

Project Deliverable 2

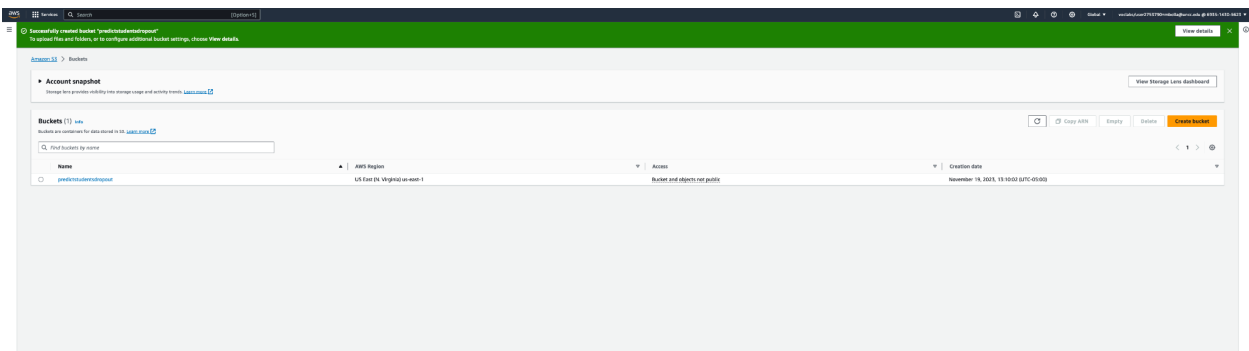
Github Repository -

<https://github.com/DelphineMeera/Predict-students-dropout-and-academic-success>

S3 Data Storage:

Create an S3 bucket to store the dataset(s)

Upload data to the S3 Bucket



Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 520.7 KB)

Remove

Add files

Add folder

All files and folders in this table will be uploaded.

< 1 >

<input checked="" type="checkbox"/>	Name	Folder	Type	Size
<input checked="" type="checkbox"/>	Predict student's dro...	-	text/csv	520.7 KB

Destination

Destination

[s3://predictstudentsdropout](#)

► Destination details

Bucket settings that impact new objects stored in the specified destination.

► Permissions

Grant public access and access to other AWS accounts.

► Properties

Specify storage class, encryption settings, tags, and more.

Cancel

Upload

Upload successful
View details below.

Upload: status

The information below will no longer be available after you navigate away from this page.

Summary

Destination	Successful	Failed
s3://predictstudentsdropout	1 file, 520.7 KB (100.0%)	0 files, 0 B (0%)

Files and folders

Configuration

Files and folders (1 Total, 520.7 KB)

Name	Folder	Type	Size	Status	Error
Predict student's dropout and academic success.csv	-	text/csv	520.7 KB	Successful	-

Data Exploration for Insight and Pre-processing

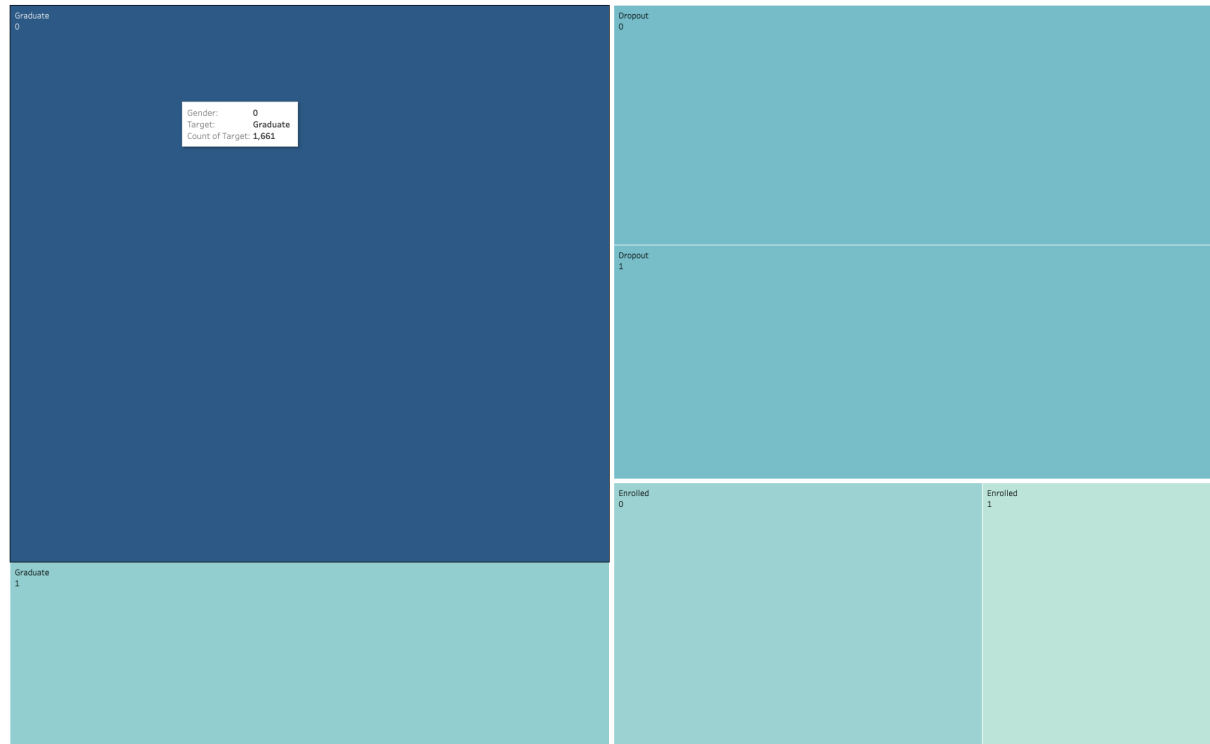
Use Amazon Athena to query the transformed data.

Use SQL queries for meaningful insights from the dataset for data exploration

Create visualizations using Amazon QuickSight.

