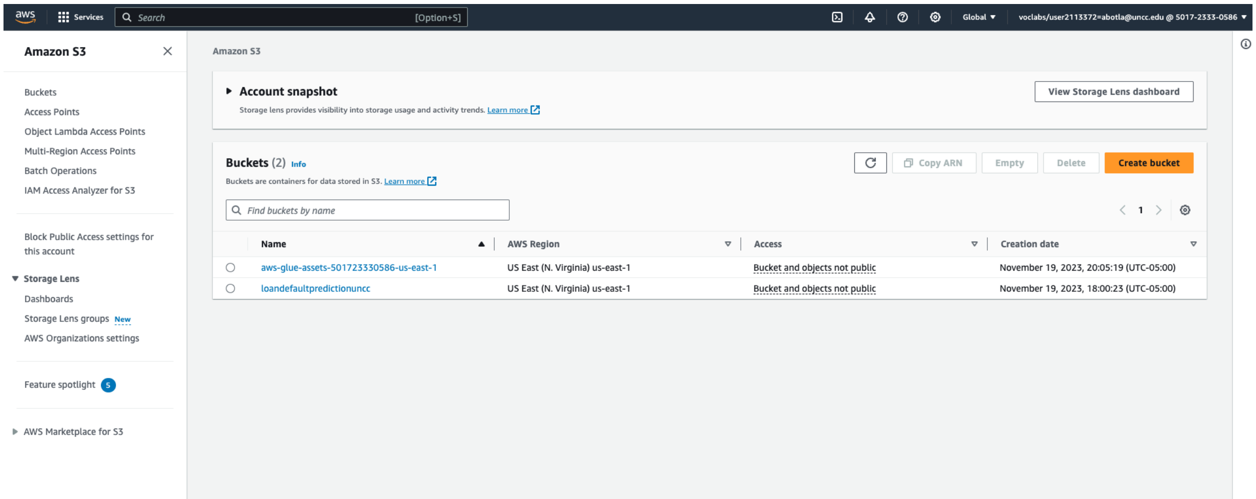
Project Deliverable 2

Team - 4
Abhinav Botla
Ajith Gannamaneni
Hrushikesh Dandge
Mitra Buggaveeti
Prateek Chanddra

Created the S3 bucket with name of **loandefaultpredictionuncc** to store the project data in the bucket


aws
Services
Search
[Option+S]
Global
voclabs/user2113372=abotla@uncc.edu @ 5017-2333-0586
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups New
AWS Organizations settings
Feature spotlight 5
AWS Marketplace for S3
View Storage Lens dashboard
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
1

Amazon S3

Account snapshot

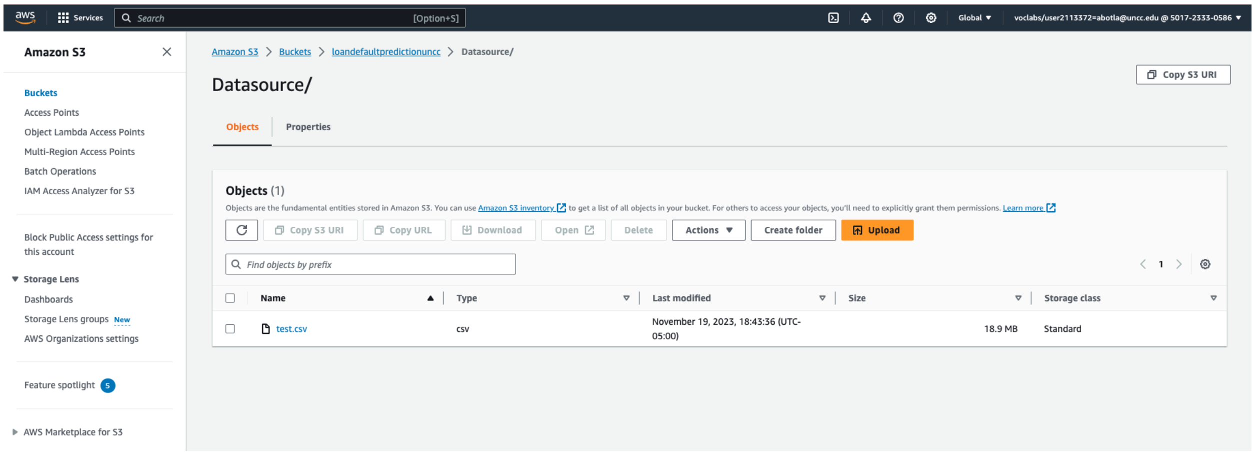
Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (2) Info

Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	aws-glue-assets-501723330586-us-east-1	US East (N. Virginia) us-east-1	Bucket and objects not public	November 19, 2023, 20:05:19 (UTC-05:00)
○	loandefaultpredictionuncc	US East (N. Virginia) us-east-1	Bucket and objects not public	November 19, 2023, 18:00:23 (UTC-05:00)

