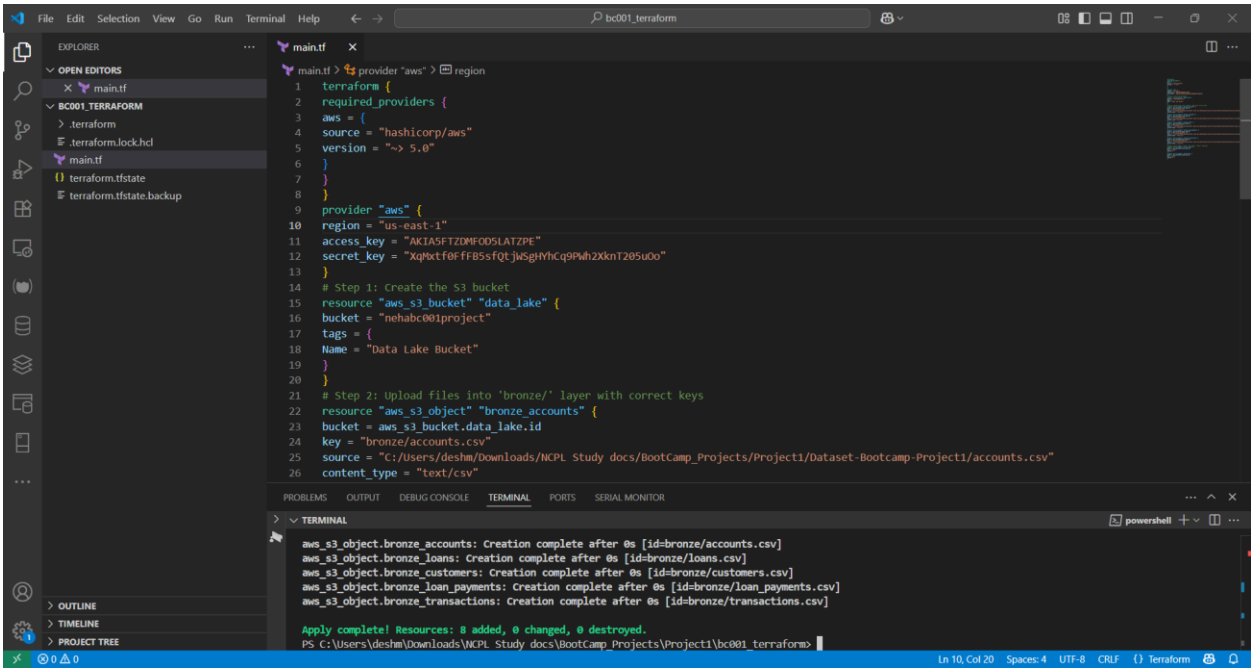
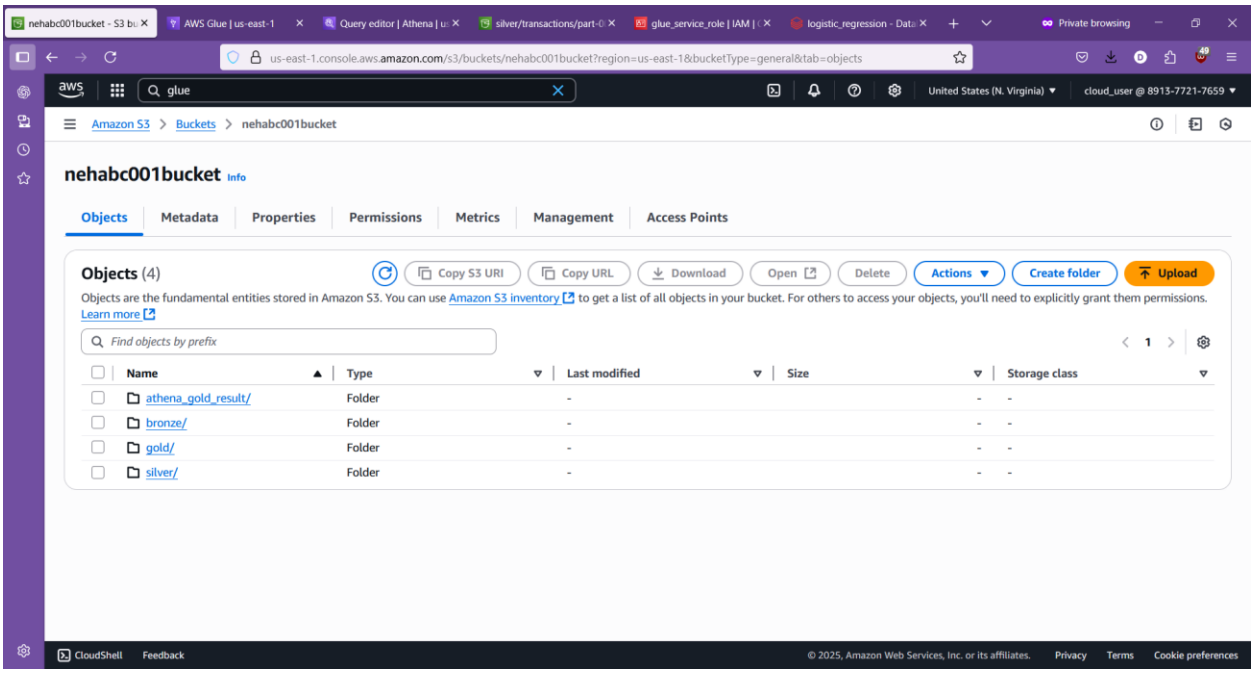
Managing Infrastructure

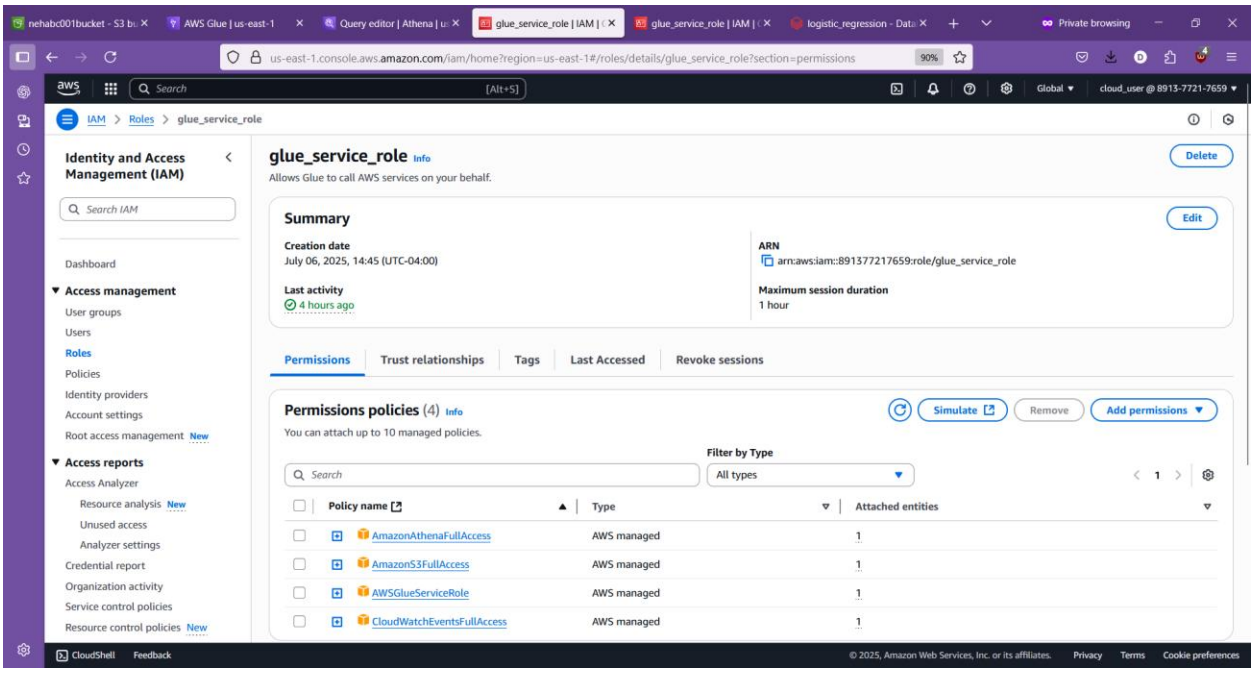
Created Resources using Terraform



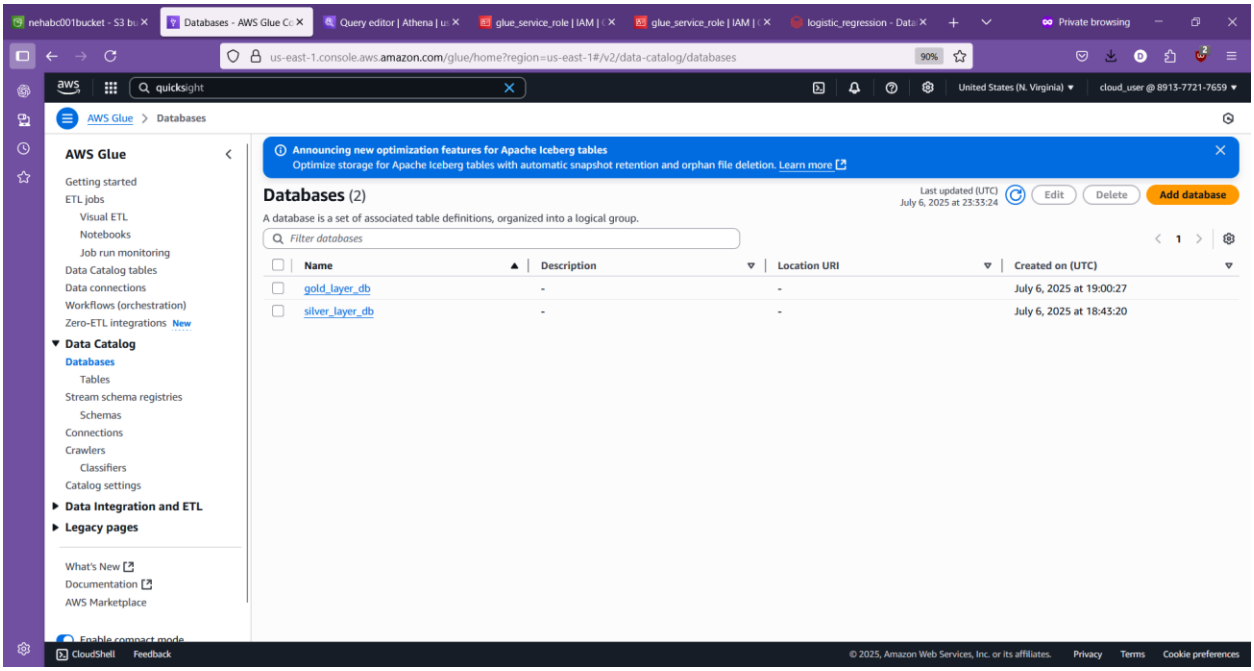
S3 Buckets



IAM Roles and Policies



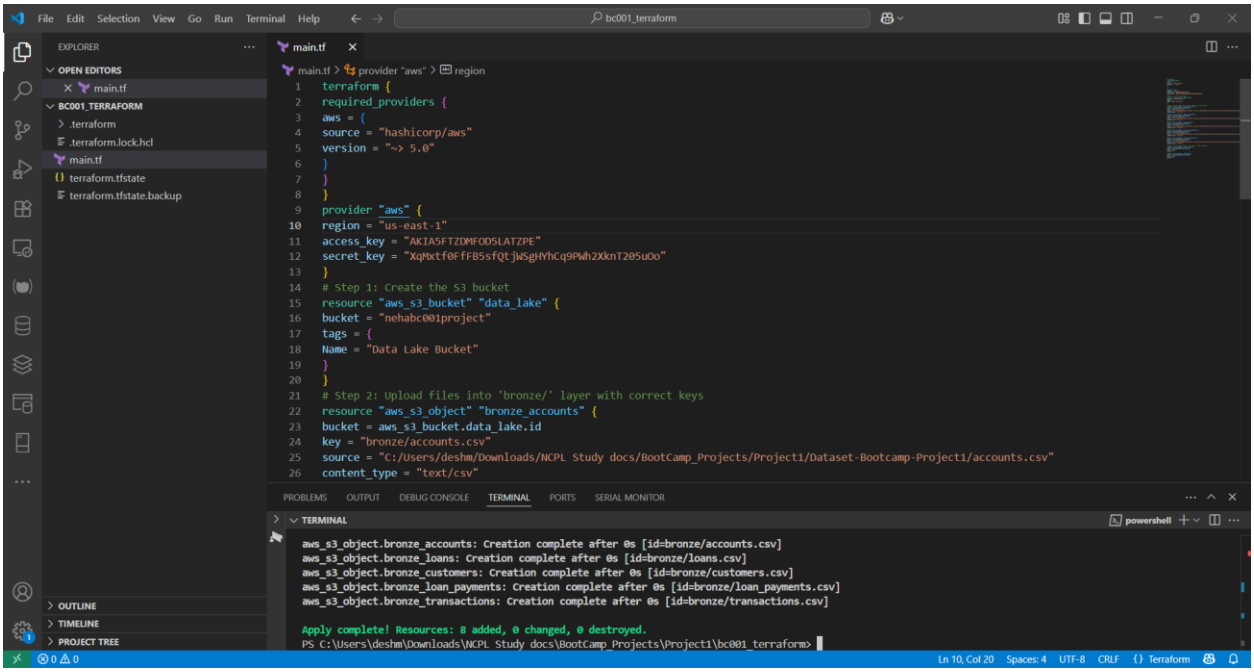
Glue Databases



Data Ingestion - Bronze Layer (Raw Data Collection)

Step 1. Create S3 Bucket using Terraform

Step 1.1. Create your Terraform configuration file.



Code:

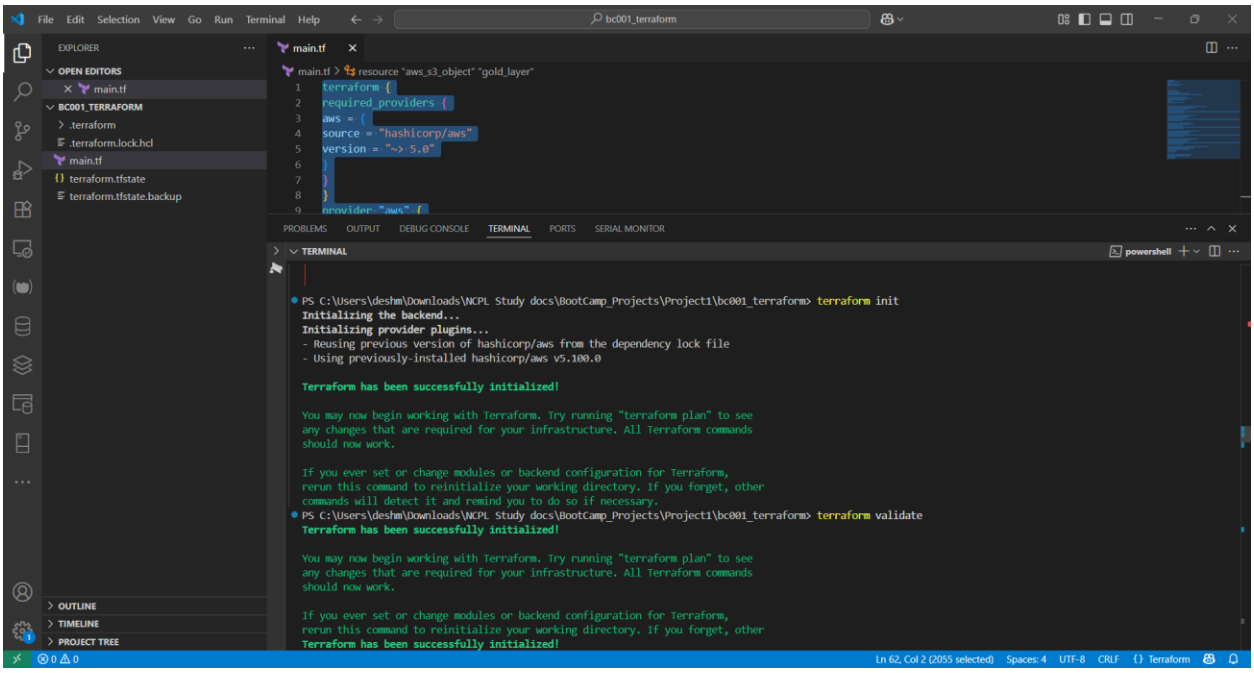
main.tf

```
terraform {
required_providers {
aws = {
source = "hashicorp/aws"
version = "~> 5.0"
}
}
}
provider "aws" {
region = "us-east-1"
```

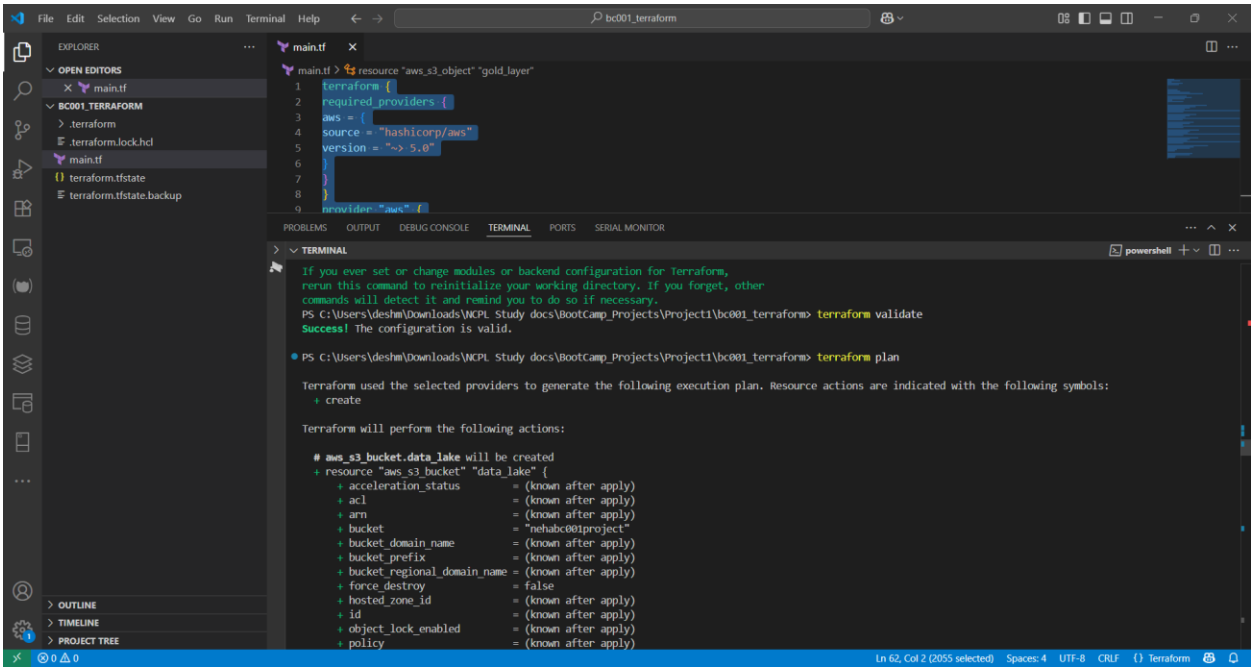
```
access_key = "AKIA5FTZDMFOD5LATZPE"
secret_key = "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"
}
# Step 1: Create the S3 bucket
resource "aws_s3_bucket" "data_lake" {
bucket = "nehabc001project"
tags = {
Name = "Data Lake Bucket"
}
}
# Step 2: Upload files into 'bronze/' layer with correct keys
resource "aws_s3_object" "bronze_accounts" {
bucket = aws_s3_bucket.data_lake.id
key = "bronze/accounts.csv"
source = "C:/Users/deshm/Downloads/NCPL Study docs/BootCamp_Projects/Project1/Dataset-Bootcamp-Project1/accounts.csv"
content_type = "text/csv"
}
resource "aws_s3_object" "bronze_customers" {
bucket = aws_s3_bucket.data_lake.id
key = "bronze/customers.csv"
source = "C:/Users/deshm/Downloads/NCPL Study docs/BootCamp_Projects/Project1/Dataset-Bootcamp-Project1/customers.csv"
content_type= "text/csv"
}
resource "aws_s3_object" "bronze_loans" {
bucket = aws_s3_bucket.data_lake.id
key = "bronze/loans.csv"
source = "C:/Users/deshm/Downloads/NCPL Study docs/BootCamp_Projects/Project1/Dataset-Bootcamp-Project1/loans.csv"
content_type = "text/csv"
```

```
}
resource "aws_s3_object" "bronze_loan_payments" {
bucket = aws_s3_bucket.data_lake.id
key = "bronze/loan_payments.csv"
source = "C:/Users/deshm/Downloads/NCPL Study docs/BootCamp_Projects/Project1/Dataset-Bootcamp-Project1/loan_payments.csv"
content_type = "text/csv"
}
resource "aws_s3_object" "bronze_transactions" {
bucket = aws_s3_bucket.data_lake.id
key = "bronze/transactions.csv"
source = "C:/Users/deshm/Downloads/NCPL Study docs/BootCamp_Projects/Project1/Dataset-Bootcamp-Project1/transactions.csv"
content_type = "text/csv"
}
# Step 3: Create empty 'silver/' and 'gold/' folders (optional)
resource "aws_s3_object" "silver_layer" {
bucket = aws_s3_bucket.data_lake.id
key = "silver/"
content = ""
}
resource "aws_s3_object" "gold_layer" {
bucket = aws_s3_bucket.data_lake.id
key = "gold/"
content = ""
}
```

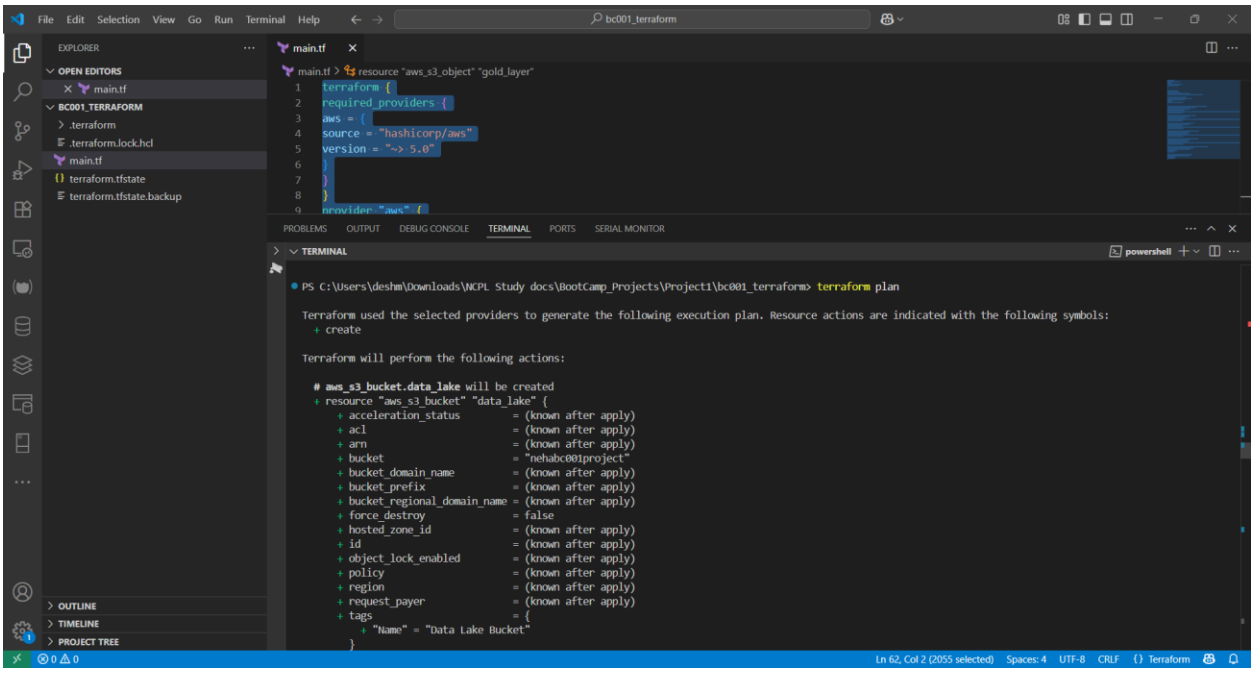
Step 1.2. To run a code and create s3 bucket run this command on terminal: terraform init



