AWS ETL Steps

Introduction

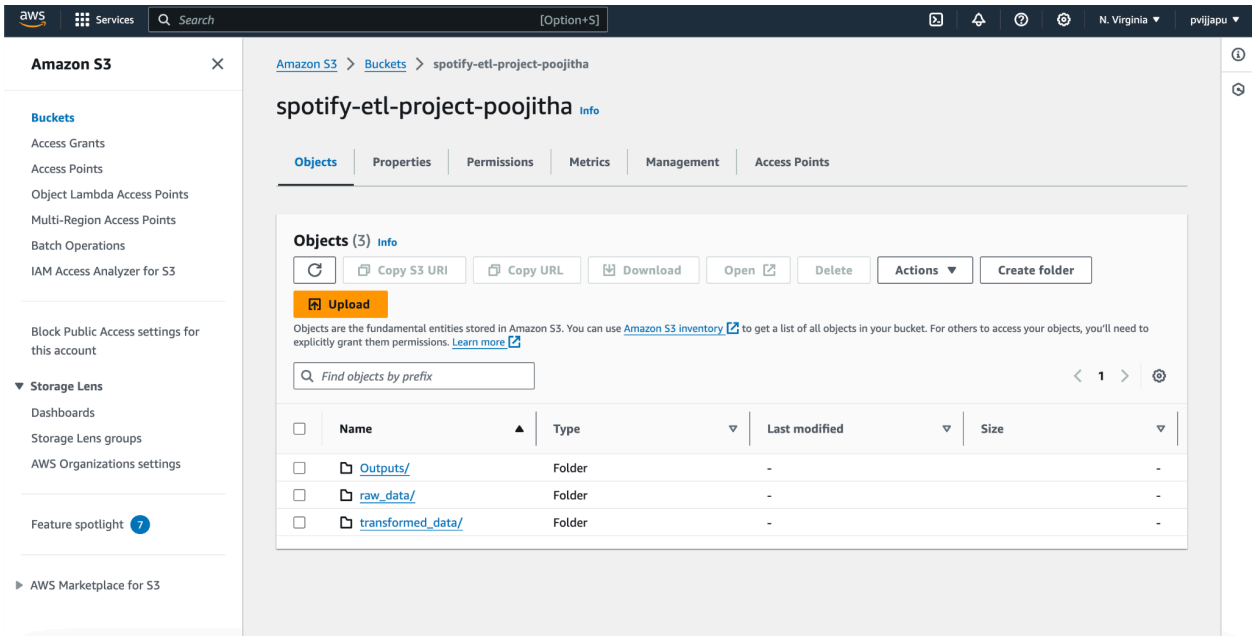
This ETL process extracts data from the Spotify API, transforms it, and loads it into Amazon S3 for further analysis using AWS Athena. The steps involve setting up S3 buckets, Lambda functions, EventBridge triggers, and crawlers to automate and streamline the data workflow.

