MILESTONE 5

PROJECT 2

GROUP – 9

TOPIC:

SPOTIFY MUSIC STREAMING TREND ANALYSIS

MEMBERS:

KAVYA PATI BANDLA: Patibandla.ka@northeastern.edu

LIKHITHA GUTTAPALLI: guttapalli.l@northeastern.edu

SPOTIFY MUSIC TREND ANALYTICS

The end-to-end ETL pipeline built for the Spotify music trend analytics is explained in this report. The pipeline processes raw data from the S3 bucket, transforms it into dimension and fact tables using Amazon glue, and then Amazon Athena is used for analytics

Dataset Information:

- Source: Kaggle
- Dataset link: Spotify Popularity Prediction-ML Practice
- Format:
- Song_data: contains detailed information regarding the audio features of the song
  - Contains 18836 rows and 10 columns of different data types
- Song_info: the information regarding the song characteristics such as the artist, album, playlist etc.
  - Contains 11836 rows and 4 columns of different data types
- Generated the song_history.json file for details regarding the number of streams and chart positions.

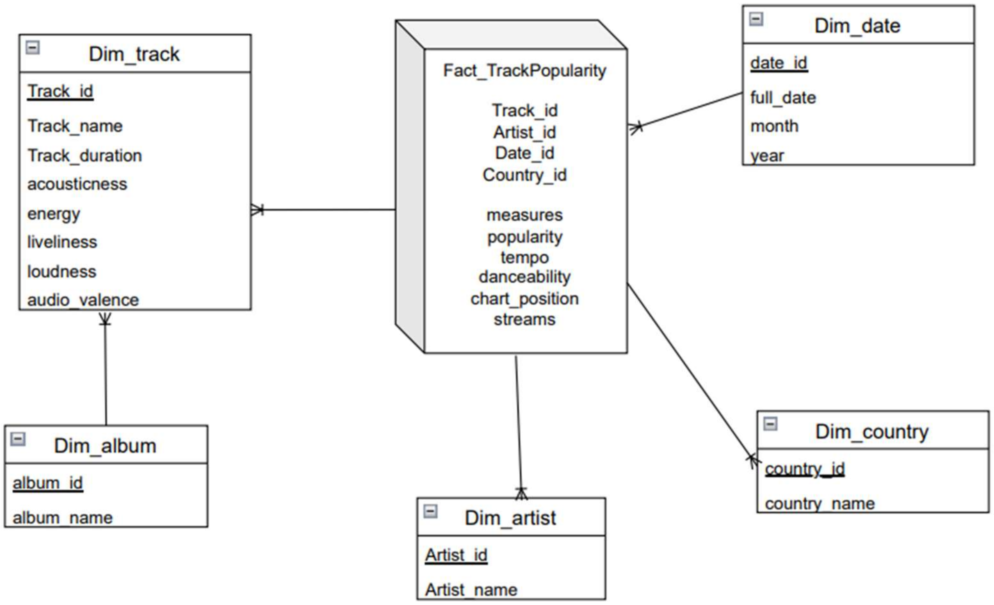
Dimension Tables:

- Dim_track: Contains information regarding the track such as track_id, track name, duration, energy, danceability, liveliness..
- Dim_artist: Contains the artist information such as the artist_id and artist name
- Dim_album: Contains the information regarding the album id and album name
- Dim_date: Contains the fulldate, month and year for further analysis
- Dim_country: contains the country_id and country name

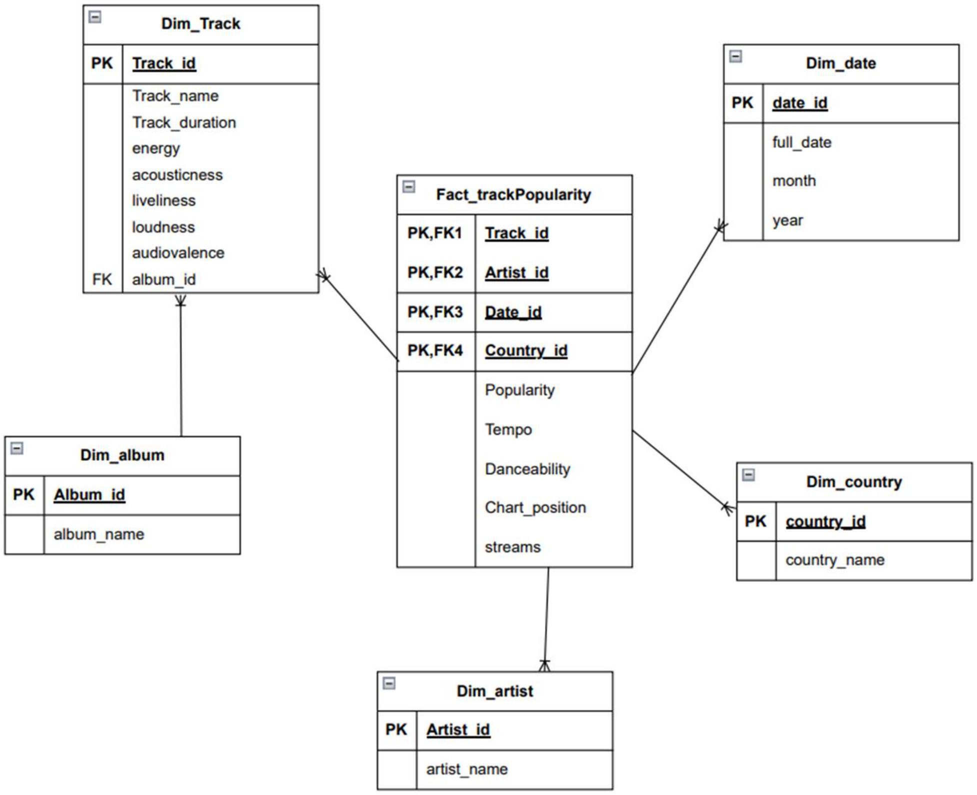
Fact Table:

- Fact_TrackPopularity: Contains the keys of the dimension table along with measures such as chart position, popularity and number of streams