Created the Datasource folder in the bucket to store the data



Amazon S3 > Buckets > loandefaultpredictionuncc > Datasource/

Datasource/

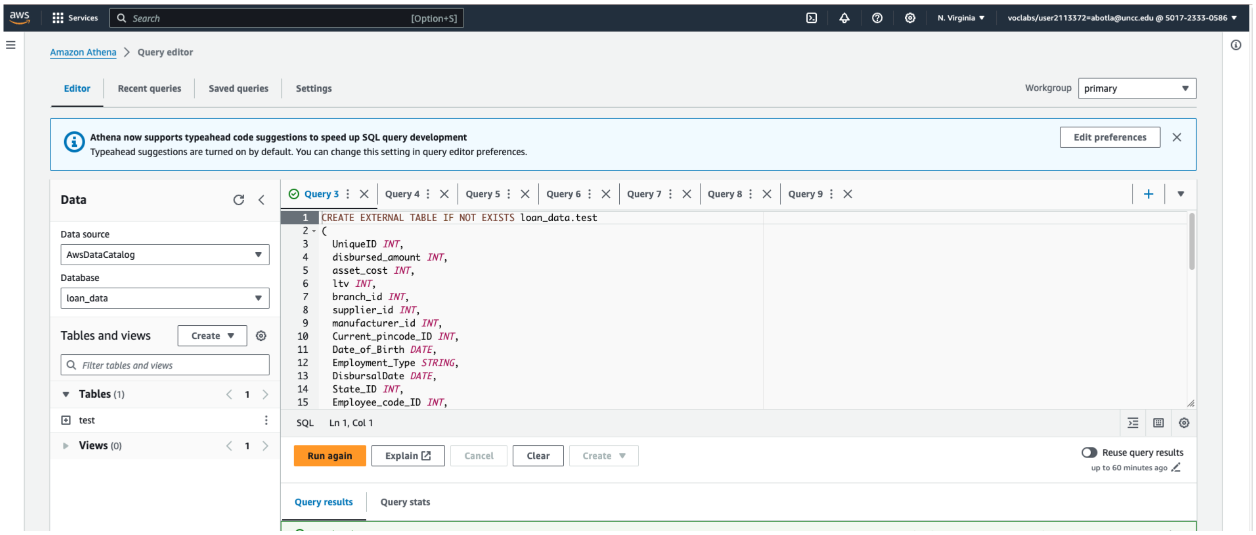
Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	test.csv	csv	November 19, 2023, 18:43:36 (UTC-05:00)	18.9 MB	Standard

Analysis of data by using the AWS Athena:

Creating table DDL in Athena with predefined attributes already available in the data source.



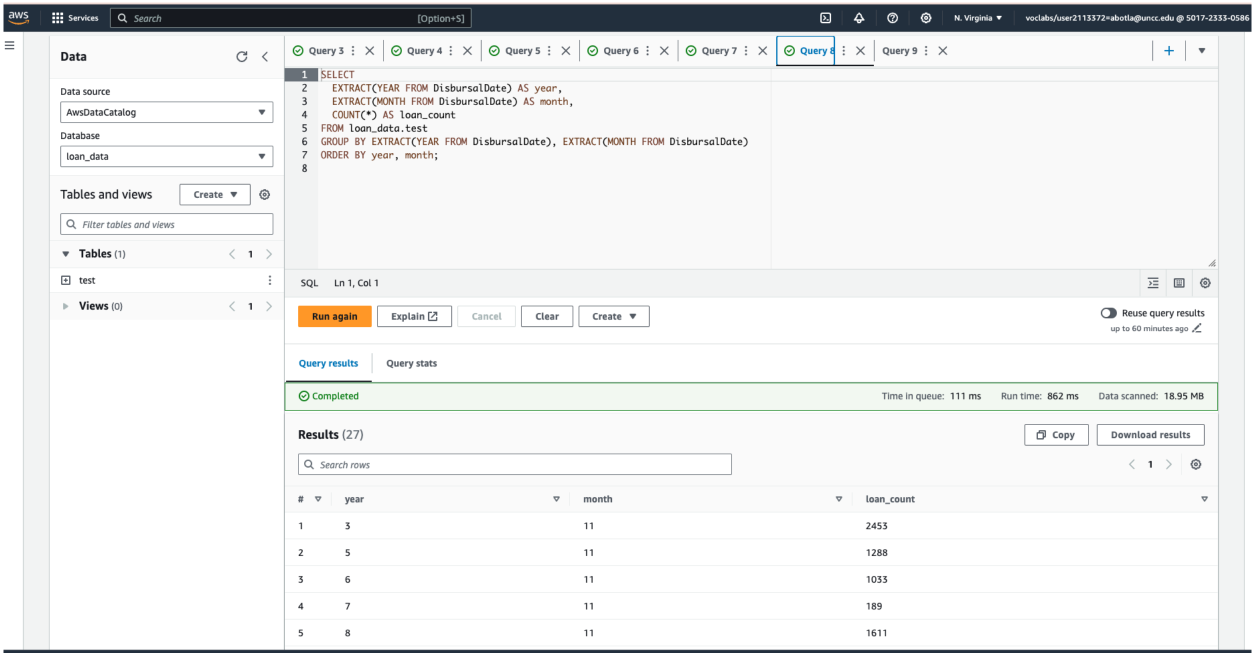
Workgroup: primary

Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

Data source: AwsDataCatalog
Database: loan_data

```
CREATE EXTERNAL TABLE IF NOT EXISTS loan_data.test
(
  UniqueID INT,
  disbursed_amount INT,
  asset_cost INT,
  ltv INT,
  branch_id INT,
  supplier_id INT,
  manufacturer_id INT,
  Current_pincode_ID INT,
  Date_of_Birth DATE,
  Employment_Type STRING,
  DisbursalDate DATE,
  State_ID INT,
  Employee_code_ID INT,
```

Time based exploratory data analysis:



Data source: AwsDataCatalog
Database: loan_data

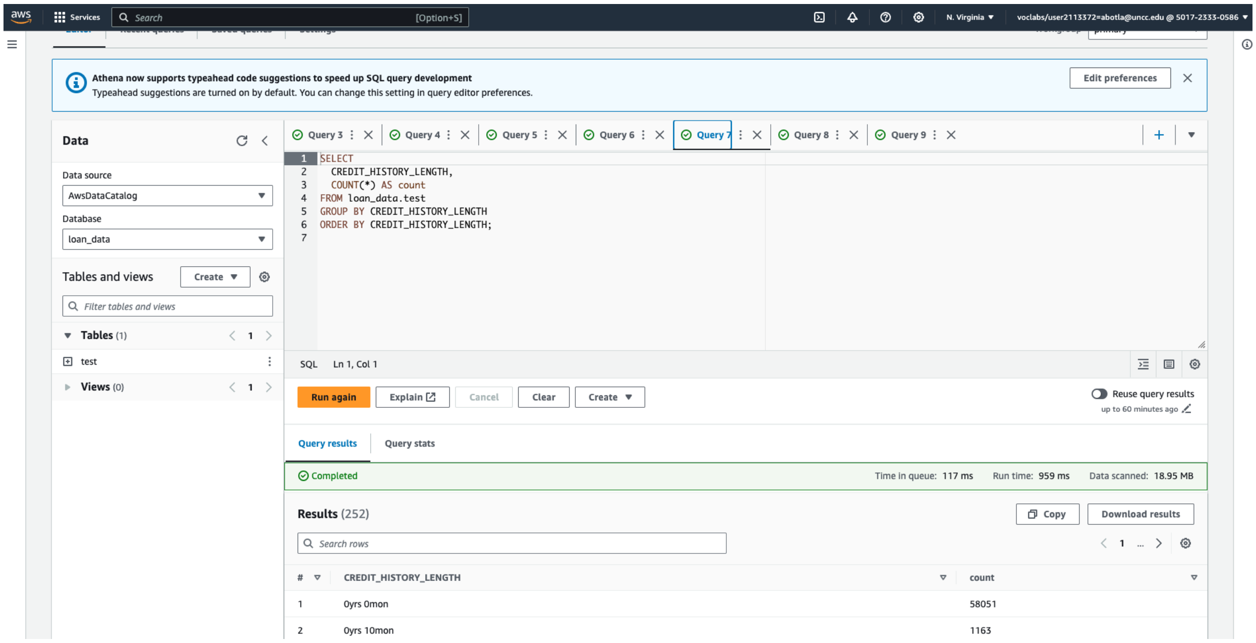
```
SELECT
  EXTRACT(YEAR FROM DisbursalDate) AS year,
  EXTRACT(MONTH FROM DisbursalDate) AS month,
  COUNT(*) AS loan_count
FROM loan_data.test
GROUP BY EXTRACT(YEAR FROM DisbursalDate), EXTRACT(MONTH FROM DisbursalDate)
ORDER BY year, month;
```

