Cloud Datapipeline

Yash Bhatnagar

Abstract

This project, titled "Cloud Data Pipeline Mastery: End-to-End Analytics with AWS," focuses on building a versatile and scalable data engineering pipeline using AWS services to analyze and visualize any dataset. For this implementation, we used the Spotify dataset, but the architecture is designed to accommodate any dataset, making it suitable for a wide range of real-world applications in data-driven decision-making. By leveraging AWS Glue, S3, Athena, CloudWatch and QuickSight, this project provides an efficient, scalable, and cost-effective solution for processing and analyzing data, offering valuable insights to stakeholders.

Introduction

In this project, we will build a comprehensive data engineering pipeline using AWS cloud services. While we demonstrate the pipeline with the Spotify dataset, the project architecture is flexible enough to handle any dataset. The focus is on processing and analyzing data using various AWS tools like S3, Glue, Athena, and QuickSight.

Project Architecture Overview

- **Staging Layer**: Raw data is stored in an S3 bucket.
- **ETL Pipeline**: AWS Glue processes and transfers data from the staging layer to the data warehouse.
- **Data Warehouse**: Processed data is stored in another S3 bucket.
- **Data Catalog**: AWS Glue Crawler creates a database and tables for the data warehouse.
- **Data Analysis**: AWS Athena queries the processed data.
- **Data Visualization**: AWS QuickSight visualizes the data.

.

AWS Services Used

- **Amazon S3**: For storing raw and processed data.
- **AWS Glue**: For building and managing ETL pipelines.
- **AWS Athena**: For querying data using SQL-like syntax.
- **AWS QuickSight**: For visualizing data.
- **AWS CloudWatch** monitoring, logging, and alerting across our data pipeline.

Data Source

The data used in this project is sourced from the Spotify Dataset 2023 available on Kaggle. The dataset, created by Tony Gordon Jr., includes detailed information about Spotify albums, artists, tracks, and various audio features like danceability, energy, loudness, and more. Although this project uses the Spotify dataset, the pipeline is designed to be dataset-agnostic.

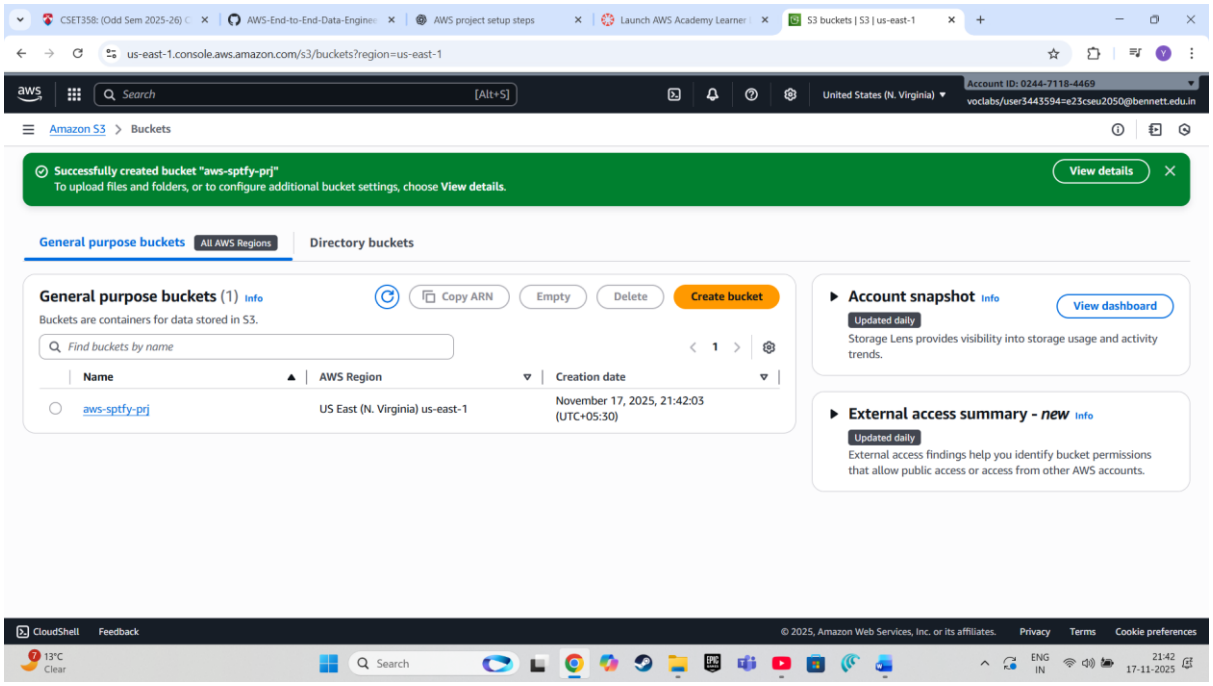
Data Description

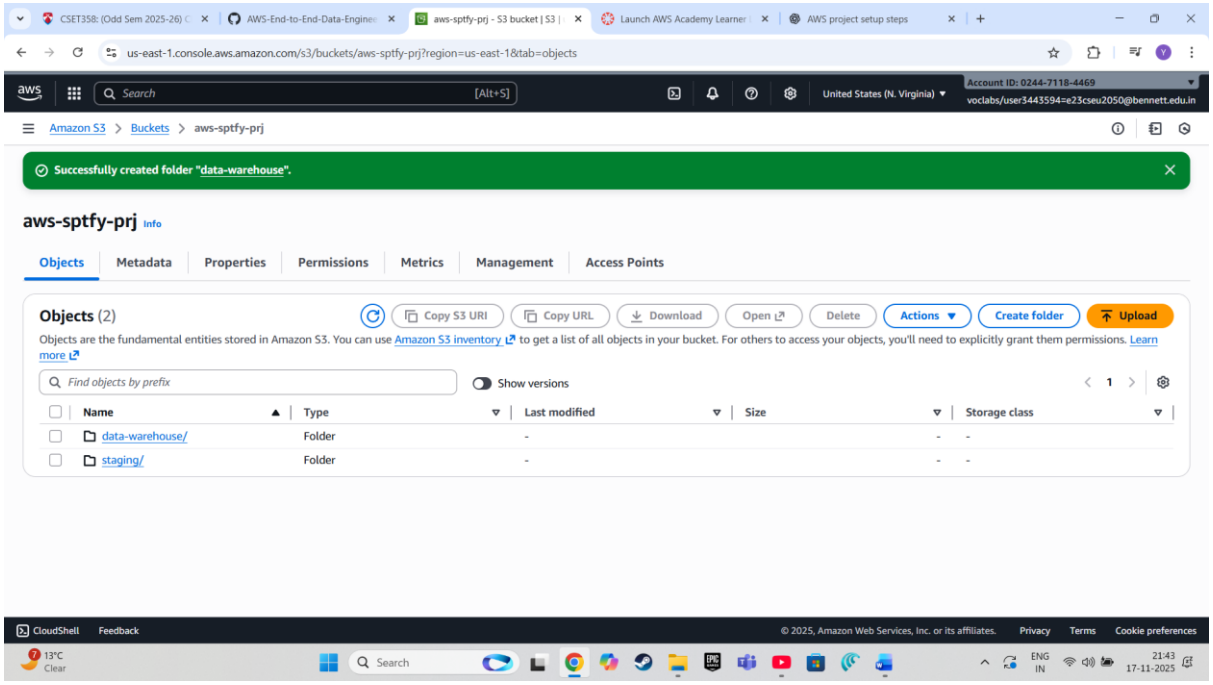
- **Albums**: Contains details of all the albums, including album ID, name, popularity, and release date.
- **Artists**: Contains information about the artists, including their names, number of followers, and genres.
- **Tracks**: Contains track-level data, including track ID, popularity, and other features like danceability and energy.
- **Spotify Features**: Contains various audio features like loudness, mode, speechiness, and valence

Project Implementation

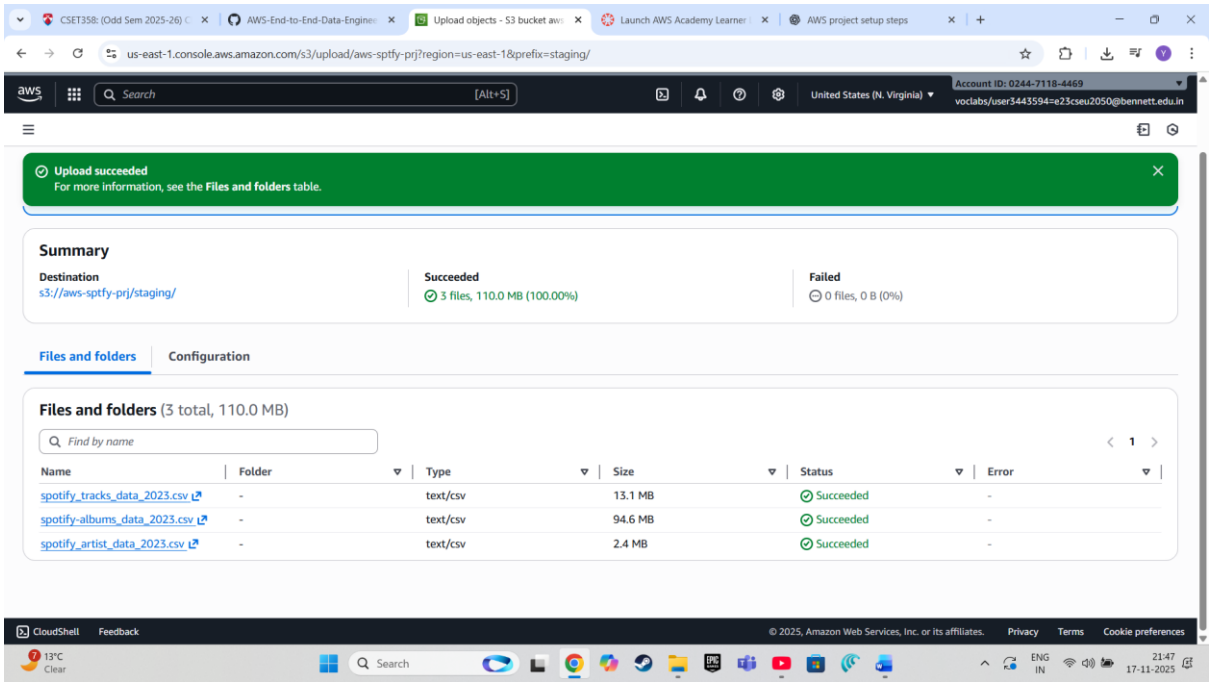
Creating S3 Buckets

We set up S3 buckets to store raw and processed data. The staging folder in the S3 bucket will hold the raw data files, while the data-warehouse folder will store the processed data. This separation ensures organized storage and easy retrieval of data at different stages of the pipeline.