Conceptual model



Logical Model

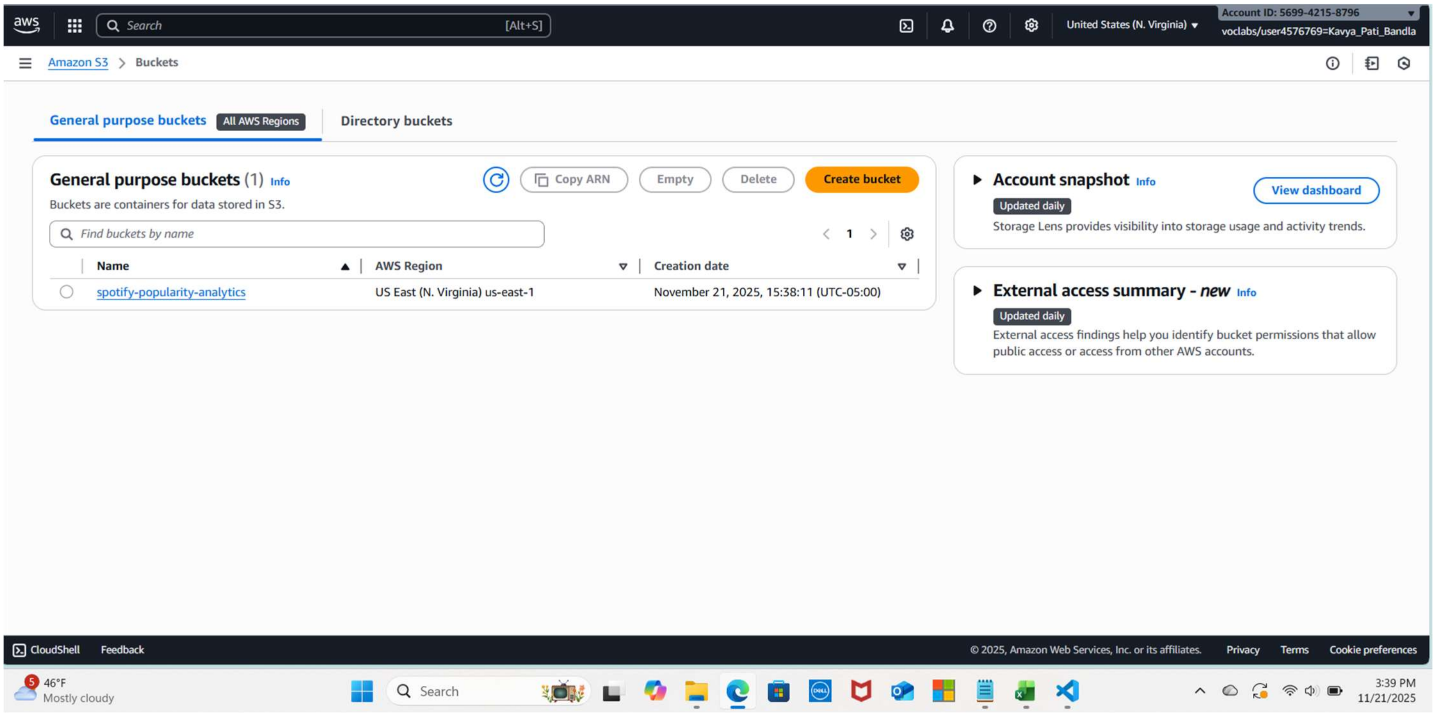


ETL PIPELINE

- **Storage layer**
  - Amazon S3 bucket:
  - Raw data path:
  - Processed data path:
- **Processing layer**
  - AWS Lambda Trigger: Activates Glue ETL jobs when all required files are present. Sequential job execution (dimension tables first, fact tables second)
  - AWS Glue ETL Jobs: Creates dimension tables from raw data and creates fact tables referencing dimension tables
- **Data Crawling**
  - Processed-data crawler: Scans the processed data files to create catalog
  - Raw-data crawler: Scans the raw data files to create catalog tables
- **Serving layer**
  - AWS Athena is used as a SQL query interface for the analytics. It enables direct querying of the processed data

Data Extraction:

- An S3 bucket is created for storing data.
- Manually uploaded the csv and the json files to the raw data folder created in the bucket
- A lambda function will be enabled on the bucket, once the files are uploaded it will trigger the ETL jobs



General purpose buckets (1) Info

Buckets are containers for data stored in S3.

	Name	AWS Region	Creation date
○	spotify-popularity-analytics	US East (N. Virginia) us-east-1	November 21, 2025, 15:38:11 (UTC-05:00)

Account snapshot Info

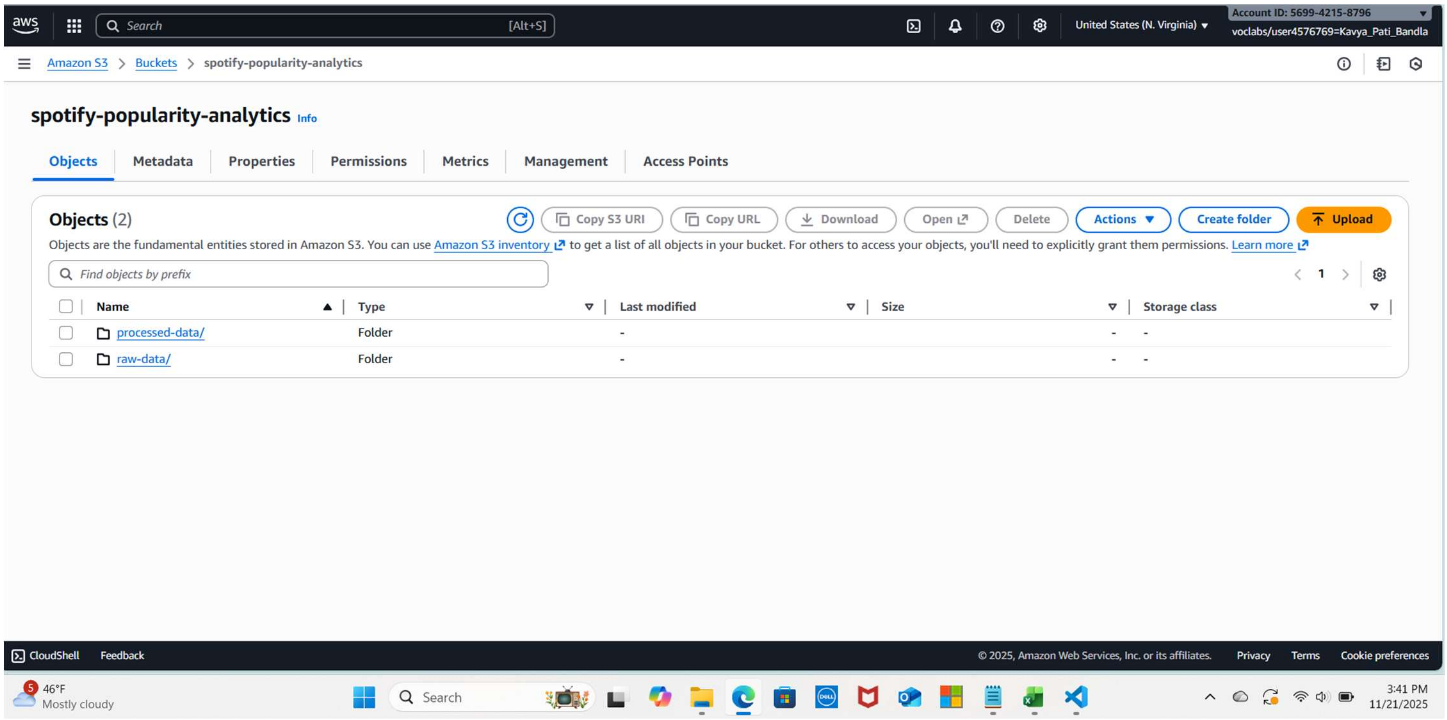
Updated daily

Storage Lens provides visibility into storage usage and activity trends.

External access summary - *new* Info

Updated daily

External access findings help you identify bucket permissions that allow public access or access from other AWS accounts.

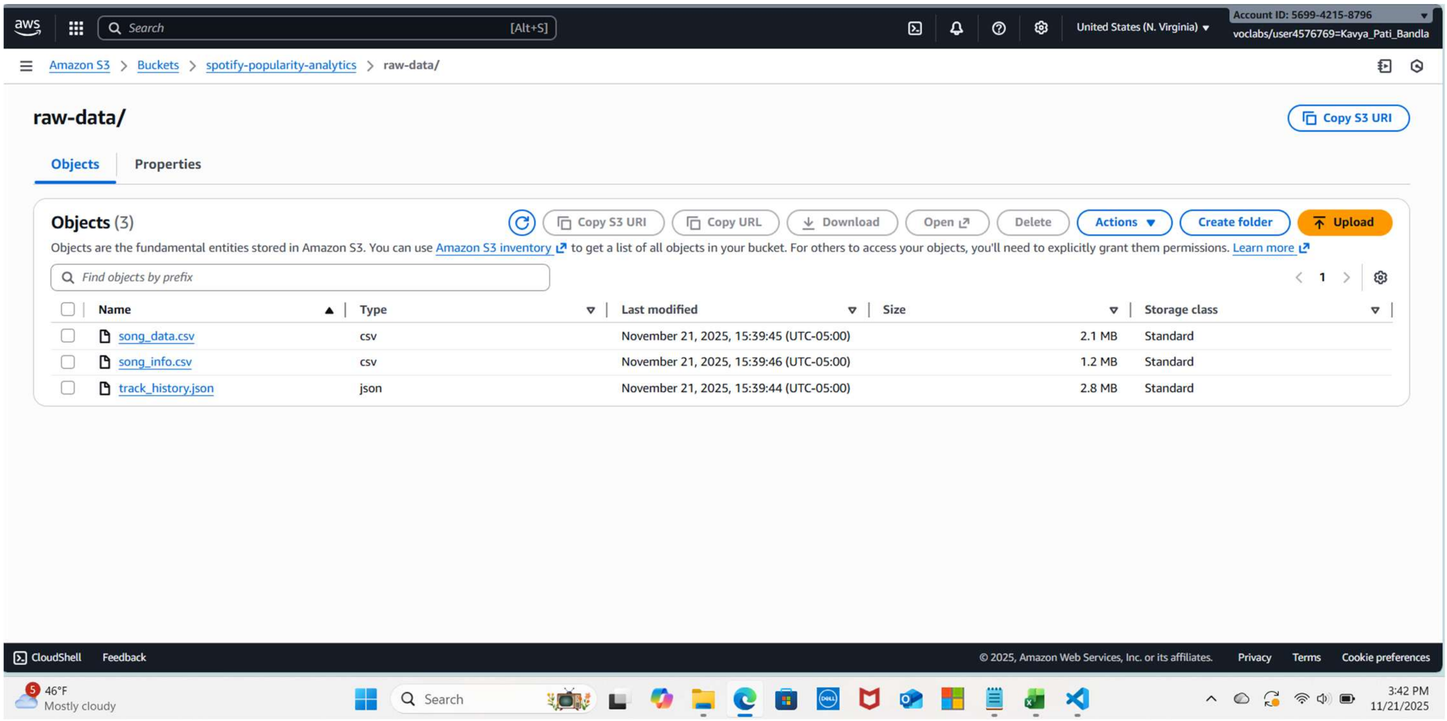


spotify-popularity-analytics Info

Objects (2)