Completed   Time in queue: 111 ms   Run time: 862 ms   Data scanned: 18.95 MB

Results (27)

#	year	month	loan_count
1	3	11	2453
2	5	11	1288
3	6	11	1033
4	7	11	189
5	8	11	1611

Analyzing the credit history of the customers:



Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

Data source: AwsDataCatalog
Database: loan_data

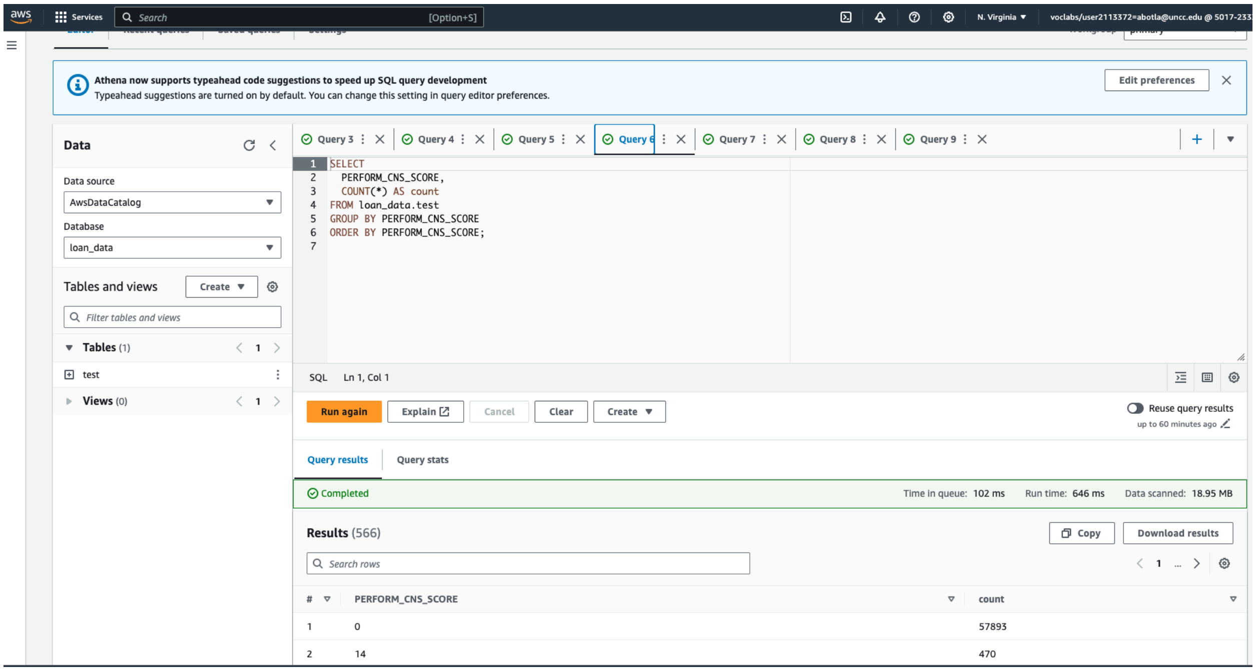
```
SELECT
  CREDIT_HISTORY_LENGTH,
  COUNT(*) AS count
FROM loan_data.test
GROUP BY CREDIT_HISTORY_LENGTH
ORDER BY CREDIT_HISTORY_LENGTH;
```

Completed   Time in queue: 117 ms   Run time: 959 ms   Data scanned: 18.95 MB

Results (252)

```
#   CREDIT_HISTORY_LENGTH   count
1   0yrs 0mon               58051
2   0yrs 10mon              1163
```

Analyzing the count of customers with specific CNS Score:



Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

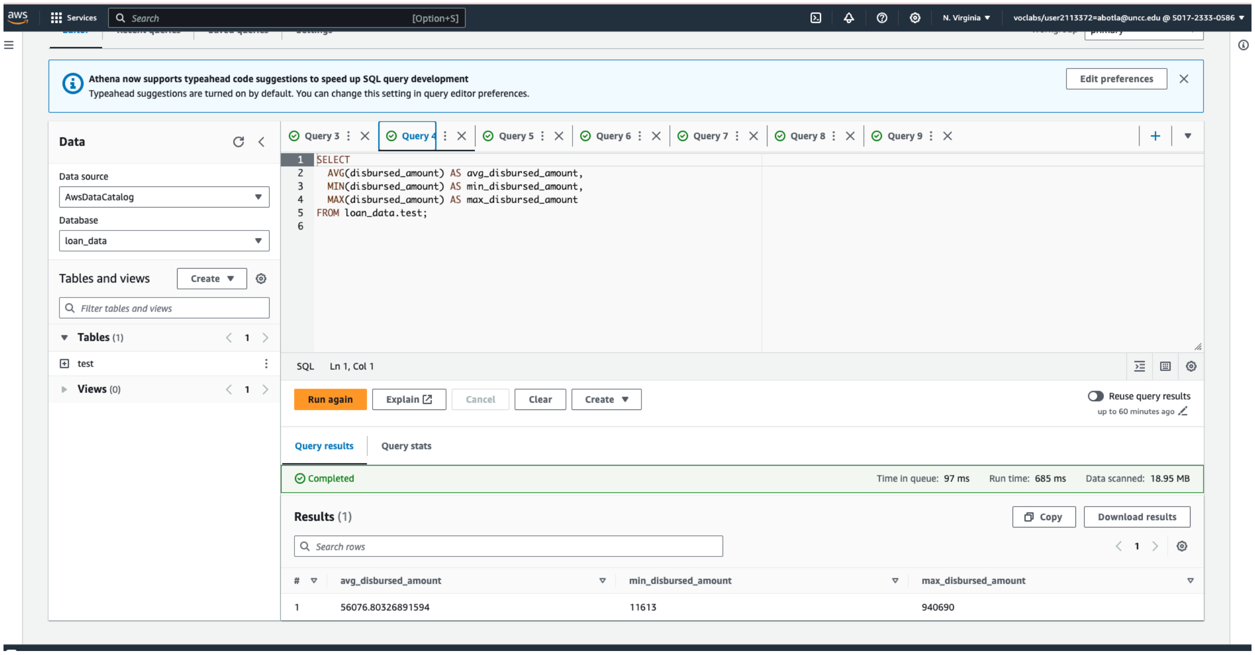
Data source: AwsDataCatalog
Database: loan_data

```
SELECT
  PERFORM_CNS_SCORE,
  COUNT(*) AS count
FROM loan_data.test
GROUP BY PERFORM_CNS_SCORE
ORDER BY PERFORM_CNS_SCORE;
```

Completed   Time in queue: 102 ms   Run time: 646 ms   Data scanned: 18.95 MB

Results (566)

```
#   PERFORM_CNS_SCORE   count
1   0                   57893
2   14                  470
```



Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

Data source: AwsDataCatalog
Database: loan_data
Tables (1): test

```
SELECT
  AVG(disbursed_amount) AS avg_disbursed_amount,
  MIN(disbursed_amount) AS min_disbursed_amount,
  MAX(disbursed_amount) AS max_disbursed_amount
FROM loan_data.test;
```

SQL Ln 1, Col 1

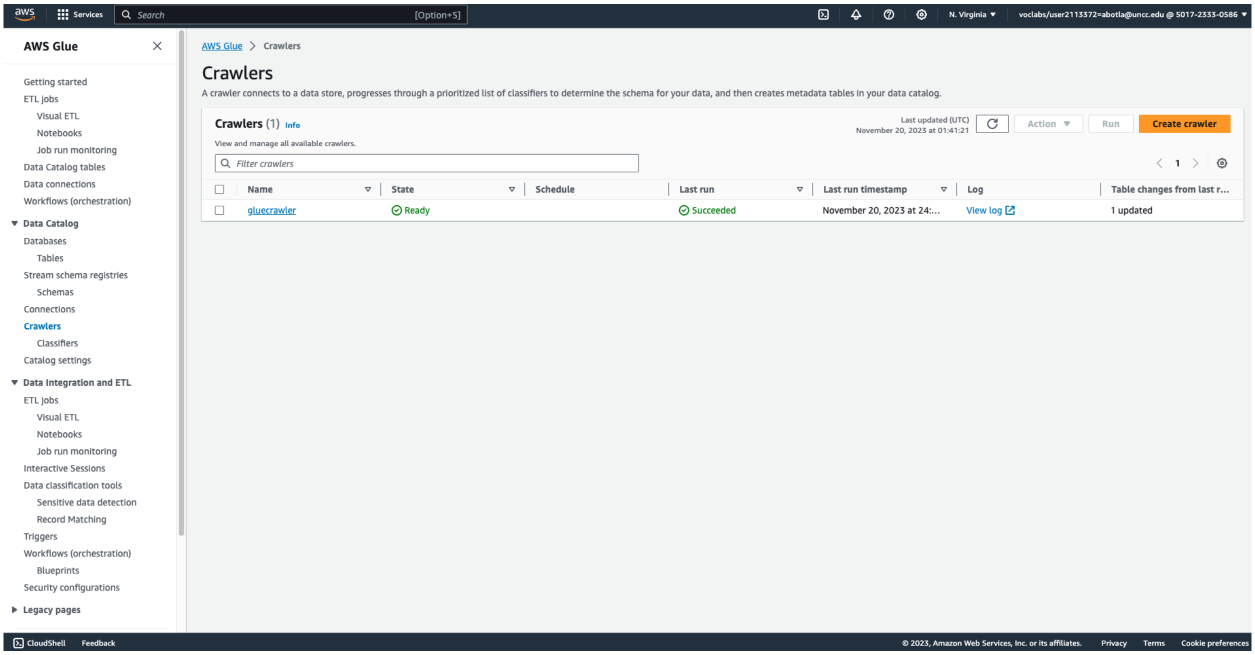
Reuse query results up to 60 minutes ago

Completed   Time in queue: 97 ms   Run time: 685 ms   Data scanned: 18.95 MB

Results (1)

#	avg_disbursed_amount	min_disbursed_amount	max_disbursed_amount
1	56076.80326891594	11613	940690

AWS GLUE:
CRAWLERS:



Crawlers

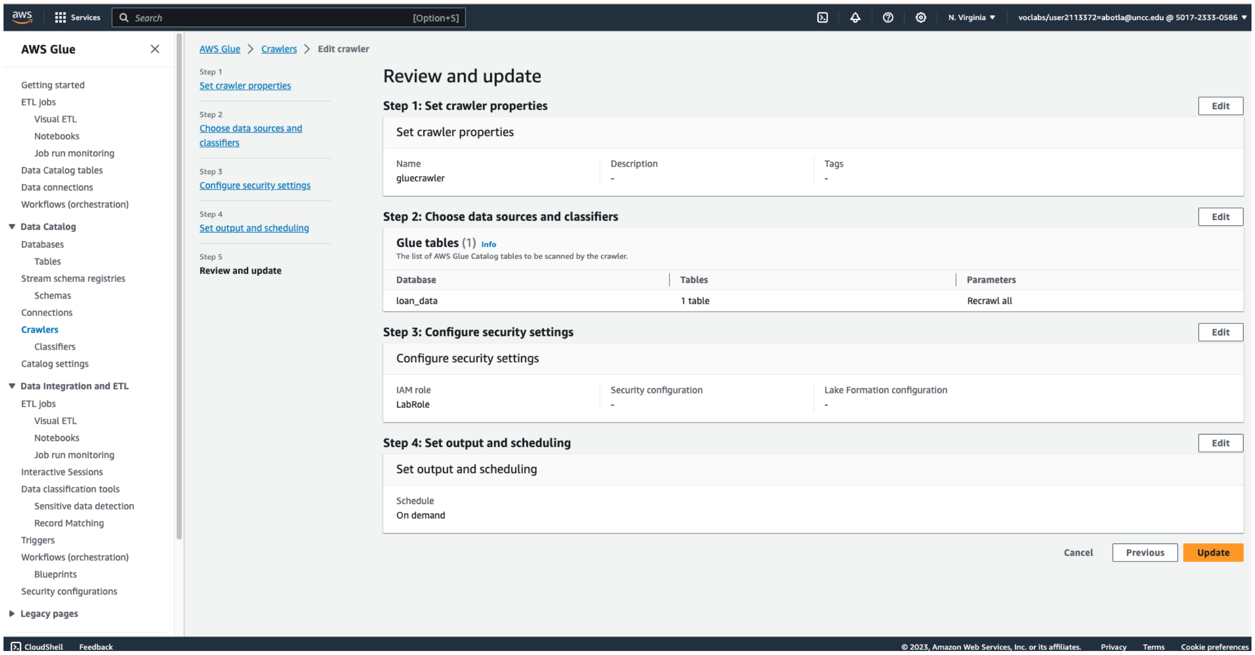
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1)

View and manage all available crawlers.

Last updated (UTC)
November 20, 2023 at 01:41:21

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last r...
gluecrawler	Ready		Succeeded	November 20, 2023 at 24:...	View log	1 updated



Review and update

Step 1: Set crawler properties

Set crawler properties

Name	Description	Tags
gluecrawler	-	-

Step 2: Choose data sources and classifiers

Glue tables (1) Info

The list of AWS Glue Catalog tables to be scanned by the crawler.

Database	Tables	Parameters
loan_data	1 table	Recrawl all

Step 3: Configure security settings

Configure security settings

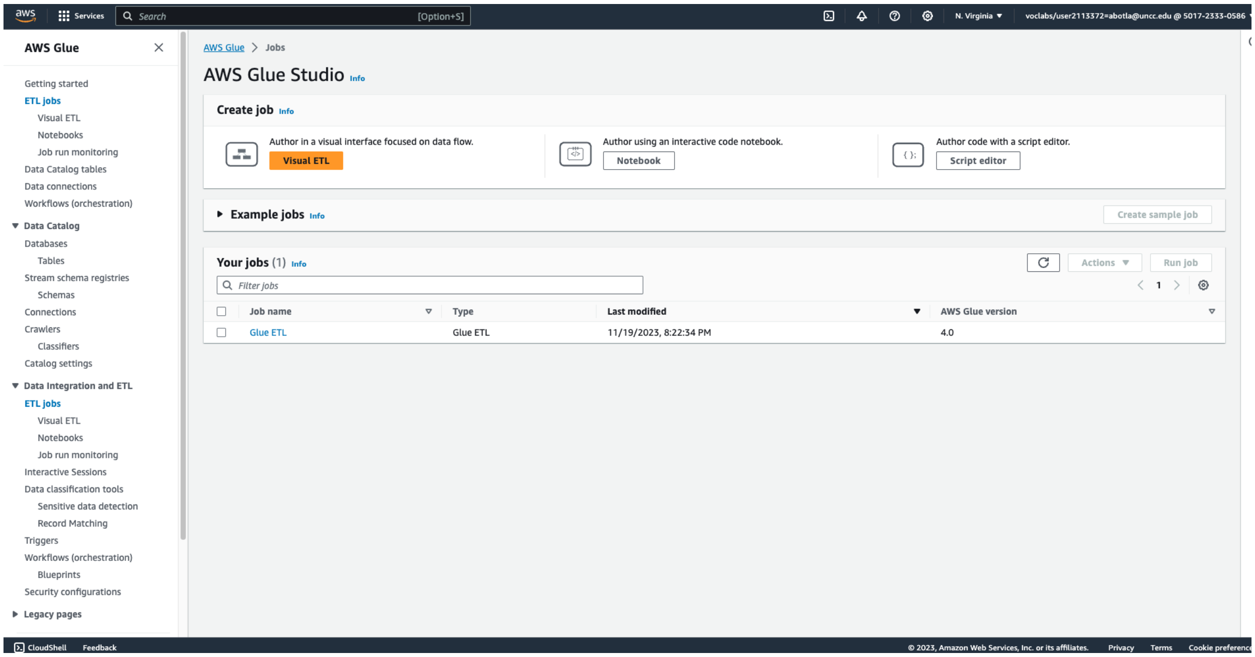
IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

Step 4: Set output and scheduling

Set output and scheduling

Schedule
On demand

Glue ETL Pipeline:



AWS Glue Studio Info

Create job Info

Author in a visual interface focused on data flow.

Author using an interactive code notebook.

