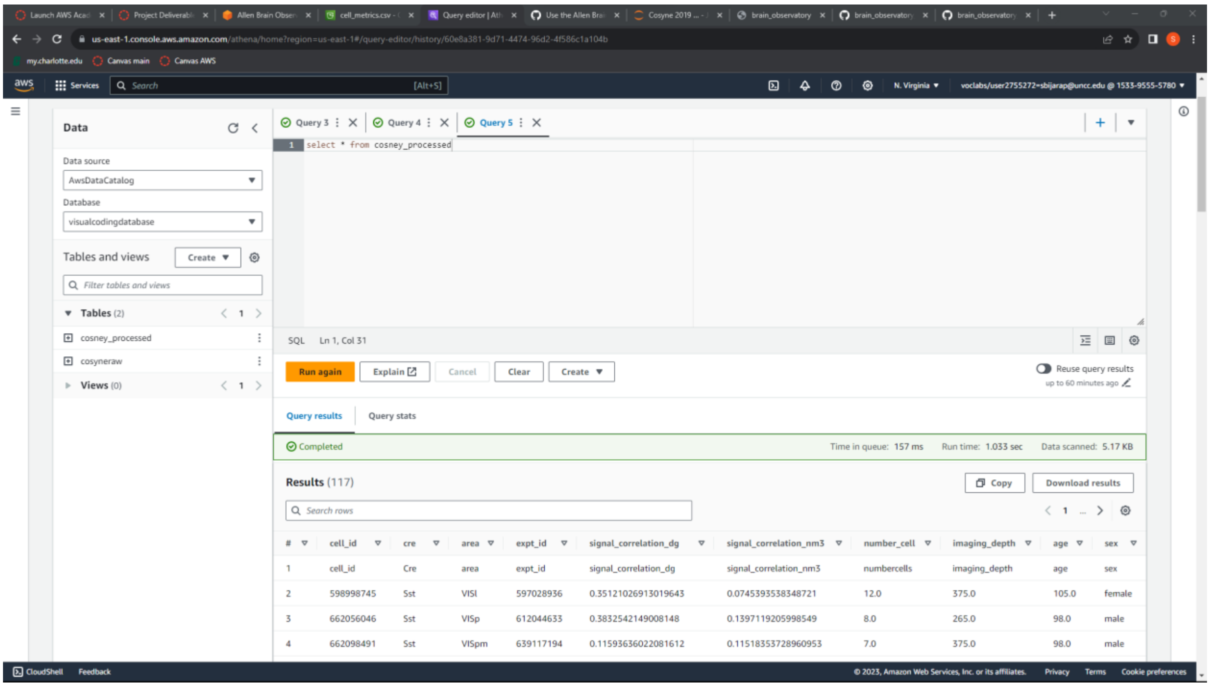


```sql
CREATE TABLE visualcodingdatabase.cosney_processed
WITH (
    external_location = 's3://cosyneraw/cosney_processed'
) AS
SELECT col1 AS cell_id,
    col2 AS cre,
    col3 AS area,
    col4 AS expt_id,
    col5 AS signal_correlation_dg,
    col6 AS signal_correlation_nm3,
    col11 AS number_cell,
    col8 AS imaging_depth,
    col9 AS age,
    col10 AS sex
    FROM visualcodingdatabase.cosyneraw;
```

SQL Ln 3, Col 55

Completed Time in queue: 58 ms Run time: 1.578 sec Data scanned: 11.21 KB

```
Query successful.
```



```sql
select * from cosney_processed
```