The descriptions to the insights are mentioned below the insights:

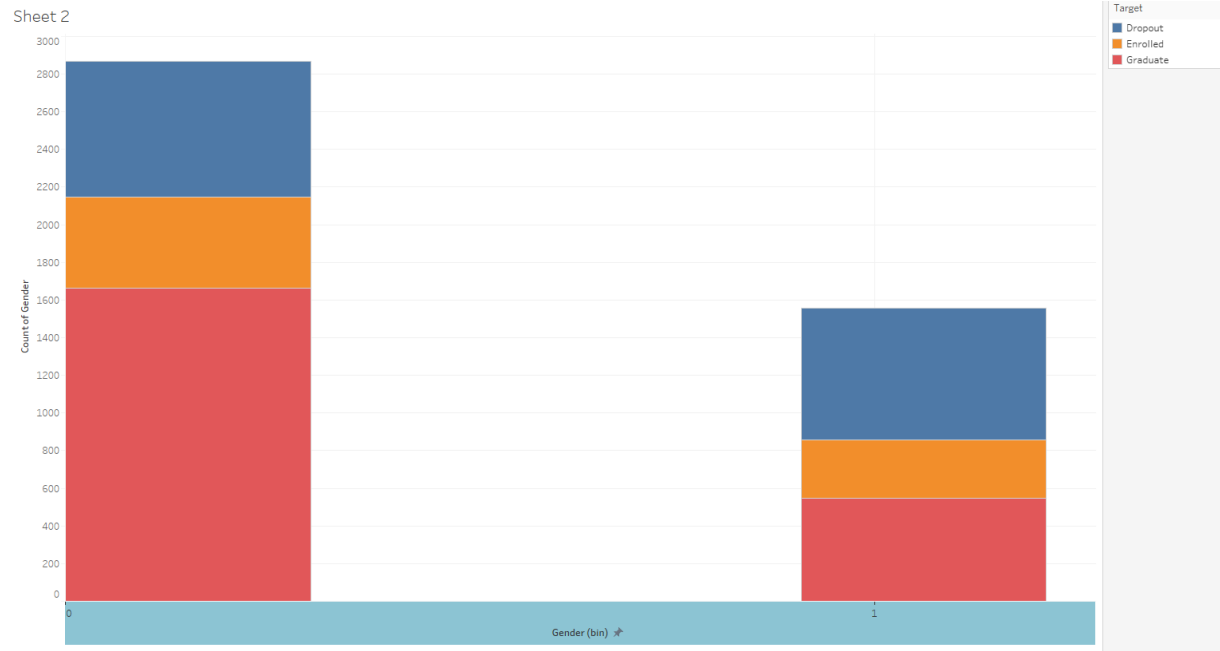
Gender and Count of Target



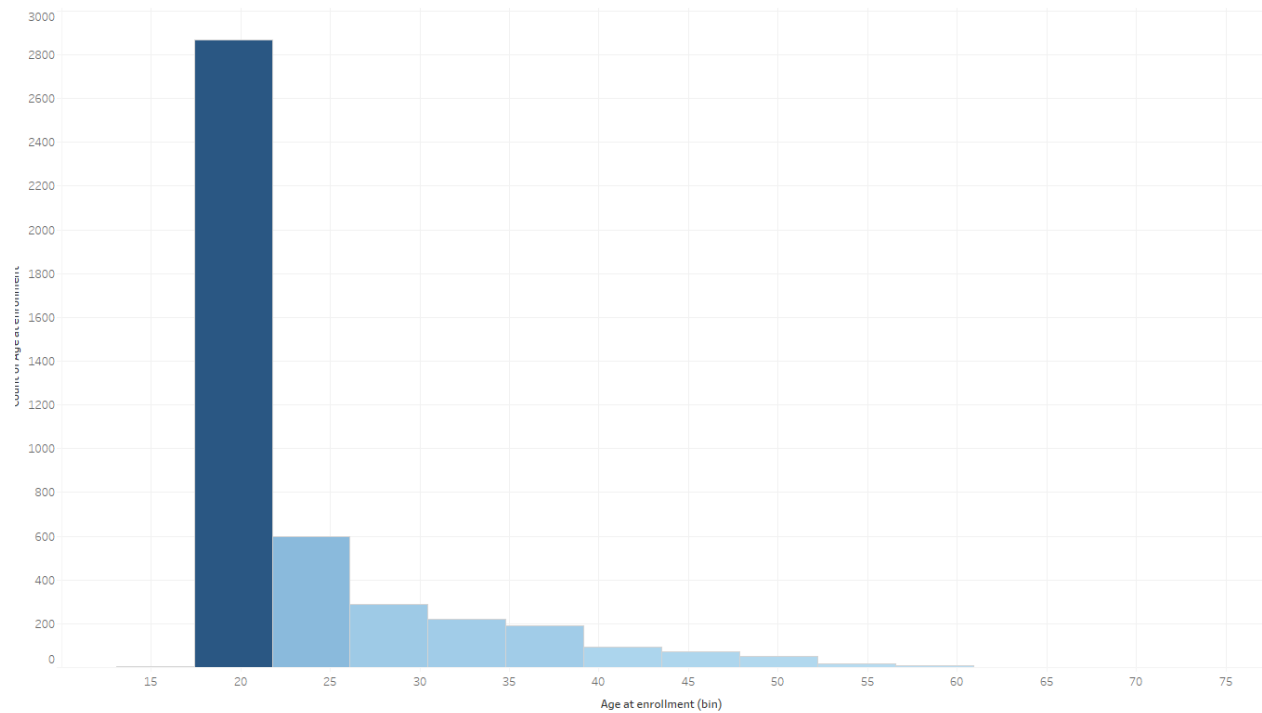
Gender and Count of Target

Gender: 1- Male, 0- Female

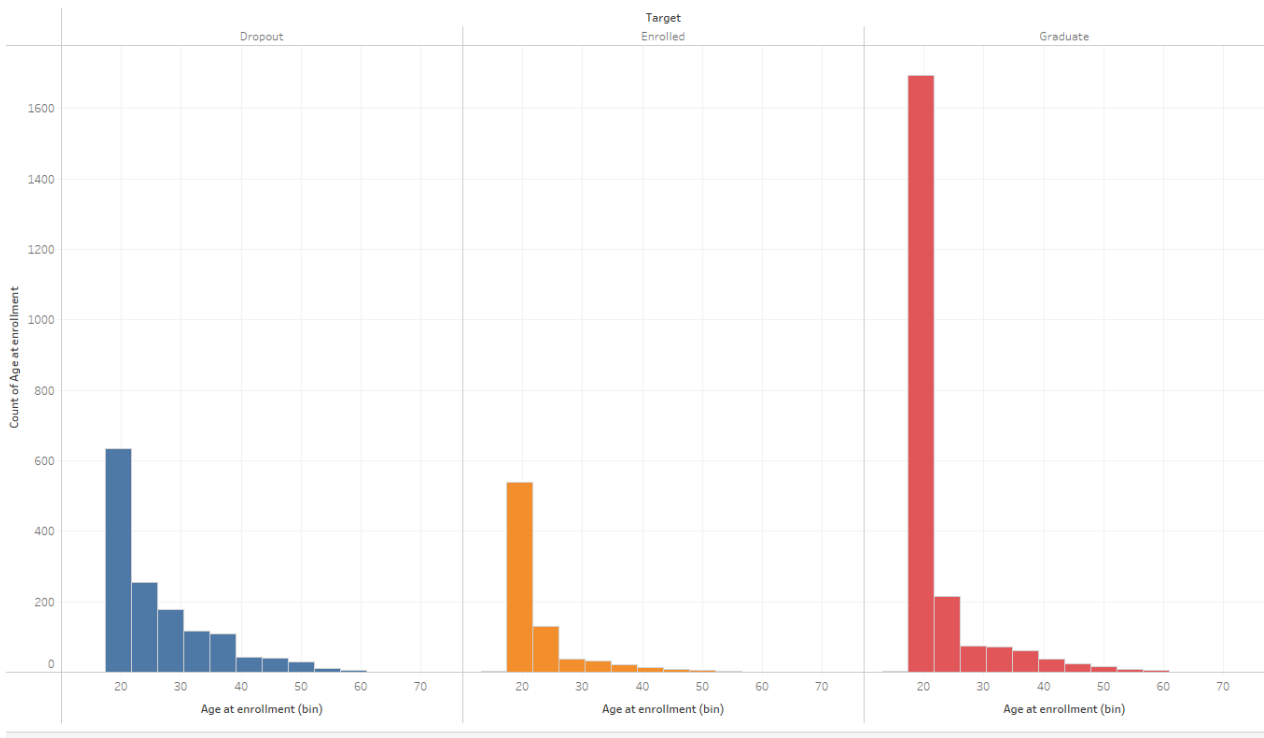
Target	1- Male (Count of target)	0- Female(Count of target)
Dropout	701	720
Enrolled	307	487
Graduate	548	1661