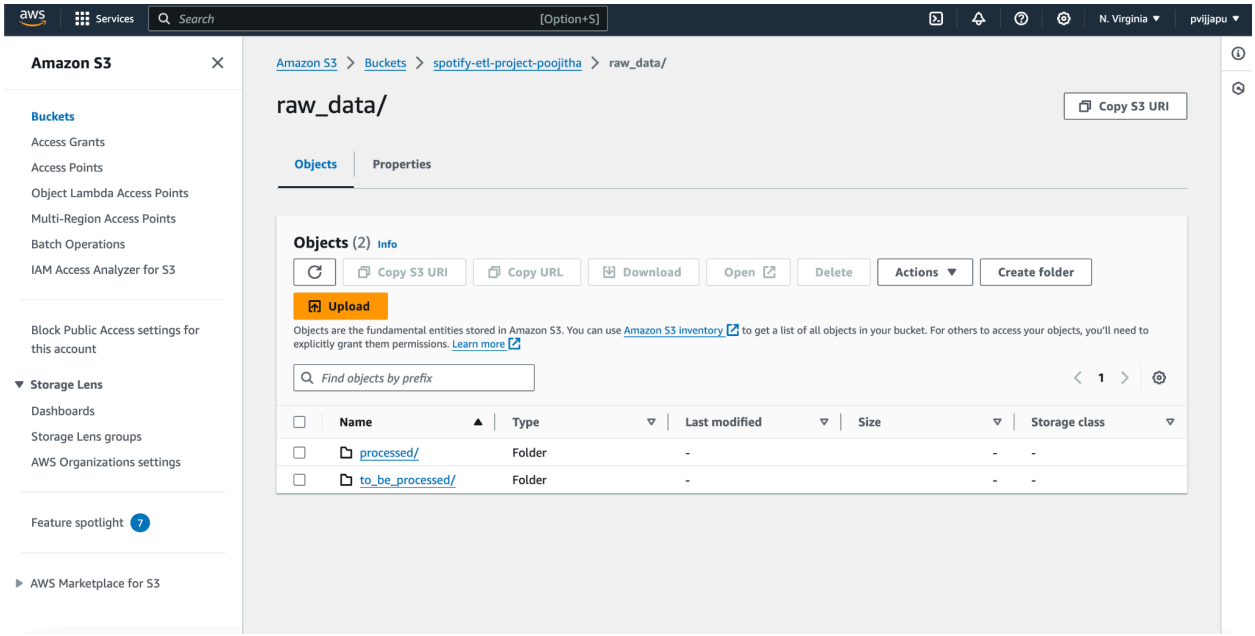
Steps

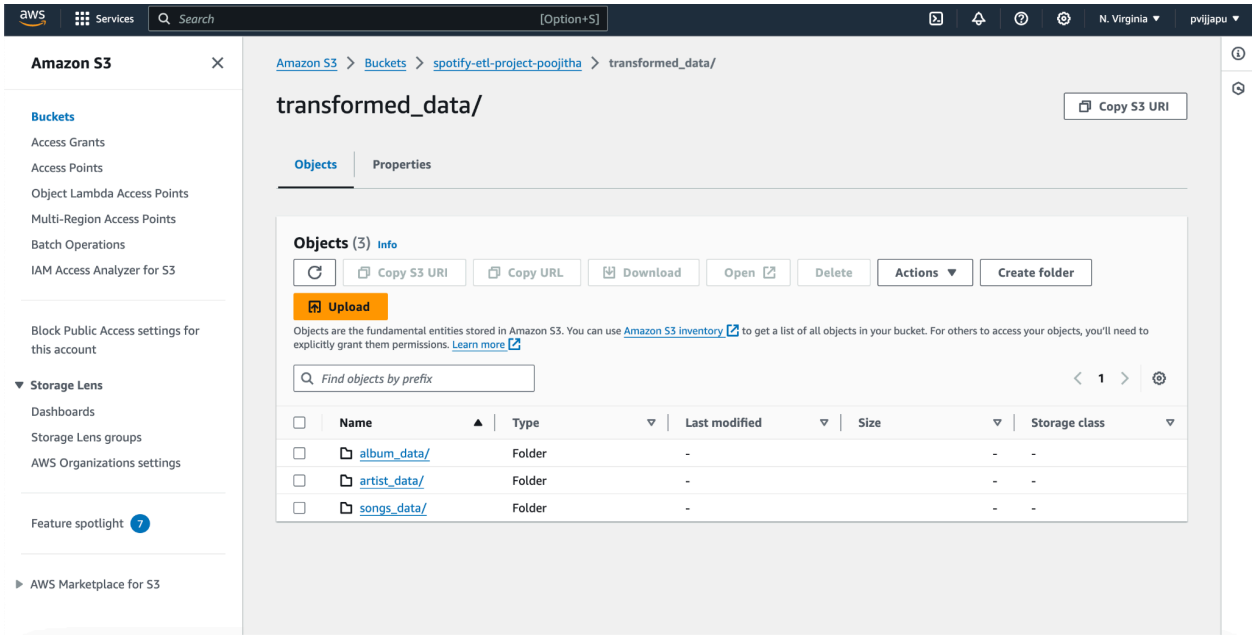
1. Creating S3 Buckets

S3 is a scalable storage service for storing and retrieving any amount of data at any time.