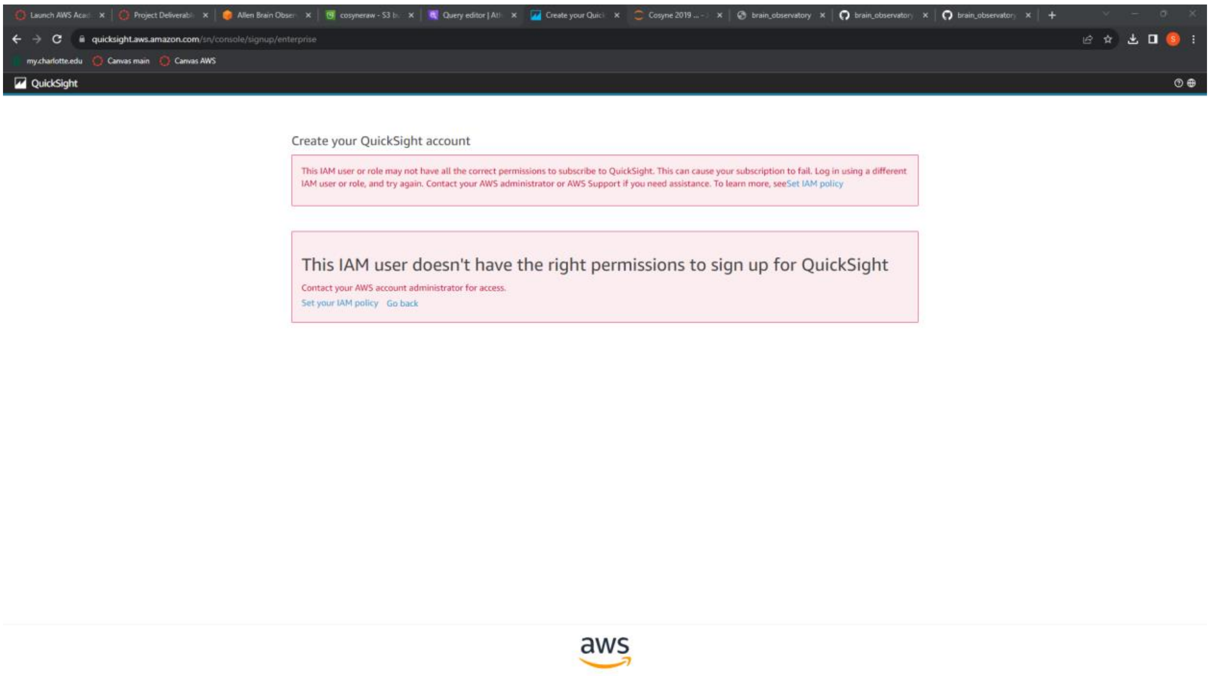
SQL Ln 1, Col 31

Completed Time in queue: 157 ms Run time: 1.033 sec Data scanned: 5.17 KB

Results (117)

#	cell_id	cre	area	expt_id	signal_correlation_dg	signal_correlation_nm3	number_cell	imaging_depth	age	sex
1	cell_id	Cre	area	expt_id	signal_correlation_dg	signal_correlation_nm3	numbercells	imaging_depth	age	sex
2	598998745	Sst	VISl	597028936	0.35121026913019643	0.0745393538348721	12.0	375.0	105.0	female
3	662056046	Sst	VISp	612044633	0.3832542149008148	0.1397119205998549	8.0	265.0	98.0	male
4	662098491	Sst	VISpm	639117194	0.11593636022081612	0.11518353728960953	7.0	375.0	98.0	male

Amazon Quicksight asked for a subscription account and even after accepting to pay it says that my role as the learner lab user restricts me from using Amazon QuickSIght



Create your QuickSight account

This IAM user or role may not have all the correct permissions to subscribe to QuickSight. This can cause your subscription to fail. Log in using a different IAM user or role, and try again. Contact your AWS administrator or AWS Support if you need assistance. To learn more, seeSet IAM policy

This IAM user doesn't have the right permissions to sign up for QuickSight

Contact your AWS account administrator for access.