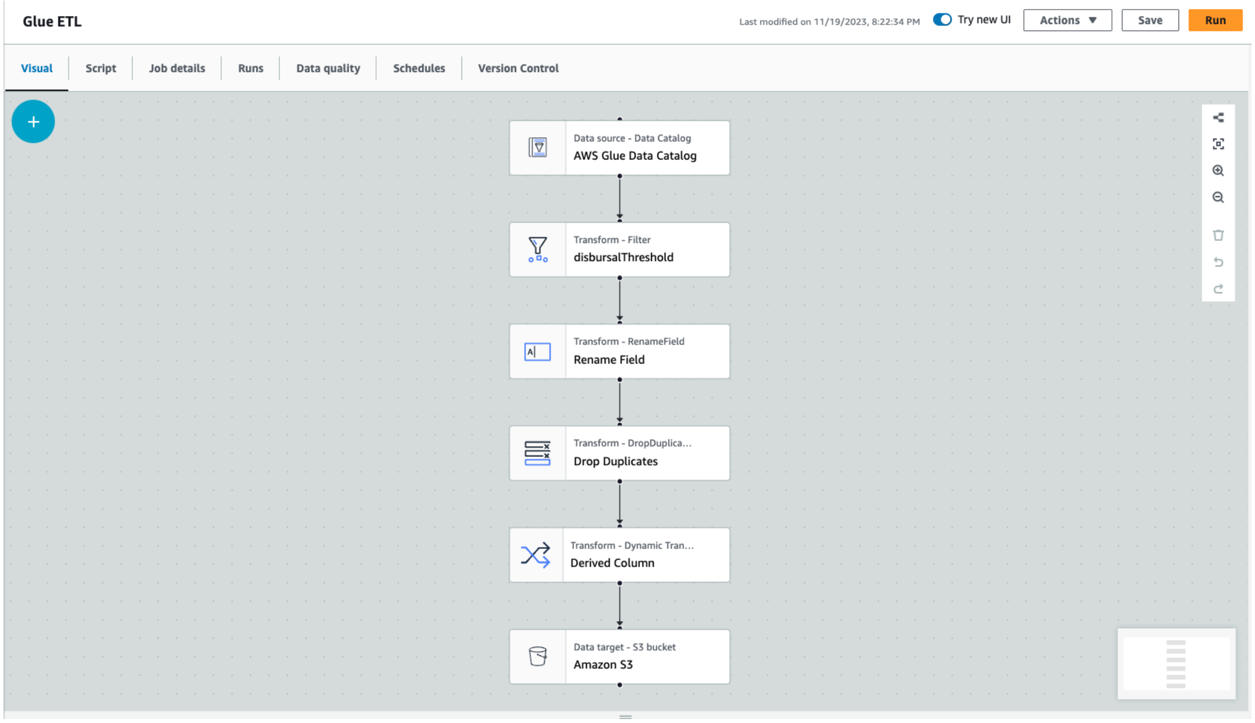
Author code with a script editor.

Example jobs Info

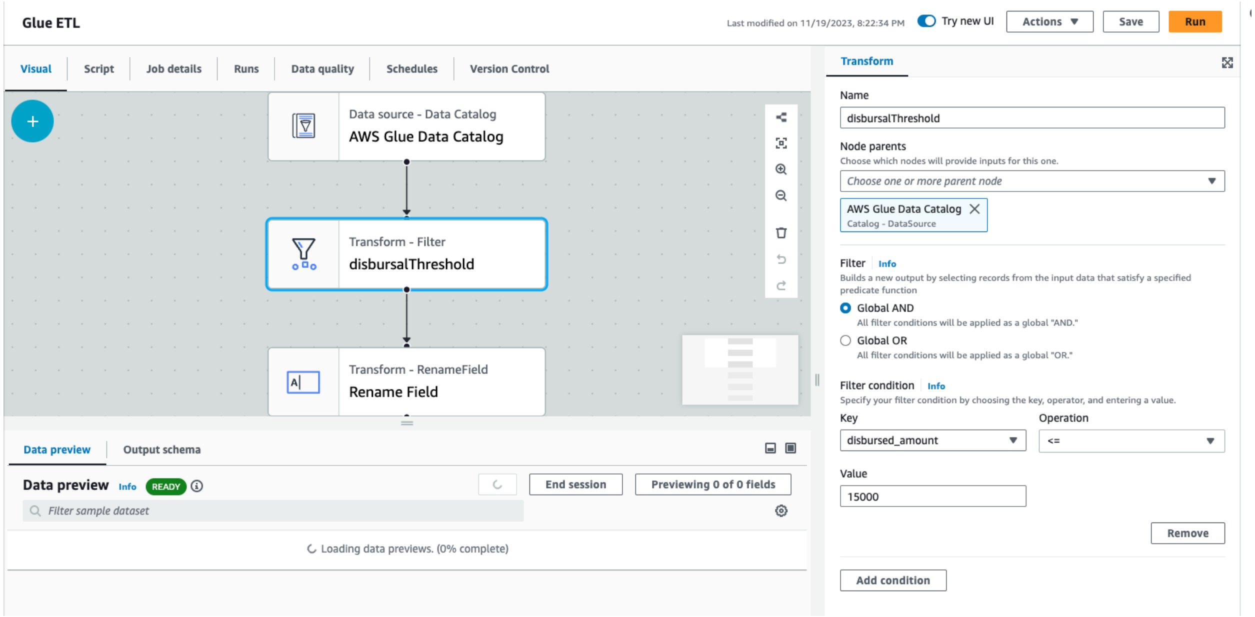
Your jobs (1) Info

Job name	Type	Last modified	AWS Glue version
Glue ETL	Glue ETL	11/19/2023, 8:22:34 PM	4.0

Visualization of the data flow



We doesnt pick the records which have a disbursed amount below 15000.



Data preview

Loading data previews. (0% complete)

Transform

Name: disbursalThreshold

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

AWS Glue Data Catalog
Catalog - DataSource

Filter Info

Builds a new output by selecting records from the input data that satisfy a specified predicate function

● Global AND
All filter conditions will be applied as a global "AND."

○ Global OR
All filter conditions will be applied as a global "OR."