- *Raw Data Bucket:* Contains subfolders `processed` and `to_be_processed`. The `to_be_processed` folder holds incoming raw data files, while `processed` stores files that have been processed.
- *Transformed Data Bucket:* Contains subfolders `albums_data`, `artists_data`, and `songs_data` to store the respective transformed datasets.




aws
Services
Search
[Option+S]
N. Virginia
pvijjapu
Amazon S3
Buckets
Access Grants
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
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets > spotify-etl-project-poojitha > raw_data/
raw_data/
Copy S3 URI
Objects
Properties
Objects (2) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
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
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
processed/
Folder
to_be_processed/
Folder

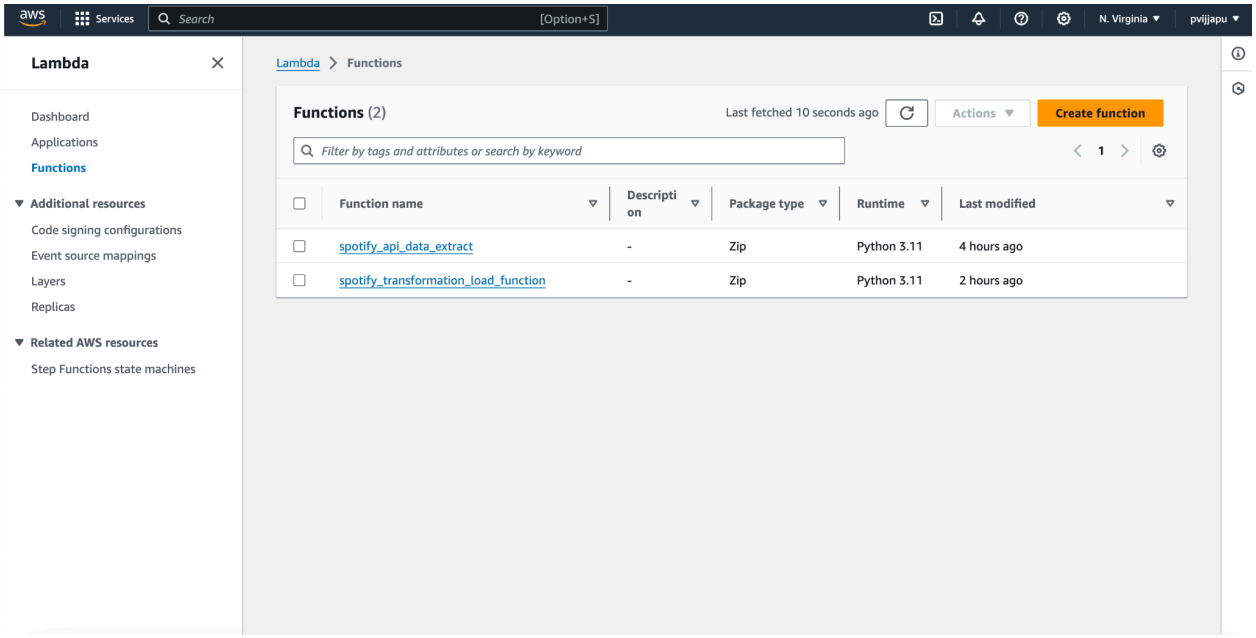

aws
Services
Search
[Option+S]
N. Virginia
pvijjapu
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets > spotify-etl-project-poojitha > transformed_data/
transformed_data/
Copy S3 URI
Objects
Properties
Objects (3) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
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
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
album_data/
Folder
artist_data/
Folder
songs_data/
Folder