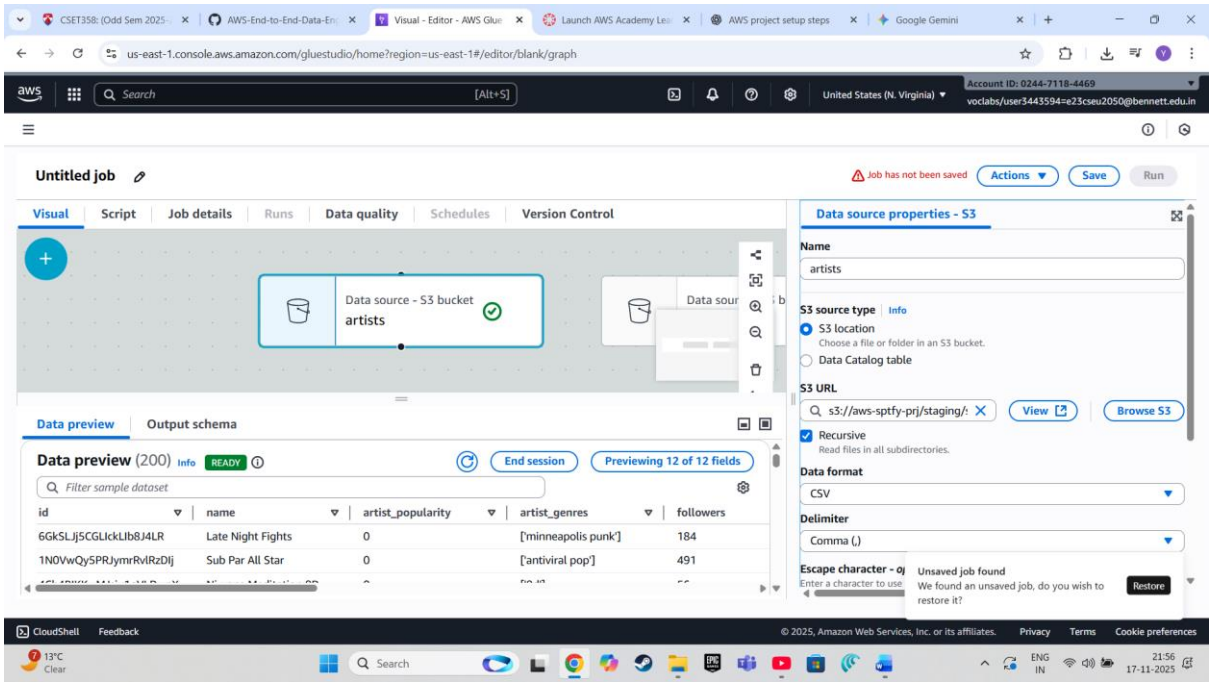


We upload the csv files in the staging folder

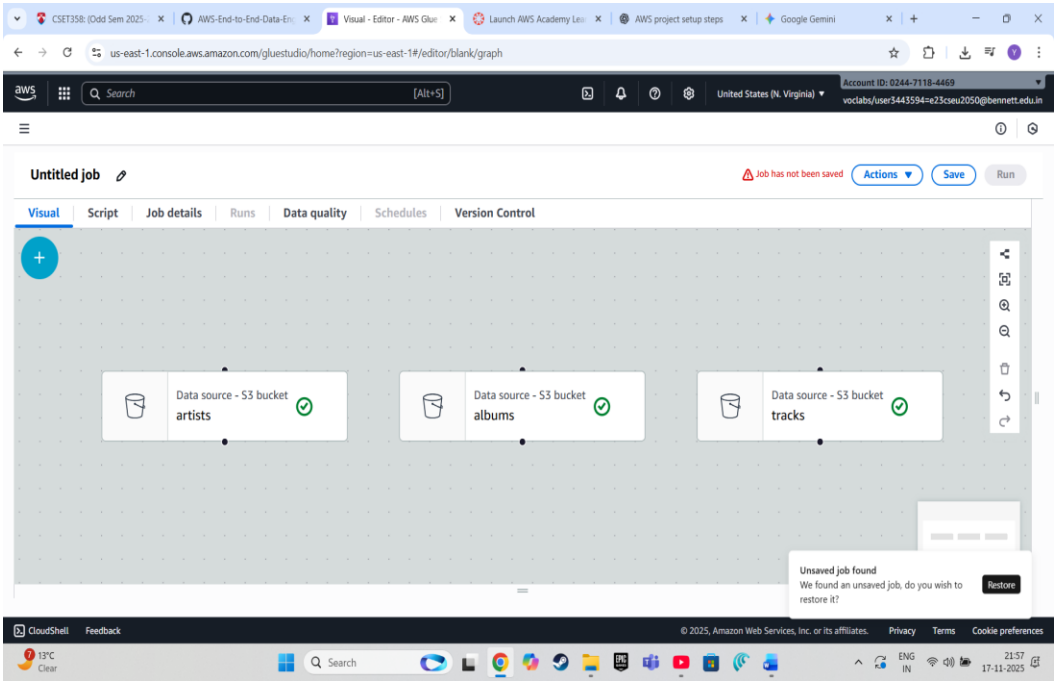


Setting Up AWS Glue for ETL

AWS Glue is used to create a data pipeline that automates the process of extracting data from the staging folder, transforming it by joining and cleaning the data, and then loading it into the data-warehouse folder. This step highlights the power of Glue in managing ETL (Extract, Transform, Load) operations seamlessly.

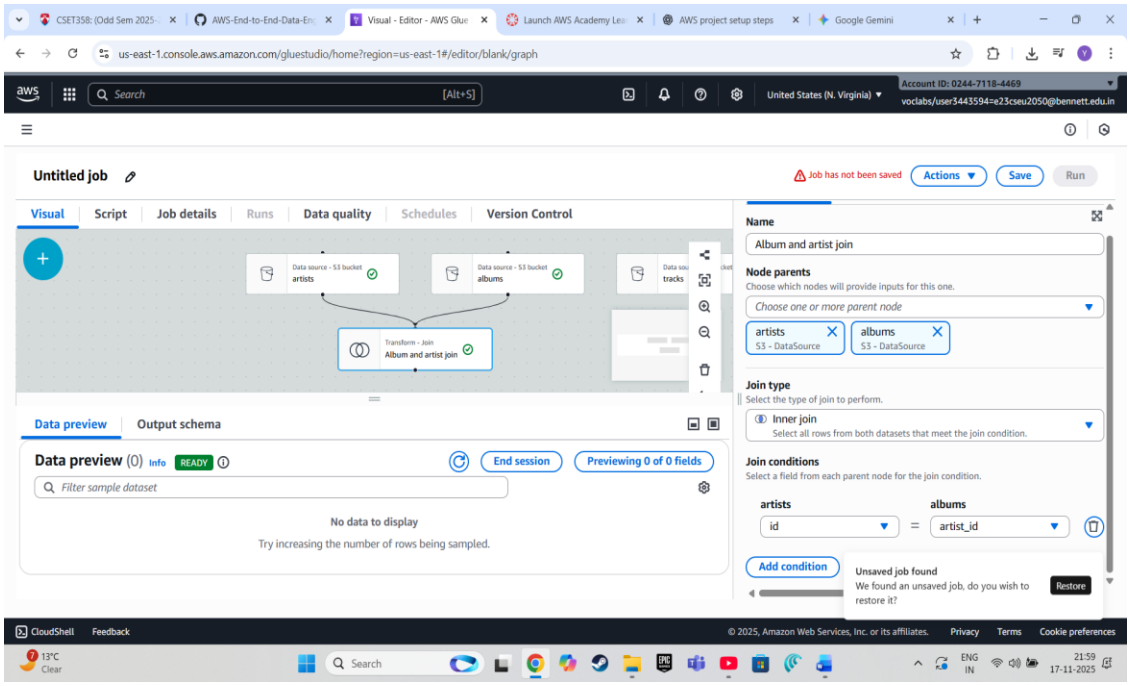


- Select the **First Amazon S3 bucket** and rename it with **"artist"** and click on **browse** and select the file from the **s3://spotify-aws-prjct/staging/spotify_artist_data_2023.csv**. Select the **Data format as CSV**.
- Repeat the step for the other two



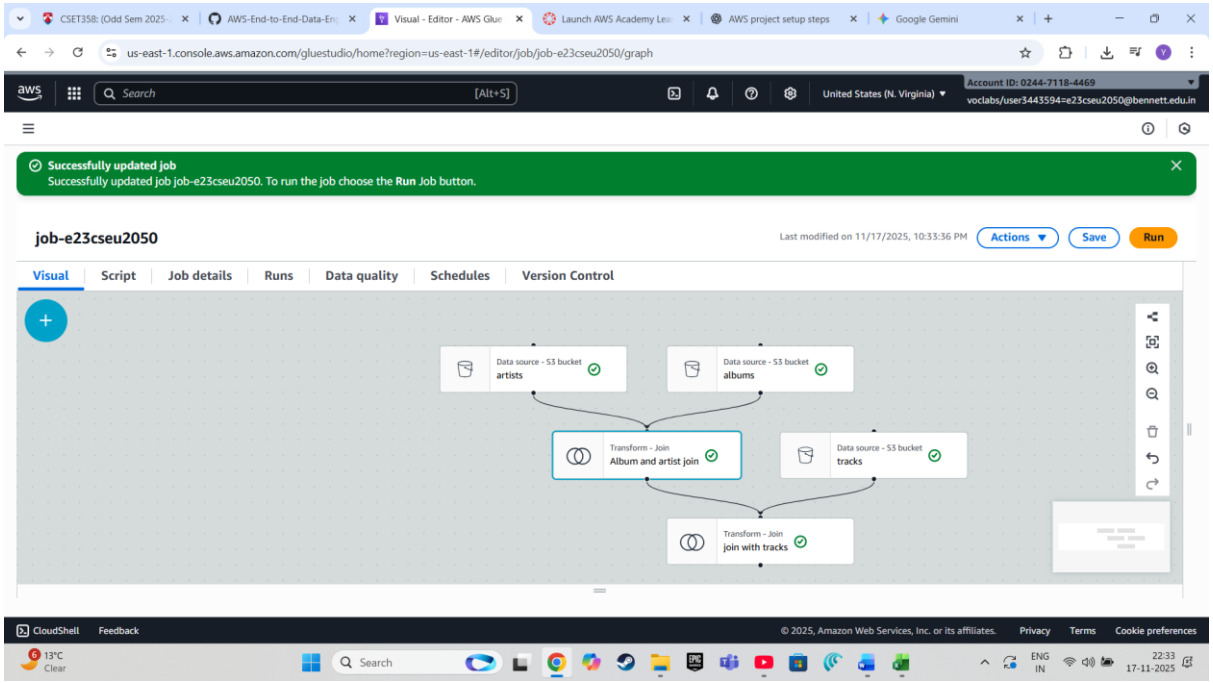
4. Configure Data Transformations:

- Now to join **album** and **artist**, click on the **ADD symbol**, select **join** from **Transforms** and connect nodes as per the image below.

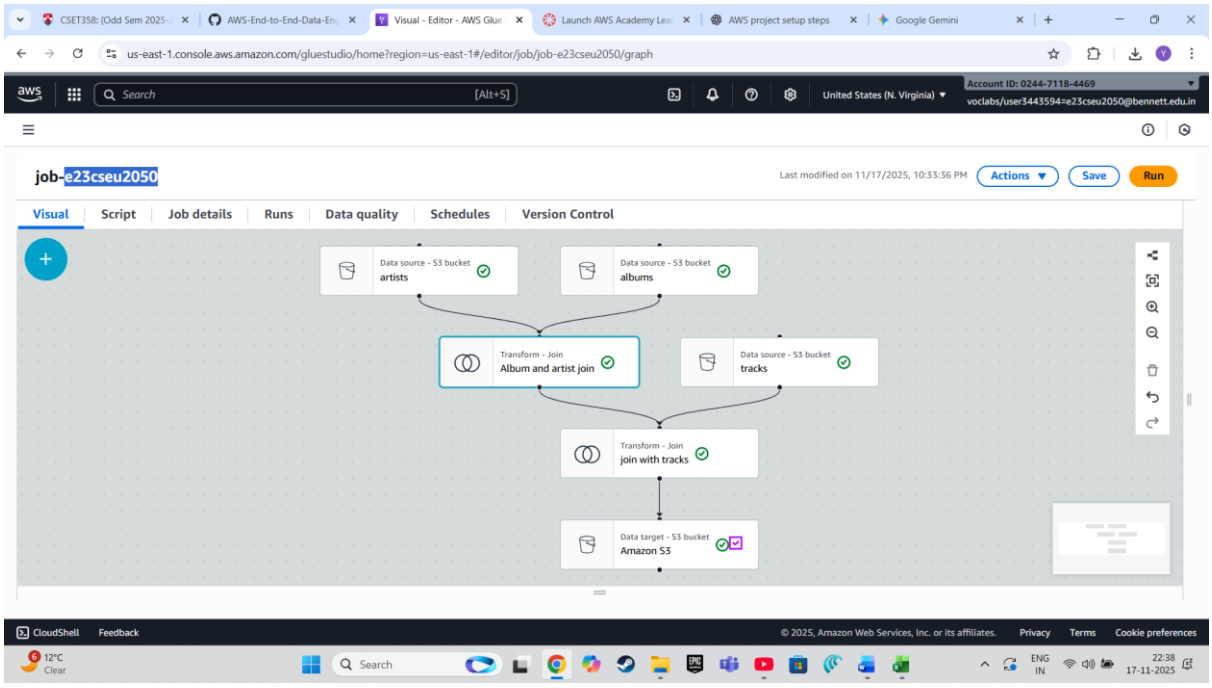


- After Joining the nodes of both the buckets to the **Join Transform**. Add a Condition where **artist 'id' = albums 'artist_id'** and rename the join **"Album and Artist Join"**.