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

□	Name	Type	Last modified	Size	Storage class
□	processed-data/	Folder	-	-	-
□	raw-data/	Folder	-	-	-



Amazon S3 > Buckets > spotify-popularity-analytics > raw-data/

raw-data/

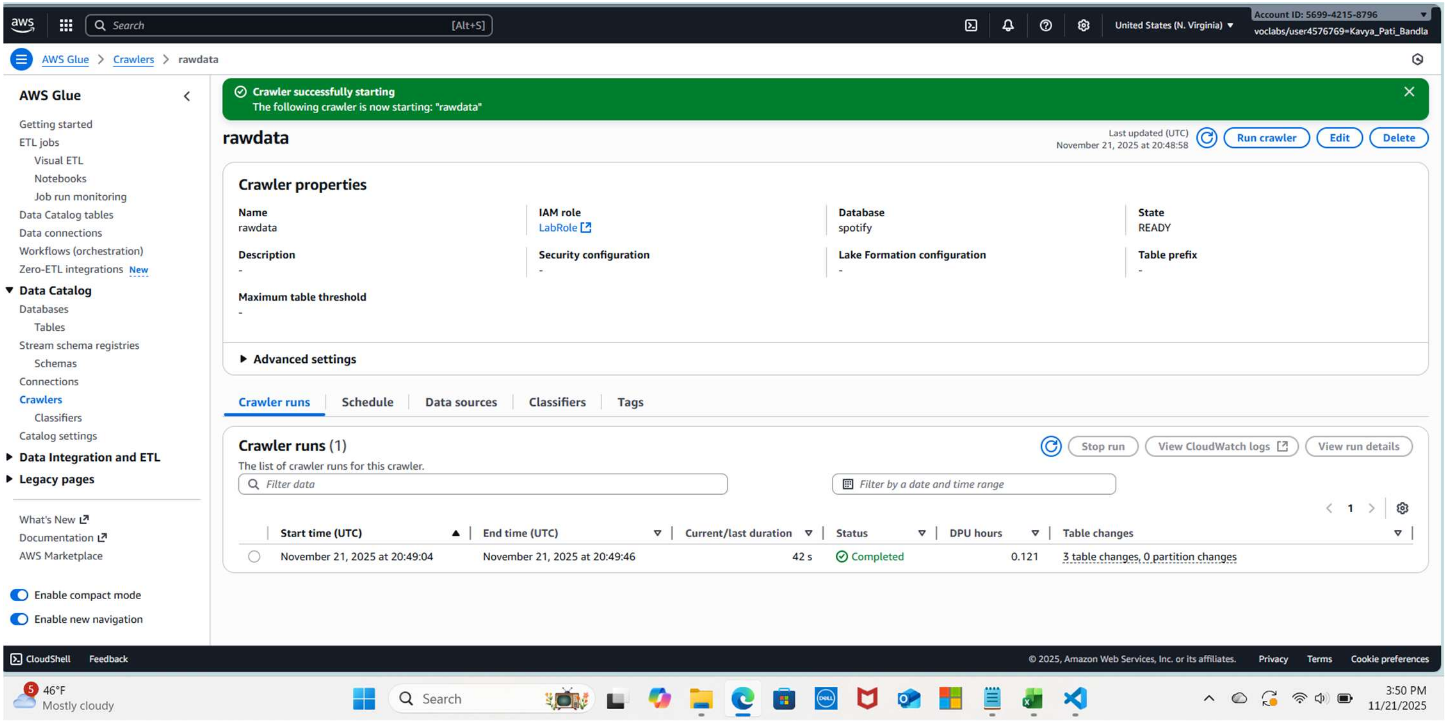
Objects (3)

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

Name	Type	Last modified	Size	Storage class
song_data.csv	csv	November 21, 2025, 15:39:45 (UTC-05:00)	2.1 MB	Standard
song_info.csv	csv	November 21, 2025, 15:39:46 (UTC-05:00)	1.2 MB	Standard
track_history.json	json	November 21, 2025, 15:39:44 (UTC-05:00)	2.8 MB	Standard

Data Crawling

- The datasets in the raw data are made available in the data catalogue using the AWS Glue Crawler. Which crawls all the sub folders present in the raw data for data.



AWS Glue > Crawlers > rawdata

Crawler successfully starting
The following crawler is now starting: "rawdata"

rawdata

Last updated (UTC)
November 21, 2025 at 20:48:58

Crawler properties

Name	IAM role	Database	State
rawdata	LabRole	spotify	READY

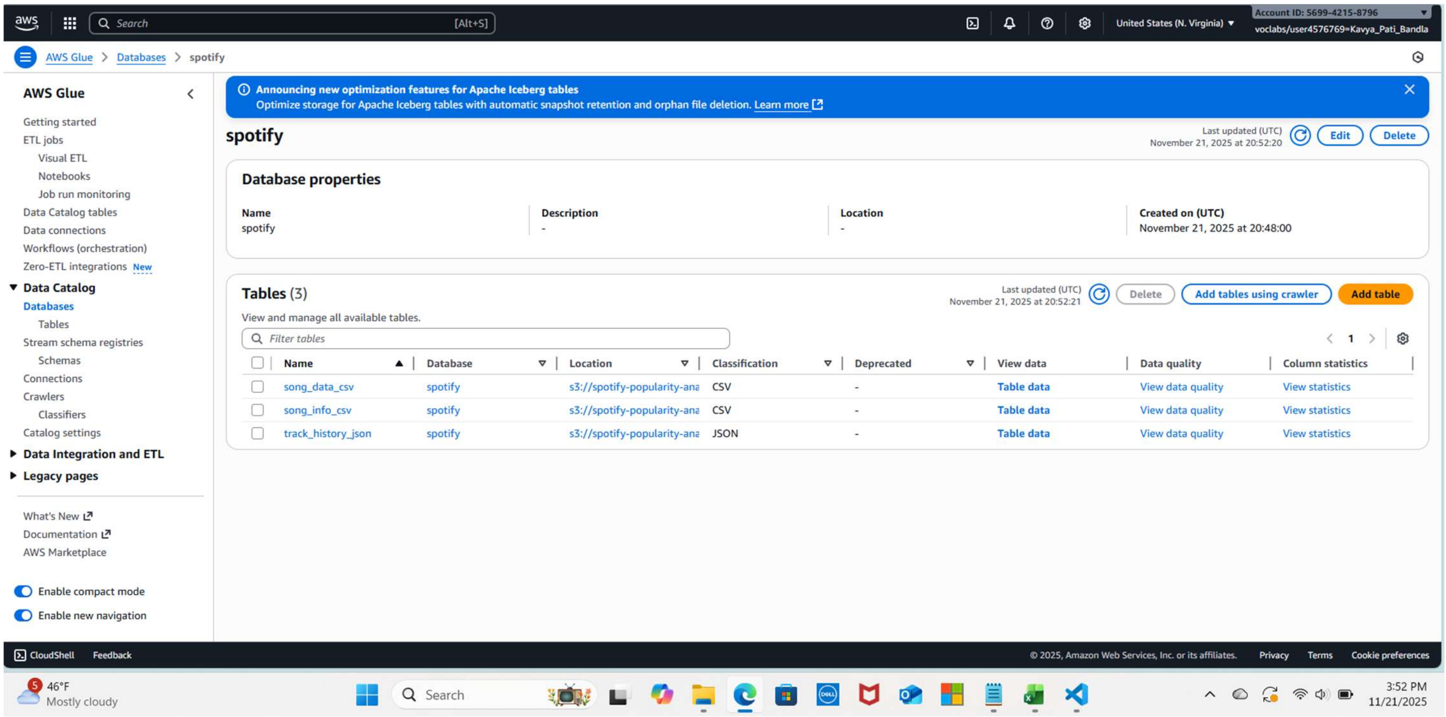
Description	Security configuration	Lake Formation configuration	Table prefix
-	-	-	-

Maximum table threshold
-

Crawler runs (1)

The list of crawler runs for this crawler.