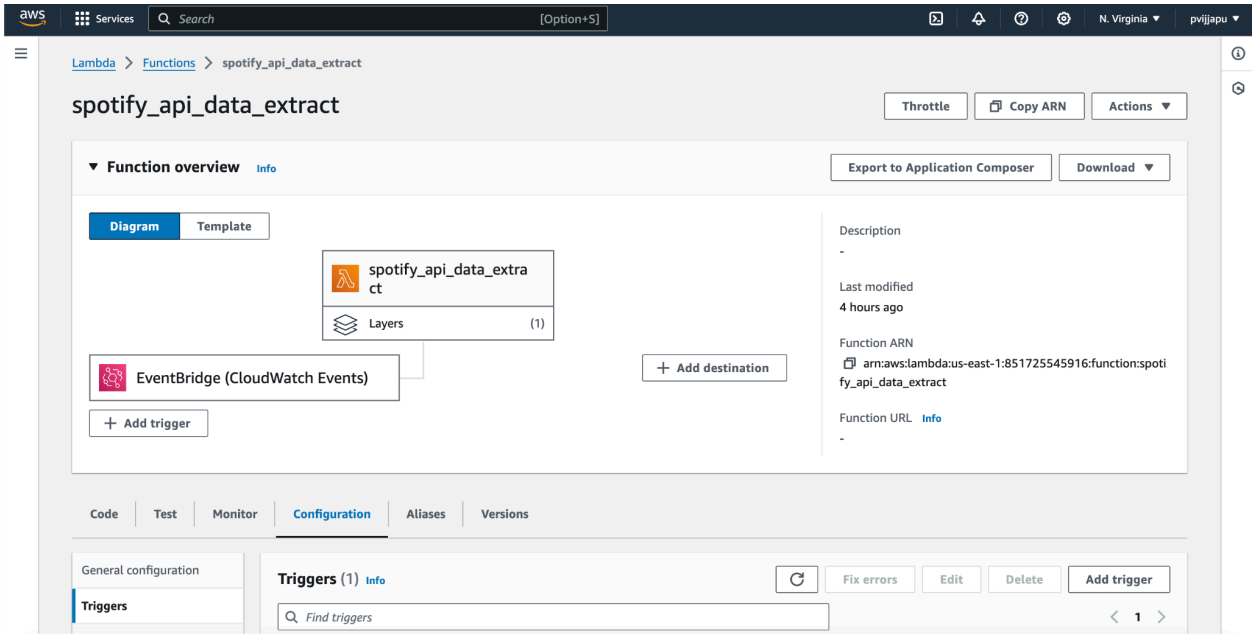
2. Lambda Functions

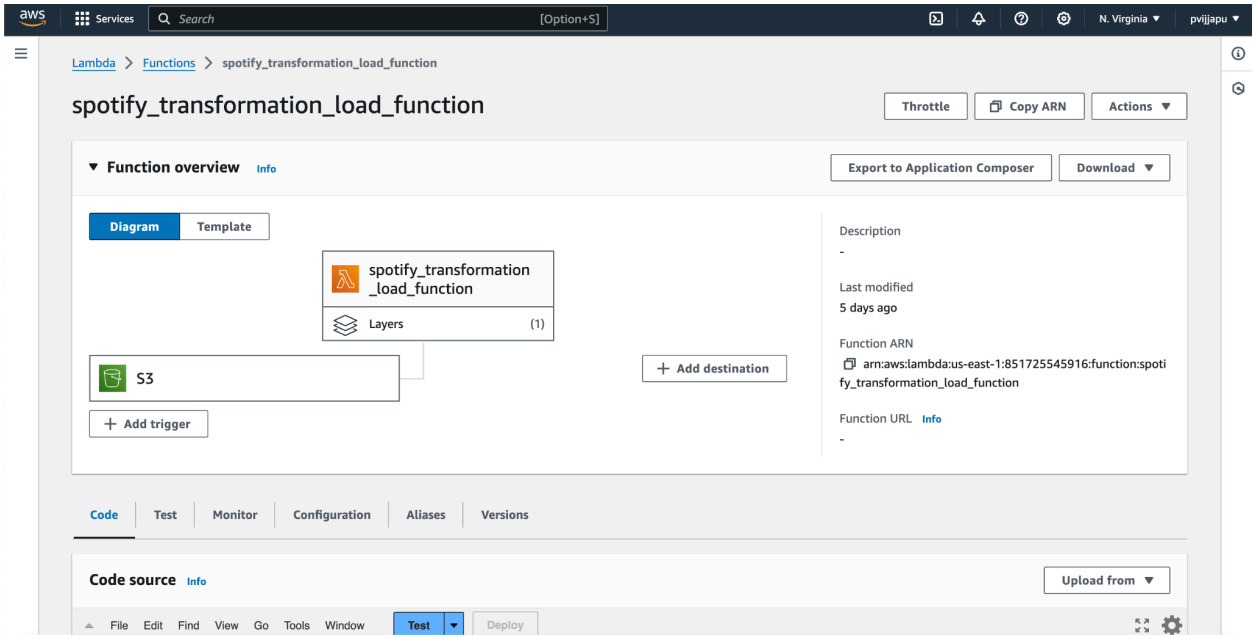
AWS Lambda is a serverless compute service that runs your code in response to events and automatically manages the underlying compute resources.