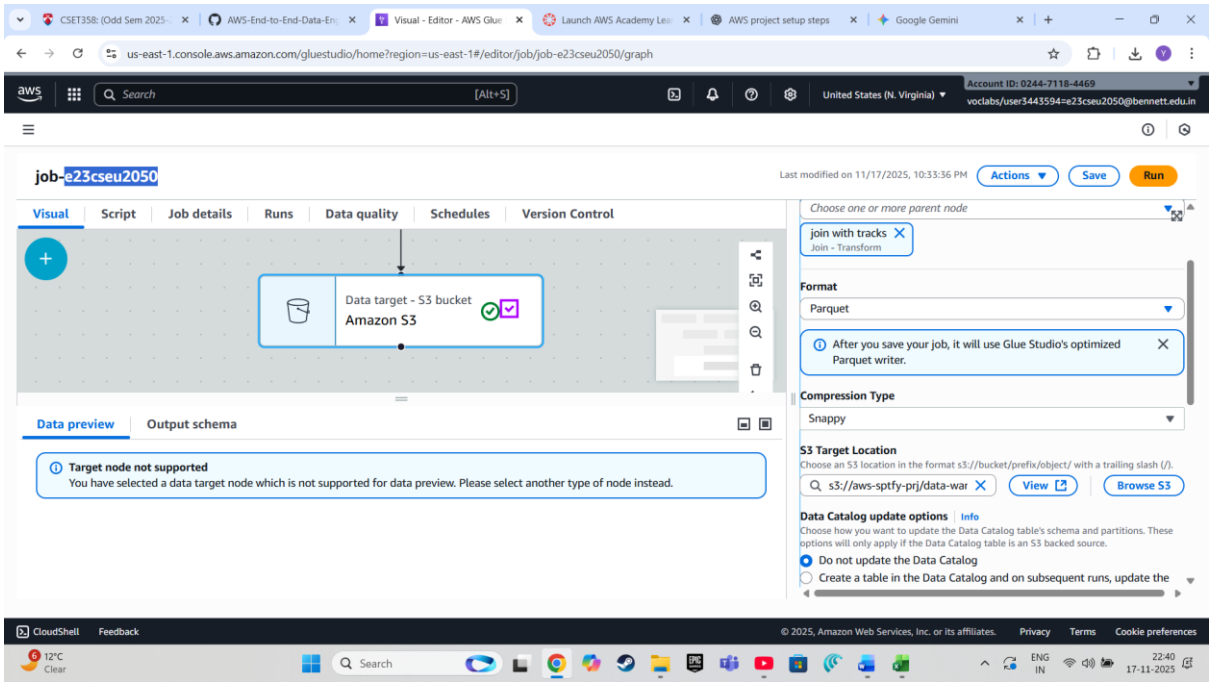
Now add another **Join Transform** to join **'track' S3 bucket** and **'Album and Artist' Join**.

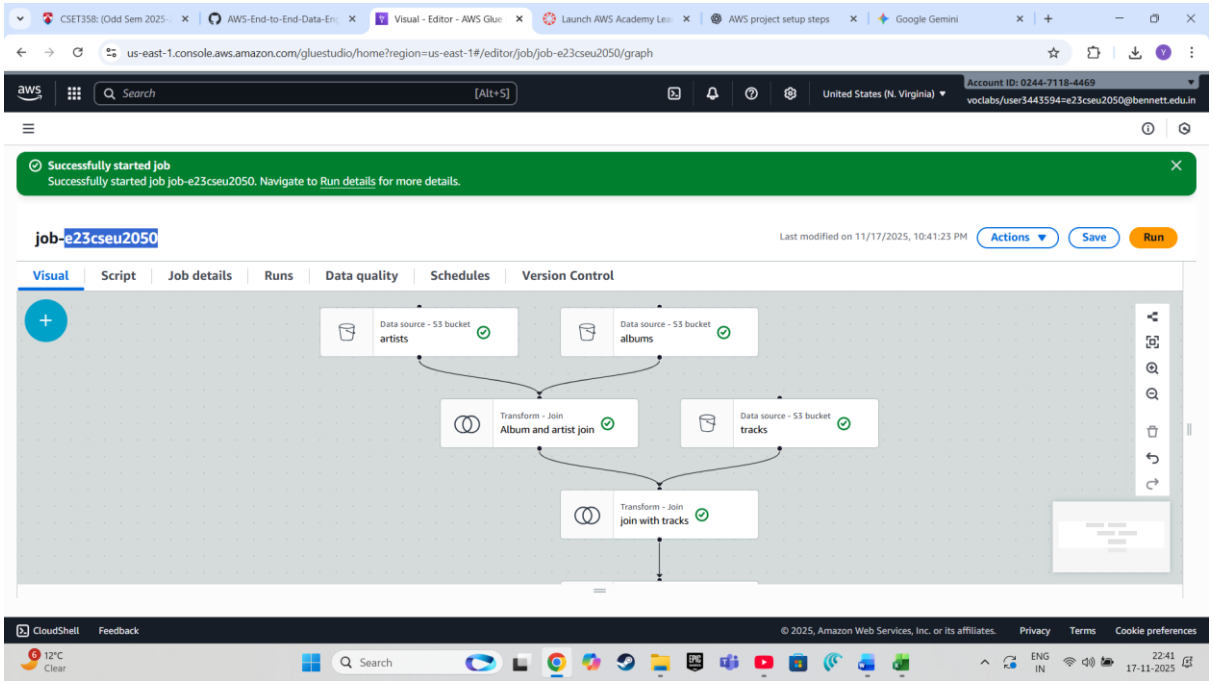


Now select the **Join** and add the condition **Album and Artist Join 'track_id' = tracks 'id'** and rename the join as **'Join with Tracks'**.



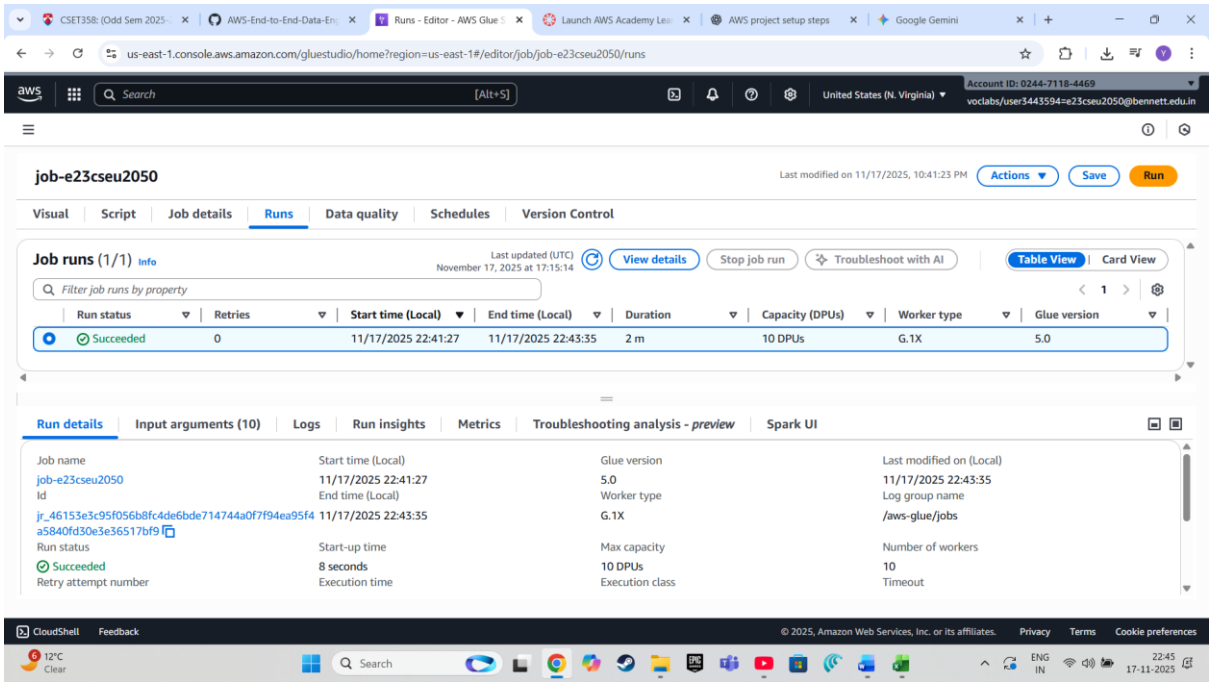
Rename the Destination node and add the Target location and make sure the compression type is Snappy.



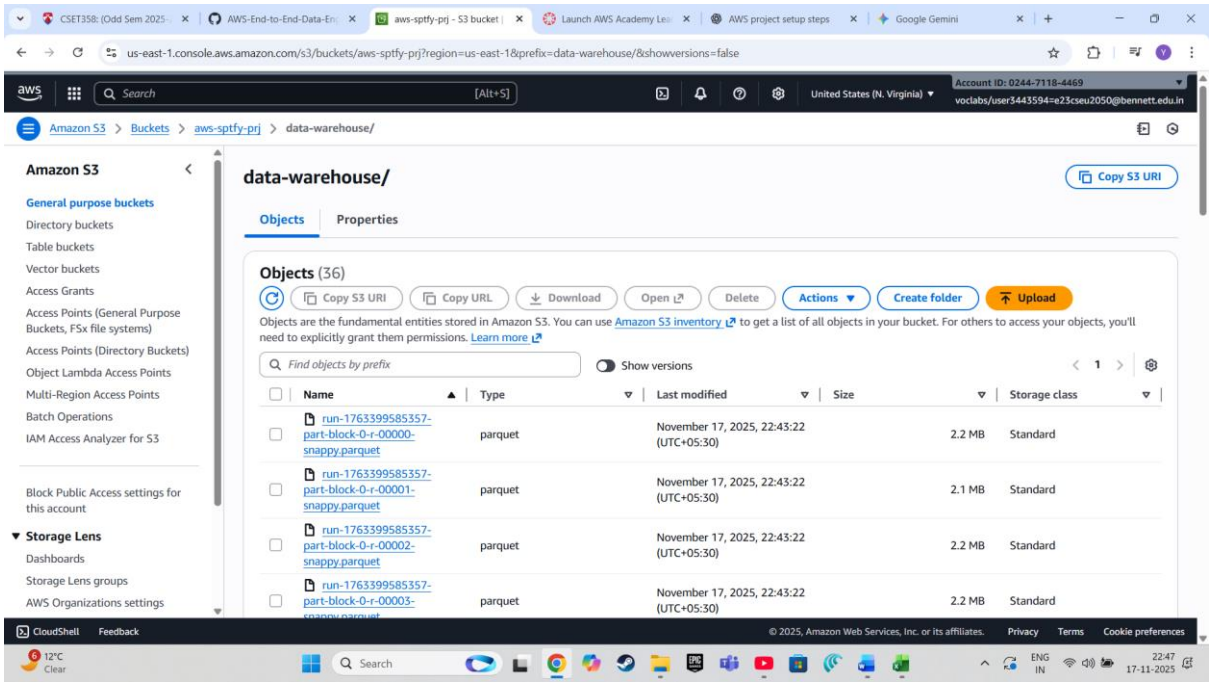


6. Run the Glue Job:

- **Review your job settings.**
- **Click "Run job" to start the ETL process, transforming and moving data from the staging layer to the data warehouse.**