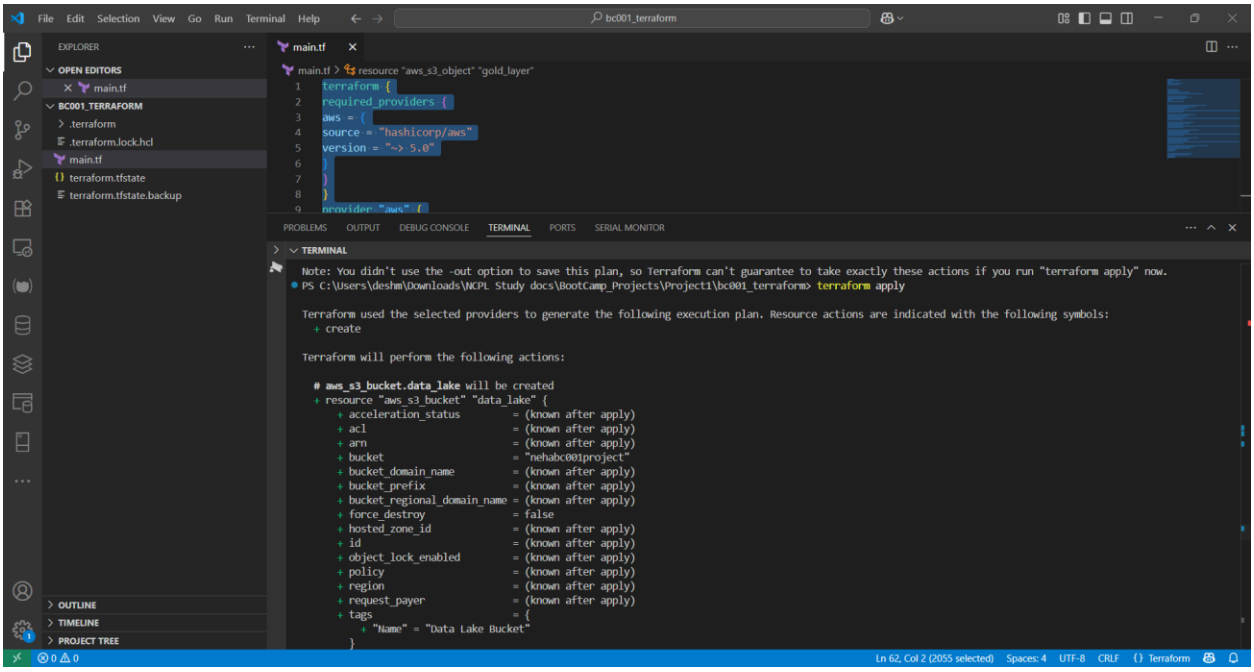
Step 1.3. Run command: terraform validate

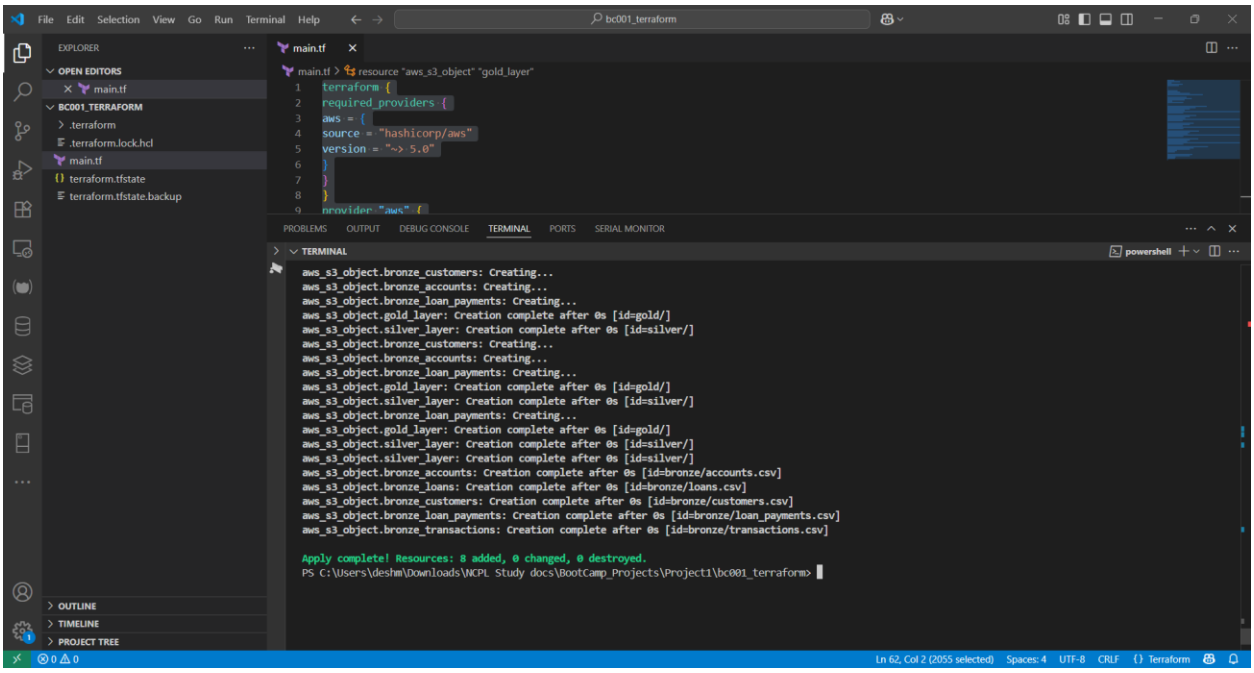


Step 1.4. Run command: terraform plan

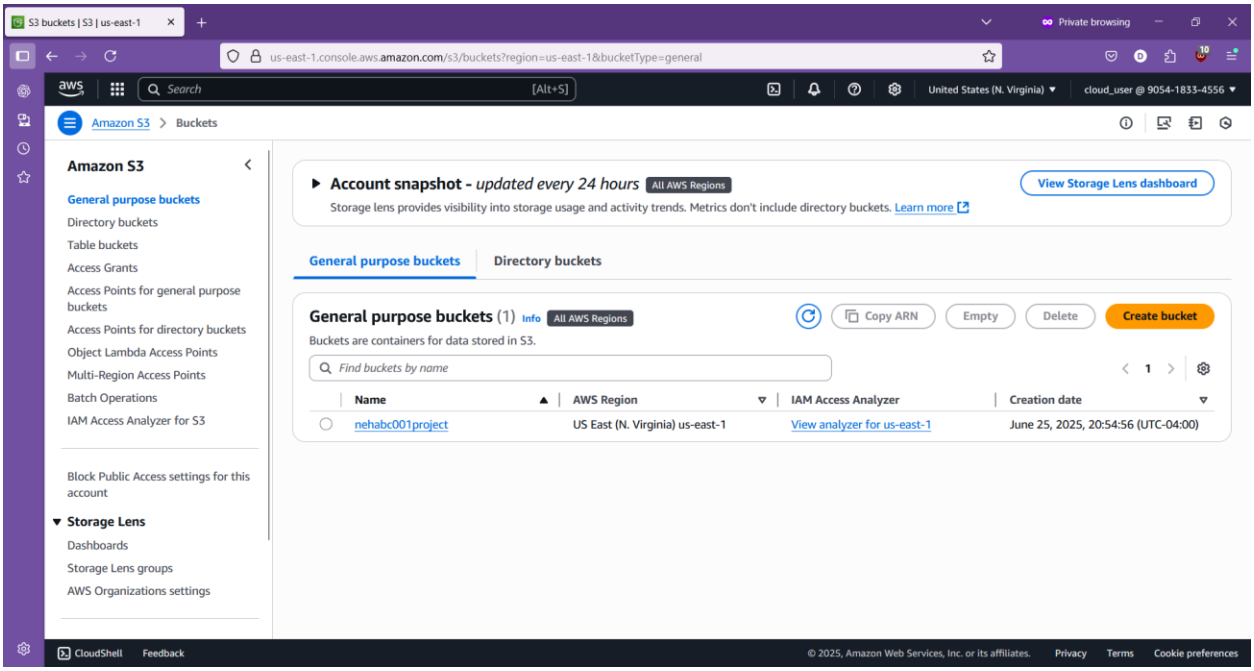


Step 1.5. Run command: terraform apply

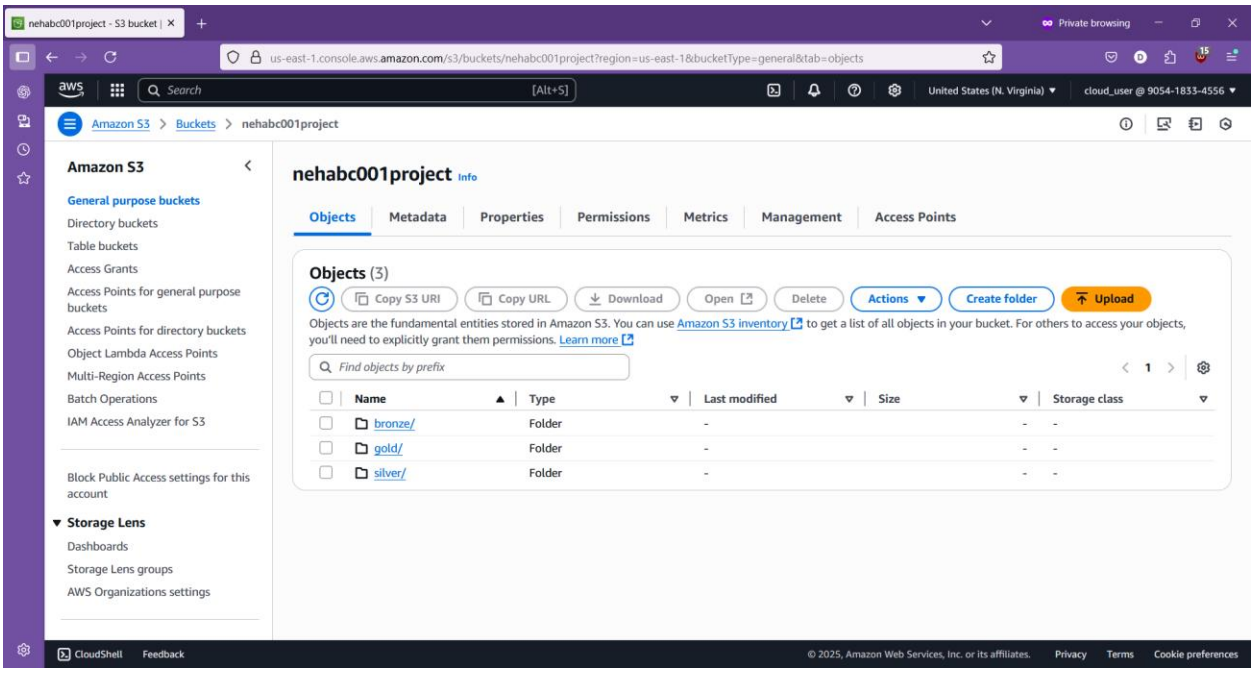




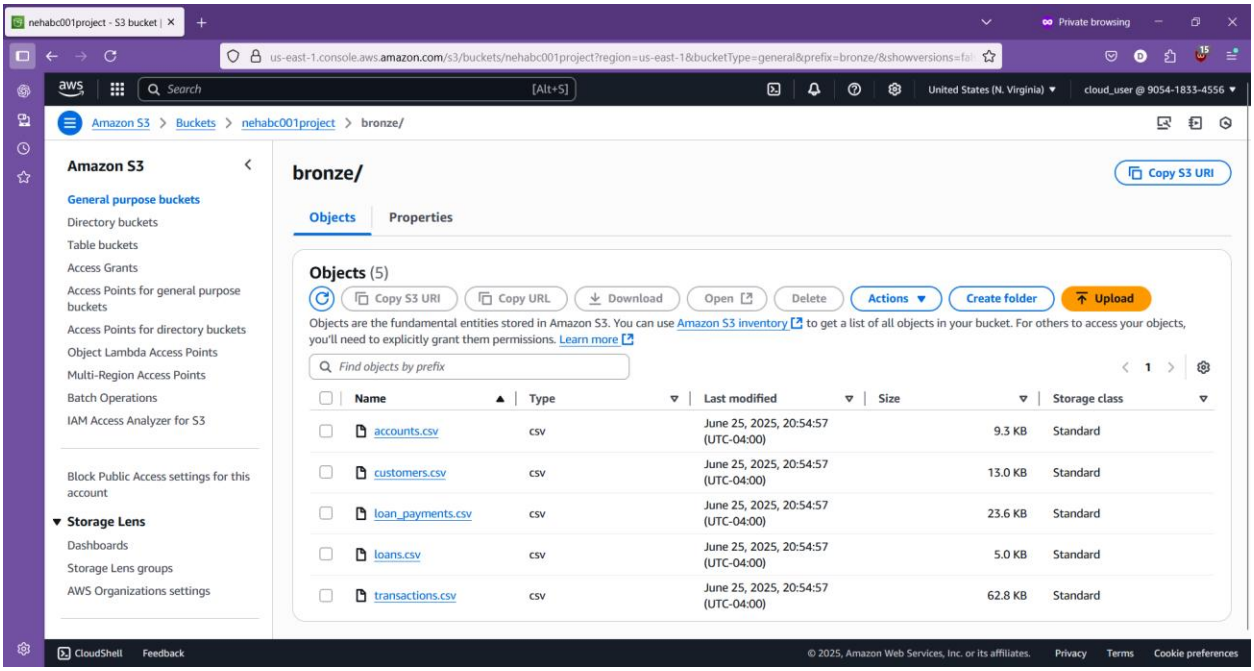
Step 2. Goto AWS console and check s3 bucket.



Subfolders in bucket:



Uploaded csv files in the folder bronze:

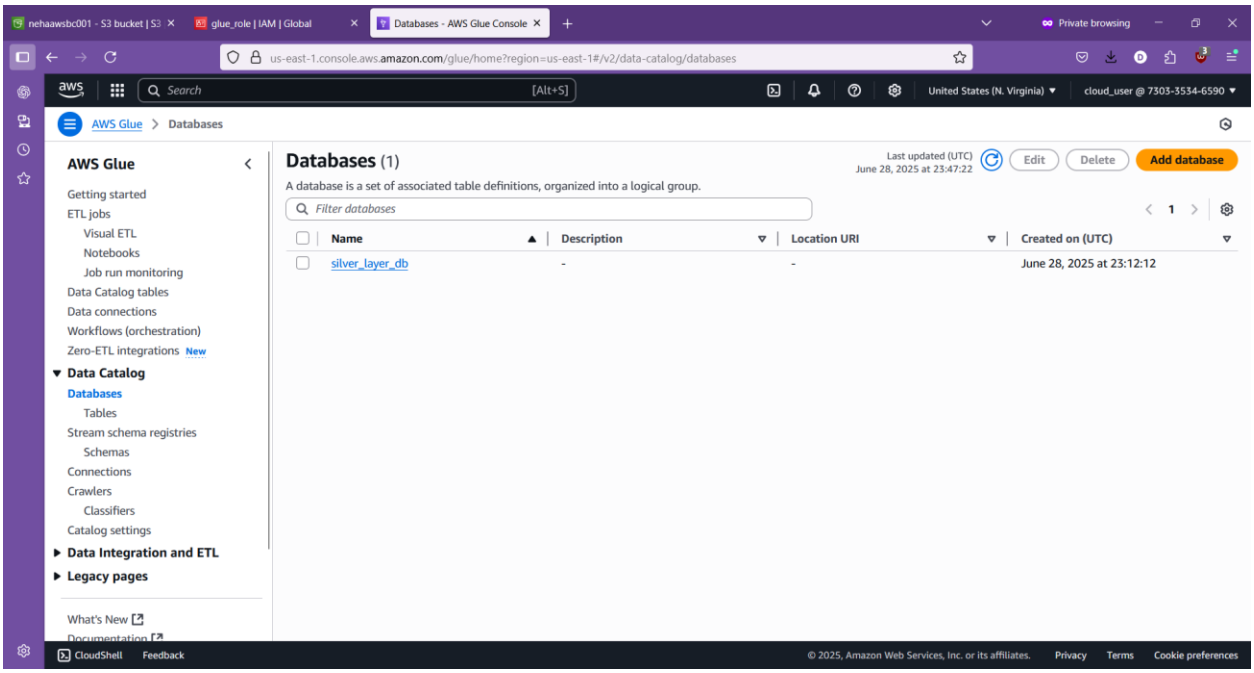


Data Processing - Silver Layer (Cleaned, Trusted Data)

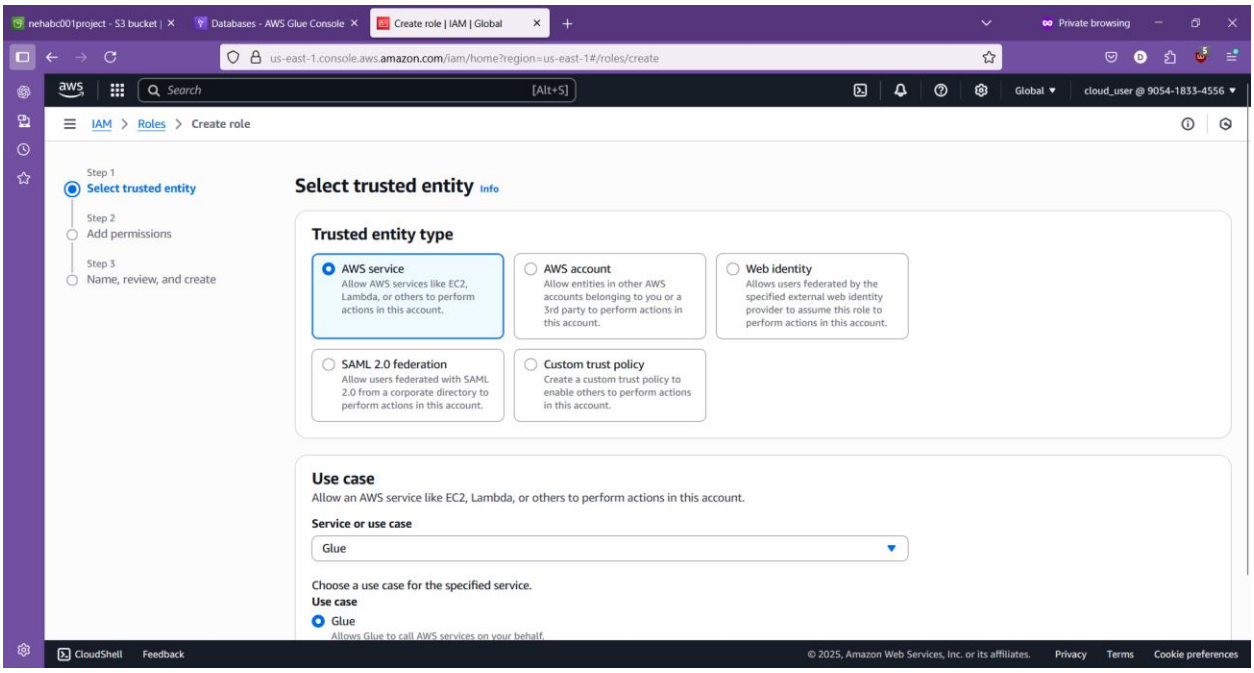
The Silver Layer, also called the Clean Layer, contains data that has been refined from the raw Bronze layer. This includes:

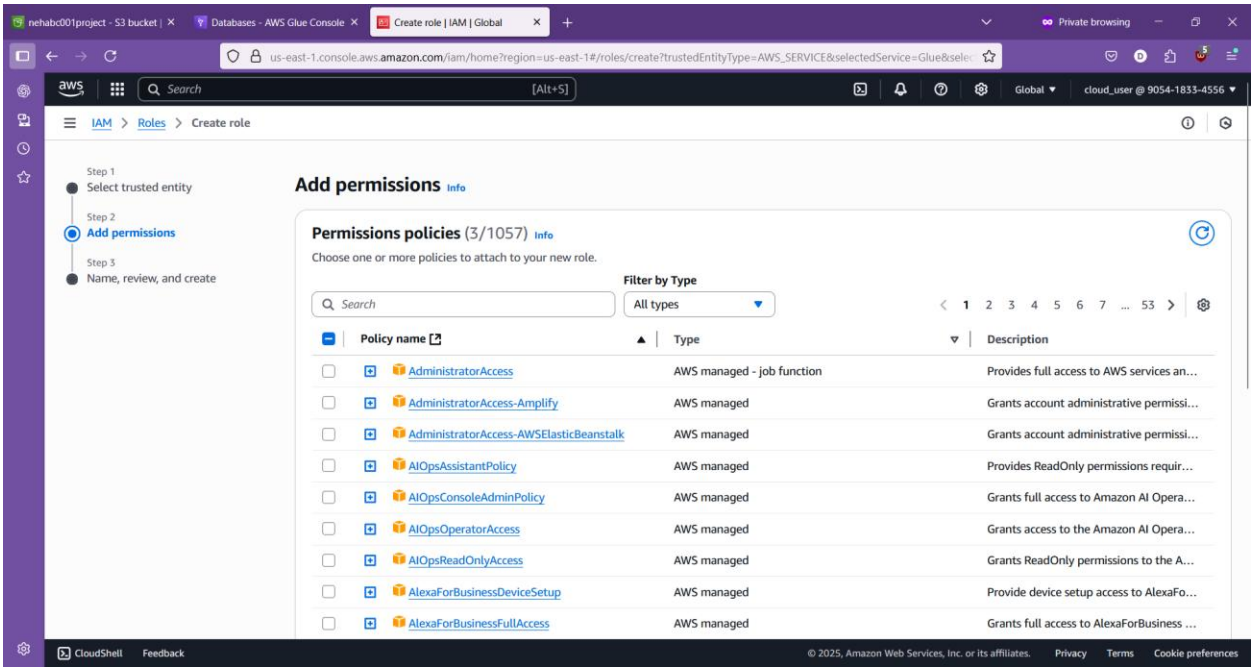
- Filtering out bad data
- Normalizing formats
- Enforcing schema
- Resolving duplicates
- Enriching with business context

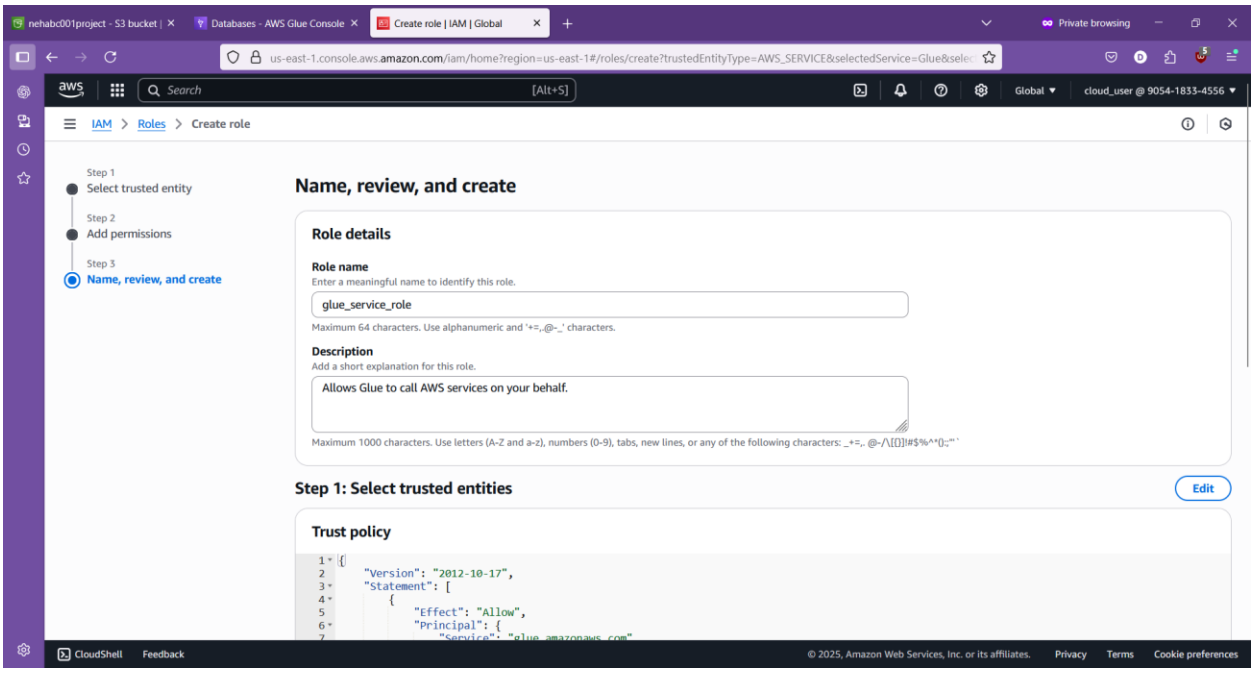
Step 1: Create a database in AWS glue.