- *spotify_api_extract*: This function extracts data from the Spotify API and stores it in the `raw_data/to_be_processed` folder in S3. It runs daily to ensure the data is up-to-date.

- *spotify_load_transform_function*: This function processes the raw data, transforming it into structured formats (albums, artists, songs) and stores the results in the appropriate subfolders in the `transformed` bucket. It also moves processed files from `to_be_processed` to `processed`.




aws
Services
Search
[Option+S]
N. Virginia
pvijjapu
Lambda > Functions > spotify_api_data_extract
spotify_api_data_extract
Throttle
Copy ARN
Actions
Function overview Info
Export to Application Composer
Download
Diagram
Template
spotify_api_data_extra ct
Layers (1)
EventBridge (CloudWatch Events)
Add trigger
Add destination
Description
-
Last modified
4 hours ago
Function ARN
arn:aws:lambda:us-east-1:851725545916:function:spotify_api_data_extract
Function URL Info
-
Code
Test
Monitor
Configuration
Aliases
Versions
General configuration
Triggers
Triggers (1) Info
Fix errors
Edit
Delete
Add trigger
Find triggers
1

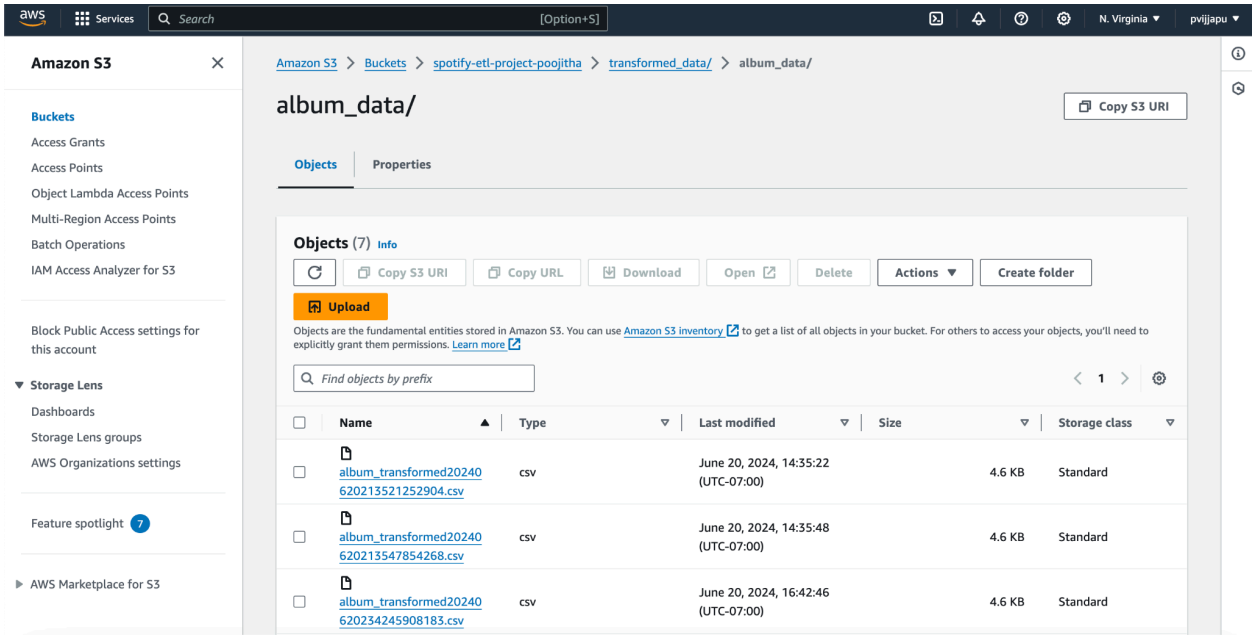

aws
Services
Search
[Option+S]
N. Virginia
pvijjapu
Lambda > Functions > spotify_transformation_load_function
spotify_transformation_load_function
Throttle
Copy ARN
Actions
Function overview Info
Export to Application Composer
Download
Diagram
Template
spotify_transformation _load_function
Layers (1)
S3
Add trigger
Add destination
Description
-
Last modified
5 days ago
Function ARN
arn:aws:lambda:us-east-1:851725545916:function:spotify_transformation_load_function
Function URL Info
-
Code
Test
Monitor
Configuration
Aliases
Versions
Code source Info
Upload from
File
Edit
Find
View
Go
Tools
Window
Test
Deploy