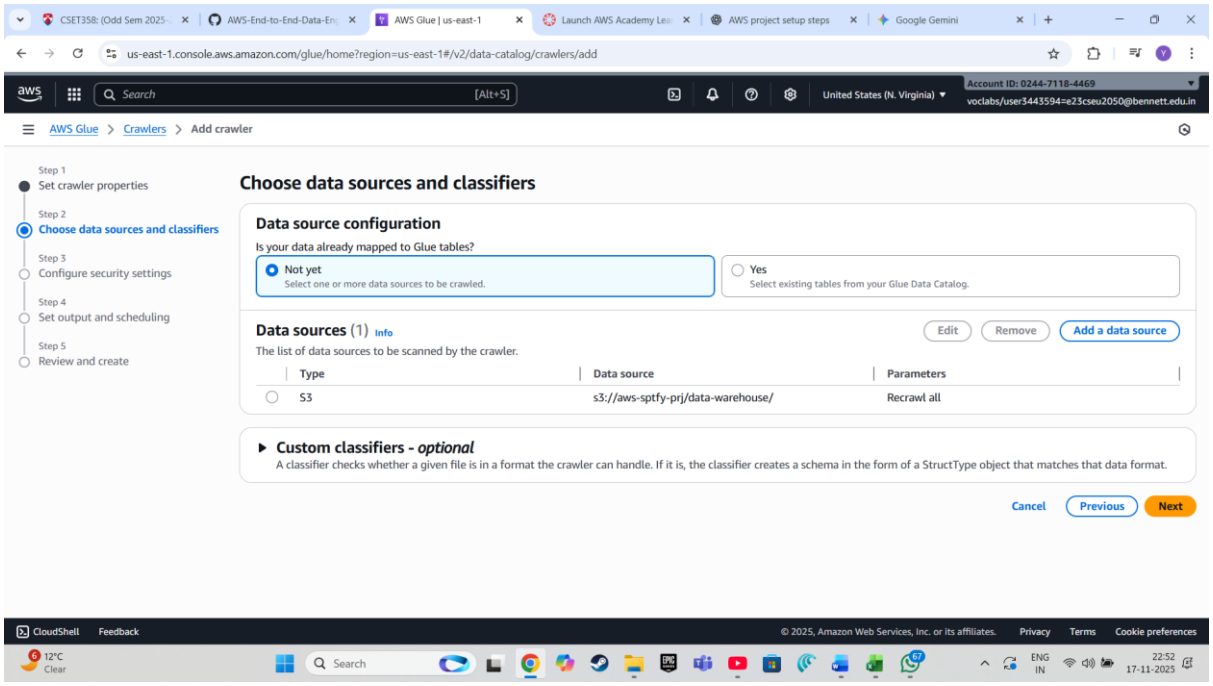
Check S3 bucket if the files are Parsed.

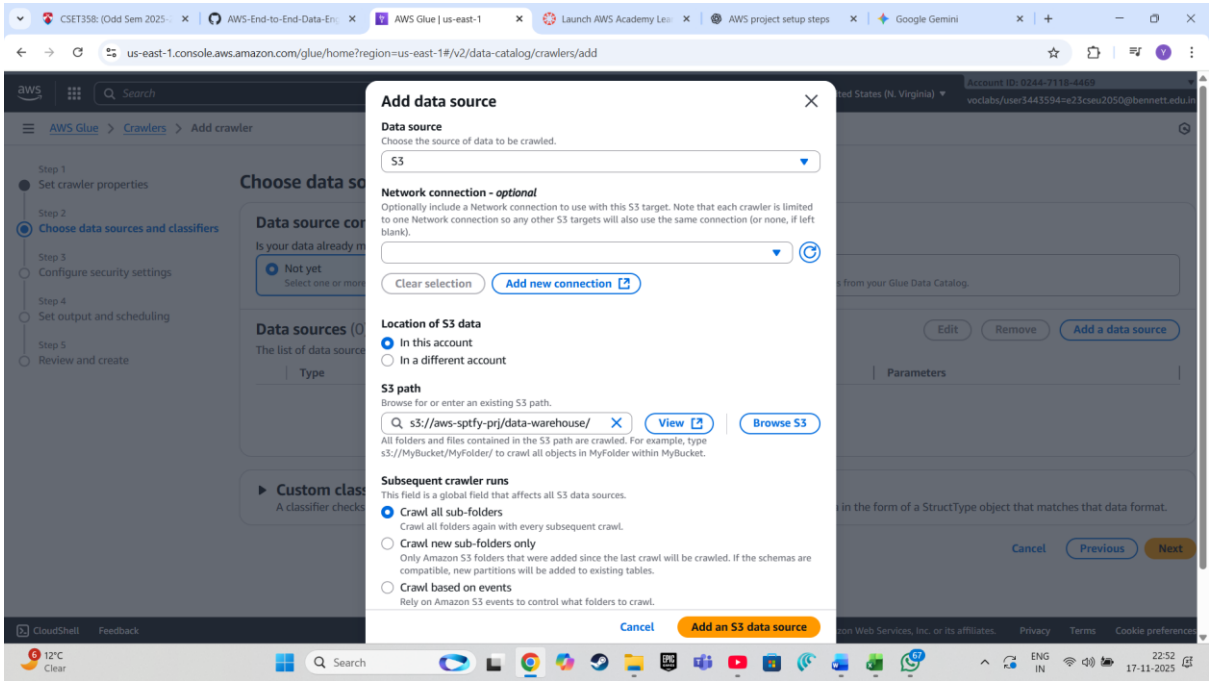


Step 4: Creating a Data Catalog with AWS Glue Crawler

1. **Create a New Crawler:**
   - **In the AWS Glue dashboard, click on "Crawlers" under the "Data catalog" section.**