Filter condition Info

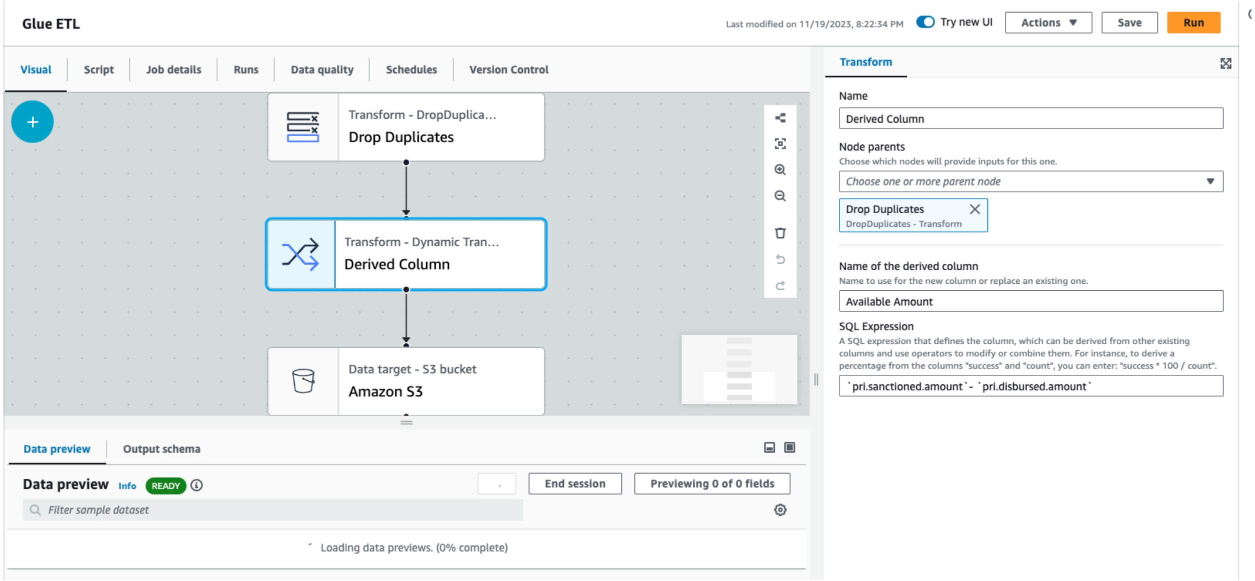
Specify your filter condition by choosing the key, operator, and entering a value.

Key: disbursed_amount

Operation: <=

Value: 15000

Dropping the duplicates from the data



Data preview READY

Loading data previews. (0% complete)

Name: Derived Column

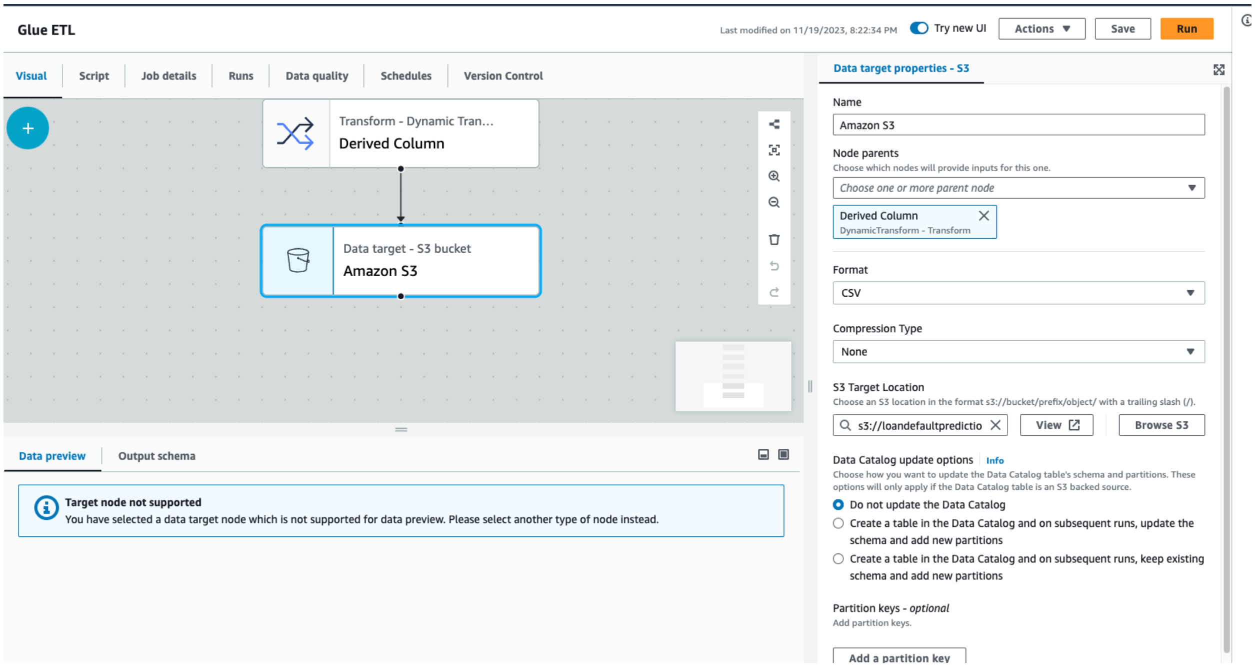
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Drop Duplicates
DropDuplicates - Transform

Name of the derived column
Name to use for the new column or replace an existing one.
Available Amount

SQL Expression
A SQL expression that defines the column, which can be derived from other existing columns and use operators to modify or combine them. For instance, to derive a percentage from the columns "success" and "count", you can enter: "success * 100 / count".

```
`pri.sanctioned.amount`- `pri.disbursed.amount`
```

Created the new column in the data set with attribute name of Available amount , which gives the information about (sanctioned amount - disbursed amount)



Target node not supported
You have selected a data target node which is not supported for data preview. Please select another type of node instead.

Name: Amazon S3

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Derived Column
DynamicTransform - Transform