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
November 21, 2025 at 20:49:04	November 21, 2025 at 20:49:46	42 s	Completed	0.121	3 table changes, 0 partition changes



Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. Learn more

spotify

Last updated (UTC)
November 21, 2025 at 20:52:20

Database properties

Name	Description	Location	Created on (UTC)
spotify	-	-	November 21, 2025 at 20:48:00

Tables (3)

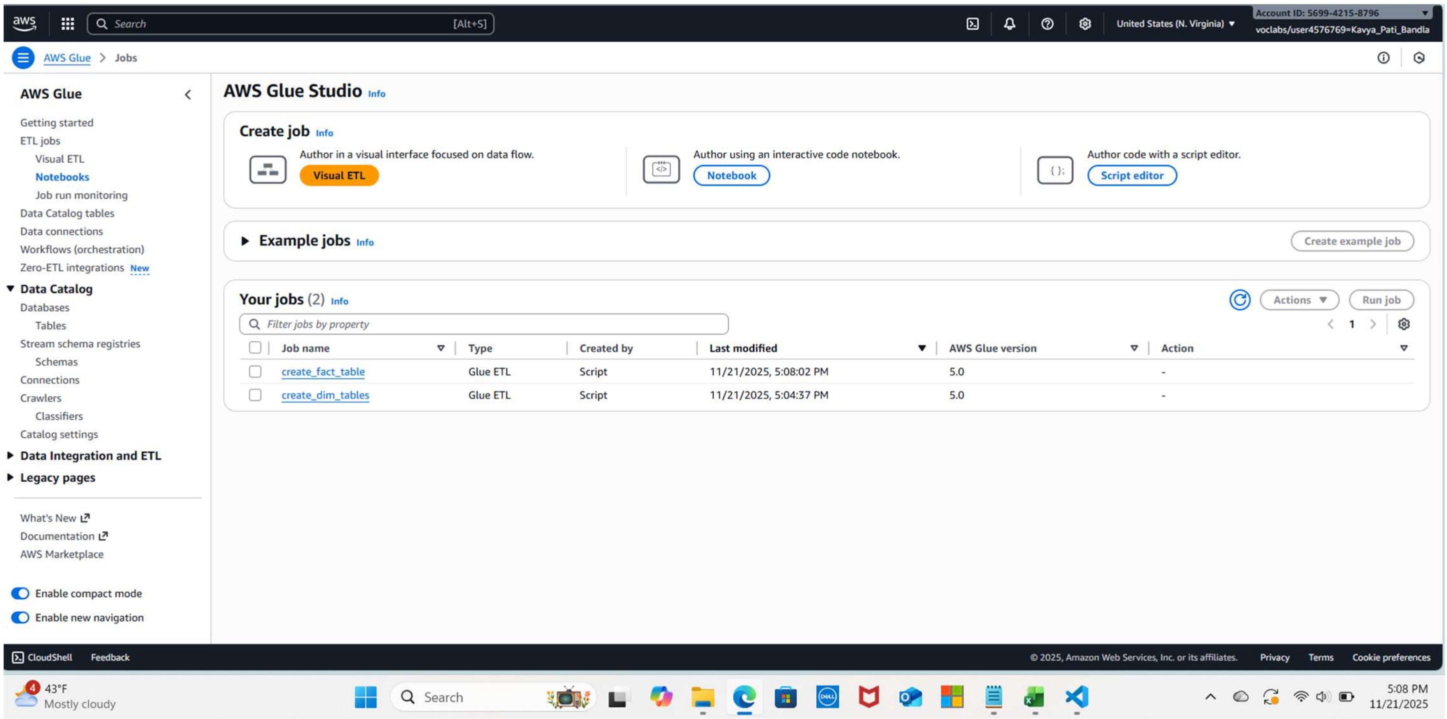
Last updated (UTC)
November 21, 2025 at 20:52:21

View and manage all available tables.

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
song_data_csv	spotify	s3://spotify-popularity-ana	CSV	-	Table data	View data quality	View statistics
song_info_csv	spotify	s3://spotify-popularity-ana	CSV	-	Table data	View data quality	View statistics
track_history_json	spotify	s3://spotify-popularity-ana	JSON	-	Table data	View data quality	View statistics

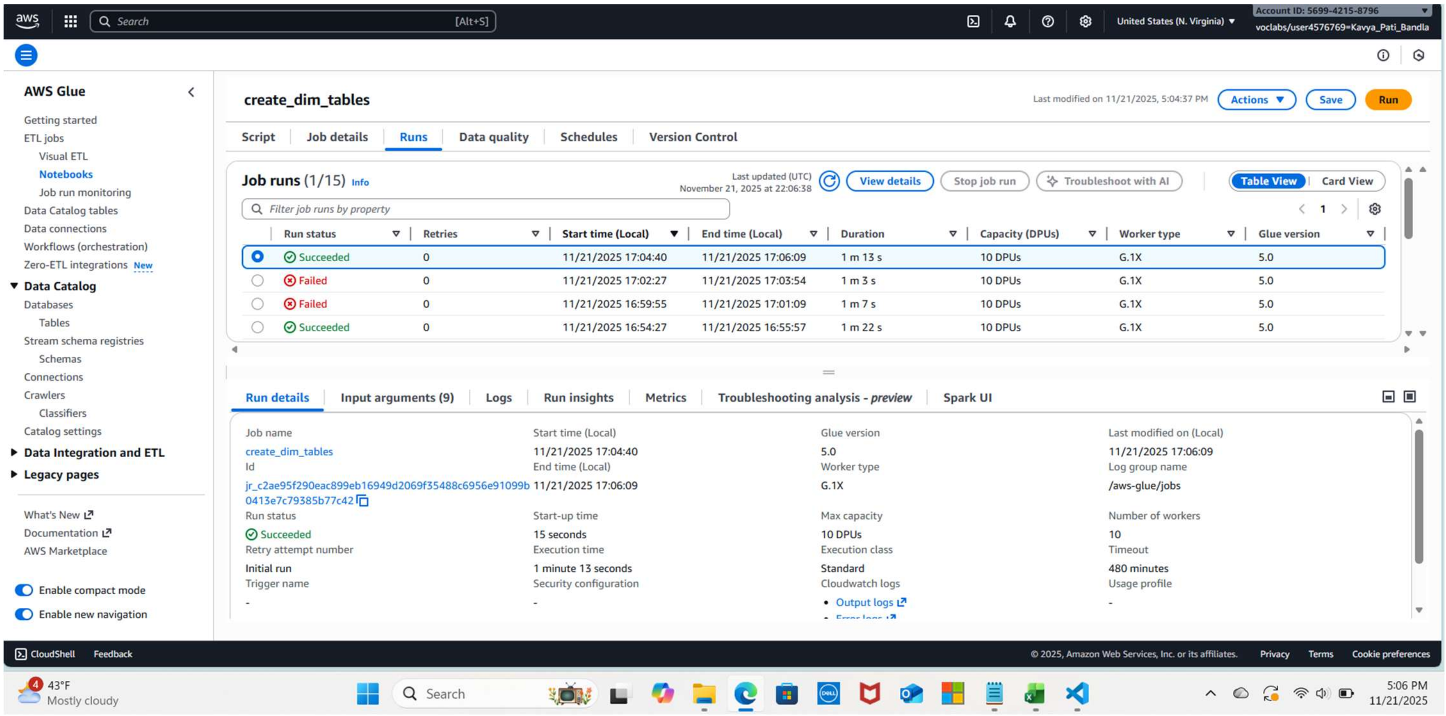
Data Transformation

- The AWS Glue python script is used to transform the raw data into processed data, that is the dimension tables and fact table
  - Dimsong
  - Dimartist
  - Dimalbum
  - Dimdate
  - Dimcountry
  - FactSongPopularity
- Created two scripts for the data transformation. Create_dim_tables and create_fact_table for the transformation of raw data to processed data.