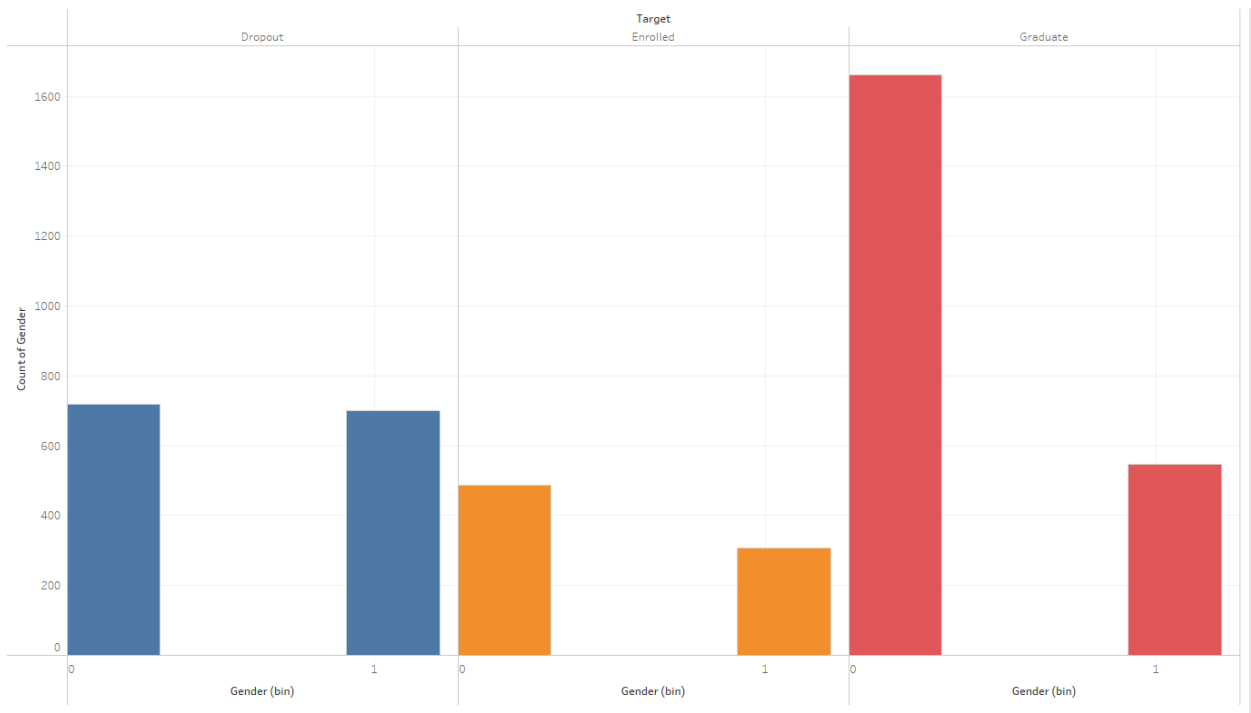
Distribution of graduate,dropout and enrolled students in each gender.



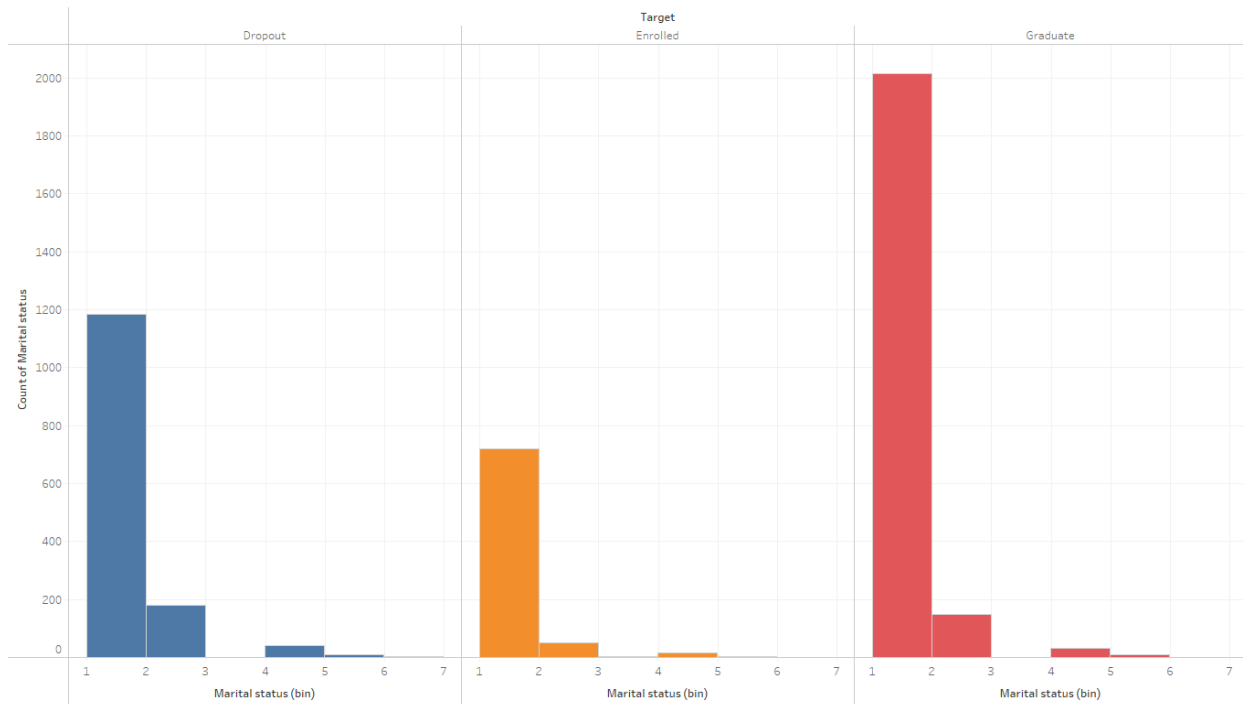
Age of enrollment of students in the dataset



Distribution of age of students enrolled with respect to dropout, enrolled and graduate value.