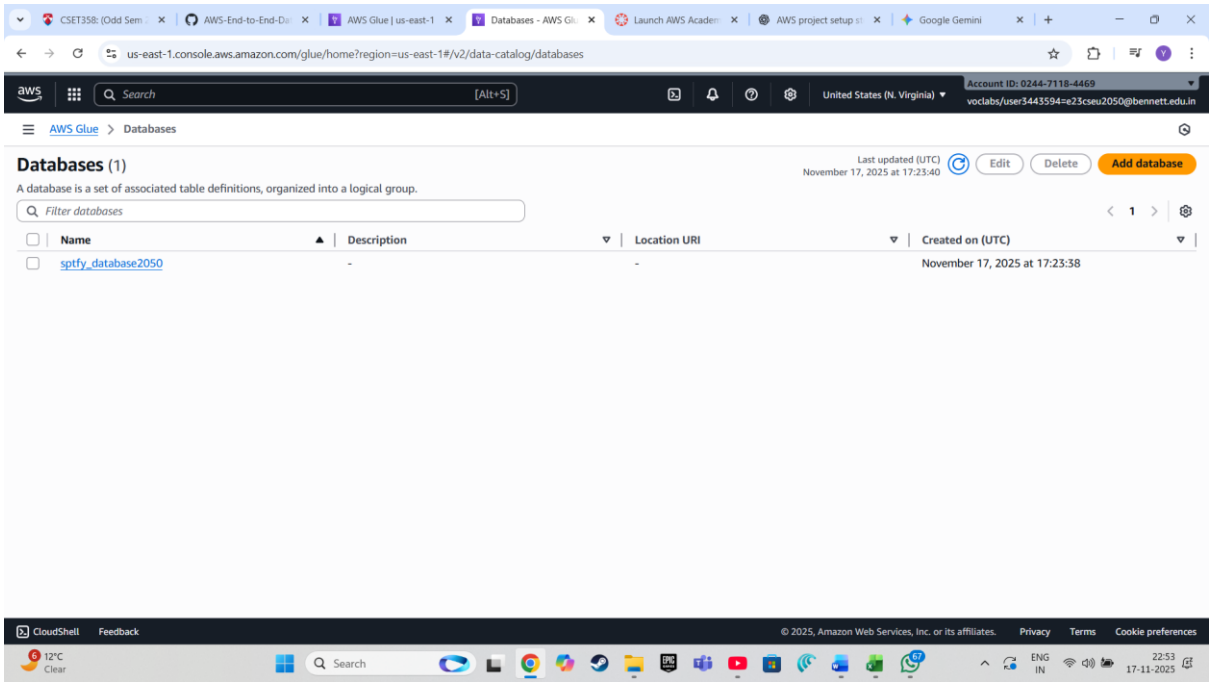
- **Click "Create crawler."**

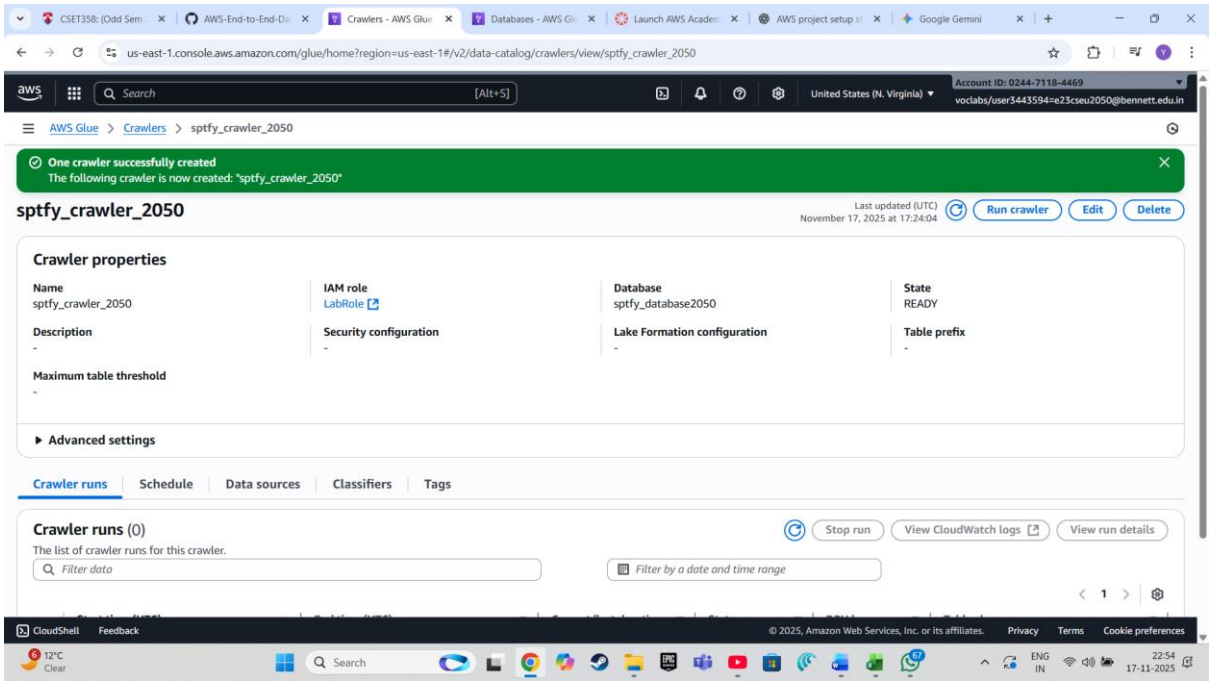


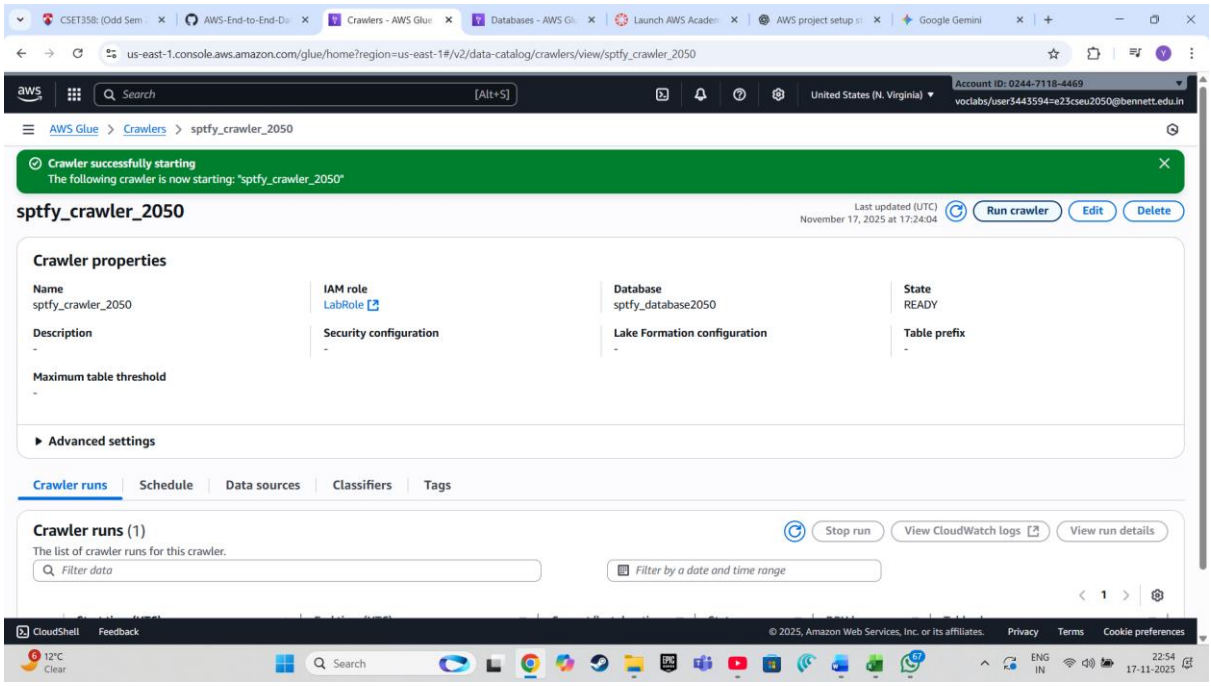


2. Create a New Database:

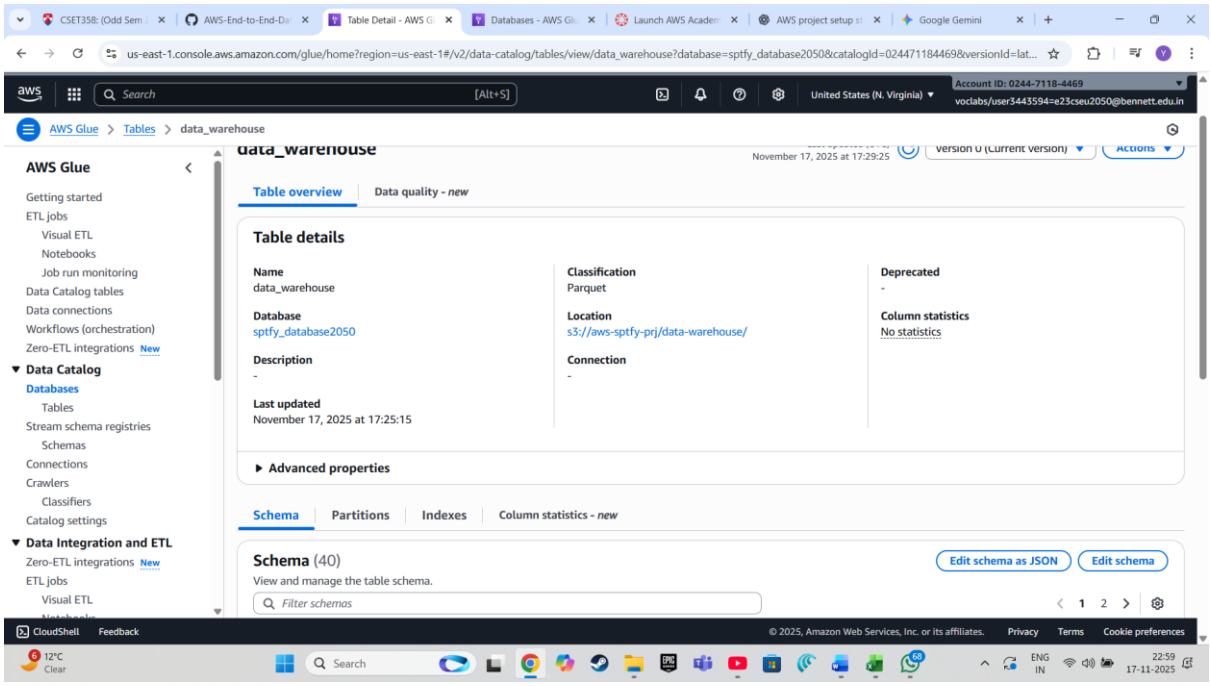
- **Open the duplicate tab. Go to the AWS Glue > Data catalog >> Databases, and create a new database.**


us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/databases
Search
[Alt+S]
United States (N. Virginia)
Account ID: 0244-7118-4469
voclabs/user3443594=e23cseu2050@bennett.edu.in
AWS Glue > Databases
Databases (1)
A database is a set of associated table definitions, organized into a logical group.
Last updated (UTC)
November 17, 2025 at 17:23:40
Edit
Delete
Add database
Filter databases
Name
Description
Location URI
Created on (UTC)
sptfy_database2050
-
-
November 17, 2025 at 17:23:38
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/view/sptfy_crawler_2050
Search
[Alt+S]
United States (N. Virginia)
Account ID: 0244-7118-4469
voclabs/user3443594=e23cseu2050@bennett.edu.in
AWS Glue > Crawlers > sptfy_crawler_2050
One crawler successfully created
The following crawler is now created: "sptfy_crawler_2050"
sptfy_crawler_2050
Last updated (UTC)
November 17, 2025 at 17:24:04
Run crawler
Edit
Delete
Crawler properties
Name
sptfy_crawler_2050
IAM role
LabRole
Database
sptfy_database2050
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
Filter data
Filter by a date and time range
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

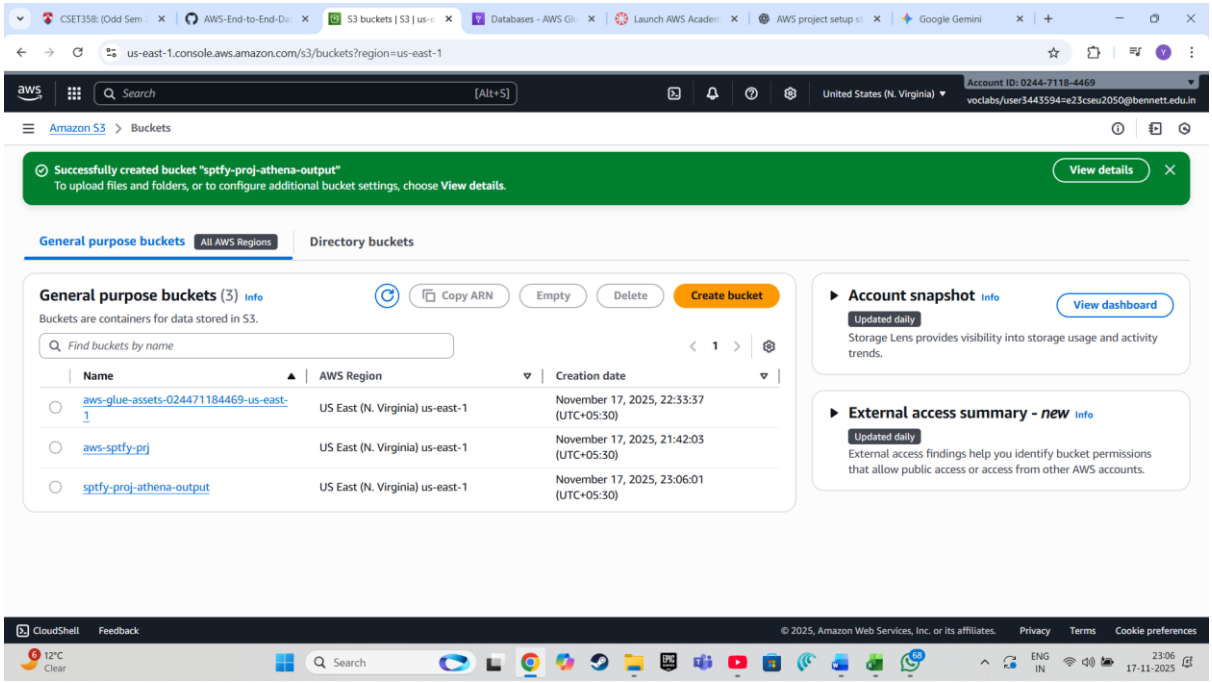


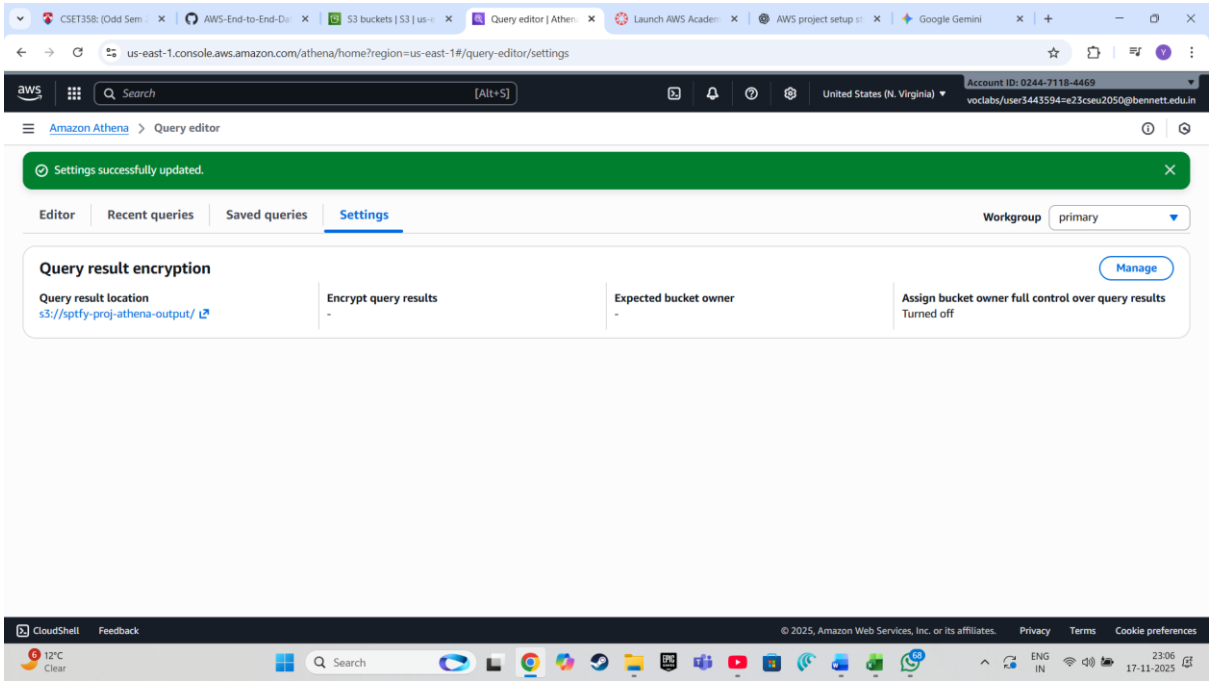
The crawler will scan the data in the **data-warehouse** folder and create corresponding tables in the **spotify_data** database.