Your jobs (2)

	Job name	Type	Created by	Last modified	AWS Glue version	Action
□	create_fact_table	Glue ETL	Script	11/21/2025, 5:08:02 PM	5.0	-
□	create_dim_tables	Glue ETL	Script	11/21/2025, 5:04:37 PM	5.0	-



create_dim_tables

Last modified on 11/21/2025, 5:04:37 PM

Job runs (1/15)

Last updated (UTC)
November 21, 2025 at 22:06:38

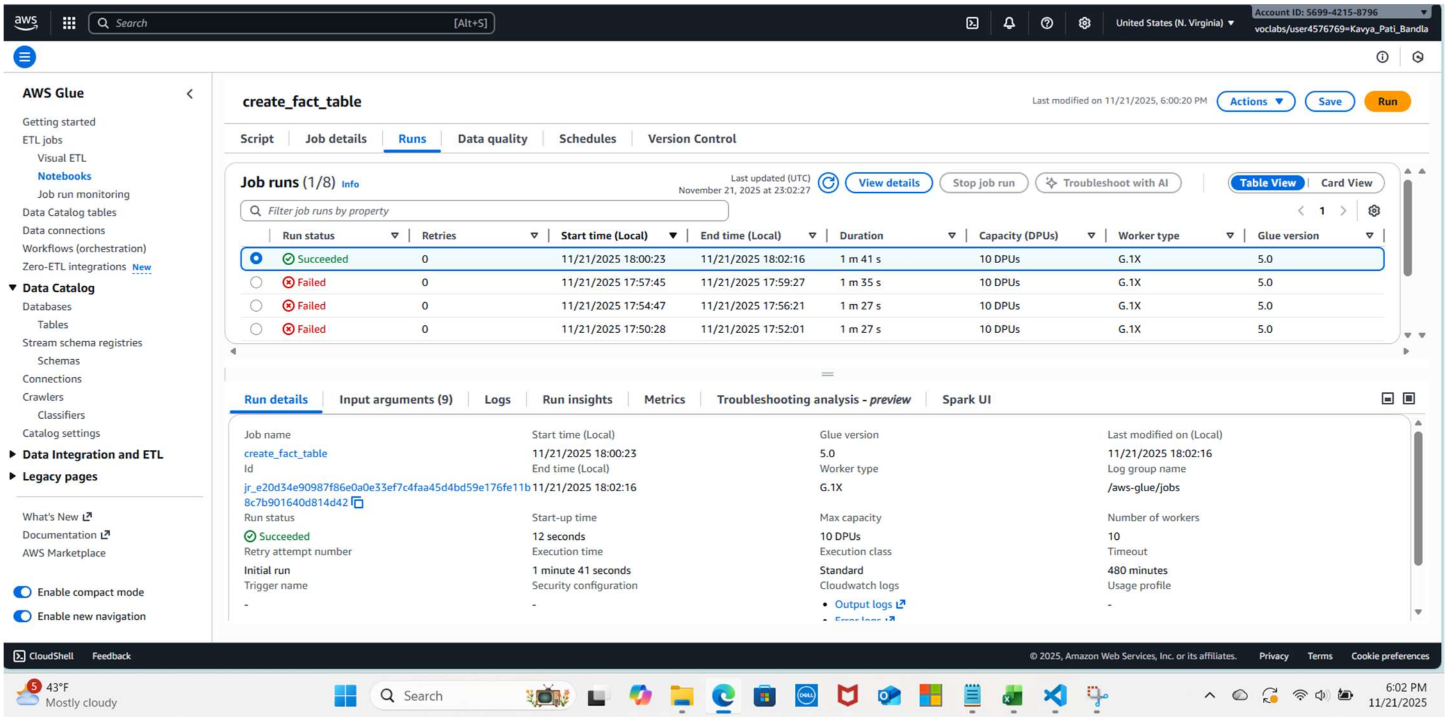
	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
●	Succeeded	0	11/21/2025 17:04:40	11/21/2025 17:06:09	1 m 13 s	10 DPUs	G.1X	5.0
○	Failed	0	11/21/2025 17:02:27	11/21/2025 17:03:54	1 m 3 s	10 DPUs	G.1X	5.0
○	Failed	0	11/21/2025 16:59:55	11/21/2025 17:01:09	1 m 7 s	10 DPUs	G.1X	5.0
○	Succeeded	0	11/21/2025 16:54:27	11/21/2025 16:55:57	1 m 22 s	10 DPUs	G.1X	5.0

Job name
create_dim_tables
Id
jr_c2ae95f290eac899eb16949d2069f35488c6956e91099b0413e7c79385b77c42
Run status
Succeeded
Retry attempt number
Initial run
Trigger name
-

Start time (Local)
11/21/2025 17:04:40
End time (Local)
11/21/2025 17:06:09
Start-up time
15 seconds
Execution time
1 minute 13 seconds
Security configuration
-

Glue version
5.0
Worker type
G.1X
Max capacity
10 DPUs
Execution class
Standard
Cloudwatch logs
- Output logs
- Error logs

Last modified on (Local)
11/21/2025 17:06:09
Log group name
/aws-glue/jobs
Number of workers
10
Timeout
480 minutes
Usage profile
-



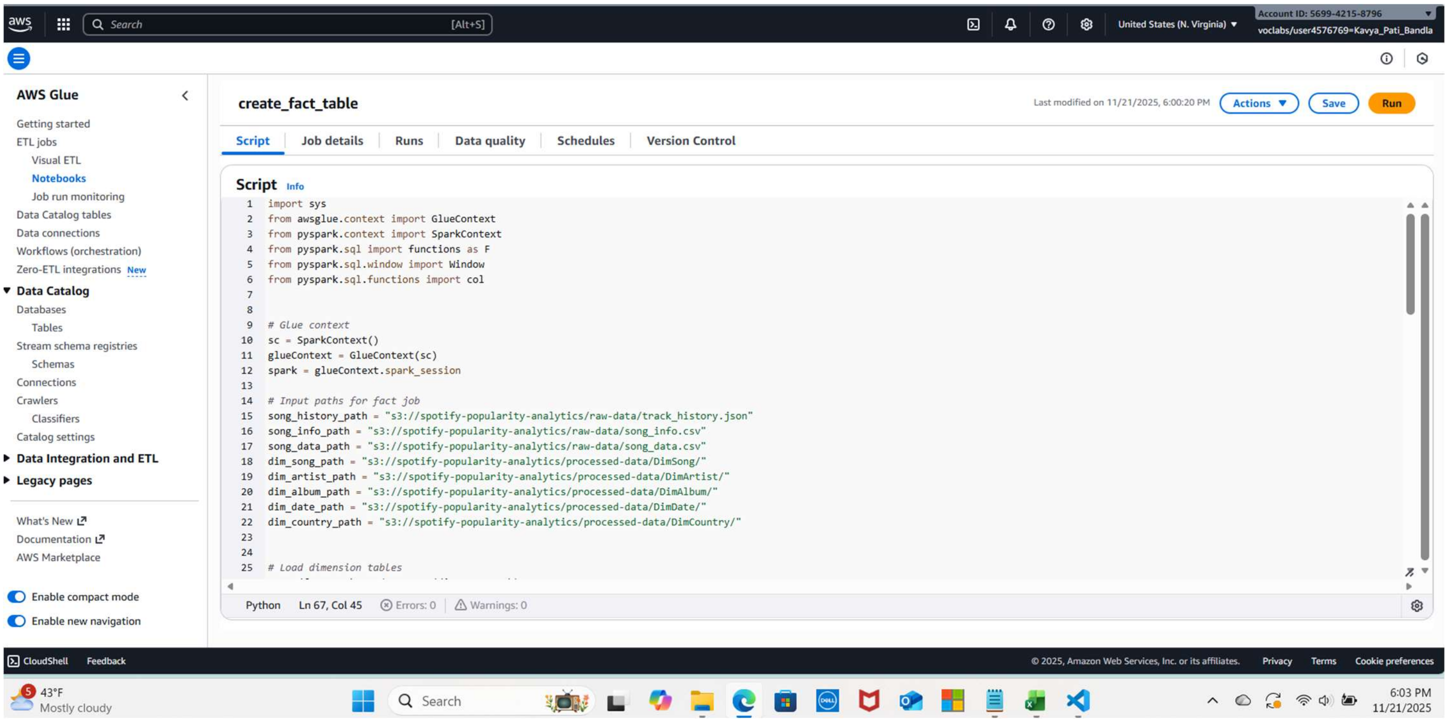
create_fact_table

Job runs (1/8)

	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
●	✓ Succeeded	0	11/21/2025 18:00:23	11/21/2025 18:02:16	1 m 41 s	10 DPUs	G.1X	5.0
○	⊗ Failed	0	11/21/2025 17:57:45	11/21/2025 17:59:27	1 m 35 s	10 DPUs	G.1X	5.0
○	⊗ Failed	0	11/21/2025 17:54:47	11/21/2025 17:56:21	1 m 27 s	10 DPUs	G.1X	5.0
○	⊗ Failed	0	11/21/2025 17:50:28	11/21/2025 17:52:01	1 m 27 s	10 DPUs	G.1X	5.0

Run details | Input arguments (9) | Logs | Run insights | Metrics | Troubleshooting analysis - *preview* | Spark UI