Distribution of male and female students with respect to their target value (Dropout, pass, fail)



Distribution of students with respect to their marital status and their target value(Dropout, pass, fail)

Running Queries using Amazon Athena

Total records:

aws

Services

Search

[Option+S]

N. Virginia

voclabs/user2754958=raditya@uncc.edu @ 0451-0997-4975

EditorRecent queriesSaved queriesSettingsworkgroupprimary

Data

↻ <

Data source

AwsDataCatalog

Database

dropout-db

Tables and views

Create

Filter tables and views

▼ Tables (1)

< 1 >

dropout_transformed_data_input

► Views (0)

< 1 >

Query 1 : X

✓ Query 2 : X

✓ Query 3 : X

+

▼

1

SELECT COUNT(*) AS total_records FROM "AwsDataCatalog"."dropout-db"."dropout_transformed_data_input";

SQL

Ln 1, Col 102

⋮

⌂

⚙

Run again

Explain

Cancel

Clear

Create

Reuse query results
up to 60 minutes ago

Query results

Query stats

✓ Completed

Time in queue: 102 ms

Run time: 858 ms

Data scanned: 466.70 KB

Results (1)

Copy

Download results

Search rows

< 1 >

⚙

#

total_records

▼

1

4424

Dropout as per gender:

Query 3 : X

Query 4 : X

Query 5 : X

+

▼

```
1 SELECT
2     "Gender",
3     "Target",
4     COUNT(*) AS number_of_records
5 FROM
6     "AwsDataCatalog"."dropout-db"."dropout_transformed_data_input"
7 GROUP BY
8     "Gender",
9     "Target"
10 ORDER BY
11     "Gender",
12     "Target";
```

Results (6)

Copy

Download results

< 1 > ⚙

# ▼	Gender ▼	Target ▼	number_of_records ▼
1	0	Dropout	720
2	0	Enrolled	487
3	0	Graduate	1661
4	1	Dropout	701
5	1	Enrolled	307
6	1	Graduate	548

Dropout as per Marital Status:

✔ Query 3 : ✕ | ✔ Query 4 : ✕ | ✔ Query 5 : ✕

```
1 SELECT
2     "Marital status",
3     "Target",
4     COUNT(*) AS number_of_records
5 FROM
6     "AwsDataCatalog"."dropout-db"."dropout_transformed_data_input"
7 GROUP BY
8     "Marital status",
9     "Target"
10 ORDER BY
11     "Marital status",
12     "Target";
```

# ▾	Marital status ▾	Target ▾	number_of_records ▾
1	1	Dropout	1184
2	1	Enrolled	720
3	1	Graduate	2015
4	2	Dropout	179
5	2	Enrolled	52
6	2	Graduate	148
7	3	Dropout	1
8	3	Enrolled	2
9	3	Graduate	1
10	4	Dropout	42
11	4	Enrolled	16
12	4	Graduate	33
13	5	Dropout	11
14	5	Enrolled	3
15	5	Graduate	11
16	6	Dropout	4
17	6	Enrolled	1
18	6	Graduate	1

Dropout according to Mother's occupation:

```
Query 3 : X | Query 4 : X | Query 5 : X | +
1 SELECT
2     "mother's occupation",
3     "Target",
4     COUNT(*) AS number_of_records
5 FROM
6     "AwsDataCatalog"."dropout-db"."dropout_transformed_data_input"
7 GROUP BY
8     "mother's occupation",
9     "Target"
10 ORDER BY
11     "mother's occupation",
12     "Target";
```

mother's occupation	Target	number_of_records
0	Dropout	99
0	Enrolled	1
0	Graduate	44
1	Dropout	39
1	Enrolled	15
1	Graduate	48
2	Dropout	102
2	Enrolled	78
2	Graduate	138
3	Dropout	95
3	Enrolled	79
3	Graduate	177
4	Dropout	248
4	Enrolled	147
4	Graduate	422
5	Dropout	156

5	Enrolled	94
5	Graduate	280
6	Dropout	26
6	Enrolled	14
6	Graduate	51
7	Dropout	80
7	Enrolled	48
7	Graduate	144
8	Dropout	15
8	Enrolled	7
8	Graduate	14
9	Dropout	490
9	Enrolled	264
9	Graduate	823
10	Dropout	1
10	Enrolled	2
10	Graduate	1
90	Dropout	51
90	Graduate	19
99	Dropout	13
99	Enrolled	2
99	Graduate	2
122	Enrolled	2
123	Dropout	2
123	Enrolled	2
123	Graduate	3
125	Graduate	1
131	Enrolled	1
132	Enrolled	2
132	Graduate	1
134	Dropout	1

134	Enrolled	2
134	Graduate	1
141	Enrolled	4
141	Graduate	4
143	Enrolled	2
143	Graduate	1
144	Enrolled	2
144	Graduate	4
151	Enrolled	1
151	Graduate	2
152	Enrolled	1
152	Graduate	1
153	Graduate	2
171	Graduate	1
173	Enrolled	1
175	Dropout	1
175	Enrolled	2
175	Graduate	2
191	Enrolled	11
191	Graduate	15
192	Enrolled	3
192	Graduate	2
193	Dropout	1
193	Graduate	3
194	Dropout	1
194	Enrolled	7
194	Graduate	3

Dropout as per Age of enrollment:

```

1  SELECT
2      "Age at enrollment",
3      "Target",
4      COUNT(*) AS number_of_records
5  FROM
6      "AwsDataCatalog"."dropout-db"."dropout_transformed_data_input"
7  GROUP BY
8      "Age at enrollment",
9      "Target"
10 ORDER BY
11     "Age at enrollment",
12     "Target";