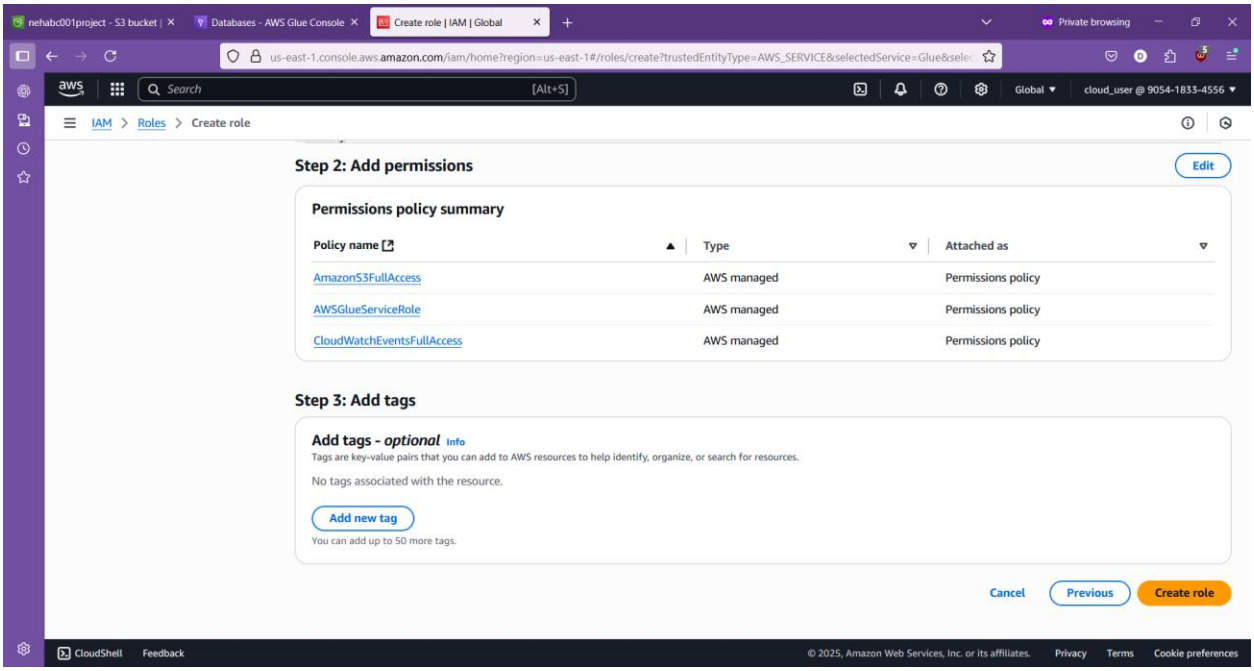


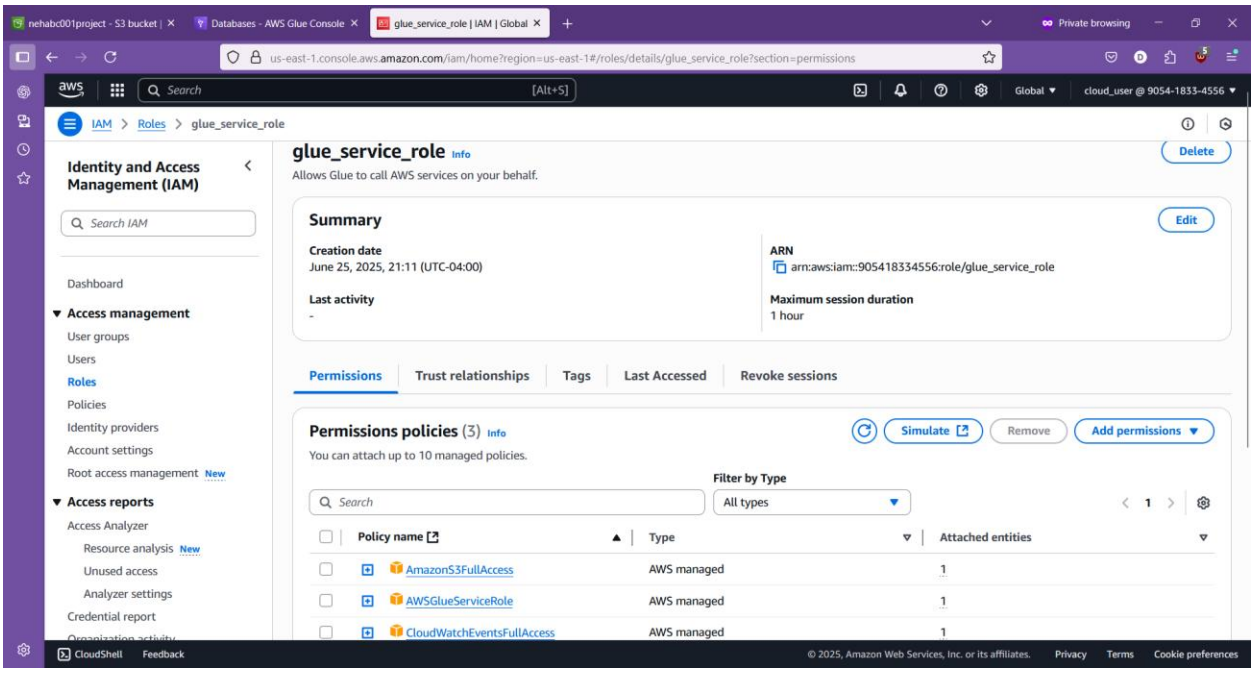
Step 2: Create IAM role for glue.


Create role | IAM | Global
us-east-1.console.aws.amazon.com/iam/home?region=us-east-1#/roles/create
IAM > Roles > Create role
Step 1
Select trusted entity
Step 2
Add permissions
Step 3
Name, review, and create
Select trusted entity Info
Trusted entity type
AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.
AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.
Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.
SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.
Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.
Use case
Allow an AWS service like EC2, Lambda, or others to perform actions in this account.
Service or use case
Glue
Choose a use case for the specified service.
Use case
Glue
Allows Glue to call AWS services on your behalf.
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

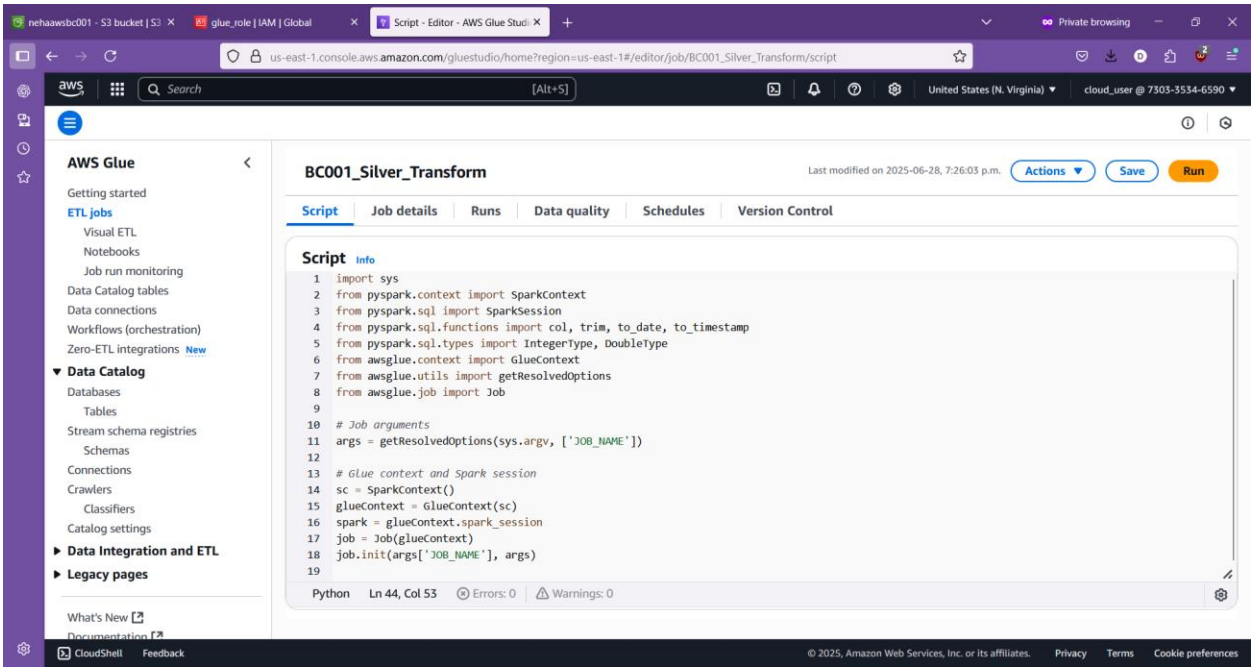

Create role | IAM | Global
IAM > Roles > Create role
Step 1
Select trusted entity
Step 2
Add permissions
Step 3
Name, review, and create
Add permissions Info
Permissions policies (3/1057) Info
Choose one or more policies to attach to your new role.
Filter by Type
All types
Policy name
Type
Description
AdministratorAccess
AWS managed - job function
Provides full access to AWS services an...
AdministratorAccess-Amplify
AWS managed
Grants account administrative permissi...
AdministratorAccess-AWSElasticBeanstalk
AWS managed
Grants account administrative permissi...
AIOpsAssistantPolicy
AWS managed
Provides ReadOnly permissions requir...
AIOpsConsoleAdminPolicy
AWS managed
Grants full access to Amazon AI Opera...
AIOpsOperatorAccess
AWS managed
Grants access to the Amazon AI Opera...
AIOpsReadOnlyAccess
AWS managed
Grants ReadOnly permissions to the A...
AlexaForBusinessDeviceSetup
AWS managed
Provide device setup access to AlexaFo...
AlexaForBusinessFullAccess
AWS managed
Grants full access to AlexaForBusiness ...
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


nehabc001project - S3 bucket |
Databases - AWS Glue Console
Create role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/iam/home?region=us-east-1#/roles/create?trustedEntityType=AWS_SERVICE&selectedService=Glue&sele
Search
[Alt+S]
Global
cloud_user @ 9054-1833-4556
IAM > Roles > Create role
Step 1
Select trusted entity
Step 2
Add permissions
Step 3
Name, review, and create
Name, review, and create
Role details
Role name
Enter a meaningful name to identify this role.
glue_service_role
Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.
Description
Add a short explanation for this role.
Allows Glue to call AWS services on your behalf.
Maximum 1000 characters. Use letters (A-Z and a-z), numbers (0-9), tabs, new lines, or any of the following characters: _+=,. @-/\[{}]!#$%^*{};:"'`
Step 1: Select trusted entities
Edit
Trust policy
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": {
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


nehabc001project - S3 bucket |
Databases - AWS Glue Console
Create role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/iam/home?region=us-east-1#/roles/create?trustedEntityType=AWS_SERVICE&selectedService=Glue&sele
Search
[Alt+S]
Global
cloud_user @ 9054-1833-4556
IAM > Roles > Create role
Step 2: Add permissions
Edit
Permissions policy summary
Policy name
Type
Attached as
AmazonS3FullAccess
AWS managed
Permissions policy
AWSGlueServiceRole
AWS managed
Permissions policy
CloudWatchEventsFullAccess
AWS managed
Permissions policy
Step 3: Add tags
Add tags - optional Info
Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.
No tags associated with the resource.
Add new tag
You can add up to 50 more tags.
Cancel
Previous
Create role
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences



Step 3: Create silver layer job to perform transformations



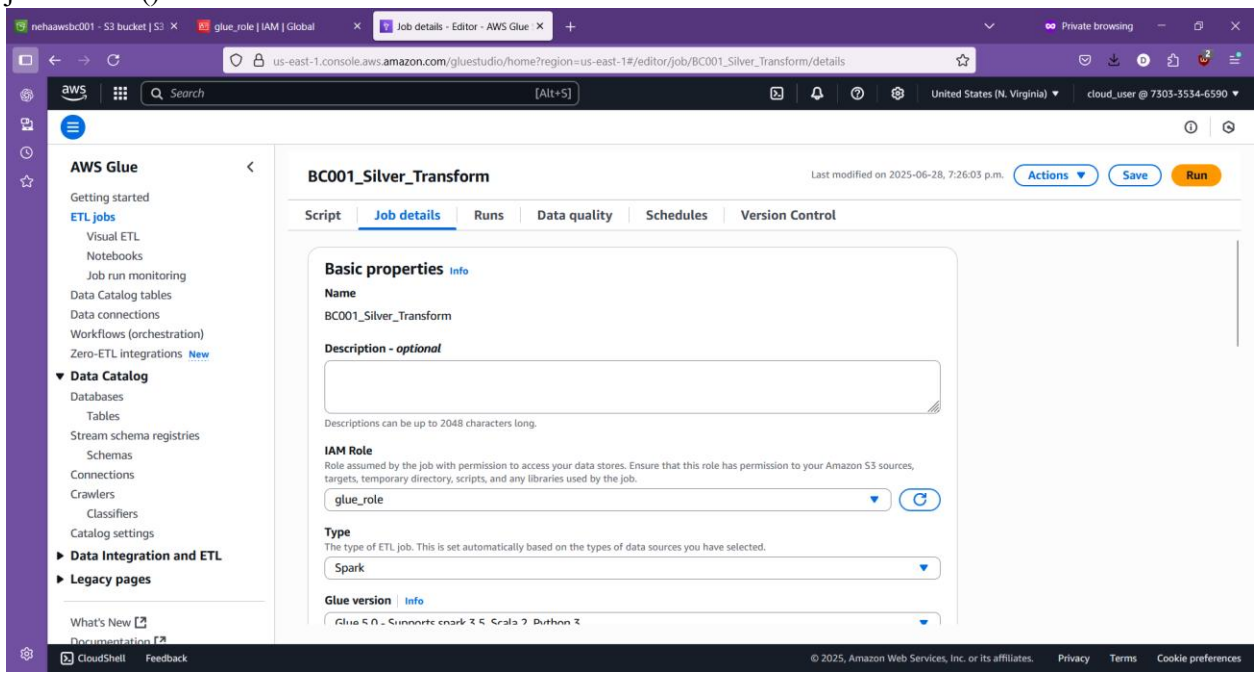
Code:

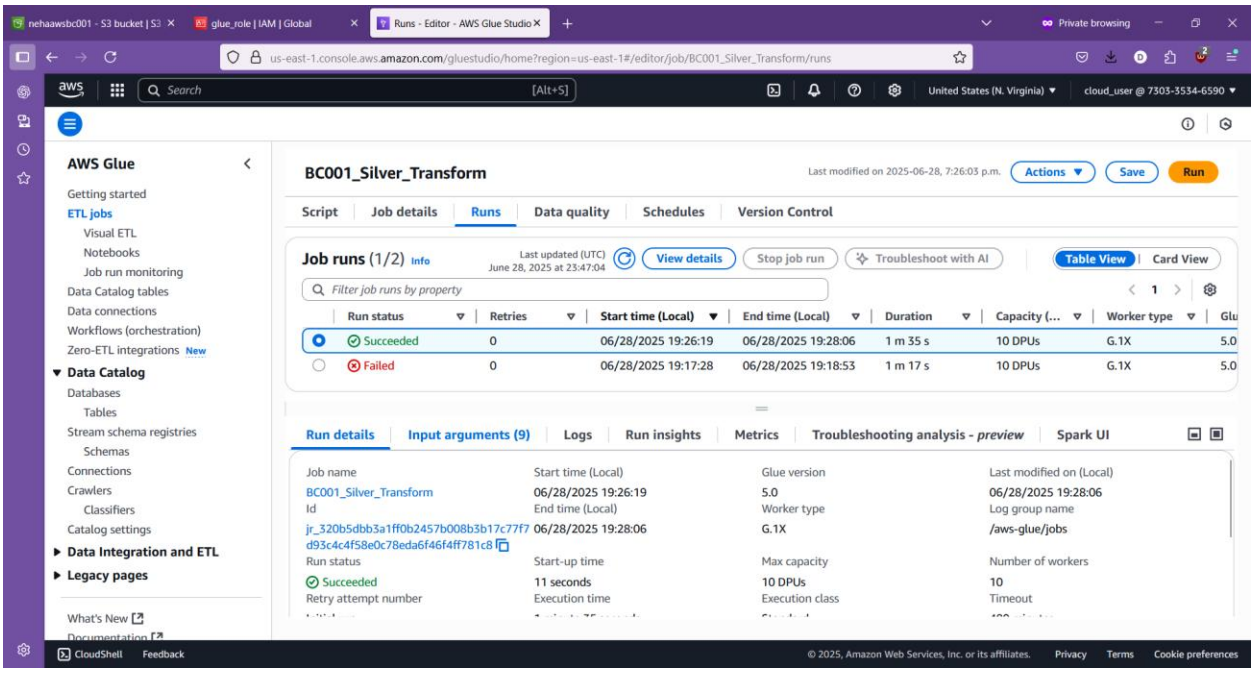
```
import sys
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, trim, to_date, to_timestamp
```

```
from pyspark.sql.types import IntegerType, DoubleType
from awsglue.context import GlueContext
from awsglue.utils import getResolvedOptions
from awsglue.job import Job

# Job arguments
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

# Glue context and Spark session
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)

# ----------------------
# Read Raw Bronze Data
# ----------------------

bronze_path = "s3://nehaawsbc001/bronze/"

# Read each CSV file
customers_df = spark.read.option("header", "true").csv(f"{bronze_path}/customers.csv")
accounts_df = spark.read.option("header", "true").csv(f"{bronze_path}/accounts.csv")
transactions_df = spark.read.option("header", "true").csv(f"{bronze_path}/transactions.csv")
loans_df = spark.read.option("header", "true").csv(f"{bronze_path}/loans.csv")
loan_payments_df = spark.read.option("header", "true").csv(f"{bronze_path}/loan_payments.csv")

# ----------------------
# Cleaning and Transformation
```

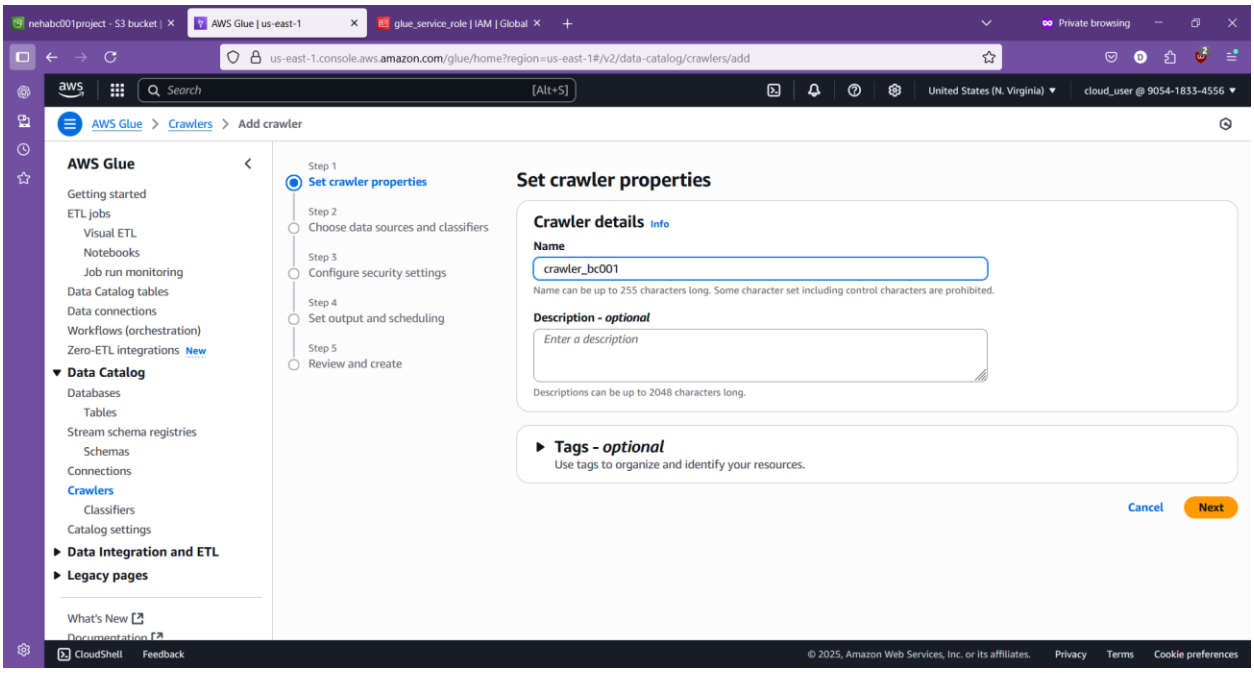
```
# ----------------------

# Clean customers
customers_clean = customers_df.dropDuplicates(["customer_id"]).na.drop(subset=["customer_id"])
customers_clean = customers_clean.withColumn("customer_id", col("customer_id").cast("string")) \
                                 .withColumn("email", trim(col("email"))) \
                                 .withColumn("date_joined", to_date("date_joined", "yyyy-MM-dd"))

# Clean accounts
accounts_clean = accounts_df.dropDuplicates(["account_id"]) \
                            .withColumn("account_id", col("account_id").cast("string")) \
                            .withColumn("customer_id", col("customer_id").cast("string")) \
                            .withColumn("balance", col("balance").cast(DoubleType()))

# Clean transactions
transactions_clean = transactions_df.dropDuplicates(["transaction_id"]) \
    .withColumn("transaction_id", col("transaction_id").cast("string")) \
    .withColumn("account_id", col("account_id").cast("string")) \
    .withColumn("amount", col("amount").cast(DoubleType()))

# Clean loans
loans_clean = loans_df.dropDuplicates(["loan_id"]) \
    .withColumn("loan_id", col("loan_id").cast("string")) \
    .withColumn("customer_id", col("customer_id").cast("string")) \
    .withColumn("loan_amount", col("loan_amount").cast(DoubleType()))

# Clean loan payments
loan_payments_clean = loan_payments_df.dropDuplicates(["payment_id"]) \
    .withColumn("payment_id", col("payment_id").cast("string")) \
    .withColumn("loan_id", col("loan_id").cast("string")) \
```