Job name	Start time (Local)	Glue version	Last modified on (Local)
create_fact_table	11/21/2025 18:00:23	5.0	11/21/2025 18:02:16
Id	End time (Local)	Worker type	Log group name
jr_e20d34e90987f86e0a0e33ef7c4faa45d4bd59e176fe11b8c7b901640d814d42	11/21/2025 18:02:16	G.1X	/aws-glue/jobs
Run status	Start-up time	Max capacity	Number of workers
✓ Succeeded	12 seconds	10 DPUs	10
Retry attempt number	Execution time	Execution class	Timeout
Initial run	1 minute 41 seconds	Standard	480 minutes
Trigger name	Security configuration	Cloudwatch logs	Usage profile
-	-	Output logs Error logs	-



create_fact_table

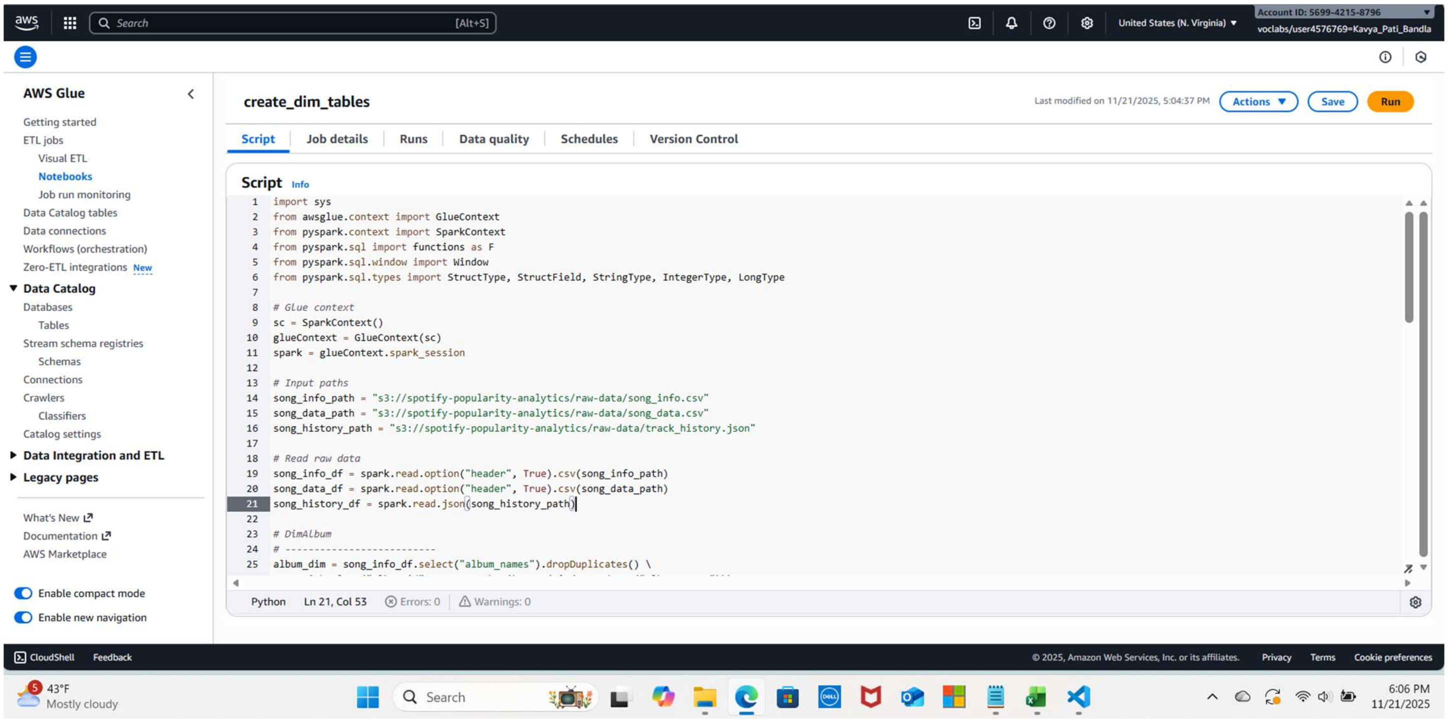
Script

```python
import sys
from awsglue.context import GlueContext
from pyspark.context import SparkContext
from pyspark.sql import functions as F
from pyspark.sql.window import Window
from pyspark.sql.functions import col


# Glue context
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

# Input paths for fact job
song_history_path = "s3://spotify-popularity-analytics/raw-data/track_history.json"
song_info_path = "s3://spotify-popularity-analytics/raw-data/song_info.csv"
song_data_path = "s3://spotify-popularity-analytics/raw-data/song_data.csv"
dim_song_path = "s3://spotify-popularity-analytics/processed-data/DimSong/"
dim_artist_path = "s3://spotify-popularity-analytics/processed-data/DimArtist/"
dim_album_path = "s3://spotify-popularity-analytics/processed-data/DimAlbum/"
dim_date_path = "s3://spotify-popularity-analytics/processed-data/DimDate/"
dim_country_path = "s3://spotify-popularity-analytics/processed-data/DimCountry/"


# Load dimension tables
```



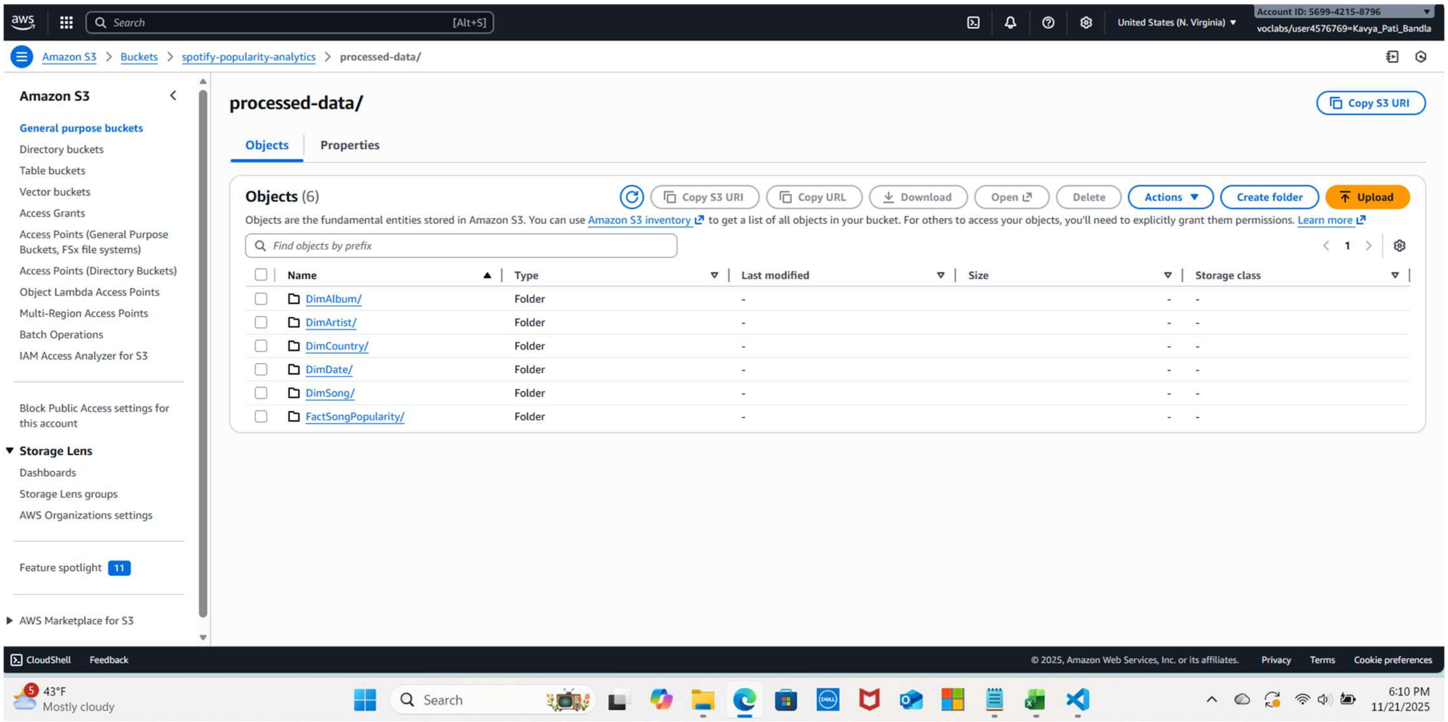
```
import sys
from awsglue.context import GlueContext
from pyspark.context import SparkContext
from pyspark.sql import functions as F
from pyspark.sql.window import Window
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, LongType

# Glue context
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

# Input paths
song_info_path = "s3://spotify-popularity-analytics/raw-data/song_info.csv"
song_data_path = "s3://spotify-popularity-analytics/raw-data/song_data.csv"
song_history_path = "s3://spotify-popularity-analytics/raw-data/track_history.json"

# Read raw data
song_info_df = spark.read.option("header", True).csv(song_info_path)
song_data_df = spark.read.option("header", True).csv(song_data_path)
song_history_df = spark.read.json(song_history_path)

# DimAlbum
# -------------------------
album_dim = song_info_df.select("album_names").dropDuplicates() \
```