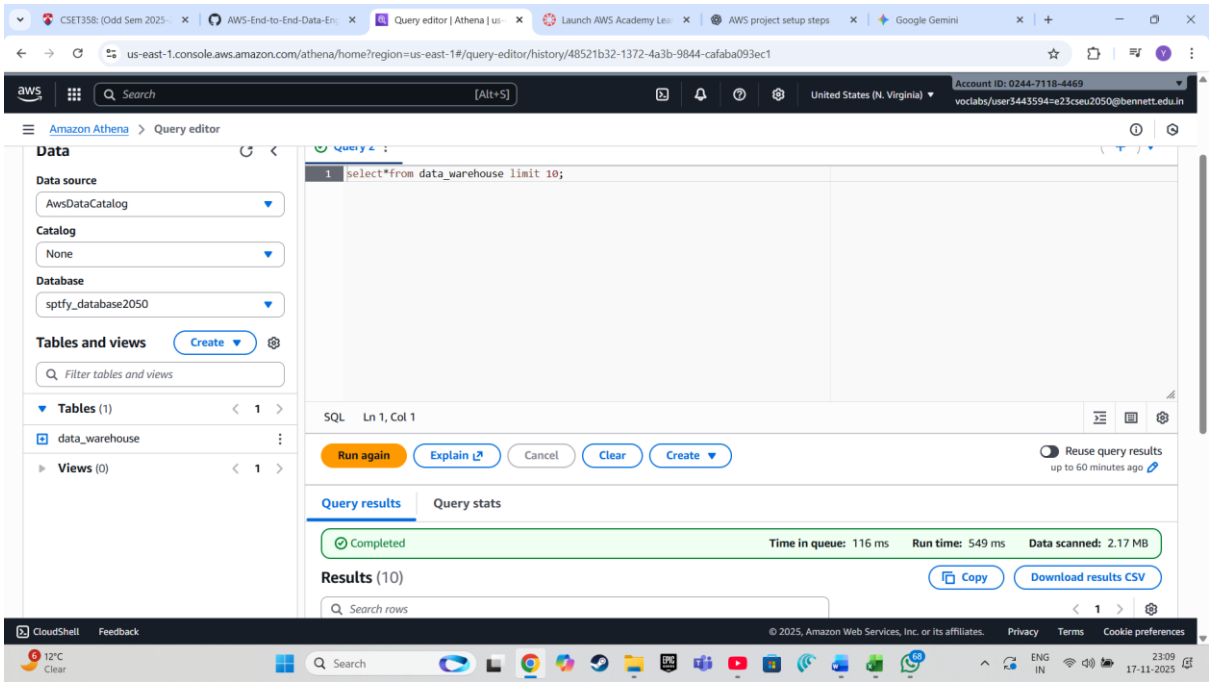
Step 5: Querying Data with AWS Athena

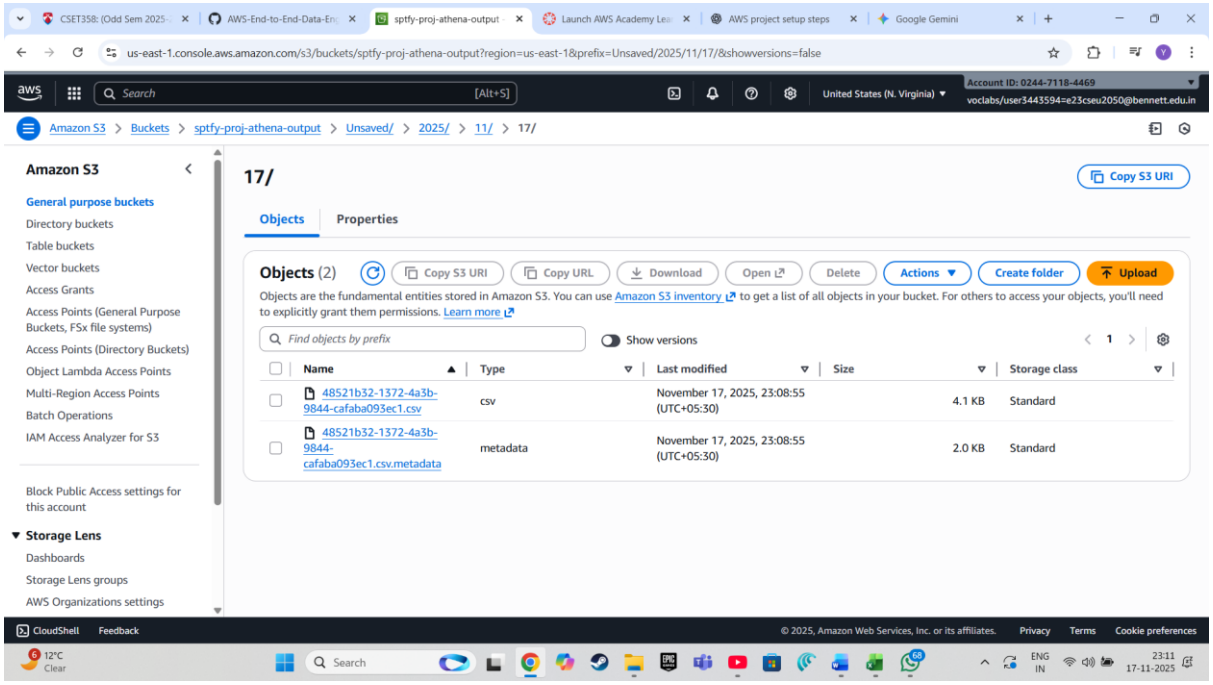
1. **Set Up Query Result Storage:**
2. Navigate to **AWS Athena** from the **AWS Management Console**
3. Before running any queries, you must specify a location for query results. Create a new **S3 bucket** named **sptfy-proj-athena-output**.





The results will be displayed in the **query editor**, and the output will be saved in the specified **S3 bucket (athena-output)**.

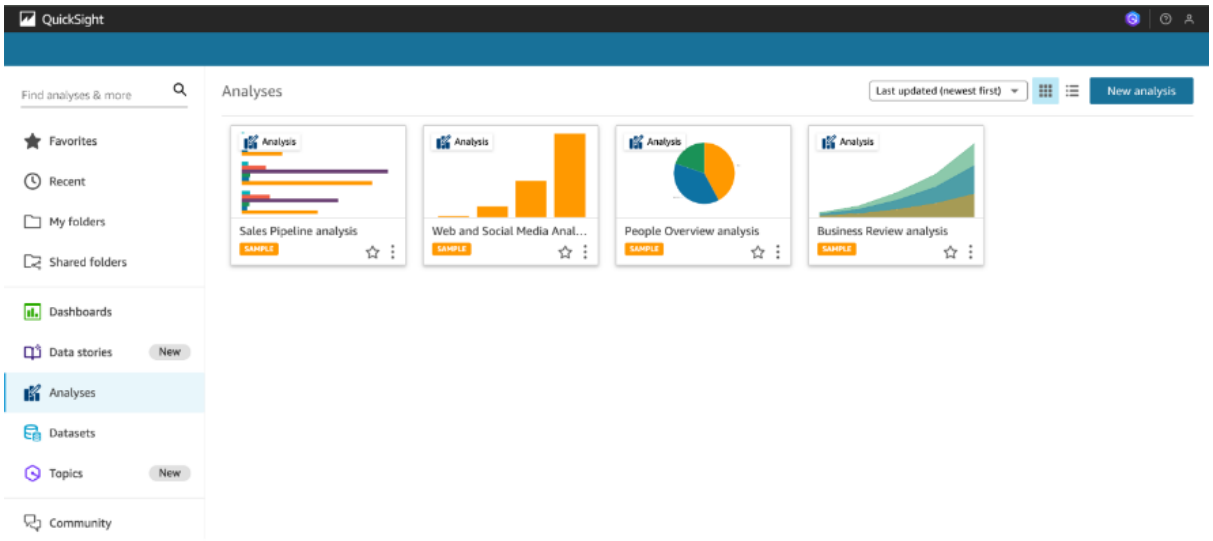




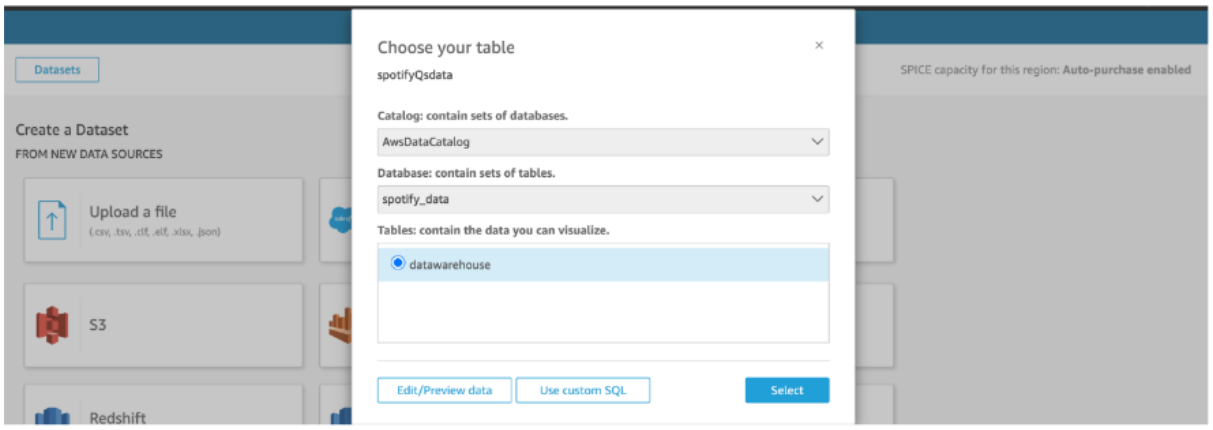
Step 6: Visualizing Data with AWS QuickSight

Connect QuickSight to Athena:

Once signed in, go to **"Datasets"** and click **"New dataset."**



- Select **"Athena"** as the data source and name it with **SpotifyQsData**
- Choose the **spotify_data** database and the **data_warehouse** table.