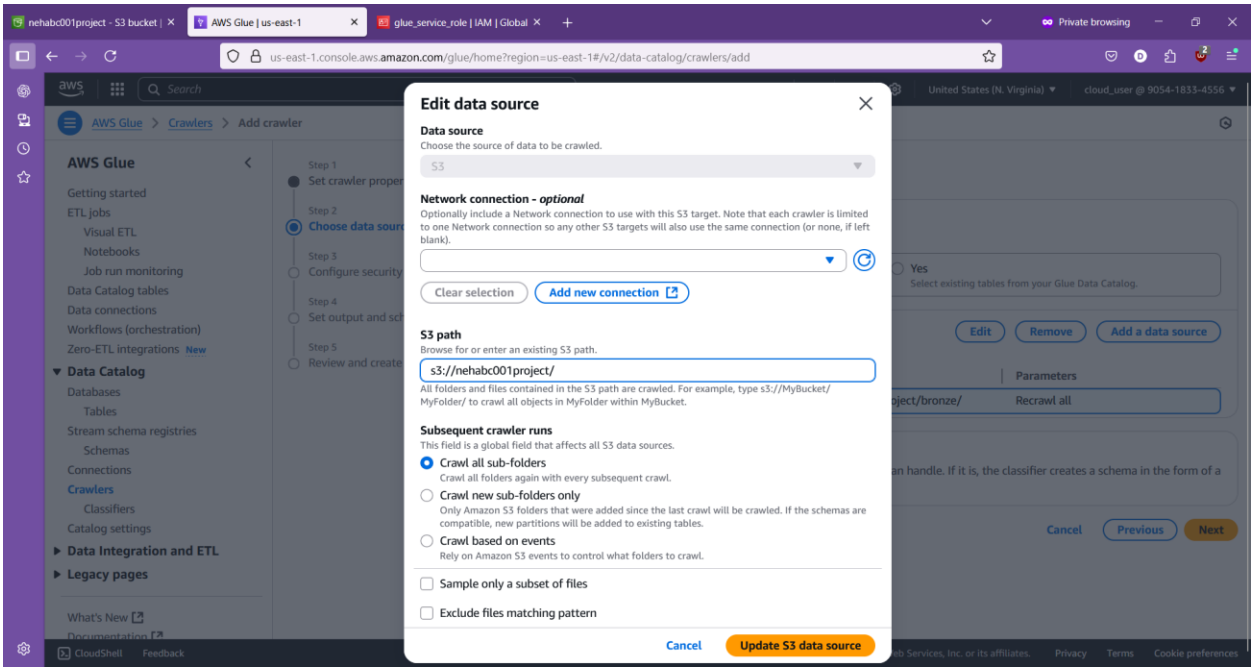
```
    .withColumn("payment_amount", col("payment_amount").cast(DoubleType()))

# ----------------------
# Write to Silver Layer
# ----------------------

silver_path = "s3://nehaawsbc001/silver/"

customers_clean.write.mode("overwrite").parquet(f"{silver_path}/customers/")
accounts_clean.write.mode("overwrite").parquet(f"{silver_path}/accounts/")
transactions_clean.write.mode("overwrite").parquet(f"{silver_path}/transactions/")
loans_clean.write.mode("overwrite").parquet(f"{silver_path}/loans/")
loan_payments_clean.write.mode("overwrite").parquet(f"{silver_path}/loan_payments/")

job.commit()
```

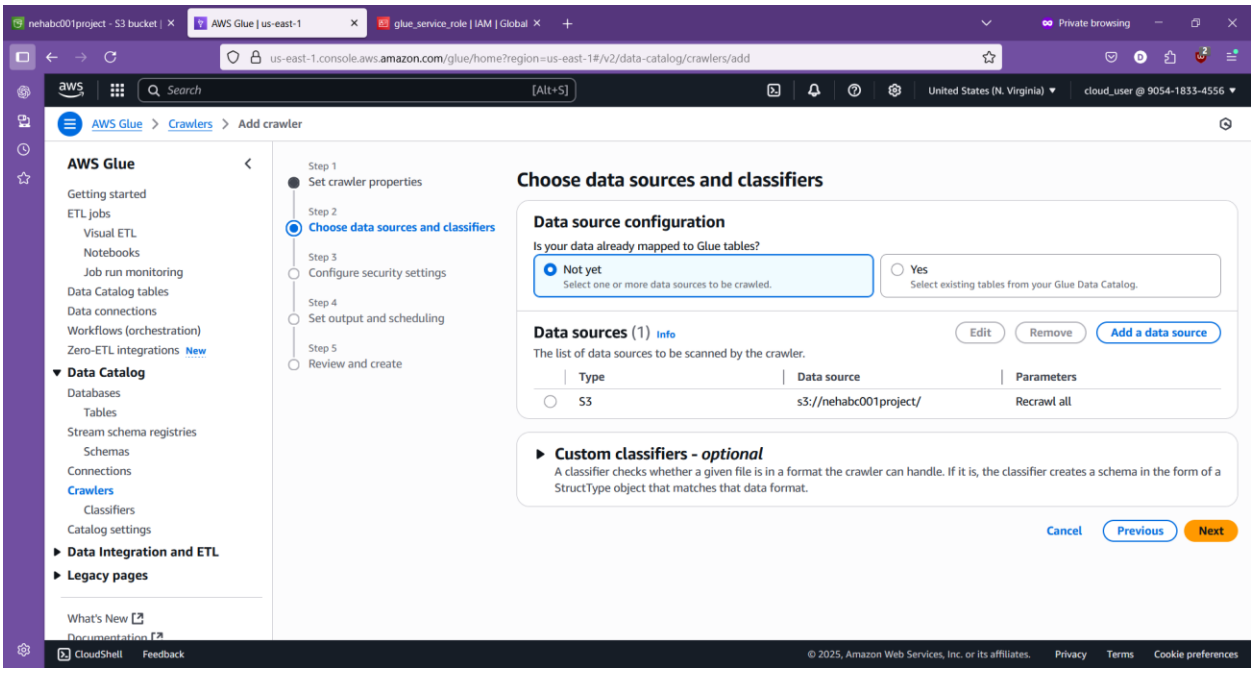


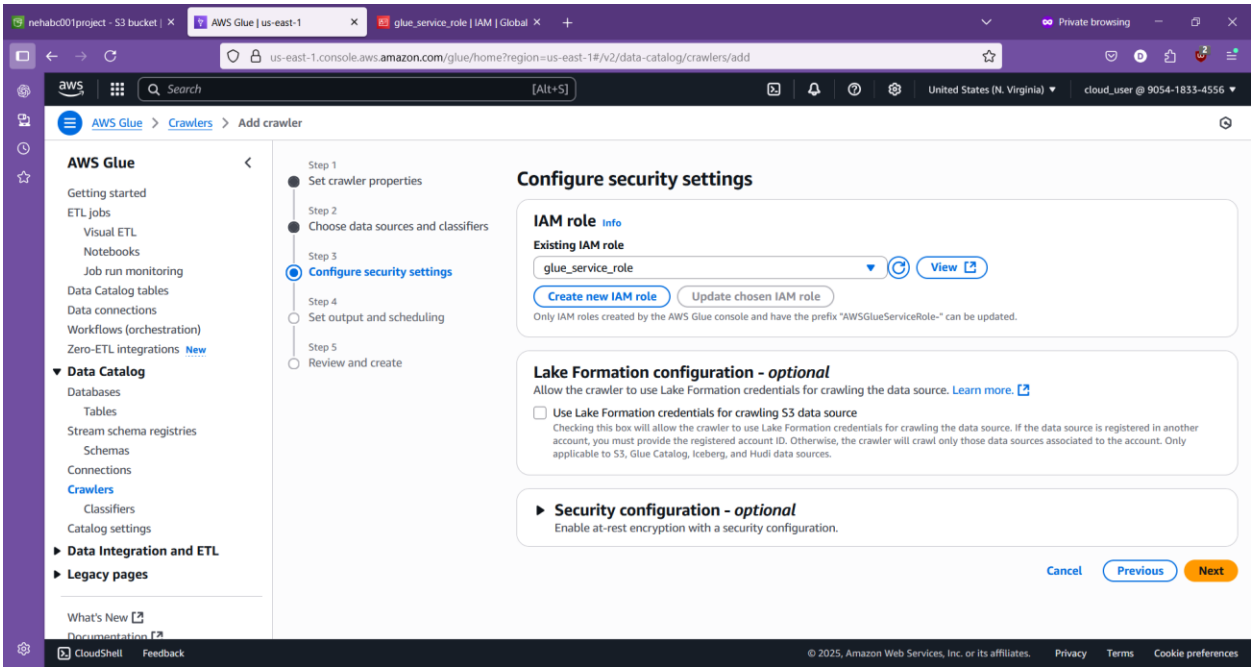


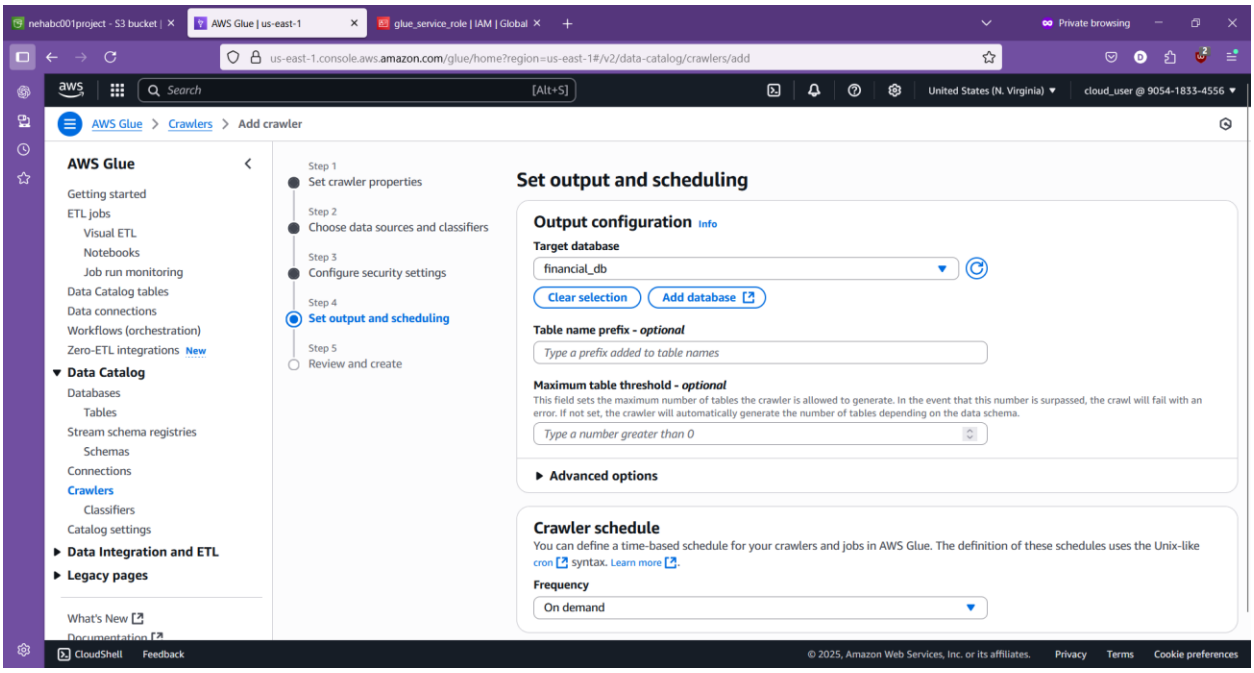
Step 4: Create a crawler and give database location.

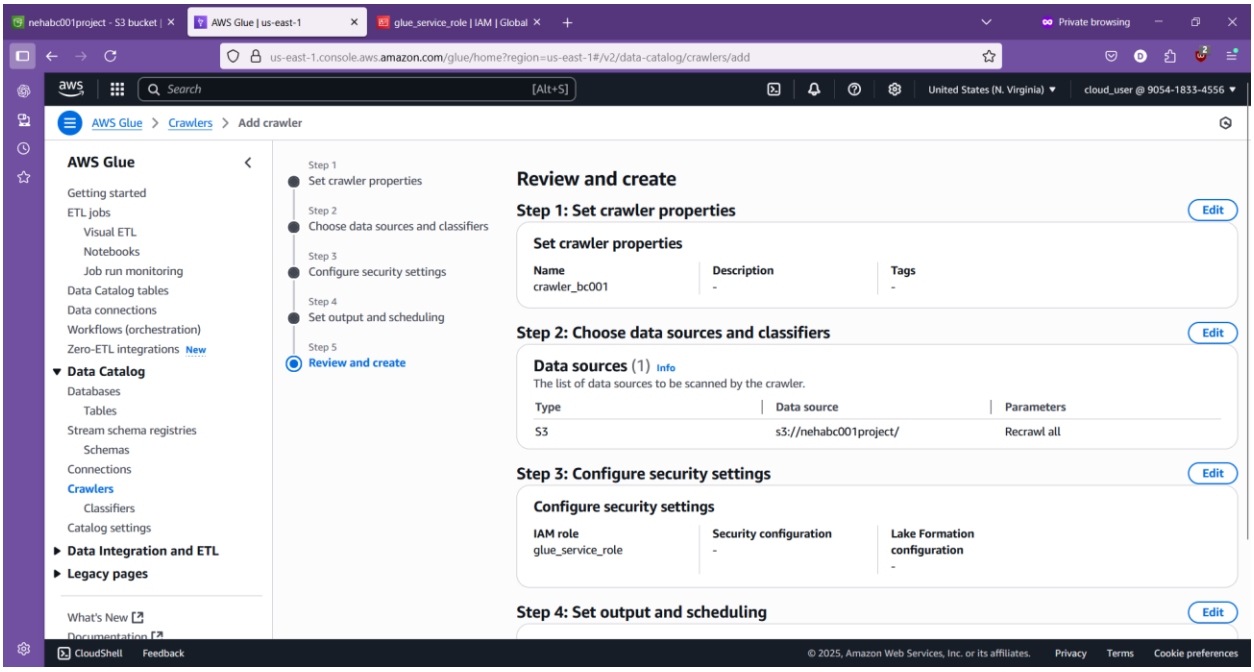

nehabc001project - S3 bucket |
AWS Glue | us-east-1
glue_service_role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 9054-1833-4556
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Set crawler properties
Crawler details Info
Name
crawler_bc001
Name can be up to 255 characters long. Some character set including control characters are prohibited.
Description - optional
Enter a description
Descriptions can be up to 2048 characters long.
Tags - optional
Use tags to organize and identify your resources.
Cancel
Next
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

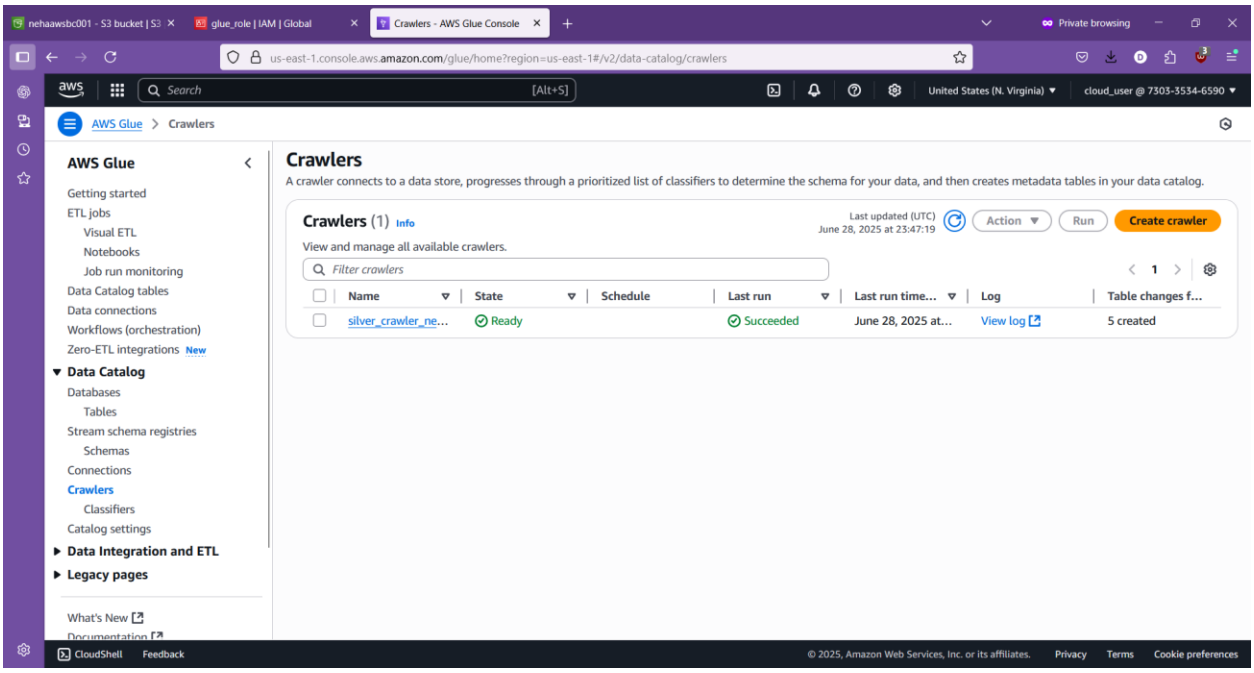

Edit data source
Data source
Choose the source of data to be crawled.
S3
Network connection - optional
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).
Clear selection
Add new connection
S3 path
Browse for or enter an existing S3 path.
s3://nehabc001project/
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.
Subsequent crawler runs
This field is a global field that affects all S3 data sources.
Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.
Sample only a subset of files
Exclude files matching pattern
Cancel
Update S3 data source

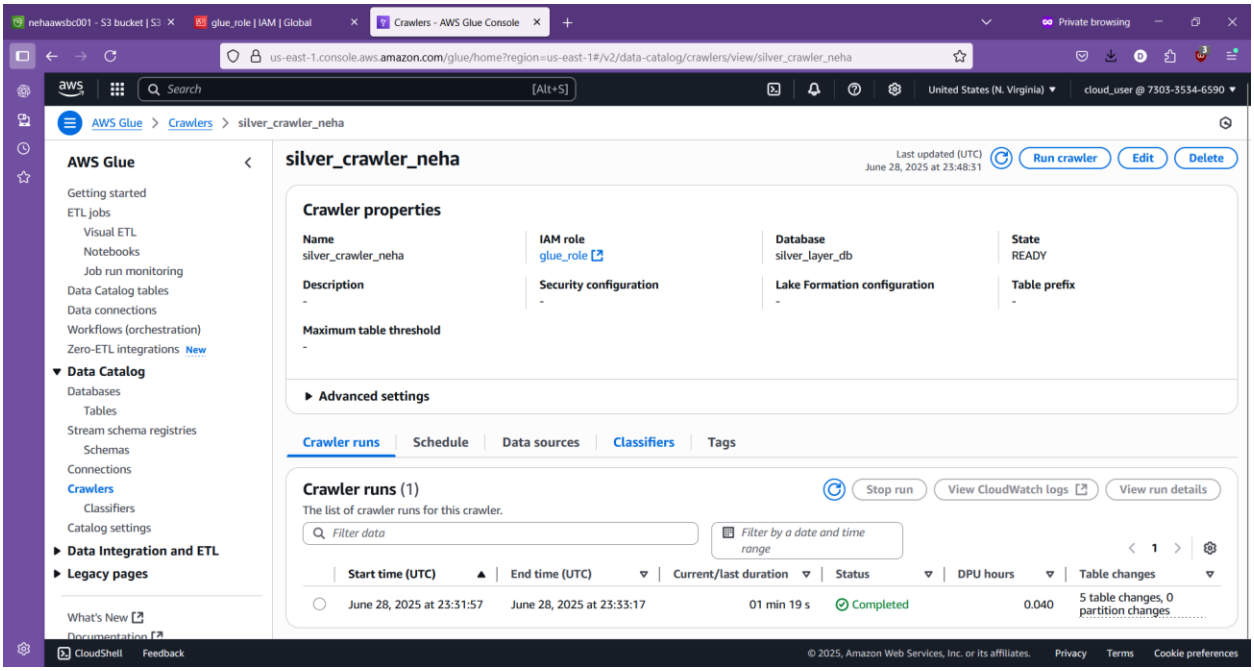

nehabc001project - S3 bucket |
AWS Glue | us-east-1
glue_service_role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 9054-1833-4556
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (1) Info
Edit
Remove
Add a data source
The list of data sources to be scanned by the crawler.
Type
Data source
Parameters
S3
s3://nehabc001project/
Recrawl all
Custom classifiers - optional
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.
Cancel
Previous
Next
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


nehabc001project - S3 bucket |
AWS Glue | us-east-1
glue_service_role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 9054-1833-4556
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Configure security settings
IAM role Info
Existing IAM role
glue_service_role
View
Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.
Lake Formation configuration - optional
Allow the crawler to use Lake Formation credentials for crawling the data source. Learn more.
Use Lake Formation credentials for crawling S3 data source
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.
Security configuration - optional
Enable at-rest encryption with a security configuration.
Cancel
Previous
Next
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

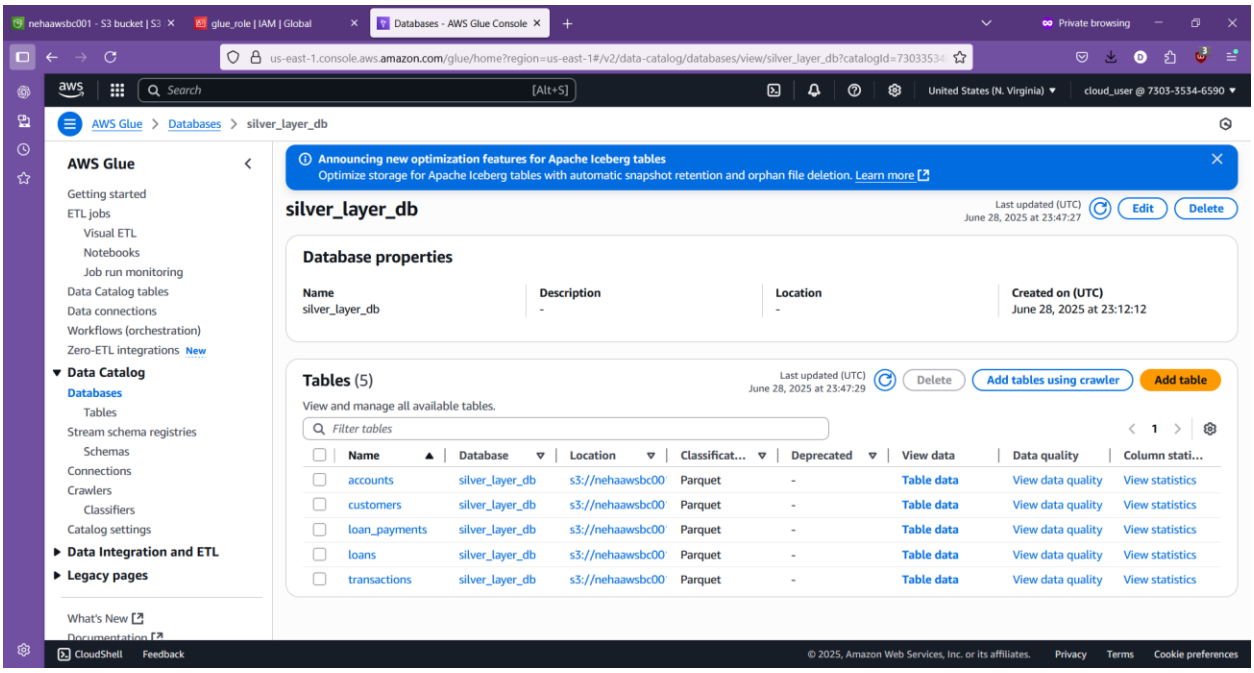

nehabc001project - S3 bucket |
AWS Glue | us-east-1
glue_service_role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 9054-1833-4556
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Set output and scheduling
Output configuration Info
Target database
financial_db
Clear selection
Add database
Table name prefix - optional
Type a prefix added to table names
Maximum table threshold - optional
This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.
Type a number greater than 0
Advanced options
Crawler schedule
You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more
Frequency
On demand
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


nehabc001project - S3 bucket |
AWS Glue | us-east-1
glue_service_role | IAM | Global
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 9054-1833-4556
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Review and create
Step 1: Set crawler properties
Edit
Set crawler properties
Name
crawler_bc001
Description
-
Tags
-
Step 2: Choose data sources and classifiers
Edit
Data sources (1) Info
The list of data sources to be scanned by the crawler.
Type
Data source
Parameters
S3
s3://nehabc001project/
Recrawl all
Step 3: Configure security settings
Edit
Configure security settings
IAM role
glue_service_role
Security configuration
-
Lake Formation configuration
-
Step 4: Set output and scheduling
Edit
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences





Step 5: Check tables under database


Databases - AWS Glue Console
AWS Glue > Databases > silver_layer_db
Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. Learn more
silver_layer_db
Last updated (UTC) June 28, 2025 at 23:47:27
Edit
Delete
Database properties
Name
silver_layer_db
Description
-
Location
-
Created on (UTC)
June 28, 2025 at 23:12:12
Tables (5)
Last updated (UTC) June 28, 2025 at 23:47:29
Delete
Add tables using crawler
Add table
View and manage all available tables.
Filter tables
Name
Database
Location
Classificat...
Deprecated
View data
Data quality
Column stati...
accounts silver_layer_db s3://nehaawsbc00 Parquet - Table data View data quality View statistics
customers silver_layer_db s3://nehaawsbc00 Parquet - Table data View data quality View statistics
loan_payments silver_layer_db s3://nehaawsbc00 Parquet - Table data View data quality View statistics
loans silver_layer_db s3://nehaawsbc00 Parquet - Table data View data quality View statistics
transactions silver_layer_db s3://nehaawsbc00 Parquet - Table data View data quality View statistics
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

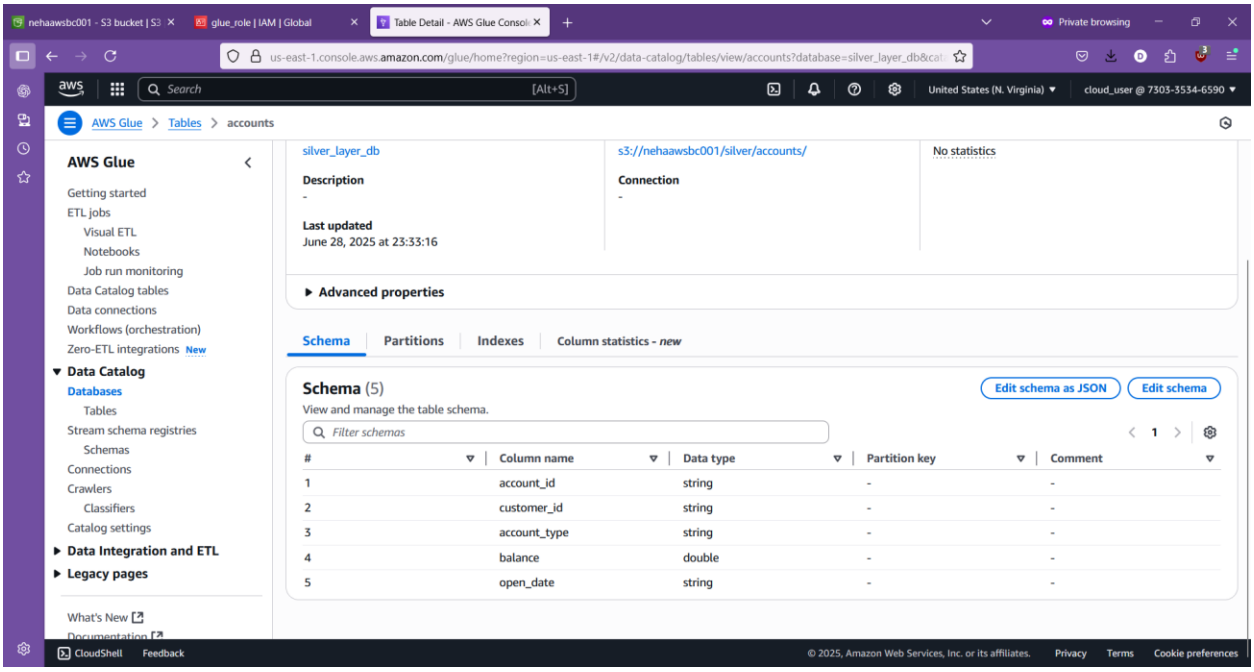
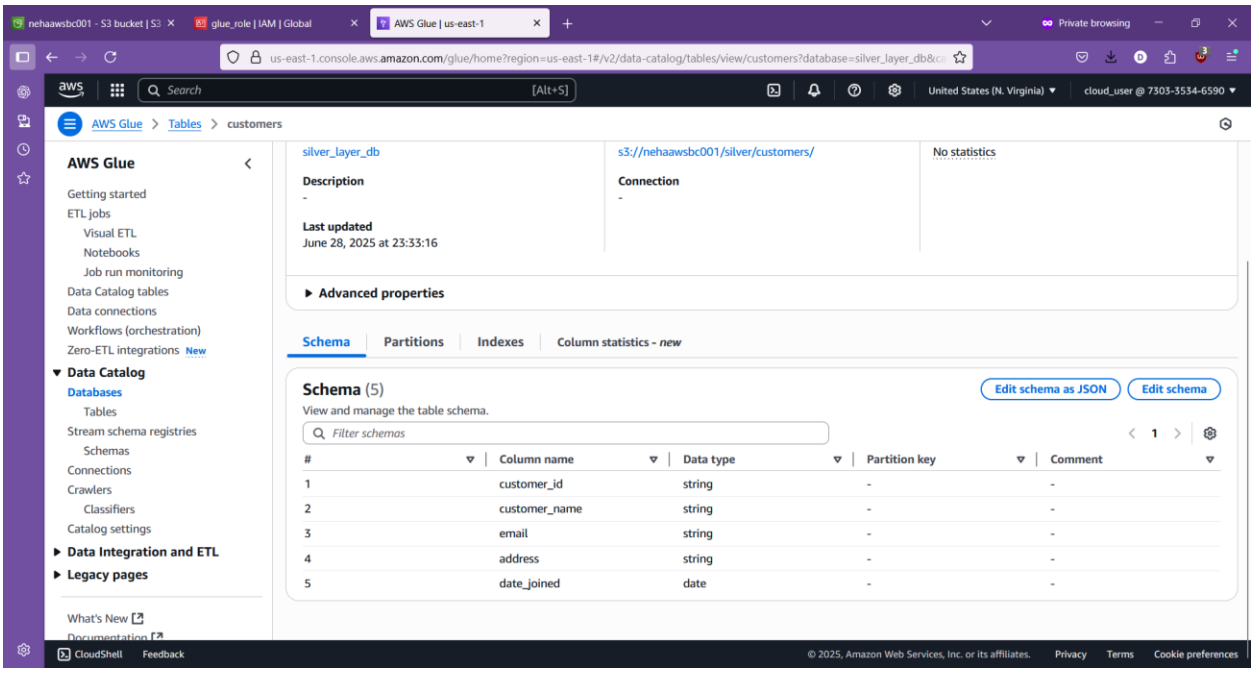
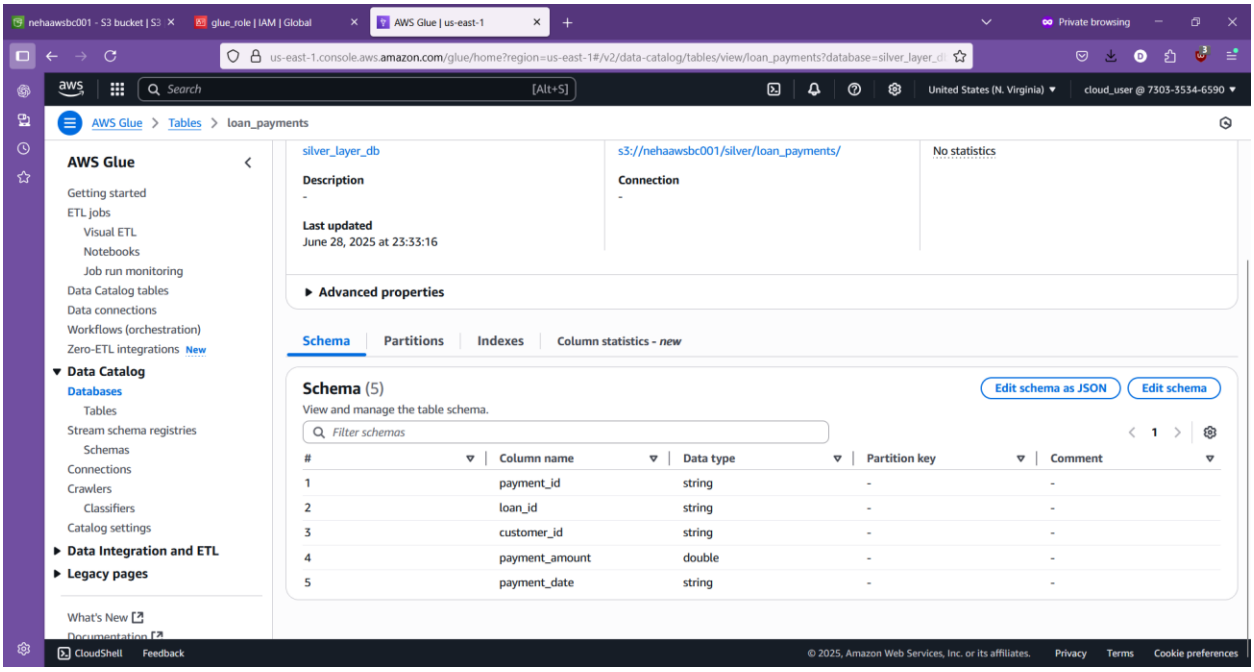
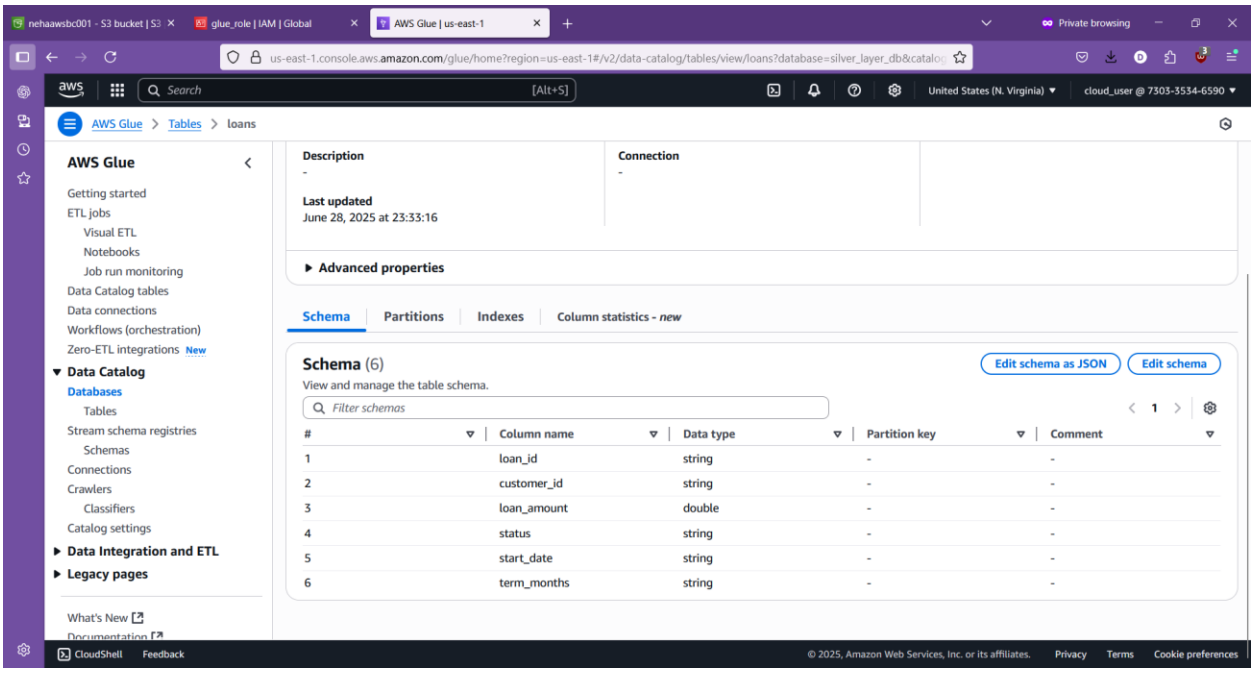
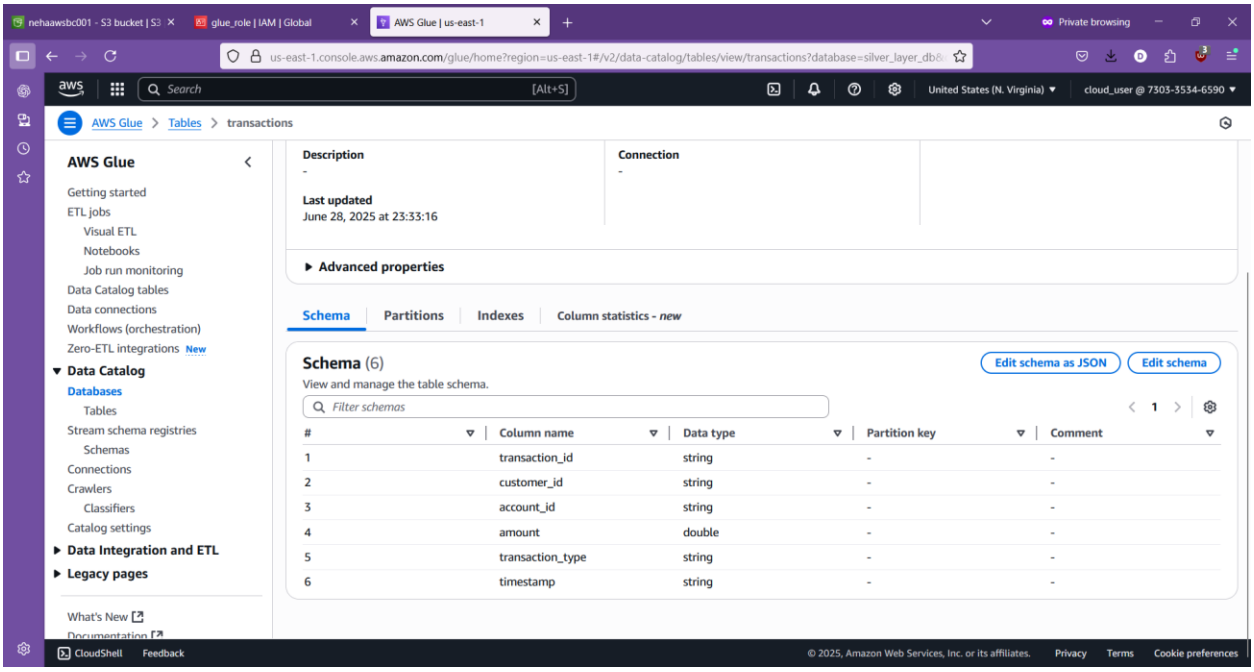


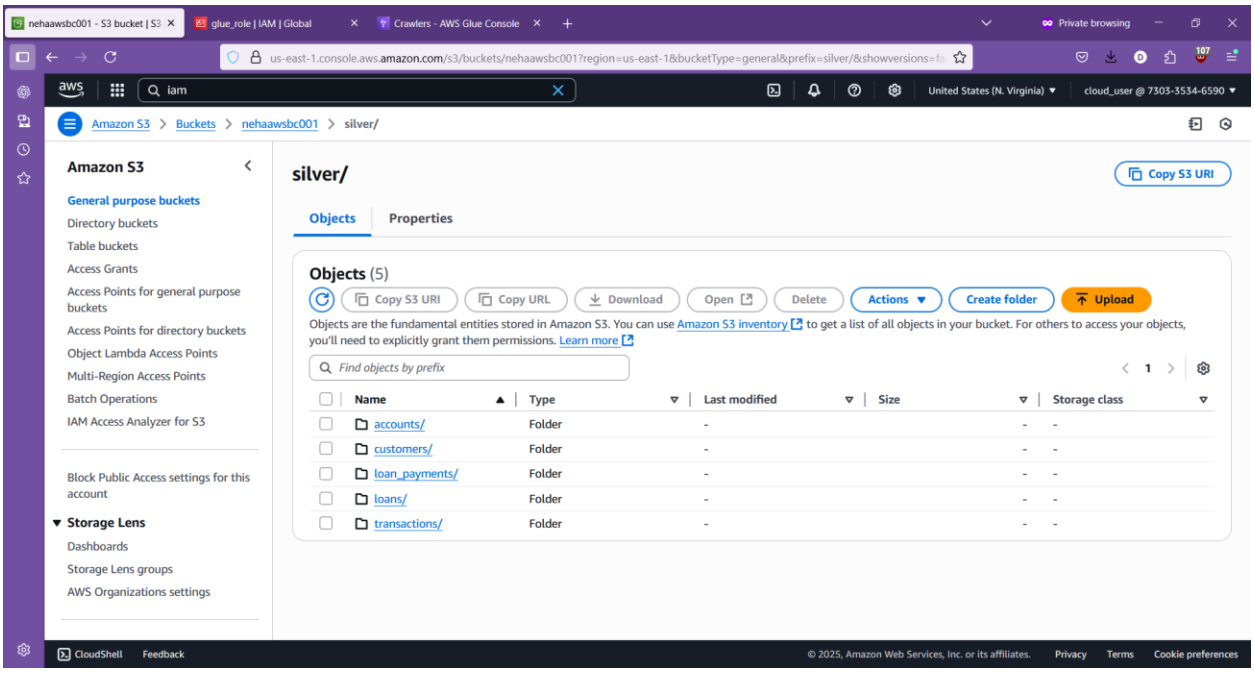
Table Detail - AWS Glue Console
AWS Glue > Tables > accounts
silver_layer_db
s3://nehaawsbc001/silver/accounts/
No statistics
Description
-
Connection
-
Last updated
June 28, 2025 at 23:33:16
Advanced properties
Schema
Partitions
Indexes
Column statistics - new
Schema (5)
Edit schema as JSON
Edit schema
View and manage the table schema.
Filter schemas
Column name Data type Partition key Comment
1 account_id string - -
2 customer_id string - -
3 account_type string - -
4 balance double - -
5 open_date string - -
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


AWS Glue | us-east-1
AWS Glue > Tables > customers
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
silver_layer_db
Description
-
Last updated
June 28, 2025 at 23:33:16
s3://nehaawsbc001/silver/customers/
Connection
-
No statistics
Advanced properties
Schema
Partitions
Indexes
Column statistics - new
Schema (5)
View and manage the table schema.
Edit schema as JSON
Edit schema
Filter schemas
| Column name | Data type | Partition key | Comment
1 | customer_id | string | - | -
2 | customer_name | string | - | -
3 | email | string | - | -
4 | address | string | - | -
5 | date_joined | date | - | -
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


AWS Glue | us-east-1
AWS Glue > Tables > loan_payments
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
silver_layer_db
Description
-
Last updated
June 28, 2025 at 23:33:16
s3://nehaawsbc001/silver/loan_payments/
Connection
-
No statistics
Advanced properties
Schema
Partitions
Indexes
Column statistics - new
Schema (5)
View and manage the table schema.
Edit schema as JSON
Edit schema
Filter schemas
| Column name | Data type | Partition key | Comment
1 | payment_id | string | - | -
2 | loan_id | string | - | -
3 | customer_id | string | - | -
4 | payment_amount | double | - | -
5 | payment_date | string | - | -
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


AWS Glue | us-east-1
AWS Glue > Tables > loans
Description
-
Connection
-
Last updated
June 28, 2025 at 23:33:16
Advanced properties
Schema
Partitions
Indexes
Column statistics - new
Schema (6)
View and manage the table schema.
Edit schema as JSON
Edit schema
Filter schemas
| Column name | Data type | Partition key | Comment
1 | loan_id | string | - | -
2 | customer_id | string | - | -
3 | loan_amount | double | - | -
4 | status | string | - | -
5 | start_date | string | - | -
6 | term_months | string | - | -


AWS Glue | us-east-1
AWS Glue > Tables > transactions
Description
-
Connection
-
Last updated
June 28, 2025 at 23:33:16
Advanced properties
Schema
Partitions
Indexes
Column statistics - new
Schema (6)
View and manage the table schema.
Edit schema as JSON
Edit schema
Filter schemas
| Column name | Data type | Partition key | Comment
1 | transaction_id | string | - | -
2 | customer_id | string | - | -
3 | account_id | string | - | -
4 | amount | double | - | -
5 | transaction_type | string | - | -
6 | timestamp | string | - | -



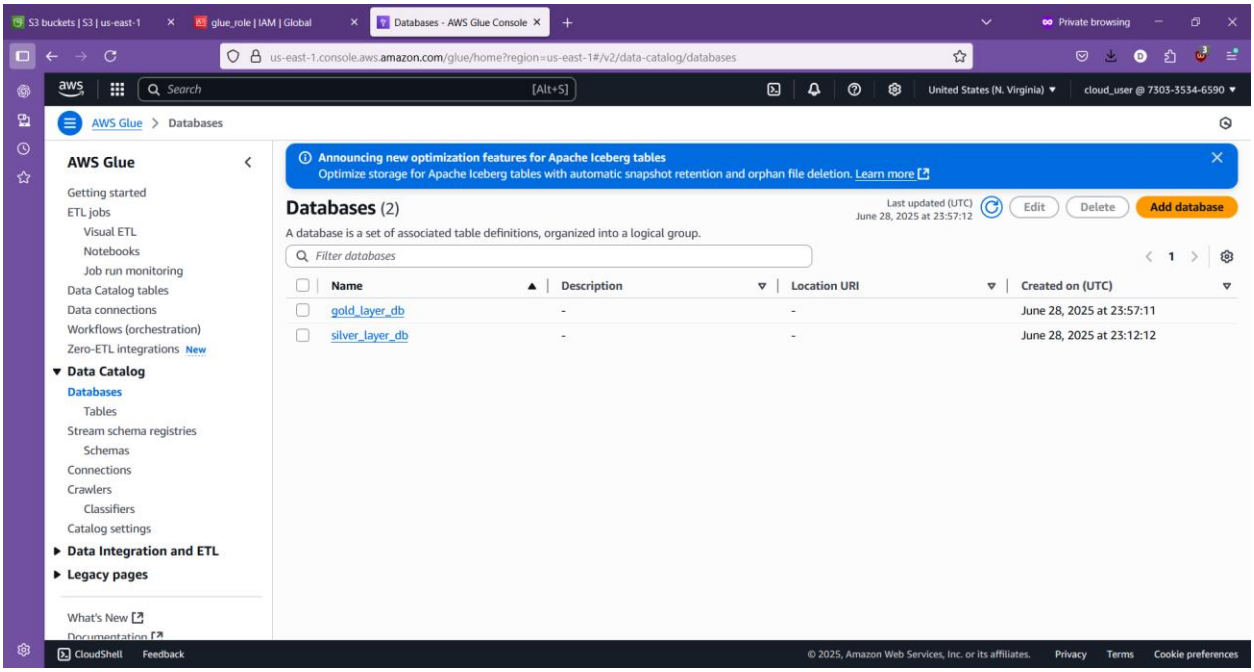
Cleaned parquet files:

part-00000-356c62ea-65ba-446c-8e22-cb2 part-00000-f76c845b-bb48-4ae6-aa76-a59 part-00000-8eb28f07-1ef6-4caa-b5f7-2877-dbbf-415f-aaa9-447 part-00000-da79e97c-edef-46f7-abc7-38a6 part-00000-0989f3a5

Data Warehouse & SCD - Gold Layer (Business-Ready Analytics)

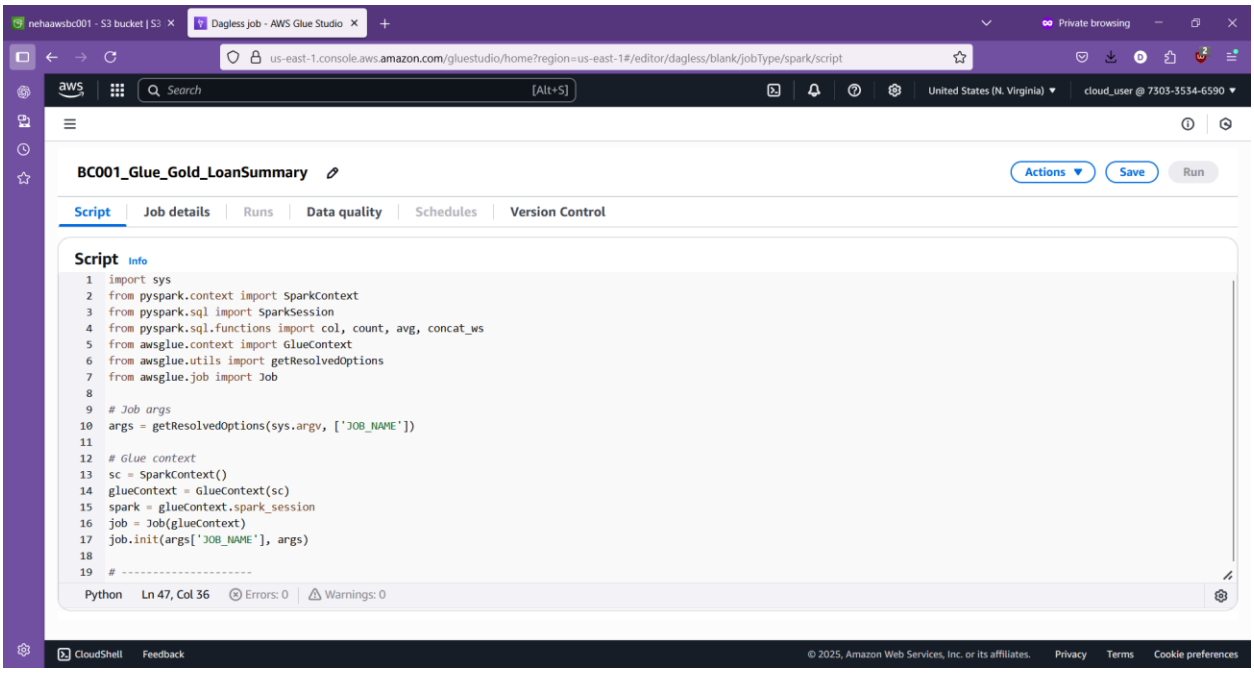
The Gold Layer transforms curated data from the Silver Layer into analytics-ready datasets that support business use cases such as: Loan performance summaries per customer. These datasets are stored in optimized formats (Parquet/Delta) for use in Athena.