```

Age at enrollment	Target	number_of_records
17	Enrolled	2
17	Graduate	3
18	Dropout	202
18	Enrolled	172
18	Graduate	662
19	Dropout	207
19	Enrolled	157
19	Graduate	547
20	Dropout	133
20	Enrolled	140
20	Graduate	326
21	Dropout	93
21	Enrolled	70
21	Graduate	159
22	Dropout	58
22	Enrolled	37
22	Graduate	79
23	Dropout	41
23	Enrolled	24

23	Graduate	43
24	Dropout	56
24	Enrolled	30
24	Graduate	45
25	Dropout	47
25	Enrolled	21
25	Graduate	25
26	Dropout	52
26	Enrolled	19
26	Graduate	23
27	Dropout	55
27	Enrolled	12
27	Graduate	24
28	Dropout	47
28	Enrolled	15
28	Graduate	21
29	Dropout	45
29	Enrolled	5
29	Graduate	16
30	Dropout	30
30	Enrolled	6
30	Graduate	13
31	Dropout	36
31	Enrolled	7
31	Graduate	12
32	Dropout	33
32	Enrolled	5
32	Graduate	23
33	Dropout	20
33	Enrolled	8
33	Graduate	17

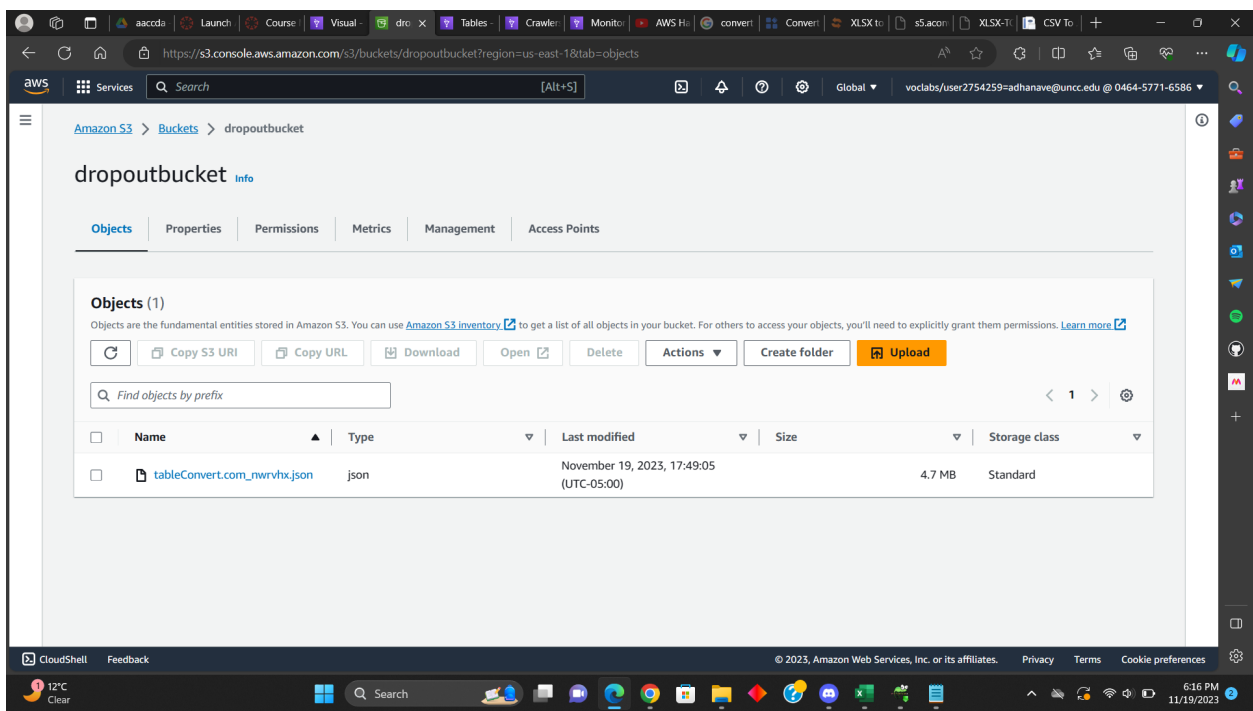
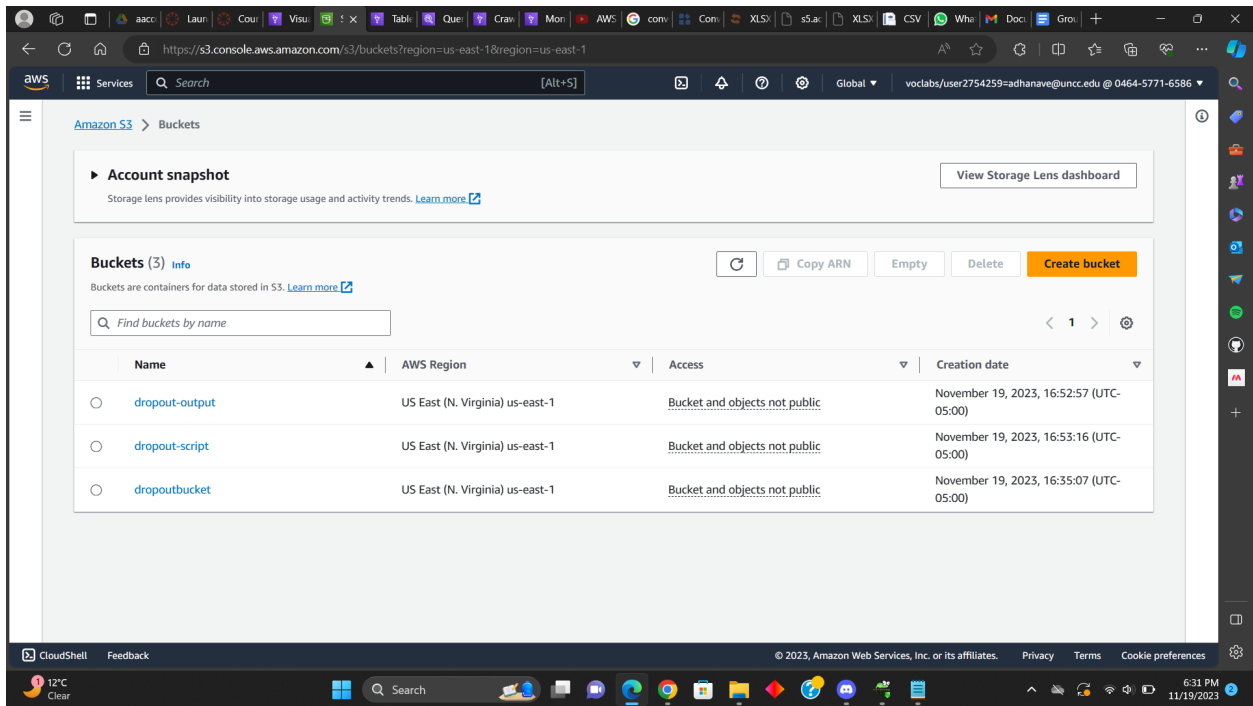
34	Dropout	29
34	Enrolled	12
34	Graduate	19
35	Dropout	29
35	Enrolled	5
35	Graduate	15
36	Dropout	21
36	Enrolled	3
36	Graduate	11
37	Dropout	23
37	Enrolled	5
37	Graduate	14
38	Dropout	19
38	Enrolled	4
38	Graduate	6
39	Dropout	18
39	Enrolled	5
39	Graduate	15
40	Dropout	13
40	Enrolled	4
40	Graduate	6
41	Dropout	11
41	Enrolled	5
41	Graduate	15
42	Dropout	7
42	Enrolled	2
42	Graduate	4
43	Dropout	11
43	Enrolled	3
43	Graduate	11
44	Dropout	10
44	Enrolled	2