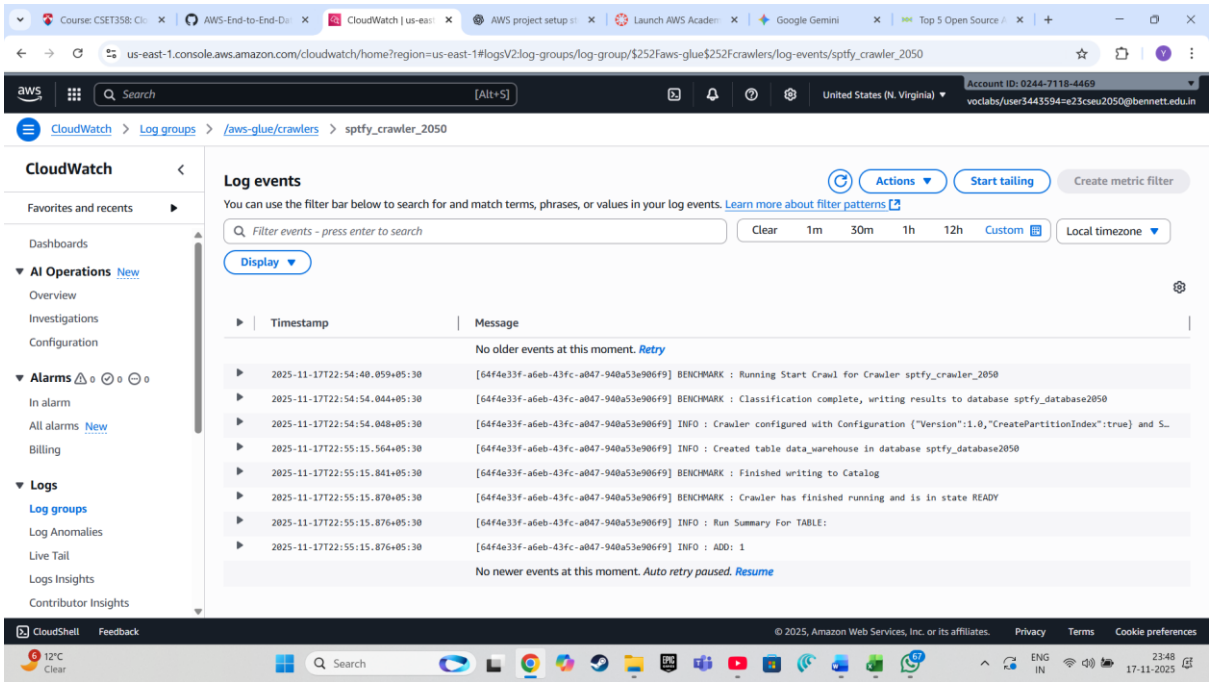
Create Visualizations:

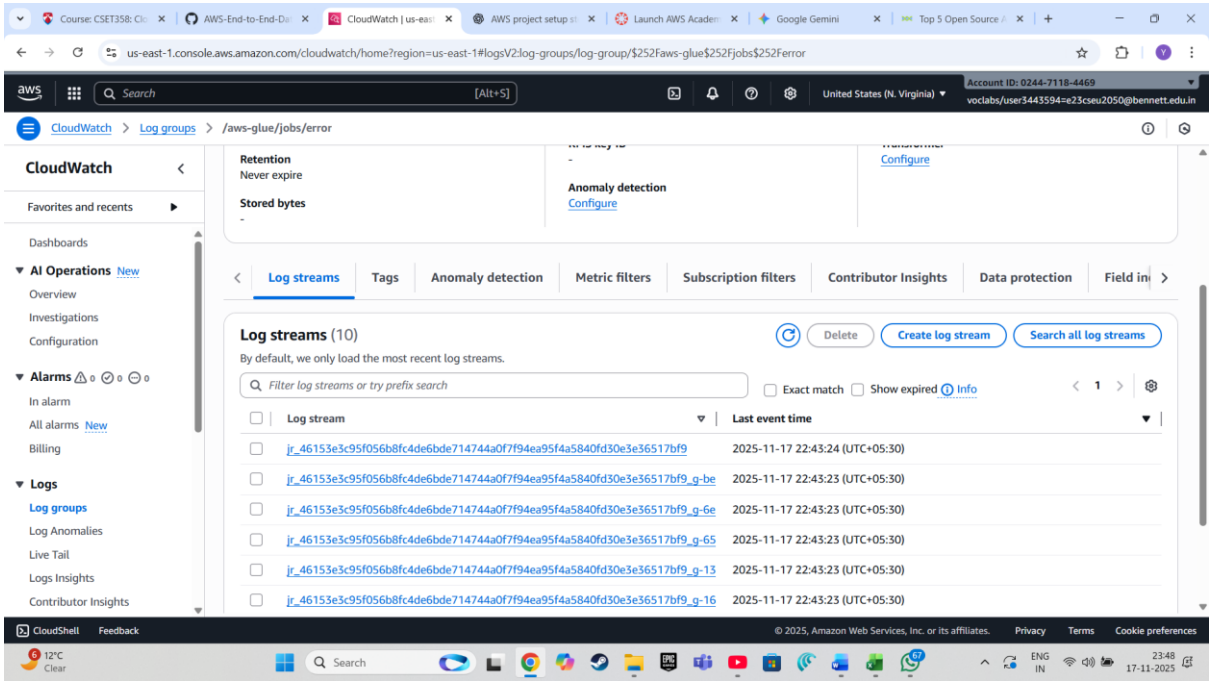
- After importing the data, you can create various types of visualizations (e.g., **bar charts**, **line charts**, **pie charts**) using the fields from the **data_warehouse** table.
- **Example:** Create a bar chart to visualize the popularity of tracks by artist.

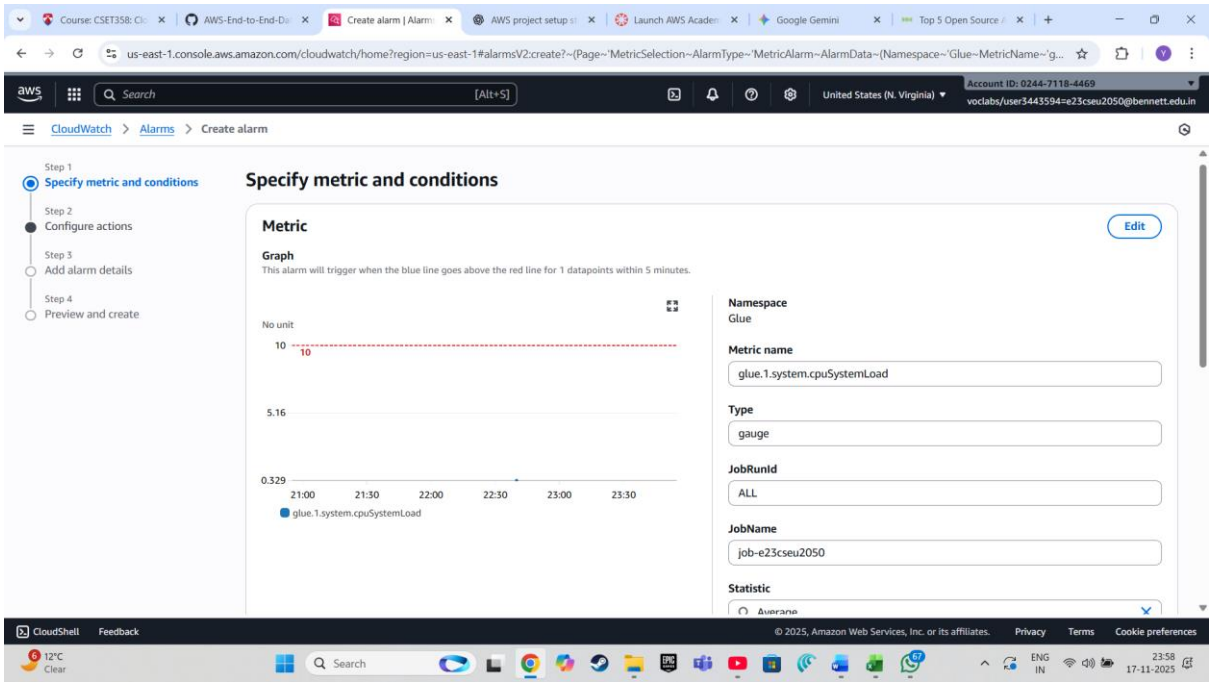
4. Publish and Share Dashboards:

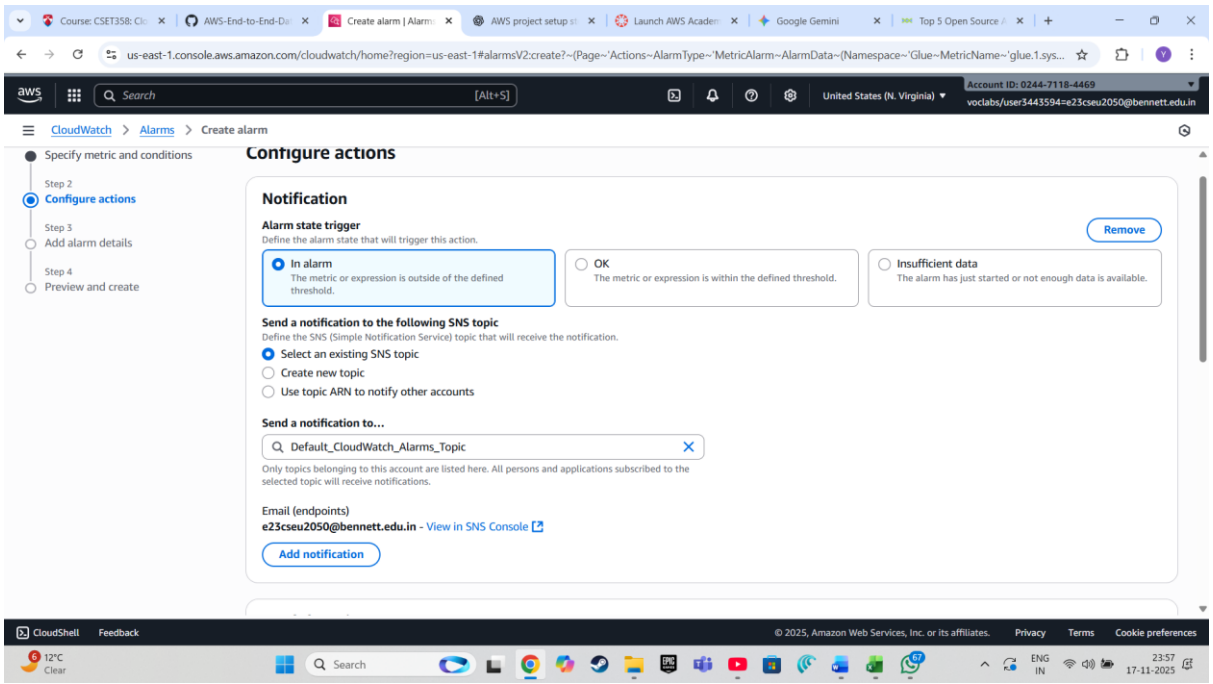
- Once your visualizations are ready, you can **publish the dashboard** and share it with stakeholders.
- Click **"Publish dashboard"** and follow the prompts to share it via email or a link.