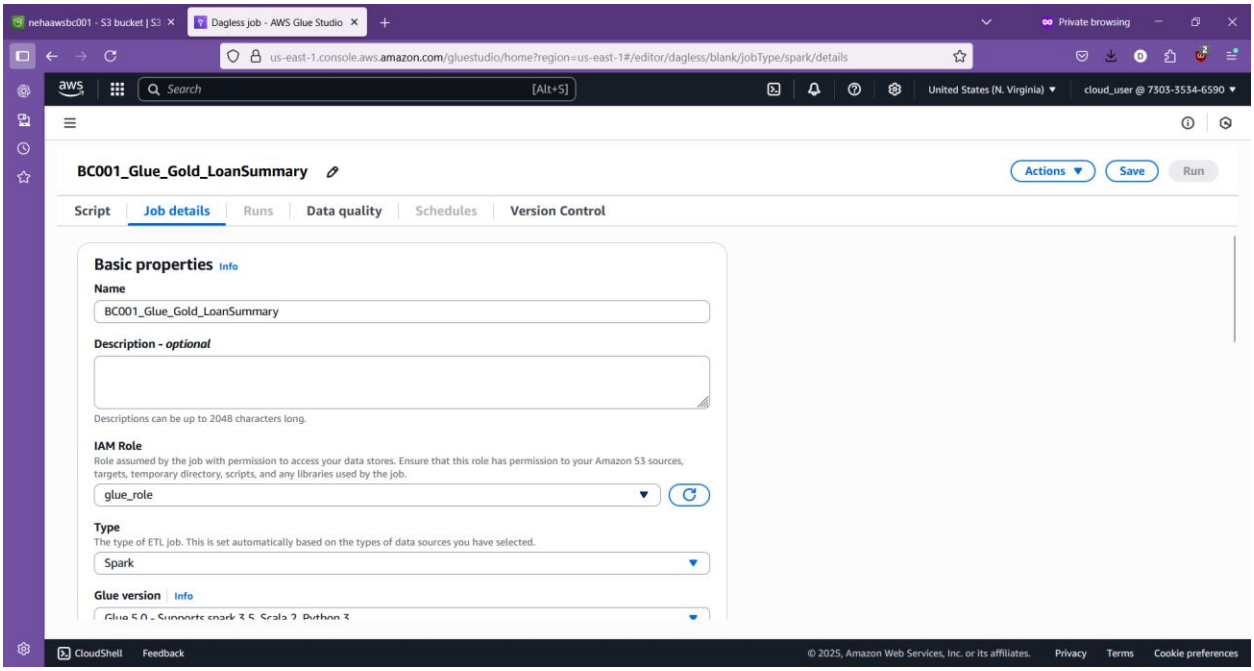
Step 1: Create Glue Database - To catalog the output of Gold-layer Glue jobs for query access in Athena and visualization tools.

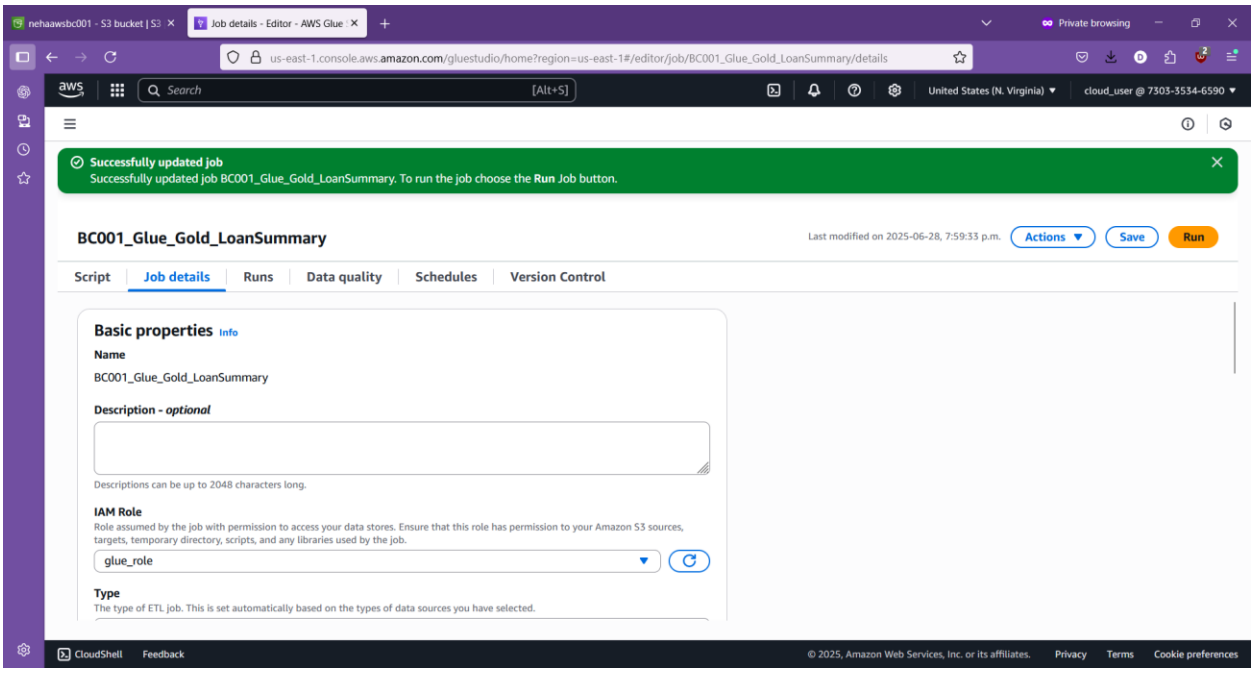


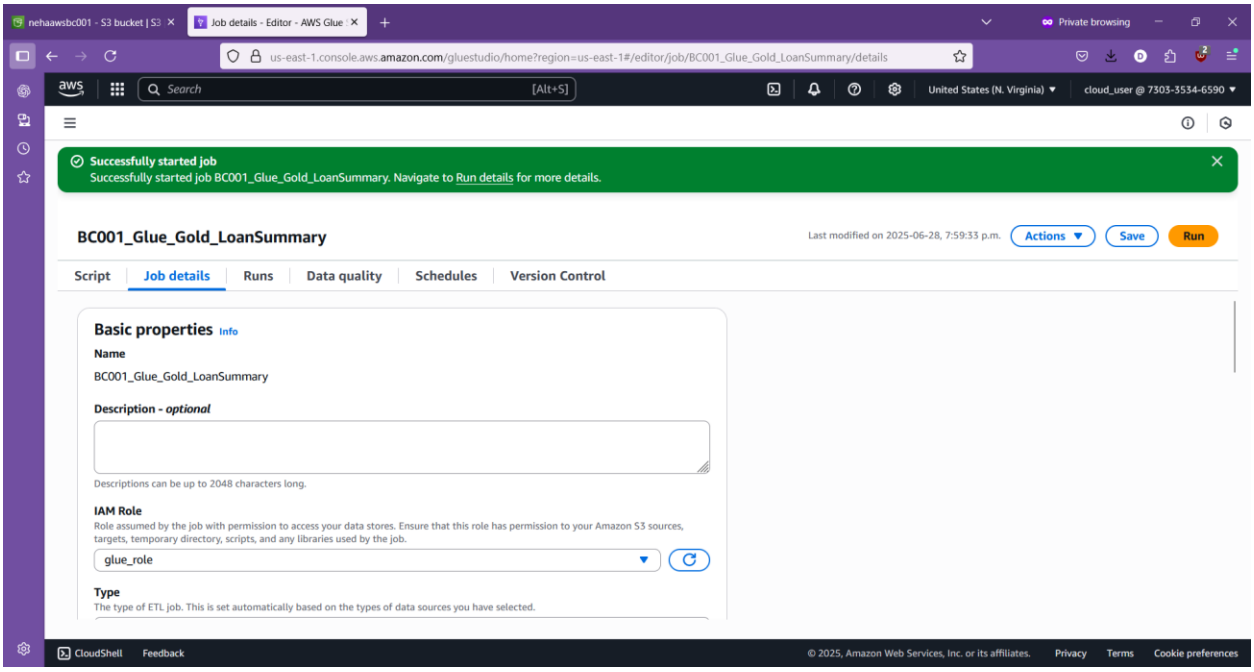
Step 2: Gold Layer Glue Job –

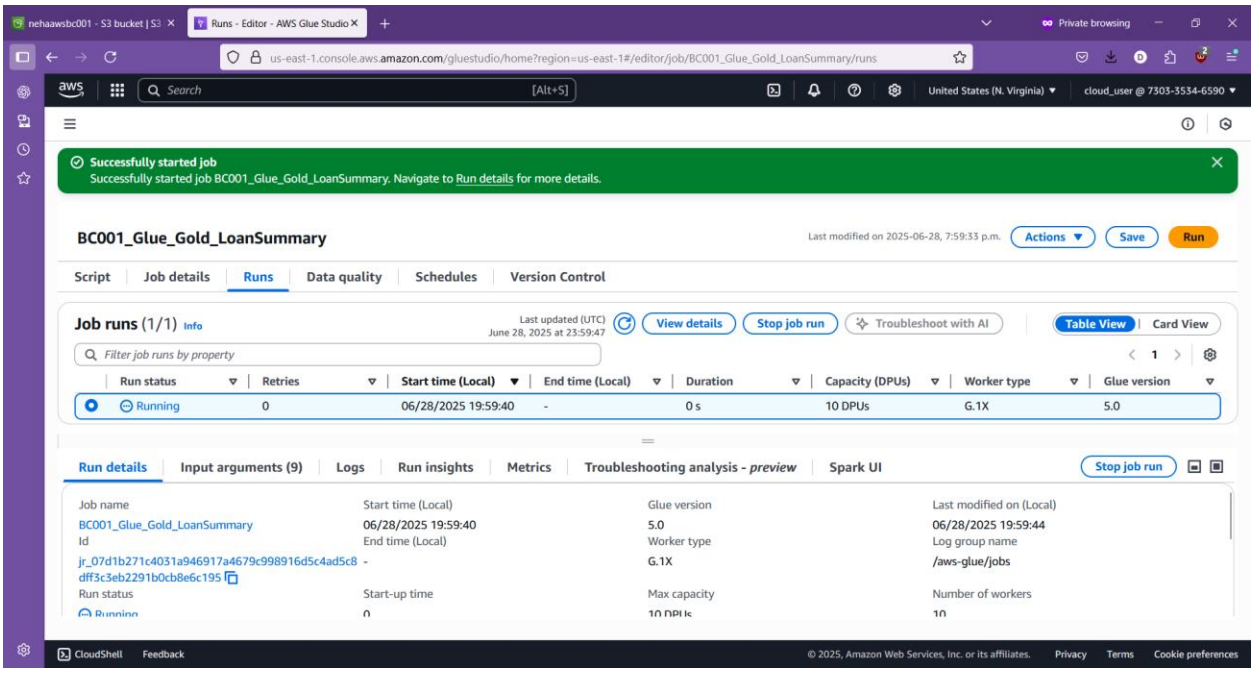
1. Join loans with customers on customer_id.
2. Aggregate per customer: total_loans: count of loans, avg_loan_amount: average loan value

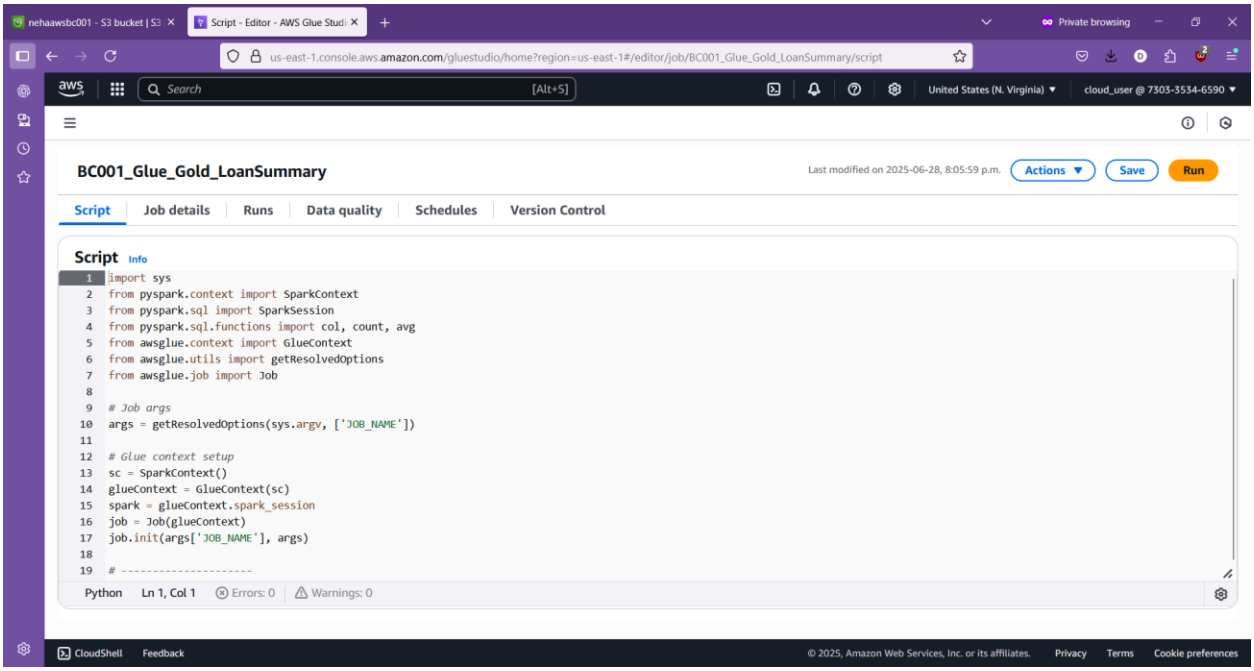

nehaawsbc001 - S3 bucket | S3
Dagless job - AWS Glue Studio
Private browsing
us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/dagless/blank/jobType/spark/script
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
BC001_Glue_Gold_LoanSummary
Actions
Save
Run
Script
Job details
Runs
Data quality
Schedules
Version Control
Script Info
import sys
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, count, avg, concat_ws
from awsglue.context import GlueContext
from awsglue.utils import getResolvedOptions
from awsglue.job import Job
# Job args
args = getResolvedOptions(sys.argv, ['JOB_NAME'])
# Glue context
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
# --------------------
Python
Ln 47, Col 36
Errors: 0
Warnings: 0
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


nehaawsbc001 - S3 bucket | S3
Dagless job - AWS Glue Studio
Private browsing
us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/dagless/blank/jobType/spark/details
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
BC001_Glue_Gold_LoanSummary
Actions
Save
Run
Script
Job details
Runs
Data quality
Schedules
Version Control
Basic properties Info
Name
BC001_Glue_Gold_LoanSummary
Description - optional
Descriptions can be up to 2048 characters long.
IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
glue_role
Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
Spark
Glue version Info
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Job details - Editor - AWS Glue
Private browsing
us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/BC001_Glue_Gold_LoanSummary/details
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
Successfully updated job
Successfully updated job BC001_Glue_Gold_LoanSummary. To run the job choose the Run Job button.
BC001_Glue_Gold_LoanSummary
Last modified on 2025-06-28, 7:59:33 p.m.
Actions
Save
Run
Script
Job details
Runs
Data quality
Schedules
Version Control
Basic properties
Info
Name
BC001_Glue_Gold_LoanSummary
Description - optional
Descriptions can be up to 2048 characters long.
IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
glue_role
Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Job details - Editor - AWS Glue
Private browsing
us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/BC001_Glue_Gold_LoanSummary/details
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
Successfully started job
Successfully started job BC001_Glue_Gold_LoanSummary. Navigate to Run details for more details.
BC001_Glue_Gold_LoanSummary
Last modified on 2025-06-28, 7:59:33 p.m.
Actions
Save
Run
Script
Job details
Runs
Data quality
Schedules
Version Control
Basic properties
Info
Name
BC001_Glue_Gold_LoanSummary
Description - optional
Descriptions can be up to 2048 characters long.
IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
glue_role
Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences



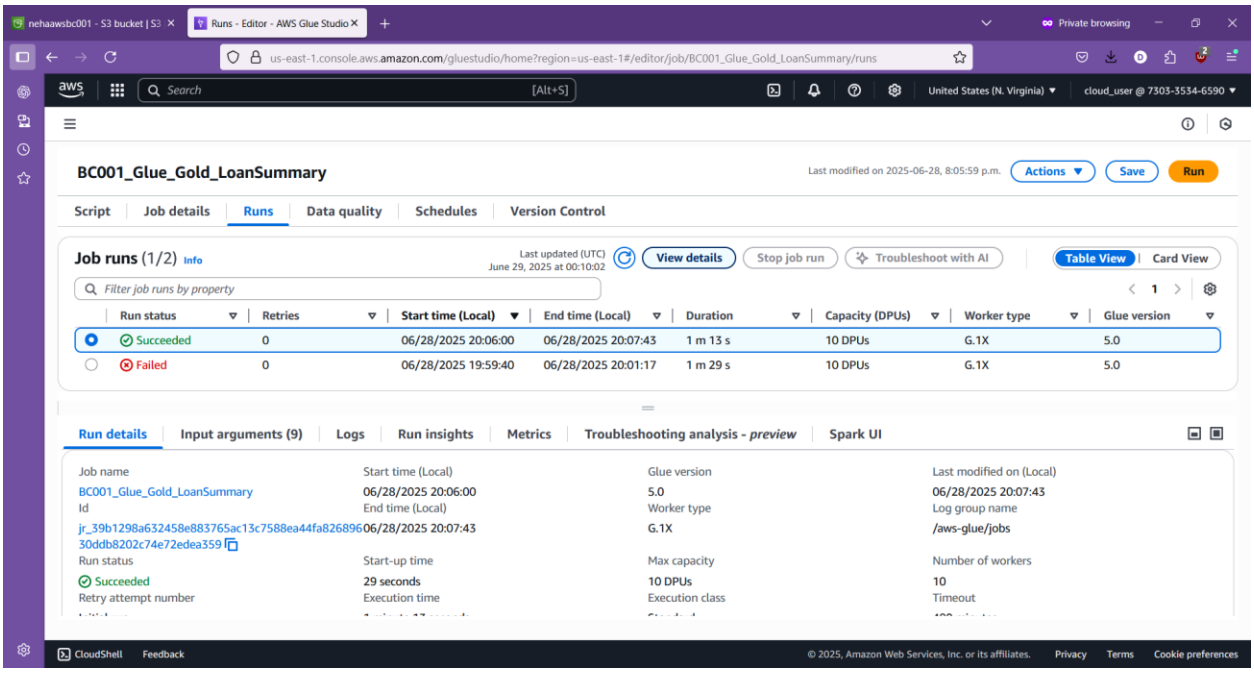


Code:

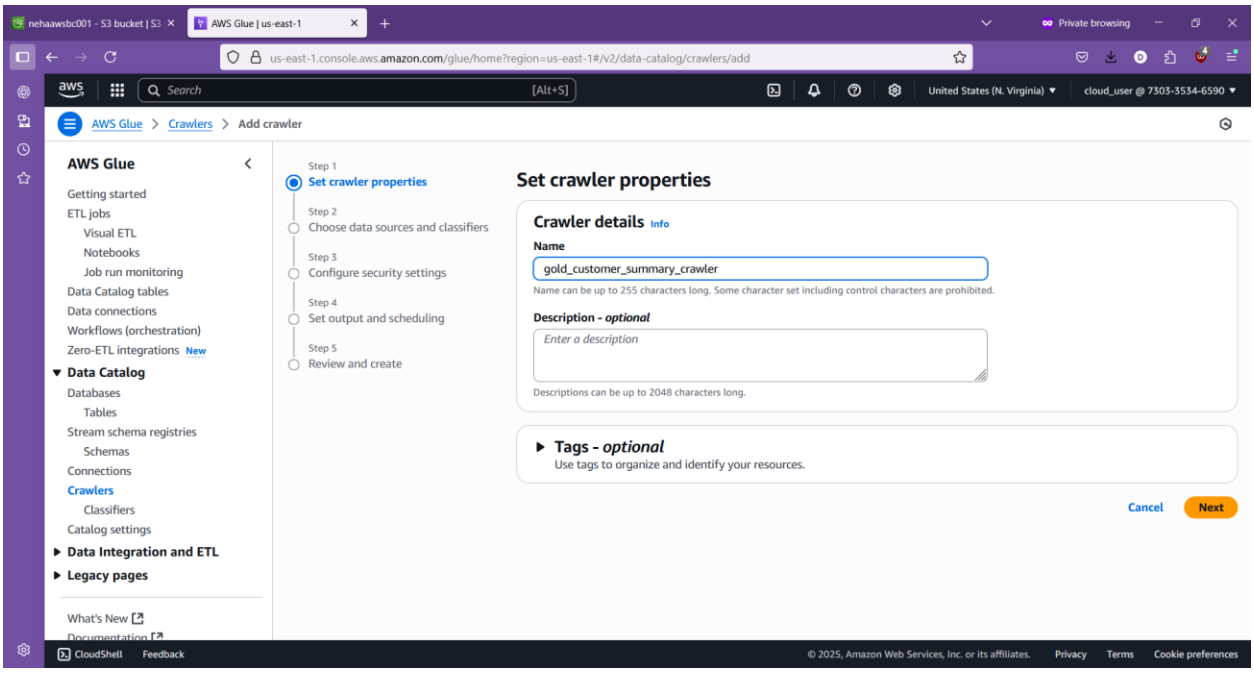
```
import sys
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, count, avg
from awsglue.context import GlueContext
```