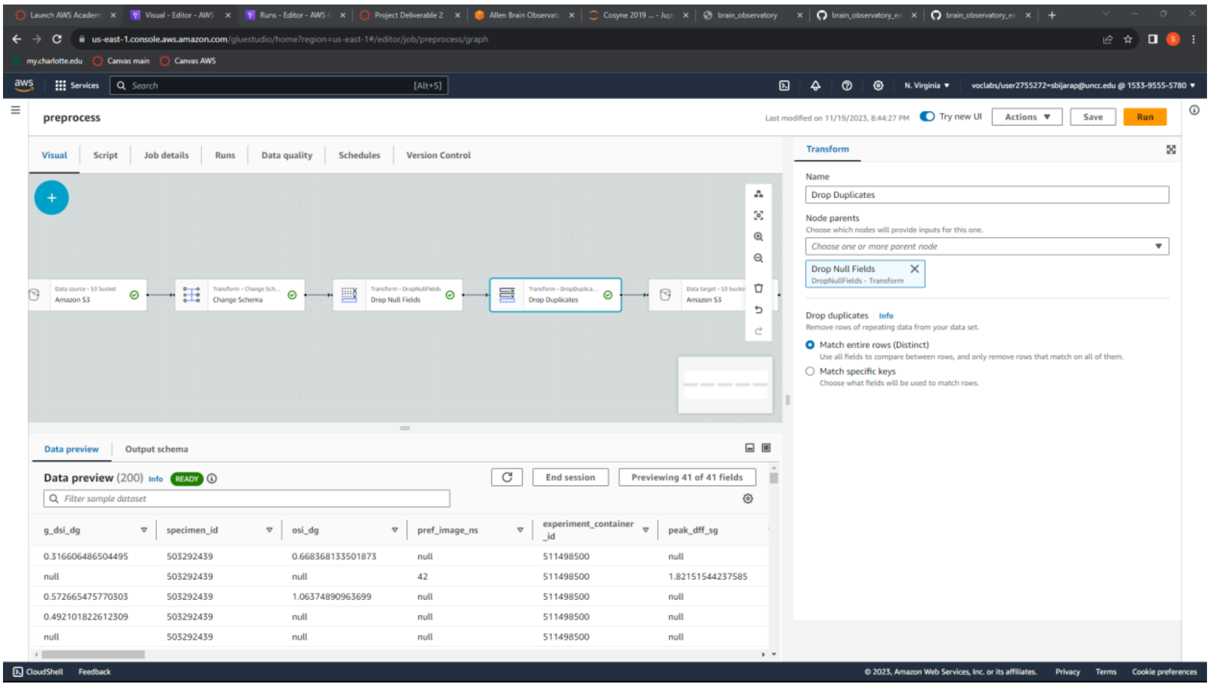
Set your IAM policy Go back

AWS Glue ETL Job:

Create an ETL job using AWS Glue to transform the dataset(s) in S3

Perform basic data transformations (e.g., filtering, aggregation, type conversions)

AWS Pipeline/Solution Chart



Name: Drop Duplicates

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Drop Null Fields
DropNullFields - Transform

Drop duplicates Info
Remove rows of repeating data from your data set.