44	Graduate	9
45	Dropout	13
45	Enrolled	4
45	Graduate	5
46	Dropout	6
46	Graduate	6
47	Dropout	11
47	Enrolled	2
47	Graduate	5
48	Dropout	7
48	Enrolled	3
48	Graduate	1
49	Dropout	7
49	Enrolled	1
49	Graduate	5
50	Dropout	9
50	Enrolled	2
50	Graduate	5
51	Dropout	5
51	Graduate	2
52	Graduate	4
53	Dropout	3
53	Enrolled	1
53	Graduate	3
54	Dropout	4
54	Enrolled	1
54	Graduate	2
55	Dropout	3
55	Graduate	2
57	Dropout	1

57	Graduate	1
58	Dropout	2
58	Graduate	1
59	Dropout	2
59	Graduate	1
60	Graduate	2
61	Dropout	1
62	Graduate	1
70	Dropout	1

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)



Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Legacy pages

What's New

Documentation

Services

Search

[Alt+S]

N. Virginia

voclabs/user2754259=adhanave@uncc.edu @ 0464-5771-6586

dropoutbucket

Last updated (UTC)
November 19, 2023 at 22:52:34

Version 0 (Current version)

Actions

Table overview

Data quality New

Table details

Advanced properties

Name dropoutbucket	Description -	Database dropout-db	Classification JSON
Location s3://dropoutbucket/	Connection -	Deprecated -	Last updated November 19, 2023 at 22:52:34
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.q1o.HiveIgnoredKeyTextOutputFormat	Serde serialization lib org.openx.data.jsonserde.JsonSerDe	

Schema

Partitions

Indexes

Column statistics - new

Schema (36)

View and manage the table schema.

Filter schemas

1 2

#	Column name	Data type	Partition key	Comment
---	-------------	-----------	---------------	---------

CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

12°C Clear

Search

6:16 PM 11/19/2023

Schema (36)

View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key	Comment
1	marital status	string	-	-
2	application mode	string	-	-
3	application order	string	-	-
4	course	string	-	-
5	previous qualification	string	-	-
6	previous qualification (grade)	string	-	-
7	nacionality	string	-	-
8	mother's qualification	string	-	-
9	father's qualification	string	-	-
10	mother's occupation	string	-	-
11	father's occupation	string	-	-
12	admission grade	string	-	-
13	displaced	string	-	-
14	educational special needs	string	-	-
15	debtor	string	-	-

Query results

Query stats

Completed Time in queue: 112 ms Run time: 727 ms Data scanned: 4.36 MB

Results (10)

Copy Download results

Search rows

#	marital status	application mode	application order	course	previous qualification	previous qualification (grade)	nacionality
1	1	17	5	171	1	122	1
2	1	15	1	9254	1	160	1
3	1	1	5	9070	1	122	1
4	1	17	2	9773	1	122	1
5	2	39	1	8014	1	100	1
6	2	39	1	9991	19	133.1	1
7	1	1	1	9500	1	142	1
8	1	18	4	9254	1	119	1

Some crawlers successfully deleted
The following crawlers are now deleted: "dropout-crawler1", "dropout-crawler"

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) Info Last updated (UTC) November 19, 2023 at 23:16:48 Action Run Create crawler

View and manage all available crawlers.

Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run time...	Log	Table changes f...
<input type="checkbox"/>	dropout-crawler	Ready		Succeeded	November 19, 20...	View log	1 created

dropout-job

Last modified on 11/19/2023, 6:11:16 PM Try new UI Actions Save Run

Visual **Script** **Job details** **Runs** **Data quality** **Schedules** **Version Control**

Transform

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
gender	gender	string	<input type="checkbox"/>
scholarship holder	scholarship	string	<input type="checkbox"/>
age at enrollment	age at enr	string	<input type="checkbox"/>
international		<input checked="" type="checkbox"/>	
curricular units 1st	curricular u	string	<input type="checkbox"/>
curricular units 1st	curricular u	string	<input type="checkbox"/>
curricular units 1st	curricular u (ns)	string	<input type="checkbox"/>
curricular units 1st	curricular u (d)	string	<input type="checkbox"/>
curricular units 1st	curricular u	string	<input type="checkbox"/>
curricular units 1st	curricular u evalu	string	<input type="checkbox"/>
curricular units 2nd	curricular u (l)	string	<input type="checkbox"/>

Data preview (200) Info READY End session Previewing 35 of 35 fields

Filter sample dataset

marital status	application mode	application order	course
1	17	5	171

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Visual

Dropout-job

Last modified on 11/19/2023, 6:11:16 PM

Try new UI

Actions

Save

Run

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Legacy pages

What's New

Documentation

CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

12°C Clear

Search

6:18 PM 11/19/2023

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Script (Locked)

Dropout-job

Last modified on 11/19/2023, 9:24:34 PM

Try new UI

Actions

Save

Run

Successfully updated job

Successfully updated job dropout-job. To run the job choose the Run Job button.