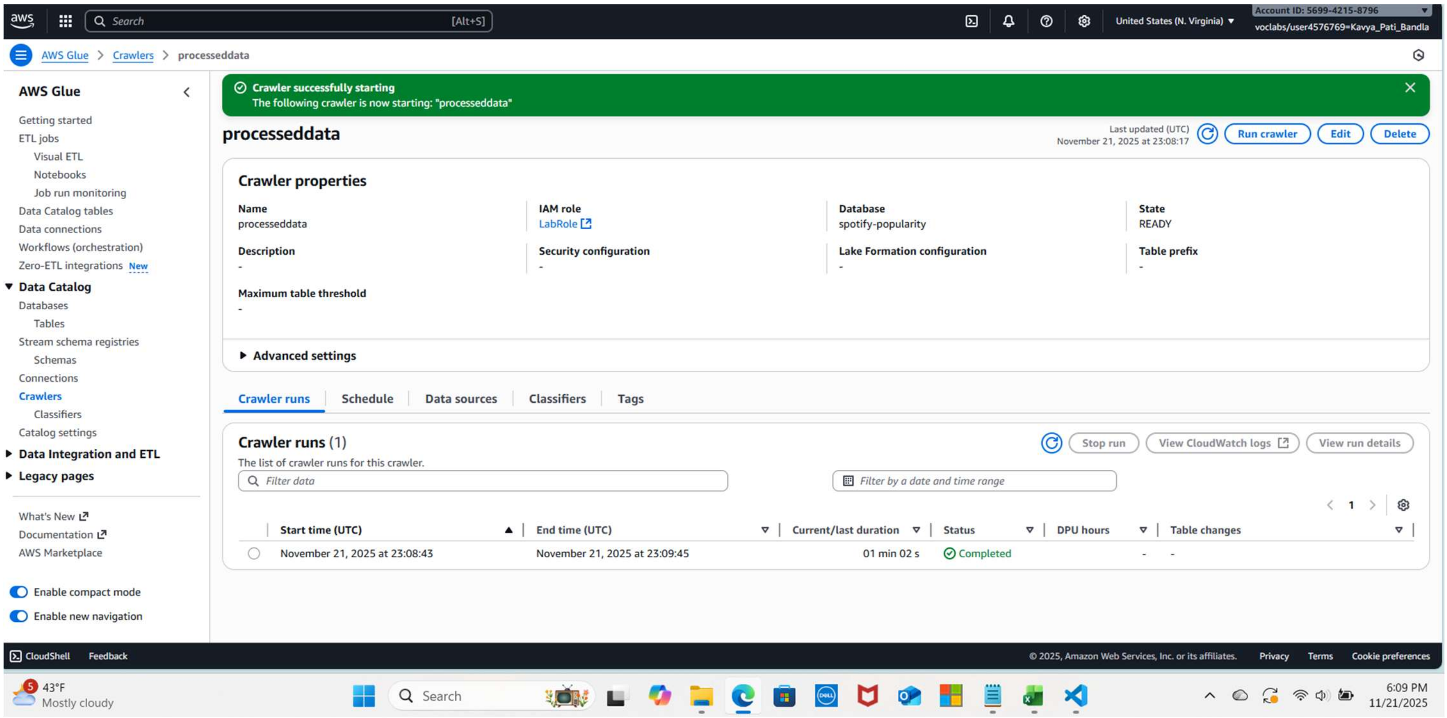


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
DimAlbum/	Folder	-	-	-
DimArtist/	Folder	-	-	-
DimCountry/	Folder	-	-	-
DimDate/	Folder	-	-	-
DimSong/	Folder	-	-	-
FactSongPopularity/	Folder	-	-	-

Data Catalog

- A glue crawler is used to make the data in the processed S3 bucket available in the Data Catalog by scanning all the subfolders within the specified bucket.



Crawler successfully starting
The following crawler is now starting: "processeddata"

processeddata

Last updated (UTC)
November 21, 2025 at 23:08:17

Crawler properties

Name	IAM role	Database	State
processeddata	LabRole	spotify-popularity	READY

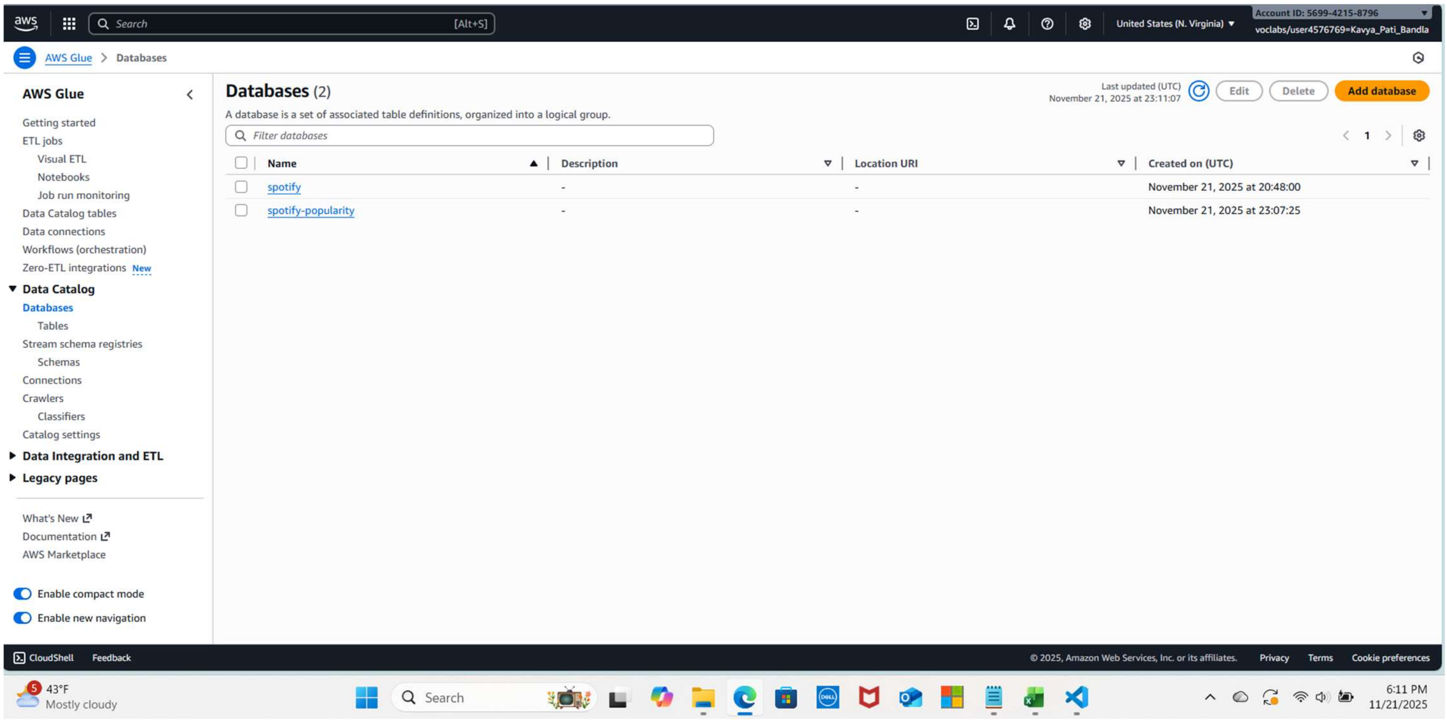
Description	Security configuration	Lake Formation configuration	Table prefix
-	-	-	-

Maximum table threshold
-

Crawler runs (1)

The list of crawler runs for this crawler.