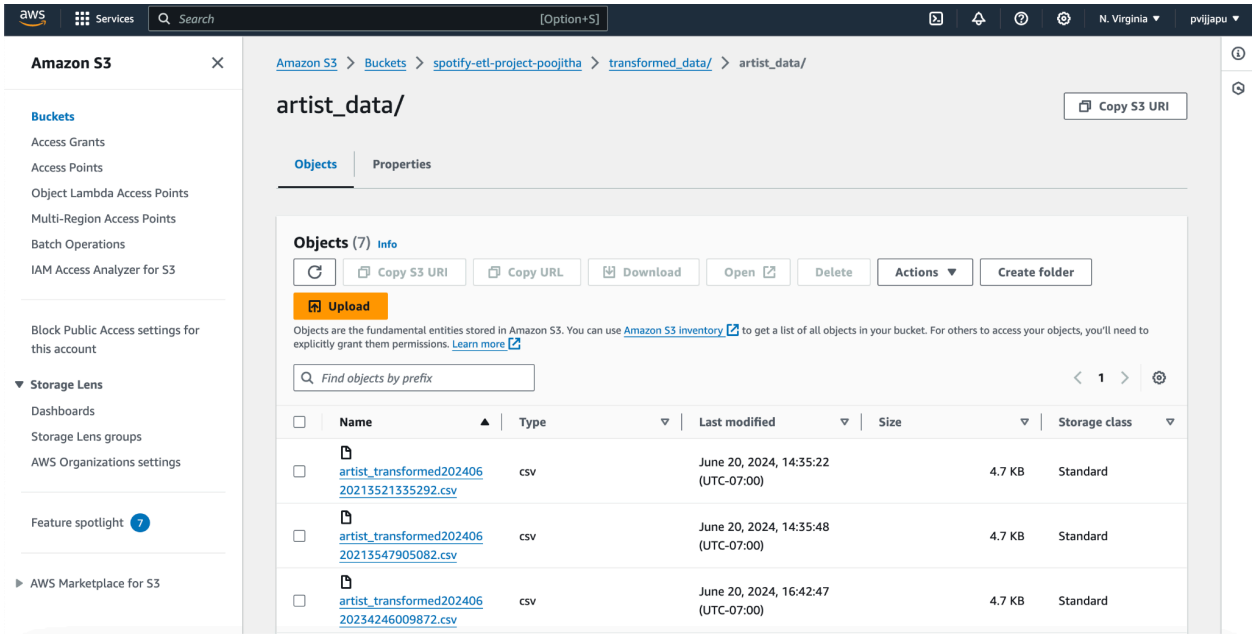
3. Triggers

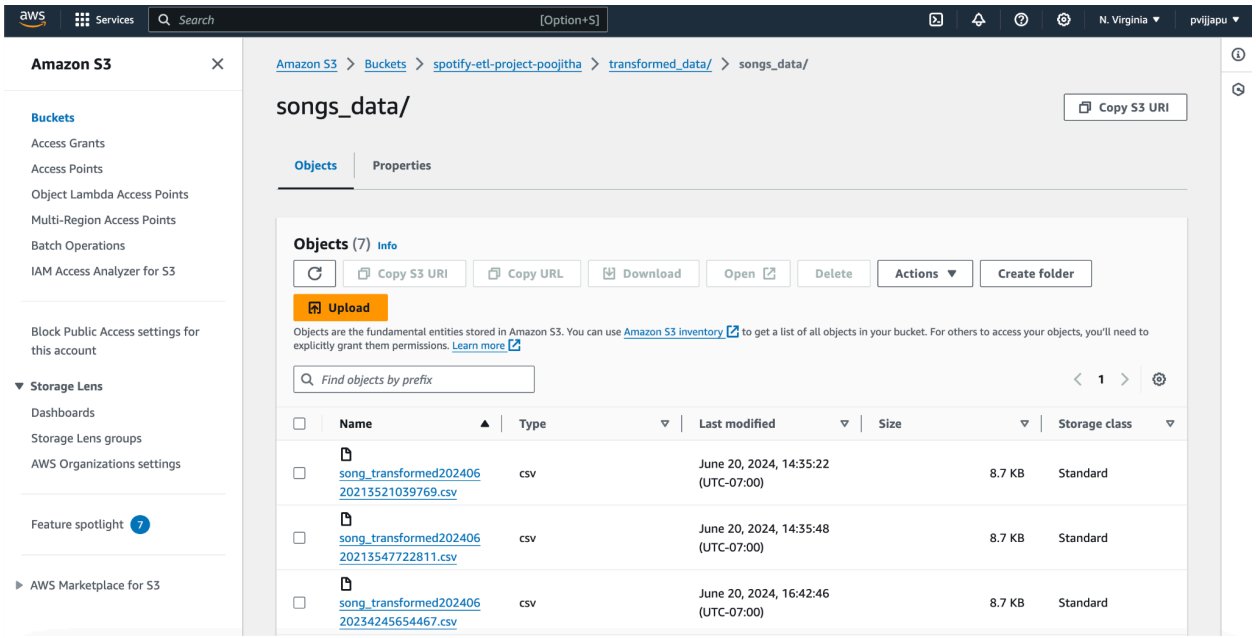
Triggers allow automation of the ETL pipeline to run at regular intervals. We utilized two triggers:

- *EventBridge Trigger*: Schedules the `spotify_api_extract` function to run daily, ensuring fresh data is collected regularly.
- *S3 Trigger*: Configured to run the `spotify_load_transform_function` whenever new data is added to the `to_be_processed` folder, ensuring timely processing of new data.

Transformed and Loaded Data :



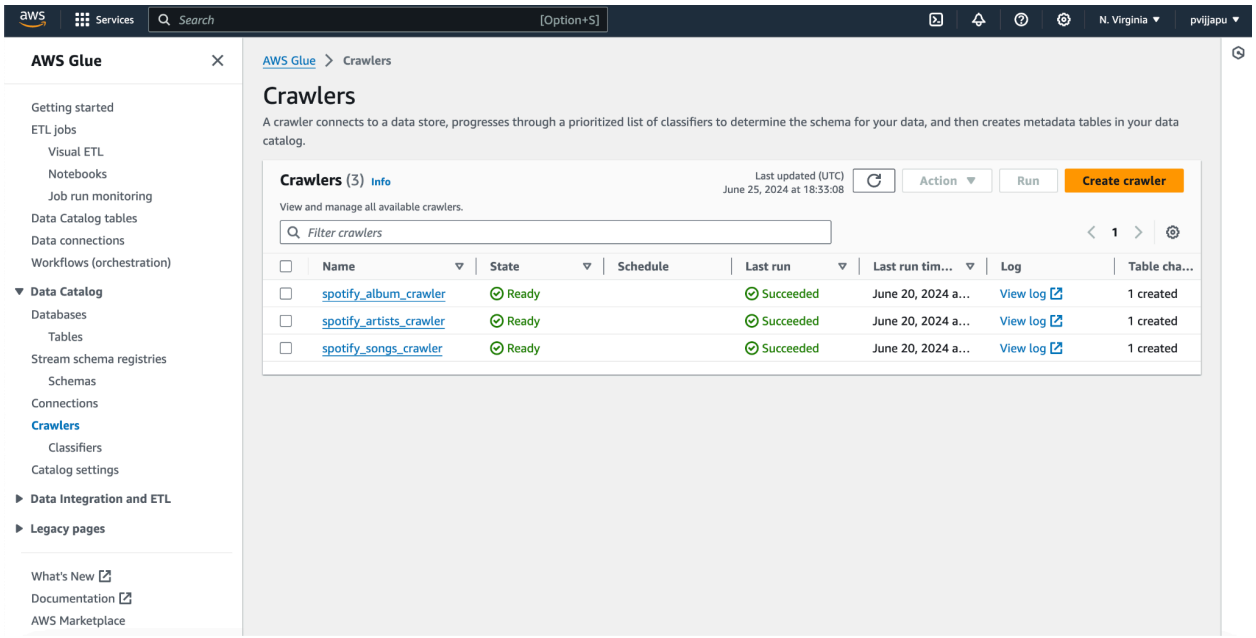

Services
Search
[Option+S]
N. Virginia
pvijjapu
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets > spotify-etl-project-poojitha > transformed_data/ > artist_data/
artist_data/
Copy S3 URI
Objects
Properties
Objects (7) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
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
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
artist_transformed20240620213521335292.csv
csv
June 20, 2024, 14:35:22 (UTC-07:00)
4.7 KB
Standard
artist_transformed20240620213547905082.csv
csv
June 20, 2024, 14:35:48 (UTC-07:00)
4.7 KB
Standard
artist_transformed20240620234246009872.csv
csv
June 20, 2024, 16:42:47 (UTC-07:00)
4.7 KB
Standard