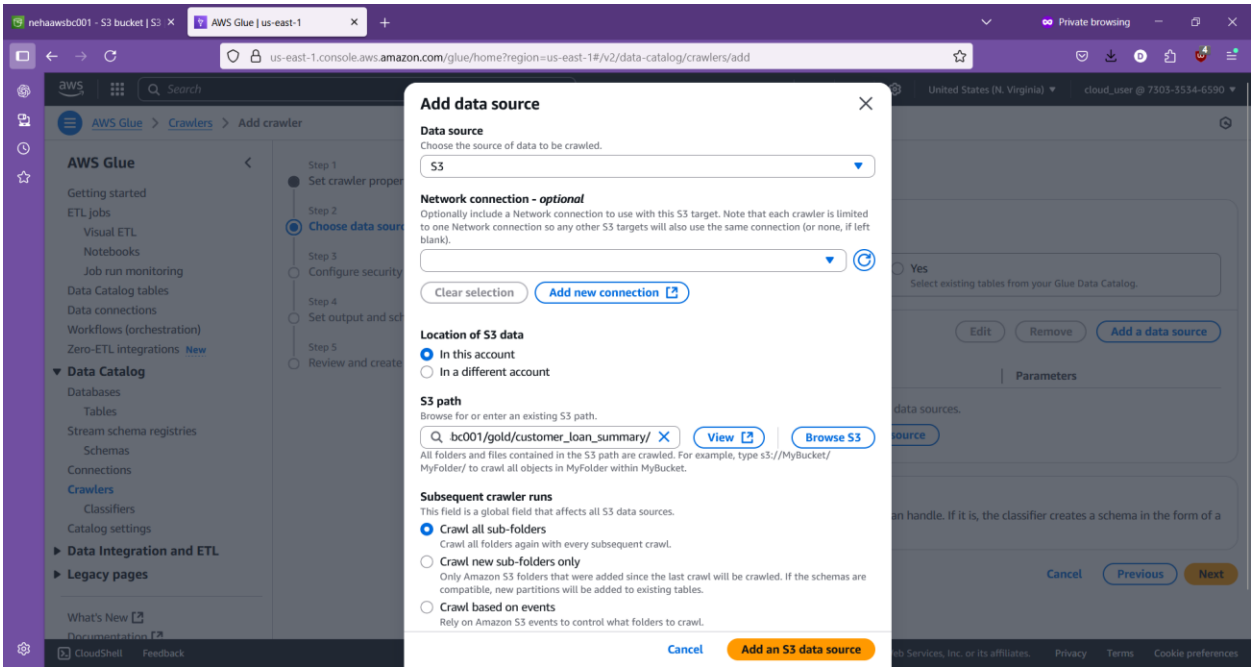
```
from awsglue.utils import getResolvedOptions
from awsglue.job import Job

# Job args
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

# Glue context setup
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)

# ----------------------
# Load Silver Layer Data
# ----------------------

silver_path = "s3://nehaawsbc001/silver/"

customers = spark.read.parquet(f"{silver_path}/customers/")
loans = spark.read.parquet(f"{silver_path}/loans/")

# ----------------------
# Join and Aggregation
# ----------------------

# Join on customer_id
joined_df = loans.join(customers, on="customer_id", how="inner")

# Use customer_name directly
```

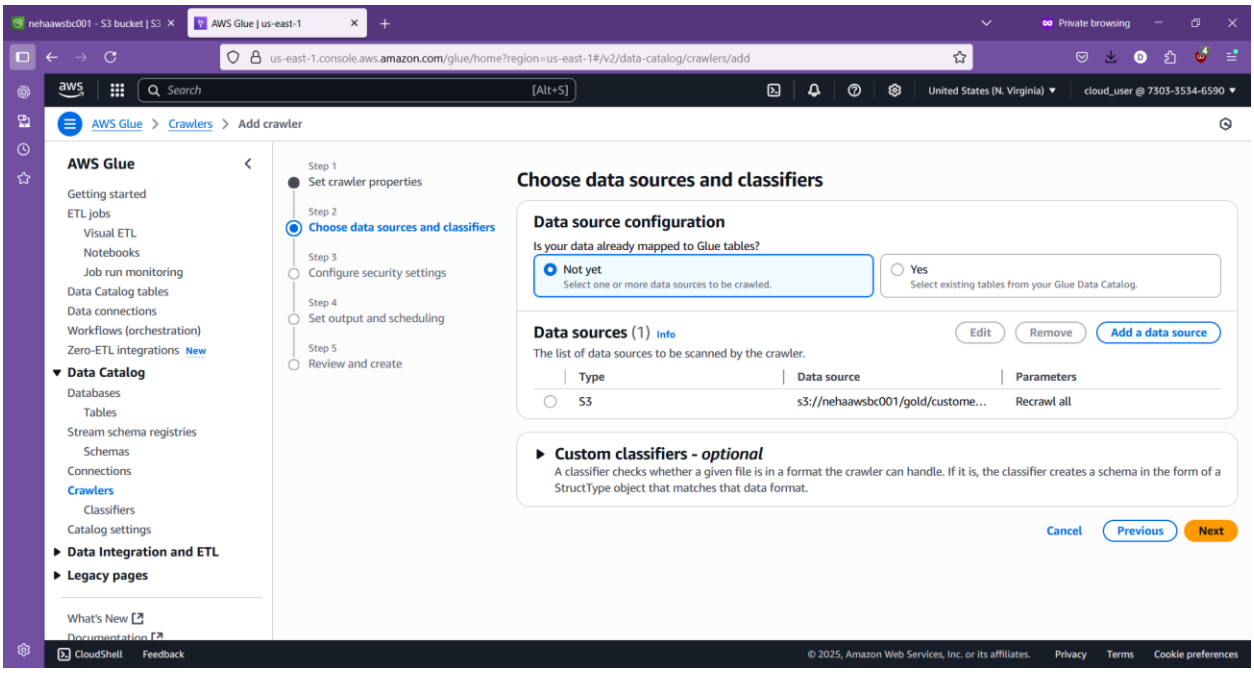
```
summary_df = joined_df.groupBy("customer_id", "customer_name") \
    .agg(
        count("loan_id").alias("total_loans"),
        avg("loan_amount").alias("avg_loan_amount")
    )

# --------------------
# Write to Gold Layer
# --------------------

gold_path = "s3://nehaawsbc001/gold/customer_loan_summary/"

summary_df.write.mode("overwrite").parquet(gold_path)

job.commit()
```

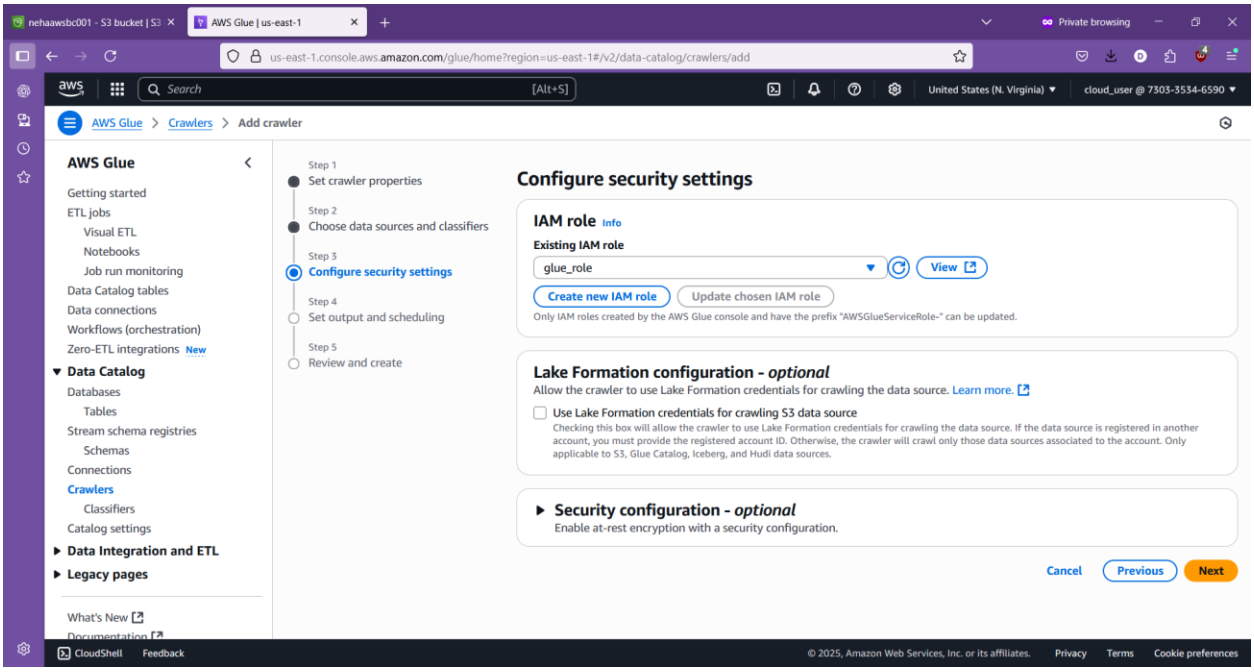


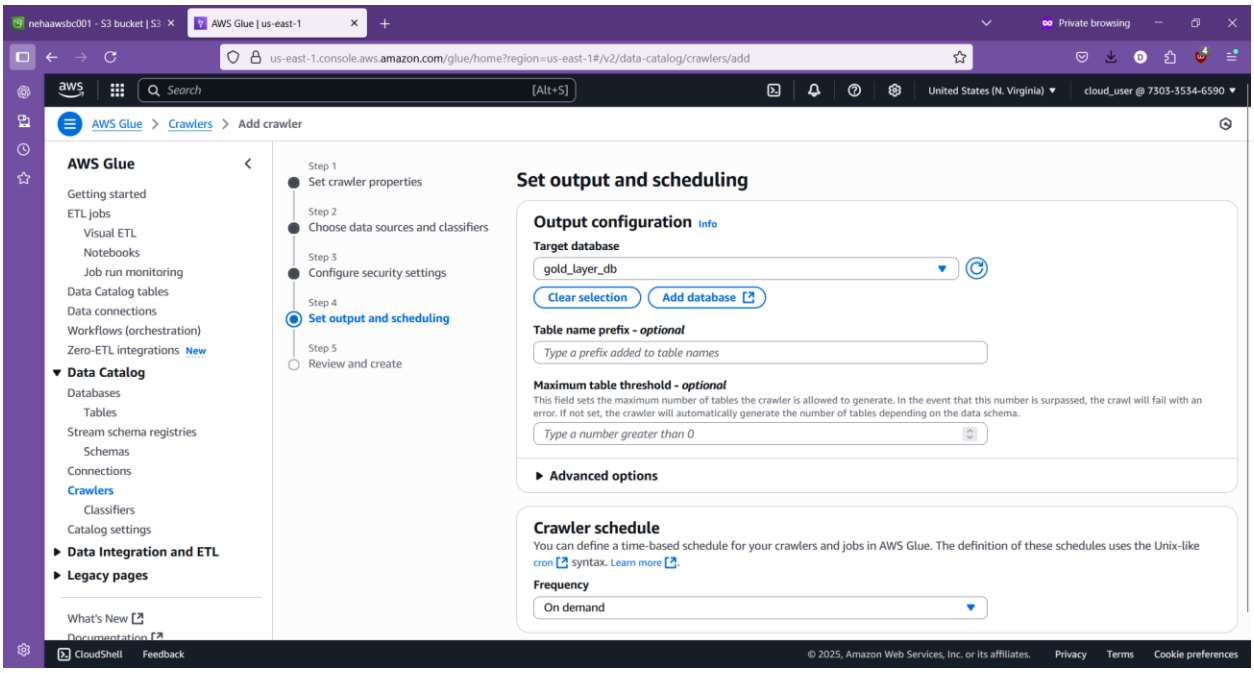
Step 3: Crawler for Gold Layer

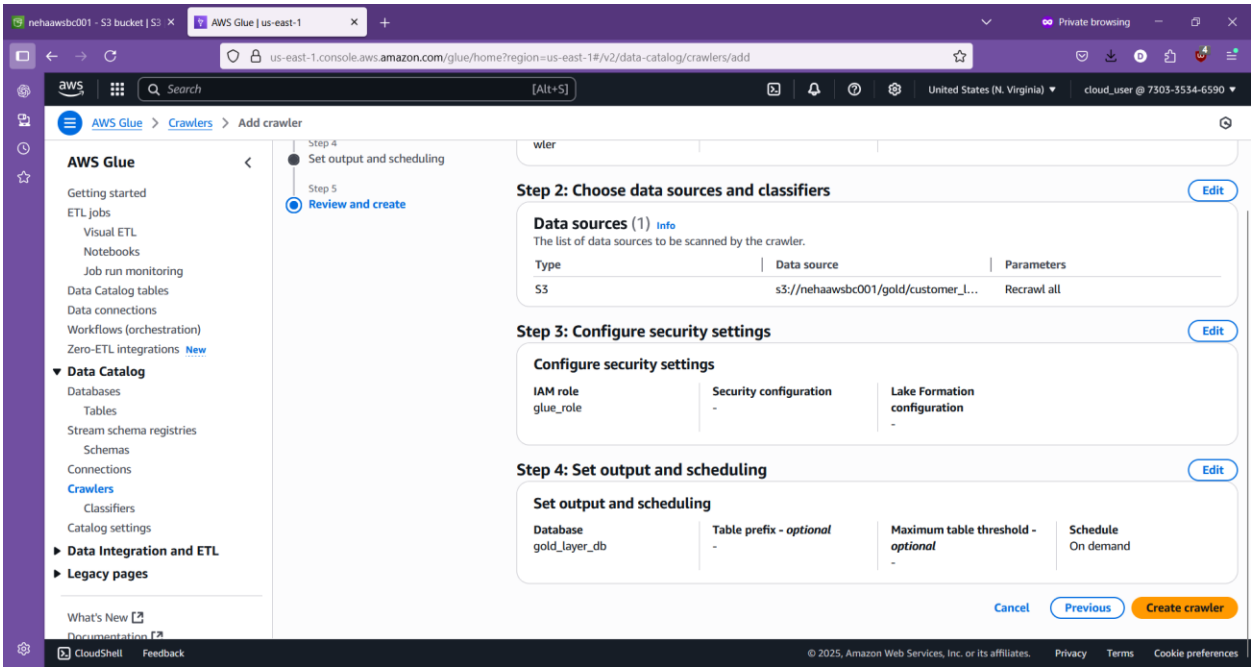

AWS Glue | us-east-1
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
AWS Glue > Crawlers > Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Set crawler properties
Crawler details Info
Name
gold_customer_summary_crawler
Name can be up to 255 characters long. Some character set including control characters are prohibited.
Description - optional
Enter a description
Descriptions can be up to 2048 characters long.
Tags - optional
Use tags to organize and identify your resources.
Cancel
Next

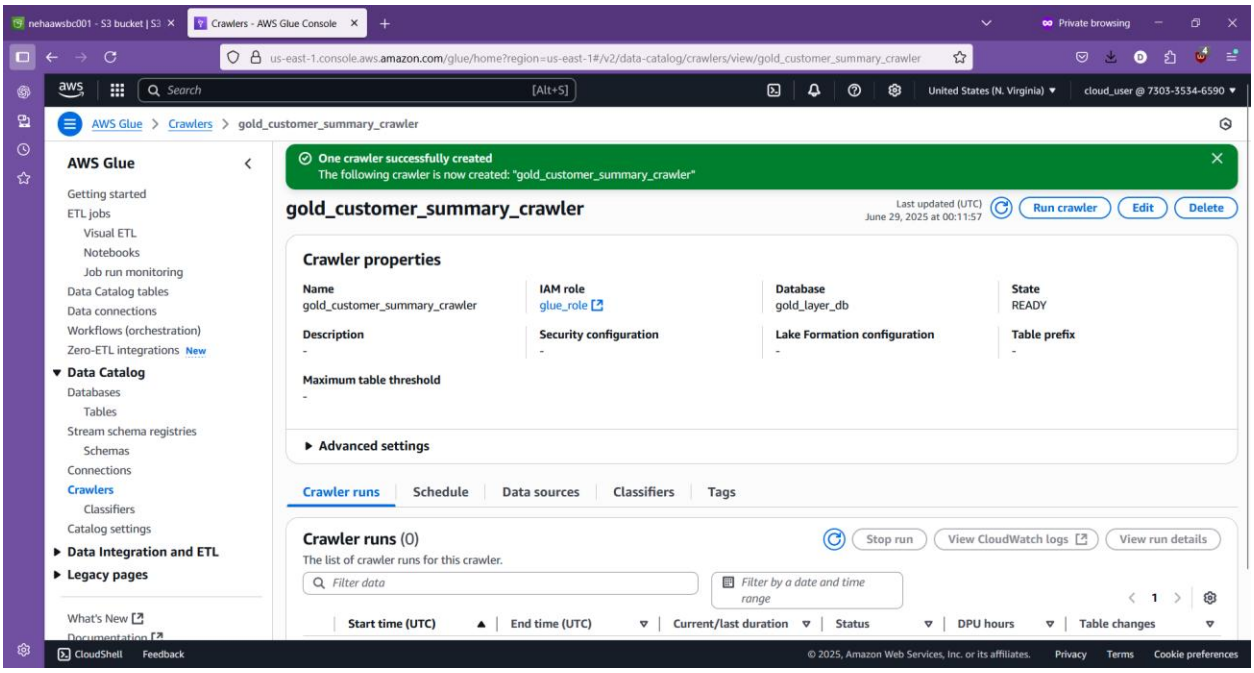

Add data source
Data source
Choose the source of data to be crawled.
S3
Network connection - optional
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).
Clear selection
Add new connection
Location of S3 data
In this account
In a different account
S3 path
Browse for or enter an existing S3 path.
bc001/gold/customer_loan_summary/
View
Browse S3
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.
Subsequent crawler runs
This field is a global field that affects all S3 data sources.
Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.
Cancel
Add an S3 data source

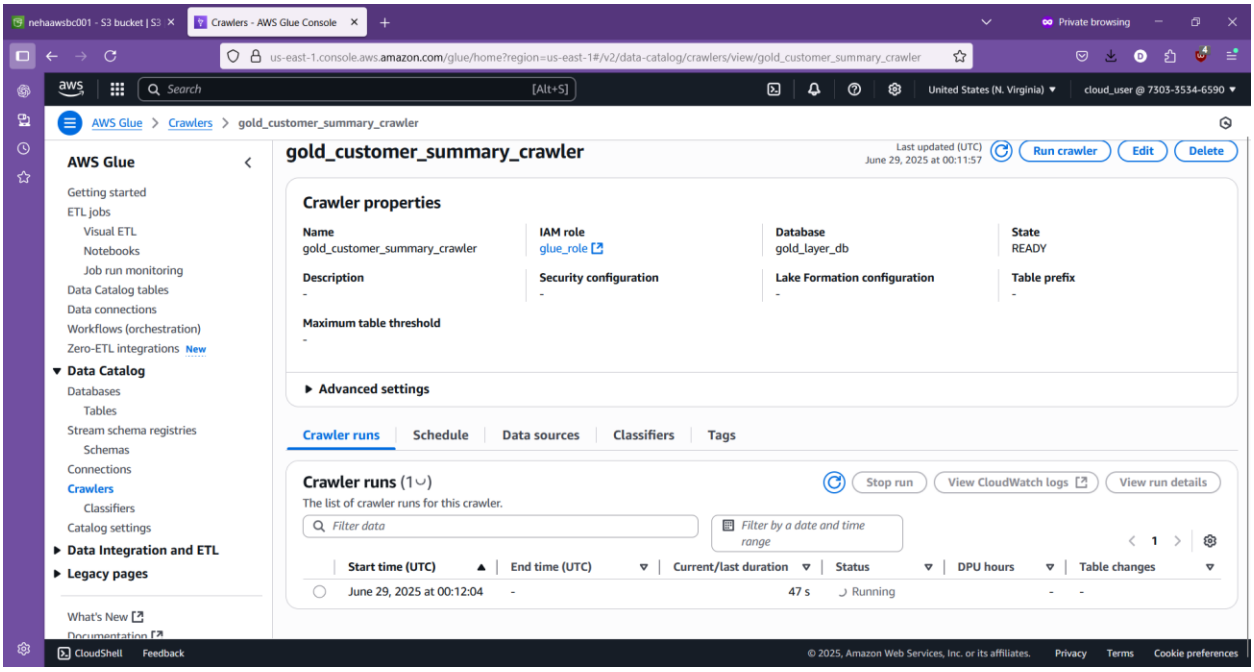

nehaawsbc001 - S3 bucket | S3
AWS Glue | us-east-1
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (1) Info
Edit
Remove
Add a data source
The list of data sources to be scanned by the crawler.
Type
Data source
Parameters
S3
s3://nehaawsbc001/gold/custome...
Recrawl all
Custom classifiers - optional
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.
Cancel
Previous
Next
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

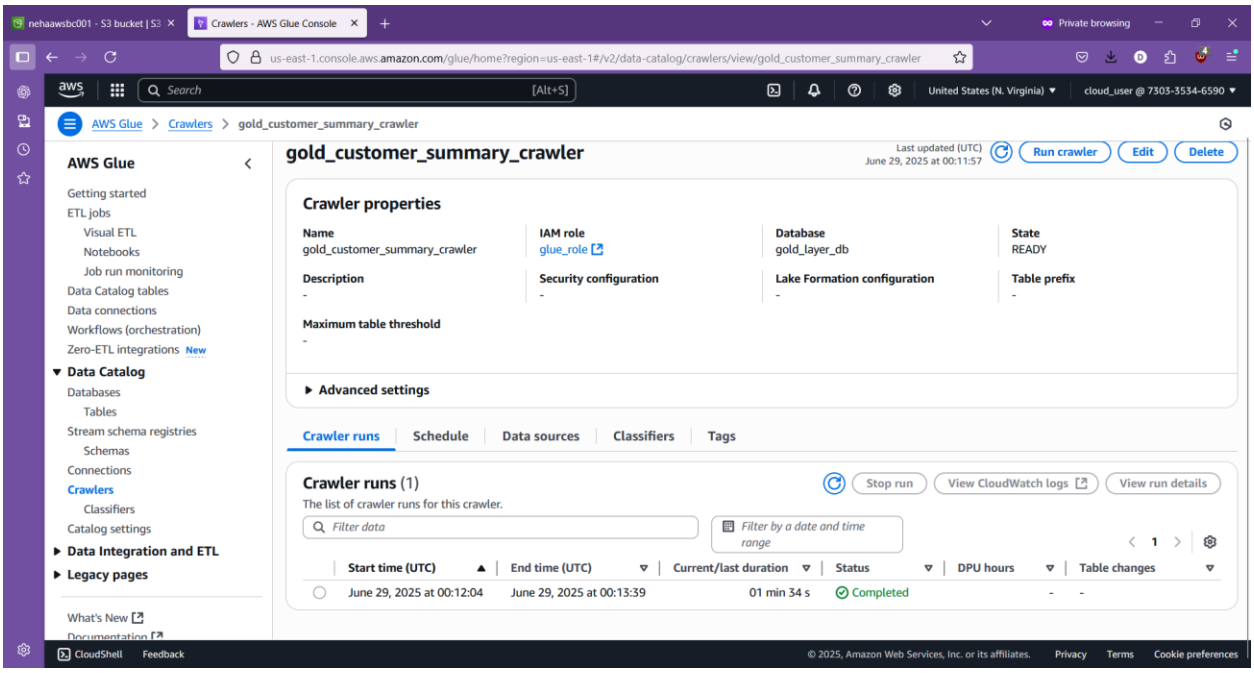

nehaawsbc001 - S3 bucket | S3
AWS Glue | us-east-1
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Configure security settings
IAM role Info
Existing IAM role
glue_role
View
Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.
Lake Formation configuration - optional
Allow the crawler to use Lake Formation credentials for crawling the data source. Learn more.
Use Lake Formation credentials for crawling S3 data source
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.
Security configuration - optional
Enable at-rest encryption with a security configuration.
Cancel
Previous
Next
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