Download script

Edit script

```
1 import sys
2 from aws glue.transforms import *
3 from aws glue.utils import getResolvedOptions
4 from pyspark.context import SparkContext
5 from aws glue.context import GlueContext
6 from aws glue.job import Job
7
8 args = getResolvedOptions(sys.argv, ["JOB_NAME"])
9 sc = SparkContext()
10 glueContext = GlueContext(sc)
11 spark = glueContext.spark_session
12 job = Job(glueContext)
13 job.init(args["JOB_NAME"], args)
14
15 # Script generated for node AWS Glue Data Catalog
16 AWSGlueDataCatalog_node1700434561434 = glueContext.create_dynamic_frame_from_catalog(
17     database="dropout-db",
18     table_name="dropoutbucket",
19     transformation_ctx="AWSGlueDataCatalog_node1700434561434",
20 )
21
22 # Script generated for node Change Schema
23 ChangeSchema_node1700447012744 = ApplyMapping.apply(
24     frame=AWSGlueDataCatalog_node1700434561434,
25     mappings=[
26         ("marital status", "string", "marital status", "string"),
```

CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

7°C Clear

Search

9:24 PM 11/19/2023

AWS Glue Monitoring

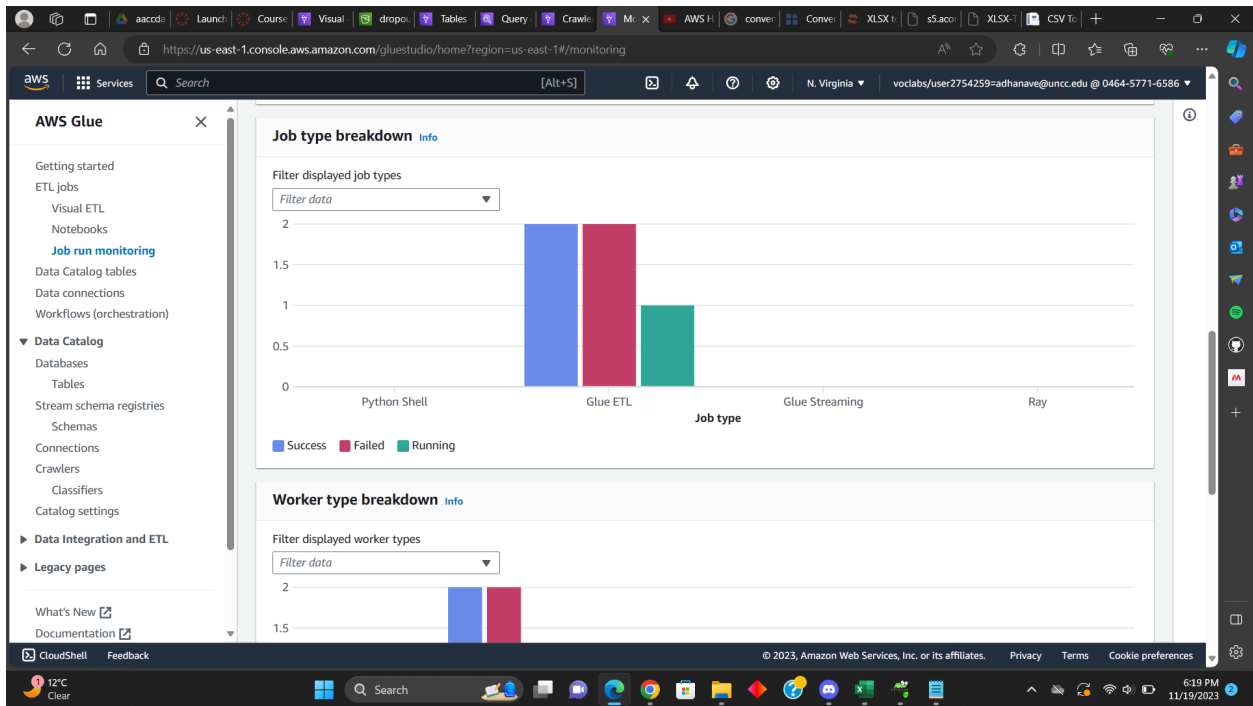
Date range: 7 Day

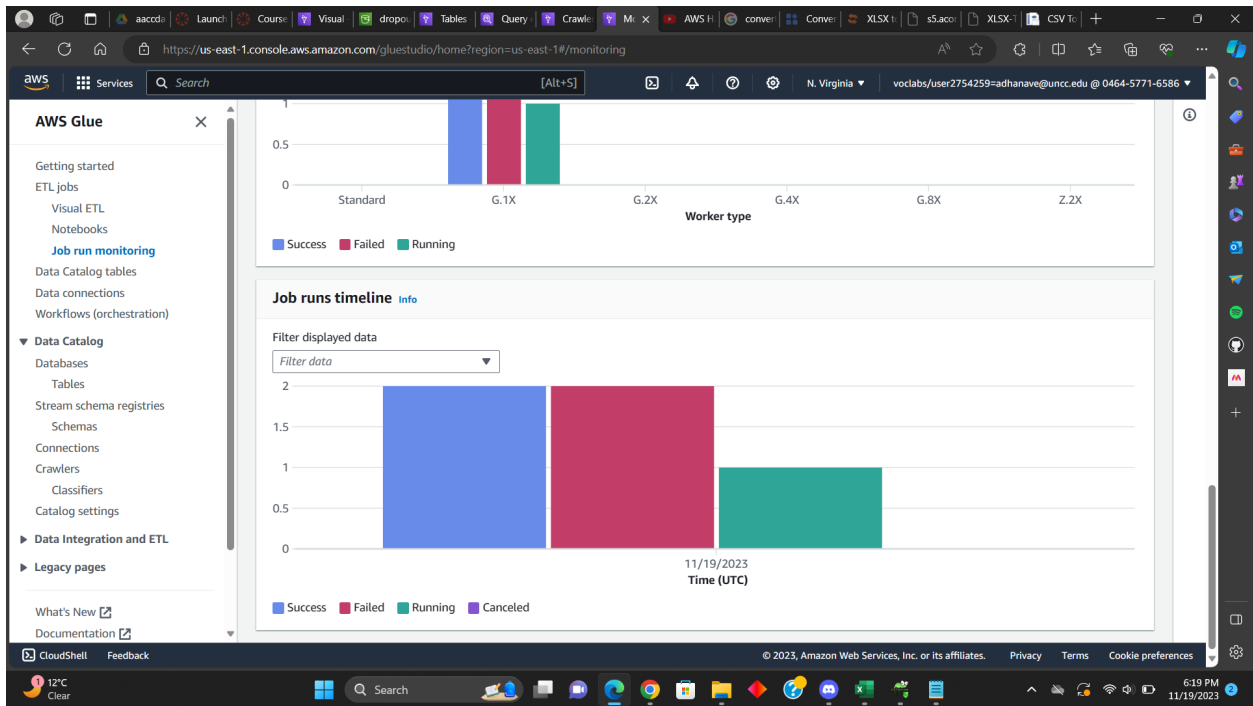
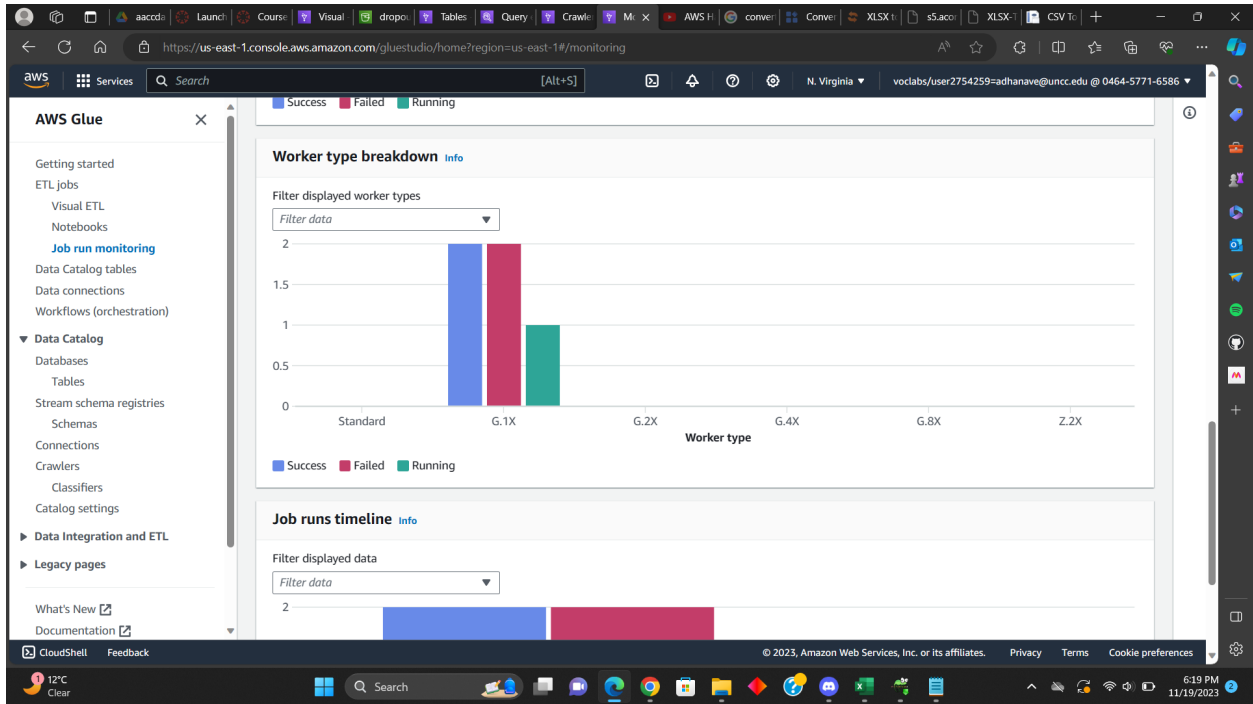
Job runs summary