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	November 21, 2025 at 23:08:43	November 21, 2025 at 23:09:45	01 min 02 s	✓ Completed	-	-

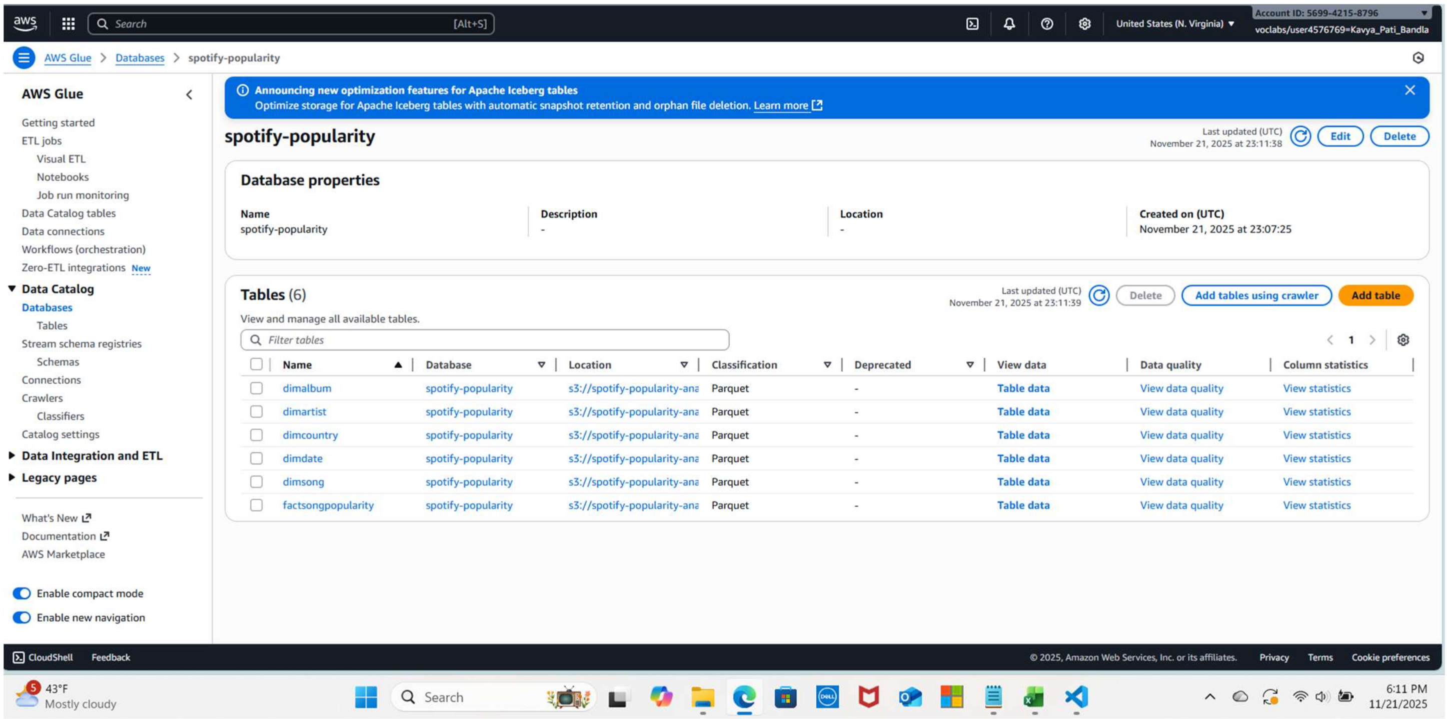


Databases (2)

Last updated (UTC)
November 21, 2025 at 23:11:07

A database is a set of associated table definitions, organized into a logical group.

□	Name	Description	Location URI	Created on (UTC)
□	spotify	-	-	November 21, 2025 at 20:48:00
□	spotify-popularity	-	-	November 21, 2025 at 23:07:25


AWS Glue > Databases > spotify-popularity
Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. Learn more

spotify-popularity

Last updated (UTC) November 21, 2025 at 23:11:38

Database properties

Name	Description	Location	Created on (UTC)
spotify-popularity	-	-	November 21, 2025 at 23:07:25

Tables (6)

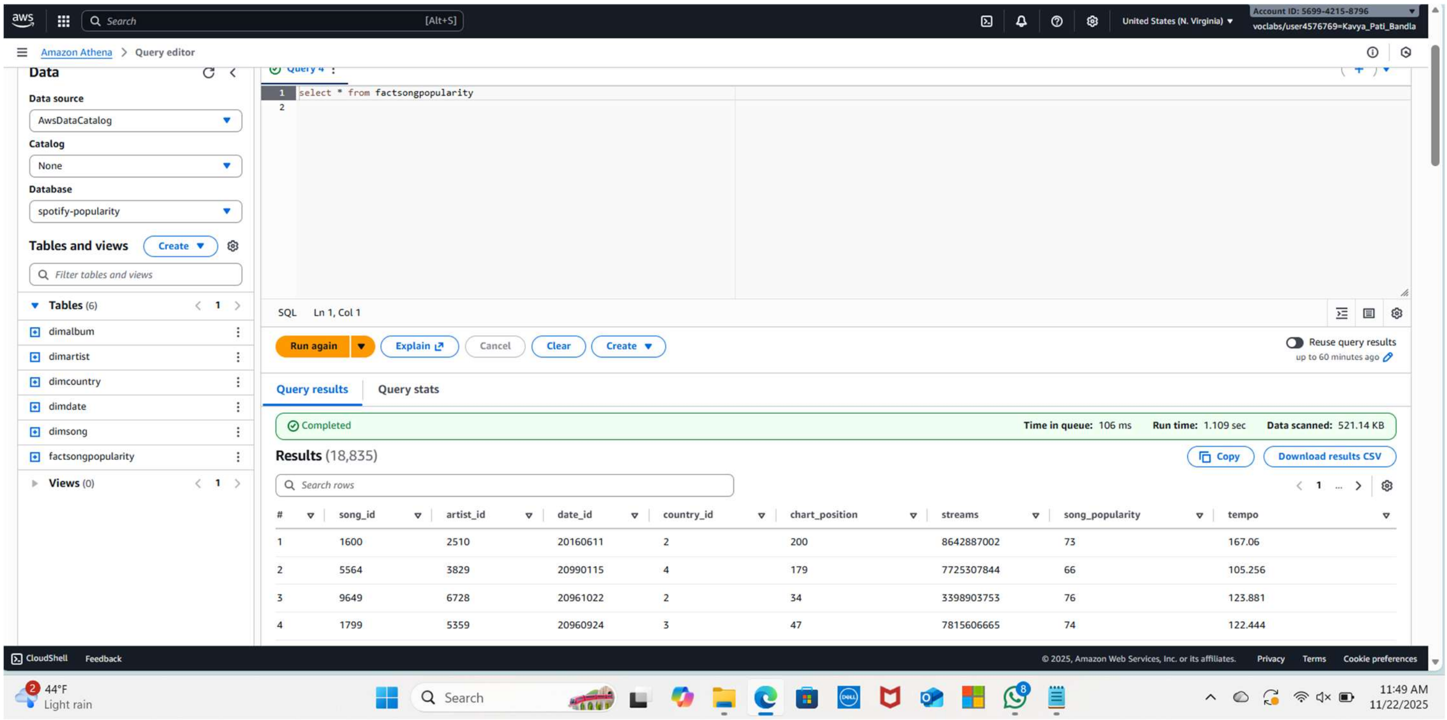
Last updated (UTC) November 21, 2025 at 23:11:39

View and manage all available tables.

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
dimalbum	spotify-popularity	s3://spotify-popularity-ana	Parquet	-	Table data	View data quality	View statistics
dimartist	spotify-popularity	s3://spotify-popularity-ana	Parquet	-	Table data	View data quality	View statistics
dimcountry	spotify-popularity	s3://spotify-popularity-ana	Parquet	-	Table data	View data quality	View statistics
dimdate	spotify-popularity	s3://spotify-popularity-ana	Parquet	-	Table data	View data quality	View statistics
dimsong	spotify-popularity	s3://spotify-popularity-ana	Parquet	-	Table data	View data quality	View statistics
factsongpopularity	spotify-popularity	s3://spotify-popularity-ana	Parquet	-	Table data	View data quality	View statistics

Data Loading:

- The target tables are then loaded into Athena for querying and for popularity analytics of the Spotify music



Data source: AwsDataCatalog

Catalog: None

Database: spotify-popularity

Tables (6)

- dimalbum
- dimartist
- dimcountry
- dimdate
- dimsong
- factsongpopularity

Views (0)

```
select * from factsongpopularity
```

Completed — Time in queue: 106 ms Run time: 1.109 sec Data scanned: 521.14 KB

Results (18,835)

#	song_id	artist_id	date_id	country_id	chart_position	streams	song_popularity	tempo
1	1600	2510	20160611	2	200	8642887002	73	167.06
2	5564	3829	20990115	4	179	7725307844	66	105.256
3	9649	6728	20961022	2	34	3398903753	76	123.881
4	1799	5359	20960924	3	47	7815606665	74	122.444