Format: CSV

Compression Type: None

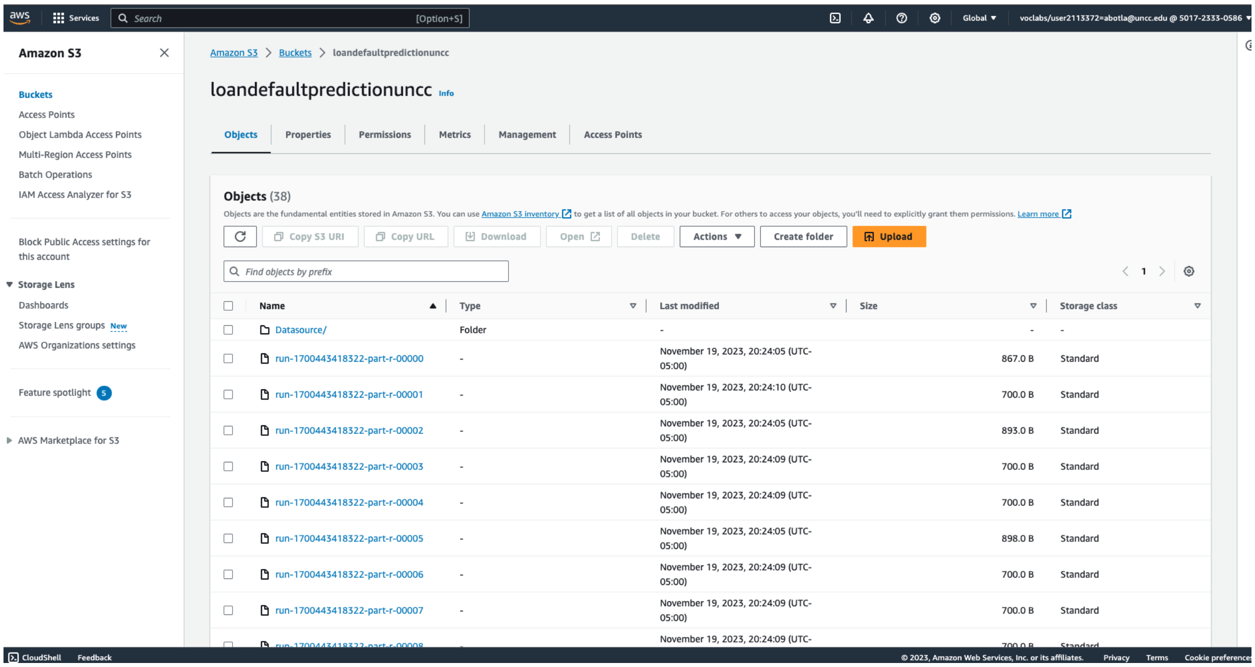
S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://loandefaultpredictio

Data Catalog update options
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.

● Do not update the Data Catalog
○ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
○ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Partition keys - *optional*
Add partition keys.

Final destination:



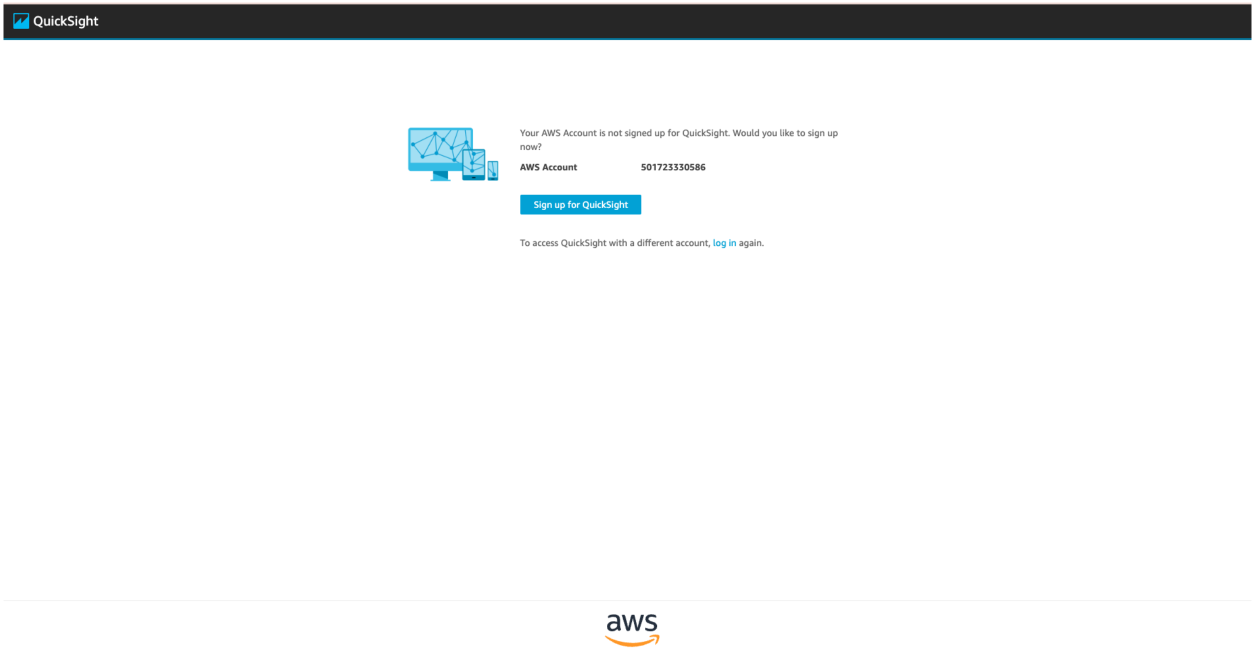
loandefaultpredictionuncc

Objects (38)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	Datasource/	Folder	-	-	-
□	run-1700443418322-part-r-00000	-	November 19, 2023, 20:24:05 (UTC-05:00)	867.0 B	Standard
□	run-1700443418322-part-r-00001	-	November 19, 2023, 20:24:10 (UTC-05:00)	700.0 B	Standard
□	run-1700443418322-part-r-00002	-	November 19, 2023, 20:24:05 (UTC-05:00)	893.0 B	Standard
□	run-1700443418322-part-r-00003	-	November 19, 2023, 20:24:09 (UTC-05:00)	700.0 B	Standard
□	run-1700443418322-part-r-00004	-	November 19, 2023, 20:24:09 (UTC-05:00)	700.0 B	Standard
□	run-1700443418322-part-r-00005	-	November 19, 2023, 20:24:05 (UTC-05:00)	898.0 B	Standard
□	run-1700443418322-part-r-00006	-	November 19, 2023, 20:24:09 (UTC-05:00)	700.0 B	Standard
□	run-1700443418322-part-r-00007	-	November 19, 2023, 20:24:09 (UTC-05:00)	700.0 B	Standard
□	run-1700443418322-part-r-00008	-	November 19, 2023, 20:24:09 (UTC-	700.0 B	Standard

Not able to access the Quick Sight.
Quick sight



Your AWS Account is not signed up for QuickSight. Would you like to sign up now?

AWS Account 501723330586

To access QuickSight with a different account, log in again.