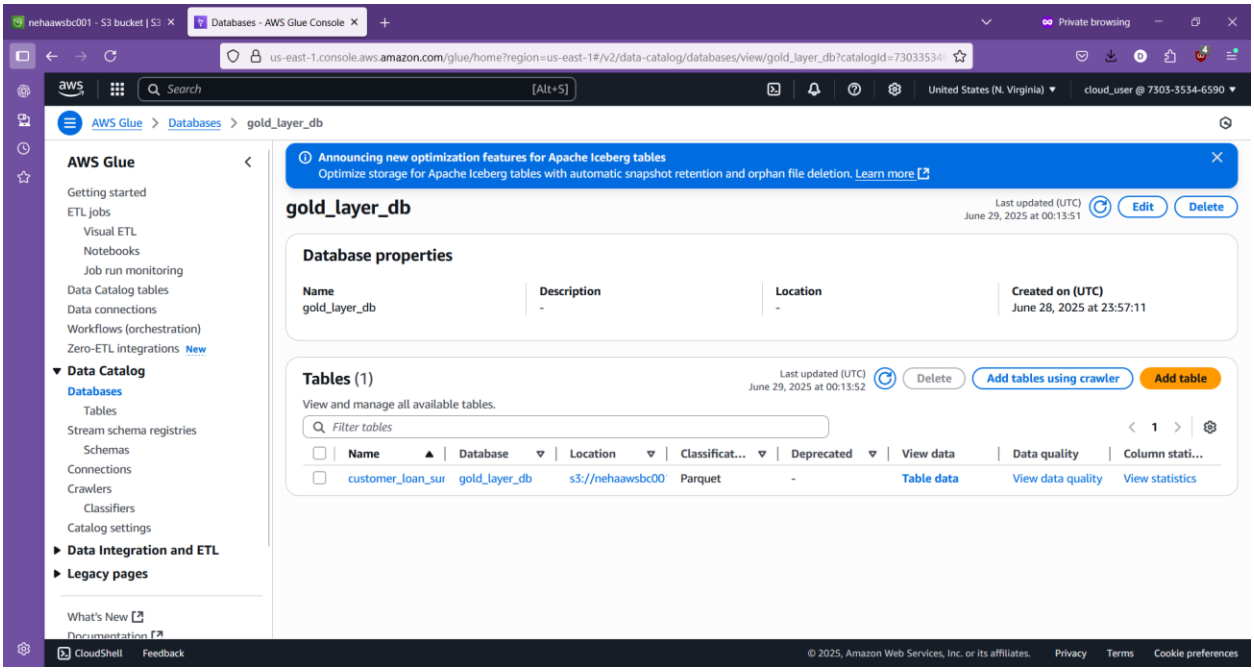

AWS Glue | us-east-1
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
AWS Glue > Crawlers > Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Set output and scheduling
Output configuration Info
Target database
gold_layer_db
Clear selection
Add database
Table name prefix - optional
Type a prefix added to table names
Maximum table threshold - optional
This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.
Type a number greater than 0
Advanced options
Crawler schedule
You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more
Frequency
On demand

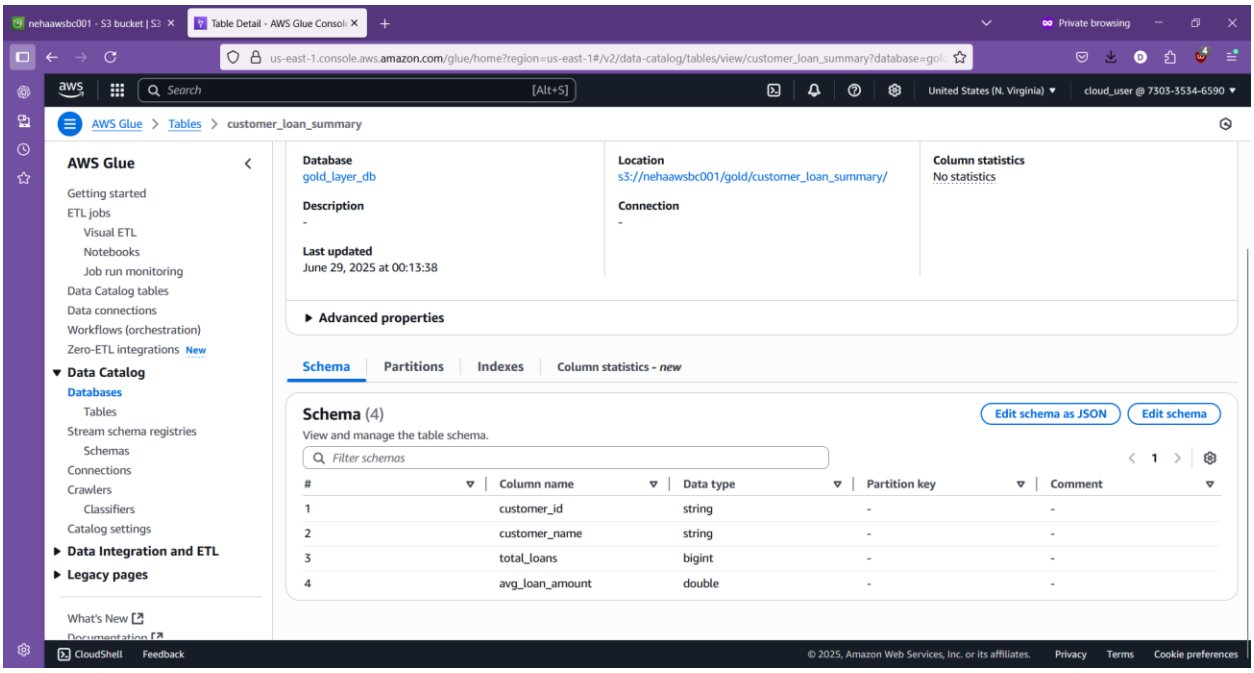

AWS Glue | us-east-1
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/add
AWS Glue > Crawlers > Add crawler
Step 4
Set output and scheduling
Step 5
Review and create
Step 2: Choose data sources and classifiers
Edit
Data sources (1) Info
The list of data sources to be scanned by the crawler.
Type
Data source
Parameters
S3
s3://nehaawsbc001/gold/customer_l...
Recrawl all
Step 3: Configure security settings
Edit
Configure security settings
IAM role
glue_role
Security configuration
-
Lake Formation configuration
-
Step 4: Set output and scheduling
Edit
Set output and scheduling
Database
gold_layer_db
Table prefix - optional
-
Maximum table threshold - optional
-
Schedule
On demand
Cancel
Previous
Create crawler


One crawler successfully created
The following crawler is now created: "gold_customer_summary_crawler"
gold_customer_summary_crawler
Run crawler
Edit
Delete
Crawler properties
Name
gold_customer_summary_crawler
IAM role
glue_role
Database
gold_layer_db
State
READY
Description
Security configuration
Lake Formation configuration
Table prefix
Maximum table threshold
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (0)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes

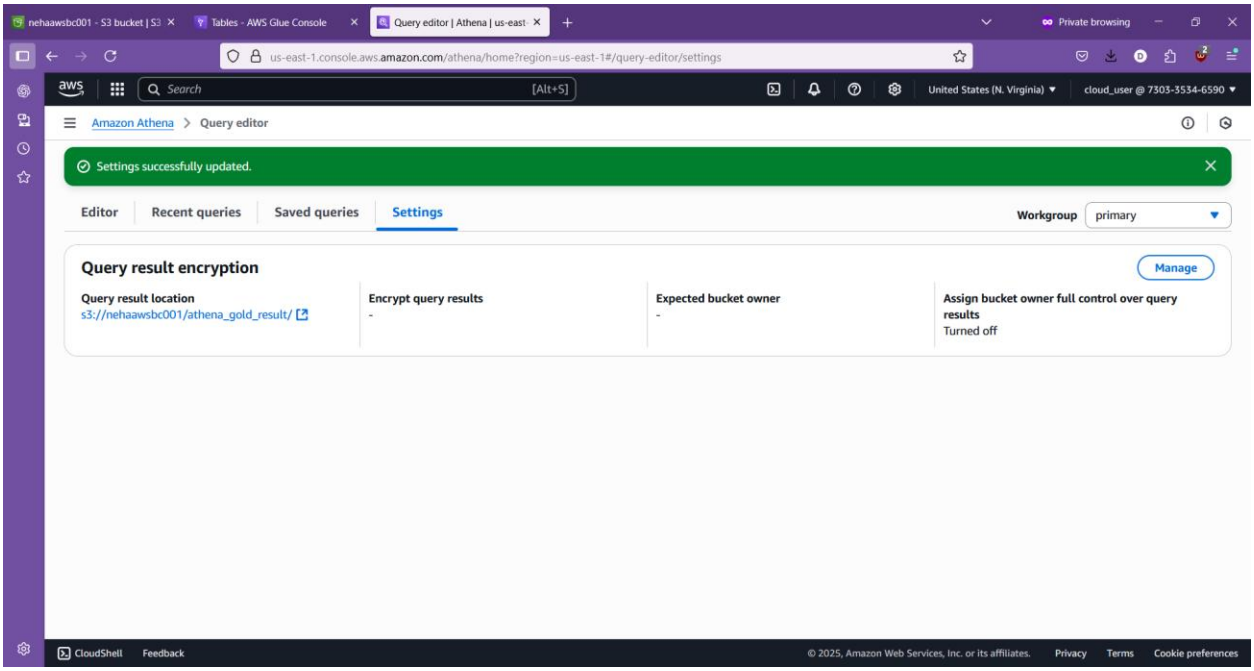

gold_customer_summary_crawler
Run crawler
Edit
Delete
Crawler properties
Name
gold_customer_summary_crawler
IAM role
glue_role
Database
gold_layer_db
State
READY
Crawler runs (1)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes
June 29, 2025 at 00:12:04
47 s
Running

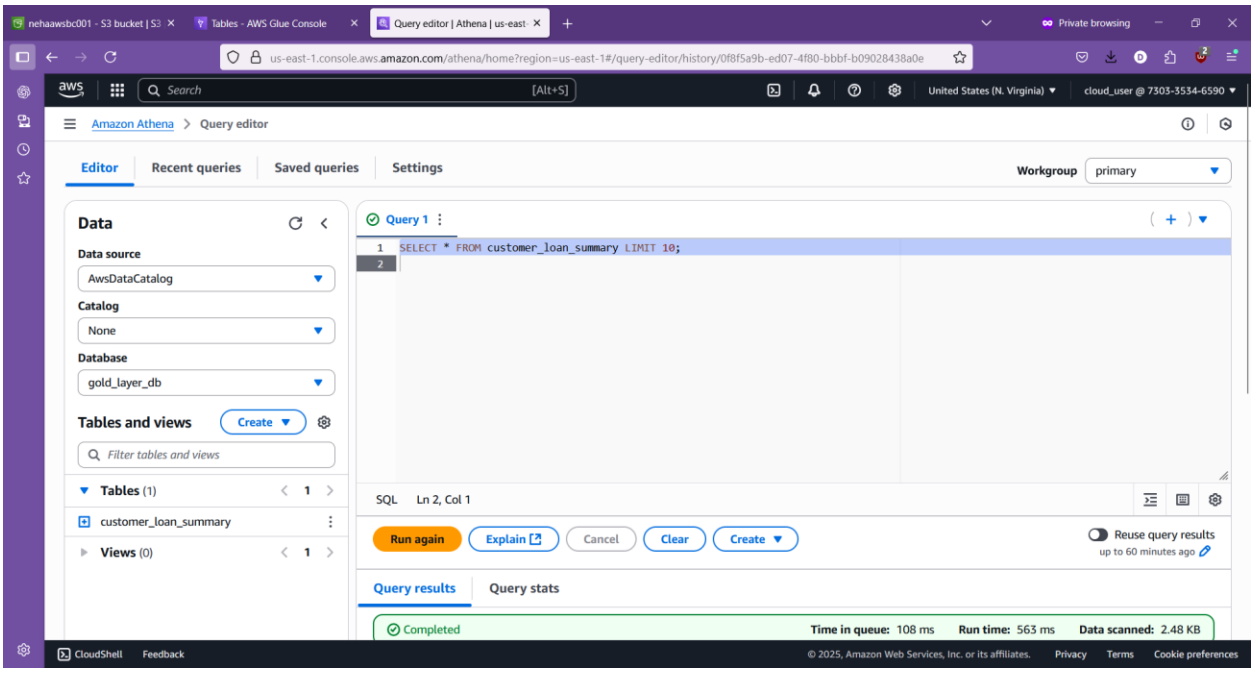

nehaawsbc001 - S3 bucket | S3
Crawlers - AWS Glue Console
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/view/gold_customer_summary_crawler
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
AWS Glue > Crawlers > gold_customer_summary_crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
gold_customer_summary_crawler
Last updated (UTC)
June 29, 2025 at 00:11:57
Run crawler
Edit
Delete
Crawler properties
Name
gold_customer_summary_crawler
IAM role
glue_role
Database
gold_layer_db
State
READY
Description
-
Security configuration
-
Lake Formation configuration
-
Table prefix
-
Maximum table threshold
-
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (1)
Stop run
View CloudWatch logs
View run details
The list of crawler runs for this crawler.
Filter data
Filter by a date and time range
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes
June 29, 2025 at 00:12:04
June 29, 2025 at 00:13:39
01 min 34 s
Completed
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


nehaawsbc001 - S3 bucket | S3
Databases - AWS Glue Console
Private browsing
us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/databases/view/gold_layer_db?catalogId=73033534
Search
[Alt+S]
United States (N. Virginia)
cloud_user @ 7303-3534-6590
AWS Glue > Databases > gold_layer_db
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. Learn more
gold_layer_db
Last updated (UTC)
June 29, 2025 at 00:13:51
Edit
Delete
Database properties
Name
gold_layer_db
Description
-
Location
-
Created on (UTC)
June 28, 2025 at 23:57:11
Tables (1)
Last updated (UTC)
June 29, 2025 at 00:13:52
Delete
Add tables using crawler
Add table
View and manage all available tables.
Filter tables
Name
Database
Location
Classificat...
Deprecated
View data
Data quality
Column stati...
customer_loan_sur
gold_layer_db
s3://nehaawsbc00
Parquet
-
Table data
View data quality
View statistics
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

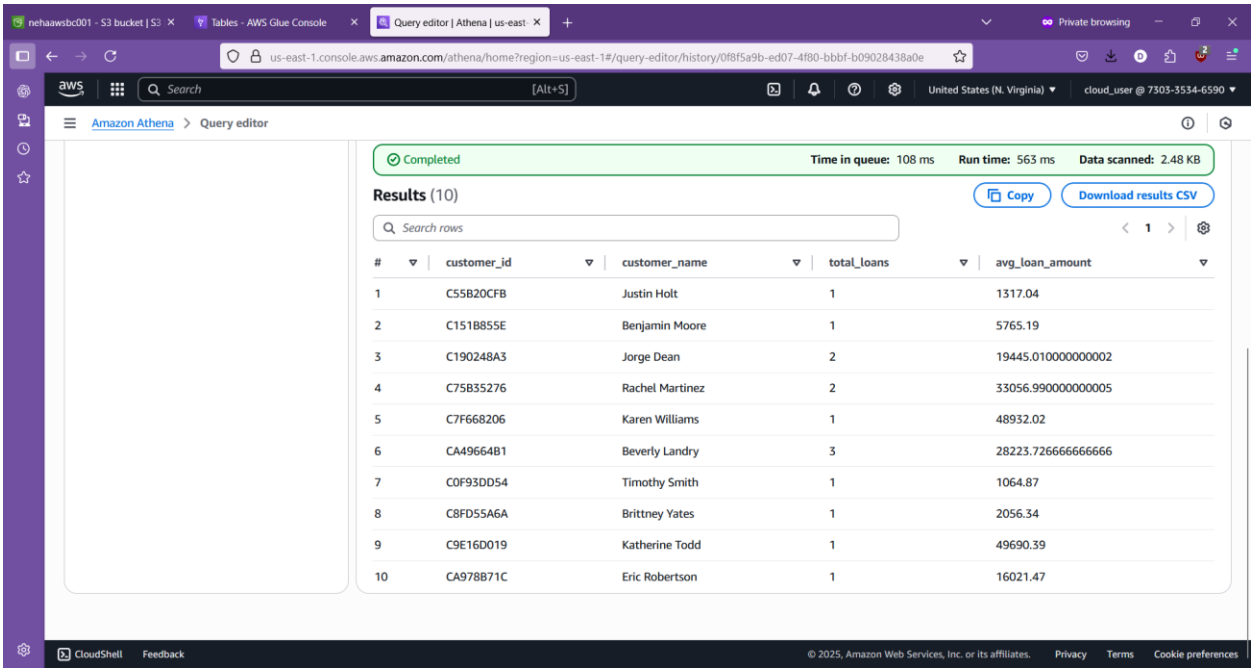


Step 4: Athena Table Verification





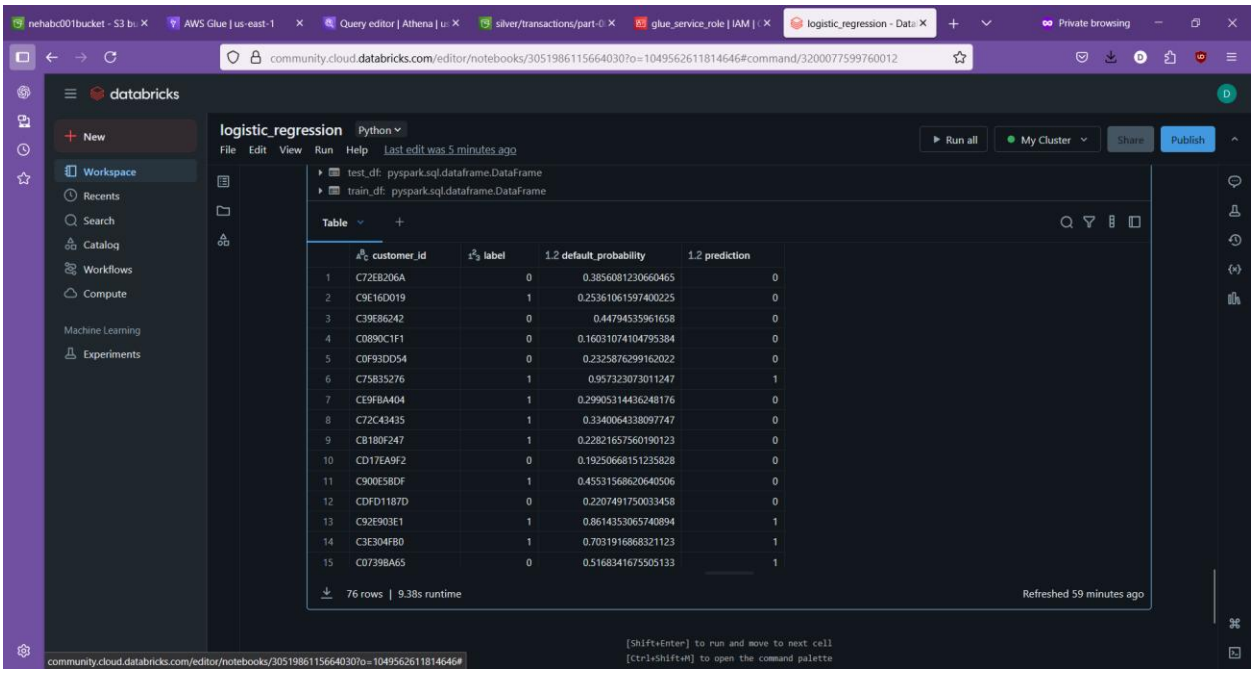
- Query: SELECT * FROM gold_layer_db.customer_loan_summary LIMIT 10;



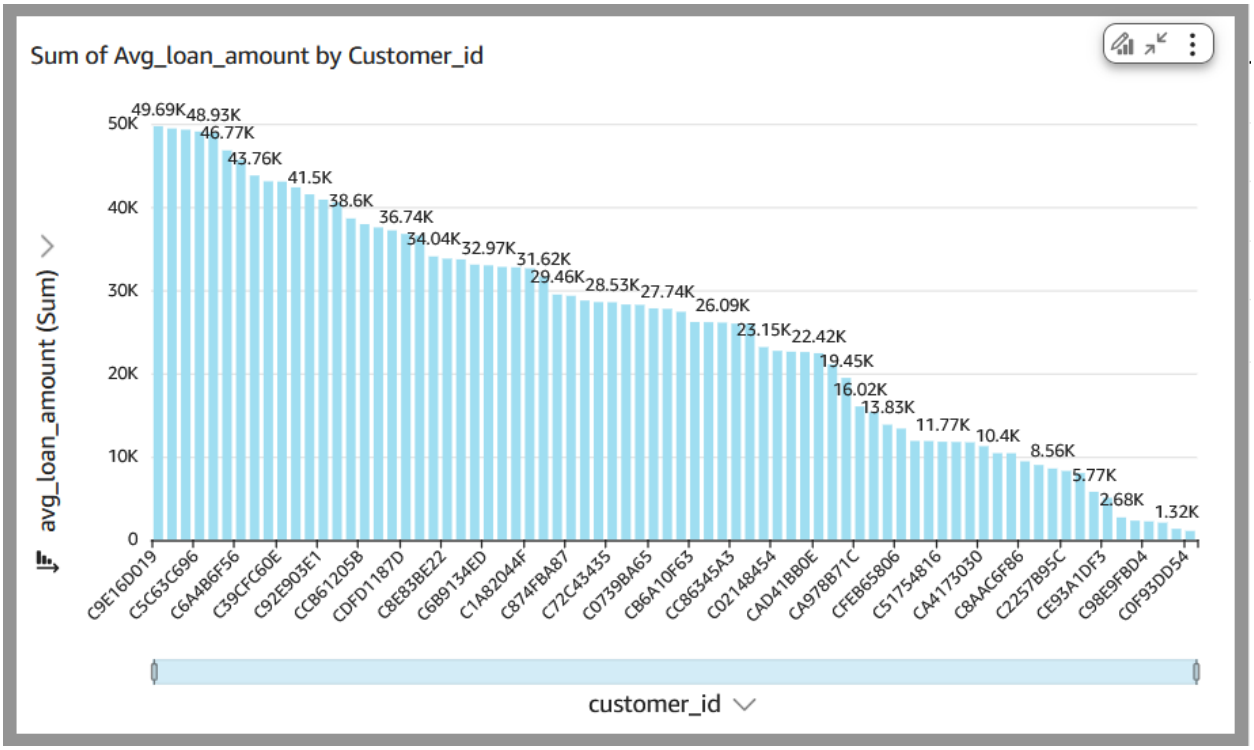
Analytics & Business Intelligence - Logistic Regression – Predict Loan Default Risk Per Customer

Objective

Train a logistic regression model using PySpark to predict the probability that a customer will default on a loan, based on their historical loan and payment behaviour. This output helps analysts identify high-risk customers for credit monitoring, risk control, or pre-emptive communication.



Dashboard in Quicksight Arena with Gold layer output



Databricks code:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1049562611814646/3051986115664030/5718004312911206/latest.html