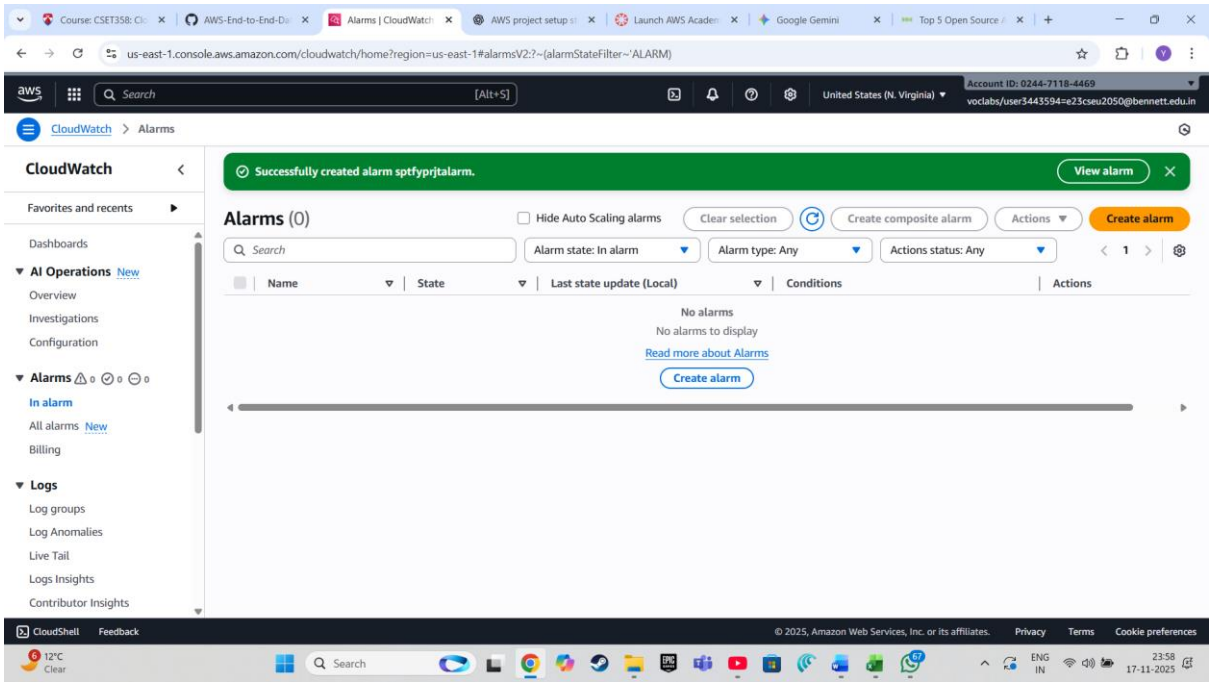
Step 7: AWS CloudWatch

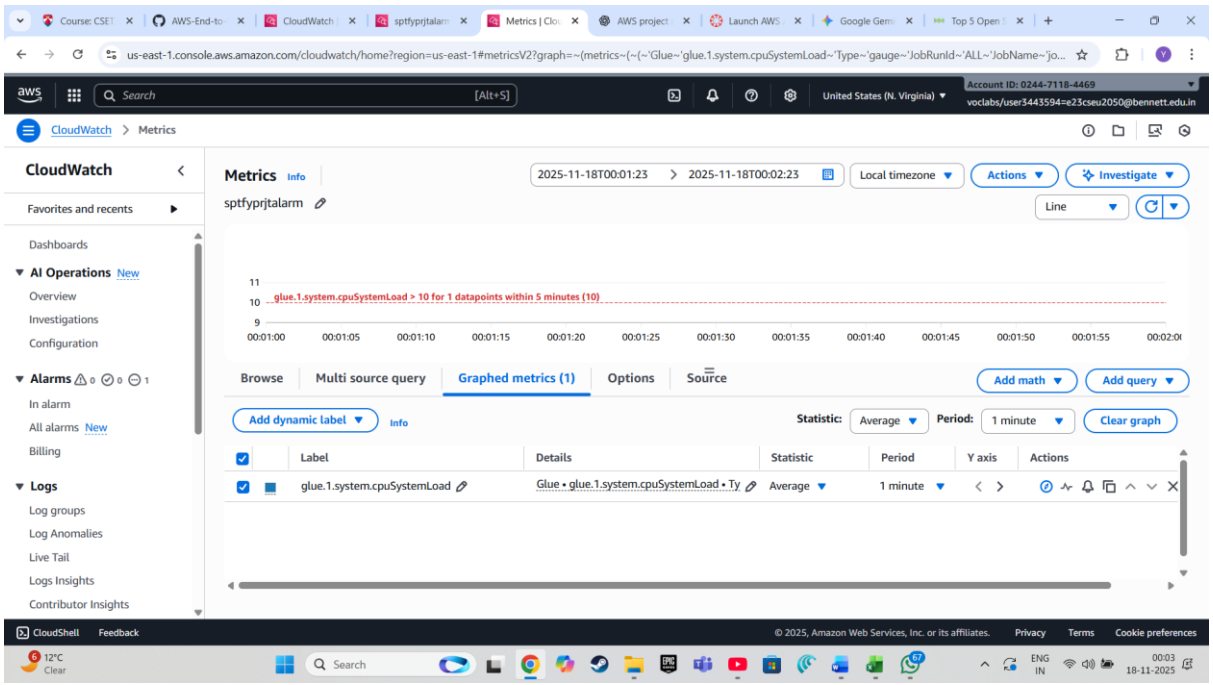

us-east-1.console.aws.amazon.com/cloudwatch/home?region=us-east-1#logsV2:log-groups/log-group/$252Faws-glue$252Fcrawlers/log-events/sptfy_crawler_2050
CloudWatch > Log groups > /aws-glue/crawlers > sptfy_crawler_2050
Log events
Actions
Start tailing
Create metric filter
You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns
Filter events - press enter to search
Clear 1m 30m 1h 12h Custom
Local timezone
Display
Timestamp
Message
No older events at this moment. Retry
2025-11-17T22:54:40.059+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] BENCHMARK : Running Start Crawl for Crawler sptfy_crawler_2050
2025-11-17T22:54:54.044+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] BENCHMARK : Classification complete, writing results to database sptfy_database2050
2025-11-17T22:54:54.048+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] INFO : Crawler configured with Configuration {"Version":1.0,"CreatePartitionIndex":true} and S…
2025-11-17T22:55:15.564+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] INFO : Created table data_warehouse in database sptfy_database2050
2025-11-17T22:55:15.841+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] BENCHMARK : Finished writing to Catalog
2025-11-17T22:55:15.870+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] BENCHMARK : Crawler has finished running and is in state READY
2025-11-17T22:55:15.876+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] INFO : Run Summary For TABLE:
2025-11-17T22:55:15.876+05:30 [64f4e33f-a6eb-43fc-a047-940a53e906f9] INFO : ADD: 1
No newer events at this moment. Auto retry paused. Resume
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


us-east-1.console.aws.amazon.com/cloudwatch/home?region=us-east-1#logsV2:log-groups/log-group/$252Faws-glue$252Fjobs$252Ferror
CloudWatch > Log groups > /aws-glue/jobs/error
Retention
Never expire
Stored bytes
Anomaly detection
Configure
Log streams
Tags
Anomaly detection
Metric filters
Subscription filters
Contributor Insights
Data protection
Log streams (10)
Delete
Create log stream
Search all log streams
By default, we only load the most recent log streams.
Filter log streams or try prefix search
Exact match
Show expired
Info
Log stream
Last event time
jr_46153e3c95f056b8fc4de6bde714744a0f7f94ea95f4a5840fd30e3e36517bf9 2025-11-17 22:43:24 (UTC+05:30)
jr_46153e3c95f056b8fc4de6bde714744a0f7f94ea95f4a5840fd30e3e36517bf9_g-be 2025-11-17 22:43:23 (UTC+05:30)
jr_46153e3c95f056b8fc4de6bde714744a0f7f94ea95f4a5840fd30e3e36517bf9_g-6e 2025-11-17 22:43:23 (UTC+05:30)
jr_46153e3c95f056b8fc4de6bde714744a0f7f94ea95f4a5840fd30e3e36517bf9_g-65 2025-11-17 22:43:23 (UTC+05:30)
jr_46153e3c95f056b8fc4de6bde714744a0f7f94ea95f4a5840fd30e3e36517bf9_g-13 2025-11-17 22:43:23 (UTC+05:30)
jr_46153e3c95f056b8fc4de6bde714744a0f7f94ea95f4a5840fd30e3e36517bf9_g-16 2025-11-17 22:43:23 (UTC+05:30)
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Specify metric and conditions
Metric
Edit
Graph
This alarm will trigger when the blue line goes above the red line for 1 datapoints within 5 minutes.
No unit
10
5.16
0.329
21:00
21:30
22:00
22:30
23:00
23:30
glue.1.system.cpuSystemLoad
Namespace
Glue
Metric name
glue.1.system.cpuSystemLoad
Type
gauge
JobRunId
ALL
JobName
job-e23cseu2050
Statistic


Configure actions
Notification
Alarm state trigger
Define the alarm state that will trigger this action.
Remove
In alarm
The metric or expression is outside of the defined threshold.
OK
The metric or expression is within the defined threshold.
Insufficient data
The alarm has just started or not enough data is available.
Send a notification to the following SNS topic
Define the SNS (Simple Notification Service) topic that will receive the notification.
Select an existing SNS topic
Create new topic
Use topic ARN to notify other accounts
Send a notification to...
Default_CloudWatch_Alarms_Topic
Only topics belonging to this account are listed here. All persons and applications subscribed to the selected topic will receive notifications.
Email (endpoints)
e23cseu2050@bennett.edu.in - View in SNS Console
Add notification


us-east-1.console.aws.amazon.com/cloudwatch/home?region=us-east-1#alarmsV2:?~(alarmStateFilter~'ALARM)
Search [Alt+S]
United States (N. Virginia)
Account ID: 0244-7118-4469
voclabs/user3443594=e23cseu2050@bennett.edu.in
CloudWatch > Alarms
CloudWatch
Favorites and recents
Dashboards
AI Operations New
Overview
Investigations
Configuration
Alarms
In alarm
All alarms New
Billing
Logs
Log groups
Log Anomalies
Live Tail
Logs Insights
Contributor Insights
Successfully created alarm sptfyprjtalarm.
View alarm
Alarms (0)
Hide Auto Scaling alarms
Clear selection
Create composite alarm
Actions
Create alarm
Search
Alarm state: In alarm
Alarm type: Any
Actions status: Any
Name
State
Last state update (Local)
Conditions
Actions
No alarms
No alarms to display
Read more about Alarms
Create alarm
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


us-east-1.console.aws.amazon.com/cloudwatch/home?region=us-east-1#metricsV2?graph=~(metrics~(~(~'Glue~'glue.1.system.cpuSystemLoad~'Type~'gauge~'JobRunId~'ALL~'JobName~'jo...
Search [Alt+S]
United States (N. Virginia)
Account ID: 0244-7118-4469
voclabs/user3443594=e23cseu2050@bennett.edu.in
CloudWatch > Metrics
CloudWatch
Favorites and recents
Dashboards
AI Operations New
Overview
Investigations
Configuration
Alarms
In alarm
All alarms New
Billing
Logs
Log groups
Log Anomalies
Live Tail
Logs Insights
Contributor Insights
Metrics Info
sptfyprjtalarm
2025-11-18T00:01:23
2025-11-18T00:02:23
Local timezone
Actions
Investigate
Line
glue.1.system.cpuSystemLoad > 10 for 1 datapoints within 5 minutes (10)
Browse
Multi source query
Graphed metrics (1)
Options
Source
Add math
Add query
Add dynamic label
Info
Statistic:
Average
Period:
1 minute
Clear graph
Label
Details
Statistic
Period
Y axis
Actions
glue.1.system.cpuSystemLoad
Glue • glue.1.system.cpuSystemLoad • Ty
Average
1 minute
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Conclusion

This document serves as a runbook for the **Spotify Data Engineering Project**. Follow each step to set up and run your end-to-end data pipeline.

Real-World Applications

The insights gained from this project can be applied in various real-world scenarios:

1. **Music Recommendations:** By analyzing the popularity of tracks and artists, platforms can improve their recommendation engines.
2. **Market Analysis:** Record labels and music producers can use the data to understand trends and make informed decisions on which genres or artists to promote.
3. **User Engagement:** Streaming services can analyze user preferences and behavior to enhance user engagement through personalized playlists and features.
4. **Business Intelligence:** The visualizations and insights derived can help business analysts make data-driven decisions to optimize marketing strategies and improve user retention