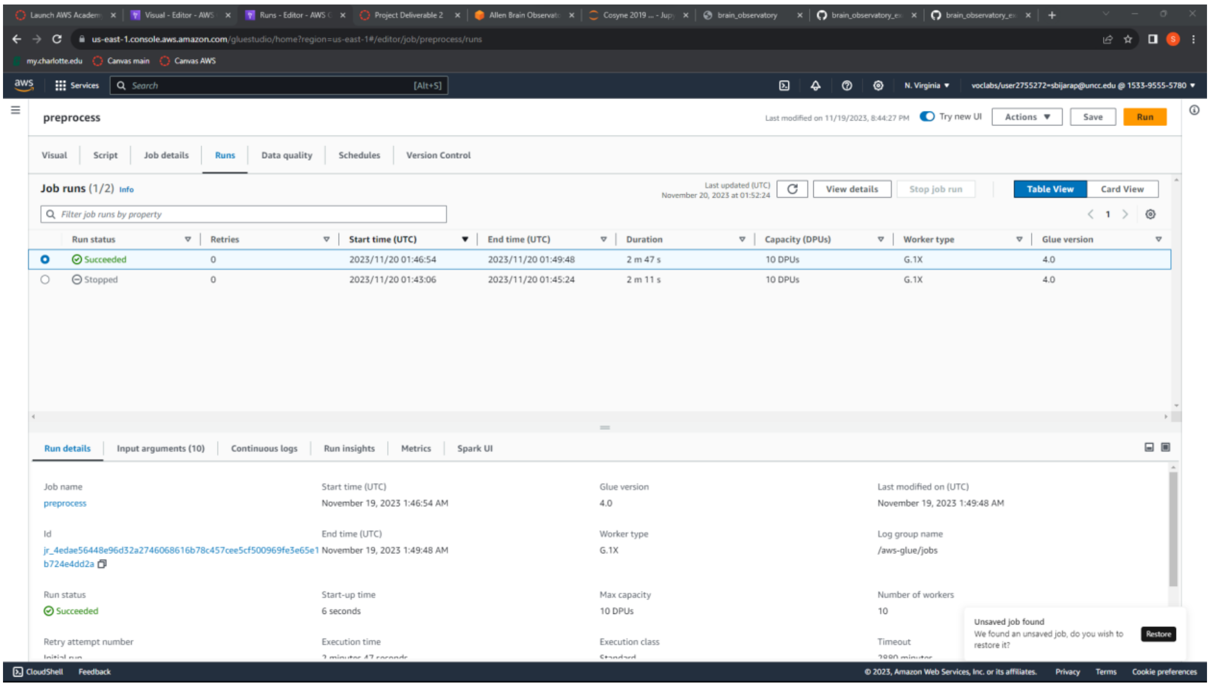
● Match entire rows (Distinct)
Use all fields to compare between rows, and only remove rows that match on all of them.

○ Match specific keys
Choose what fields will be used to match rows.

Data preview (200) Info READY

Filter sample dataset

g_dsi_dg	specimen_id	osi_dg	pref_image_ns	experiment_container_id	peak_dff_sg
0.316606486504495	503292439	0.668368133501873	null	511498500	null
null	503292439	null	42	511498500	1.82151544237585
0.572665475770303	503292439	1.06374890963699	null	511498500	null
0.492101822612309	503292439	null	null	511498500	null
null	503292439	null	null	511498500	null



preprocess

Last modified on 11/19/2023, 8:44:27 PM

Job runs (1/2)

Last updated (UTC)
November 20, 2023 at 01:52:24

Filter job runs by property

	Run status	Retries	Start time (UTC)	End time (UTC)	Duration	Capacity (DPUs)	Worker type	Glue version
●	Succeeded	0	2023/11/20 01:46:54	2023/11/20 01:49:48	2 m 47 s	10 DPUs	G.1X	4.0
○	Stopped	0	2023/11/20 01:43:06	2023/11/20 01:45:24	2 m 11 s	10 DPUs	G.1X	4.0

Job name preprocess	Start time (UTC) November 19, 2023 1:46:54 AM	Glue version 4.0	Last modified on (UTC) November 19, 2023 1:49:48 AM
Id jr_4edae56448e96d32a2746068616b78c457cee5cf500969fe3e65e1b724e4dd2a	End time (UTC) November 19, 2023 1:49:48 AM	Worker type G.1X	Log group name /aws-glue/jobs
Run status Succeeded	Start-up time 6 seconds	Max capacity 10 DPUs	Number of workers 10
Retry attempt number Initial run	Execution time 2 minutes 47 seconds	Execution class Standard	Timeout 2880 minutes

Unsaved job found
We found an unsaved job, do you wish to restore it?

© 2023, Amazon Web Services, Inc. or its affiliates.