Total runs	Running	Canceled	Success	Failed	Success rate	DPU hours
5	1	0	2	2	50%	1

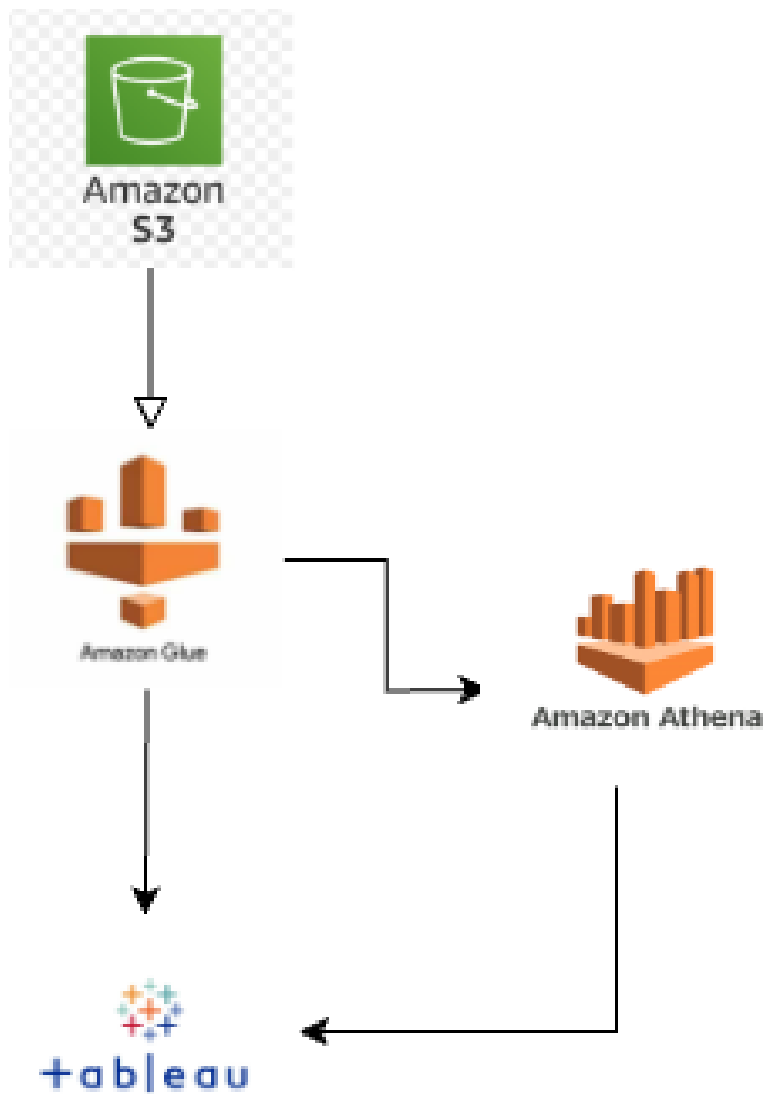
Job runs (5)

Job name	Run status	Type	Start time (UTC)	End time (UTC)	Run time	Capacity	Worker type	DPU hours
dropout-job	Succeeded	Glue ETL	2023/11/19 23:11:20	2023/11/19 23:12:16	1 minute	10	G.1X	0.1
dropout-job	Failed	Glue ETL	2023/11/19 23:06:35	2023/11/19 23:08:08	1 minute	10	G.1X	0.2
dropout-job	Failed	Glue ETL	2023/11/19 23:02:43	2023/11/19 23:04:54	2 minutes	10	G.1X	0.2
dropout-job	Succeeded	Glue ETL	2023/11/19 22:20:38	2023/11/19 22:21:34	1 minute	10	G.1X	0.1
dropout-job	Succeeded	Glue ETL	2023/11/19 22:04:09	2023/11/19 22:05:11	1 minute	10	G.1X	0.1





AWS Pipeline/Solution Chart (Important)



Phase 2 Deliverable:

Project Repository on GitHub (Updated Table of Contents)

Deliverable 2 Document accessible in Github (include screenshots)
After Completion please share the github repository link

Submitted by Group 14:

Ramit Aditya

Delphine Antony Muthu

Mounisha Bolla

Anirudh Cheruvu

Amirthavarshini Dhanavel