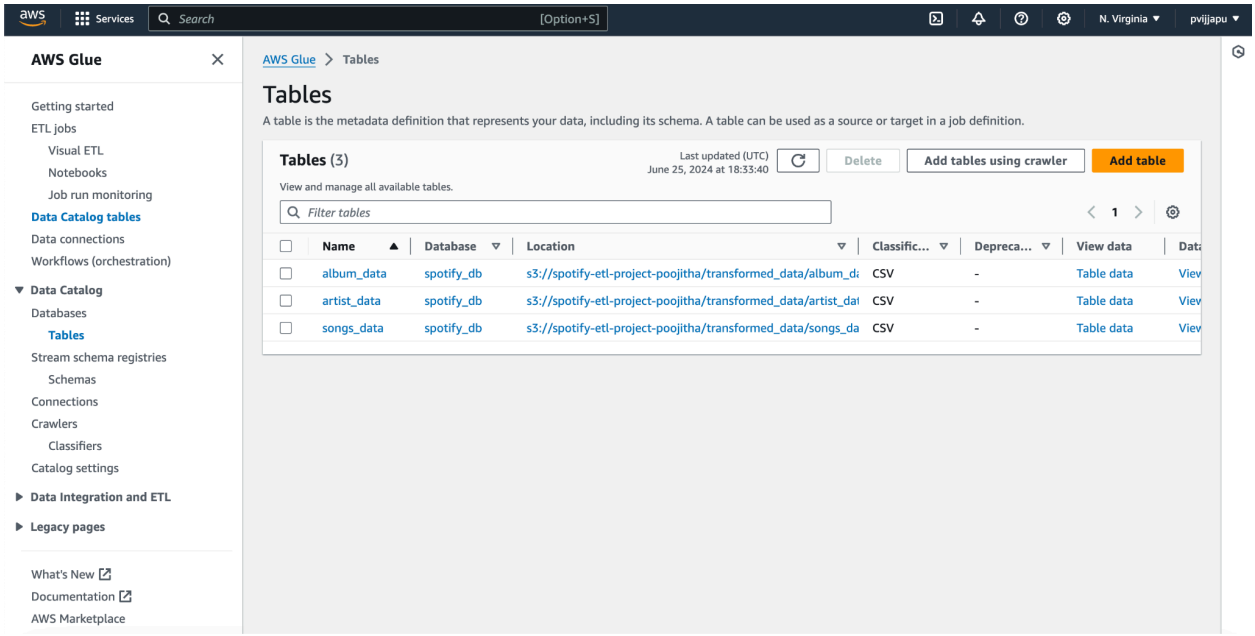

Services
Search
[Option+S]
N. Virginia
pvijjapu
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets > spotify-etl-project-poojitha > transformed_data/ > songs_data/
songs_data/
Copy S3 URI
Objects
Properties
Objects (7) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
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
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
song_transformed20240620213521039769.csv
csv
June 20, 2024, 14:35:22 (UTC-07:00)
8.7 KB
Standard
song_transformed20240620213547722811.csv
csv
June 20, 2024, 14:35:48 (UTC-07:00)
8.7 KB
Standard
song_transformed20240620234245654467.csv
csv
June 20, 2024, 16:42:46 (UTC-07:00)
8.7 KB
Standard

4. AWS Glue Crawlers

AWS Glue is a fully managed ETL service that makes it easy to prepare and load data for analytics.

- *Album Crawler:* Scans the `albums_data` folder and updates the AWS Glue Data Catalog with the latest schema and data.
- *Artist Crawler:* Scans the `artists_data` folder to ensure the Glue Data Catalog reflects the current structure and content.
- *Song Crawler:* Updates the Glue Data Catalog with the schema and data from the `songs_data` folder, keeping it ready for querying.





5. Using AWS Athena

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena queries the data stored in S3 using the metadata cataloged by AWS Glue. This allows for querying and analysis of the Spotify data.