AWS-Based Spotify End-to-End Analysis Project

Project Overview

This project leverages various AWS services to create a data pipeline for analyzing Spotify data. Data is ingested into an S3 bucket, processed through AWS Glue, queried using AWS Athena, and visualized in Amazon QuickSight. This document outlines the steps taken to implement this solution, the technologies used, and the architecture followed.

Technologies Used

- **Amazon S3**: For storage of raw and processed data.
- **AWS Glue**: For ETL (Extract, Transform, Load) and cataloging data.
- **AWS Athena**: For querying the processed data.
- **Amazon QuickSight**: For data visualization and dashboard creation.

Step-by-Step Process

1. Create an IAM User

To ensure secure access to the AWS services, we first create an IAM (Identity and Access Management) user.

Steps:

- Log in to the AWS Management Console.
- Navigate to **IAM** and click on **Users**.
- Create a new user, providing necessary credentials.
- Assign the required roles and policies to allow access to the S3, Glue, Athena, and QuickSight services.

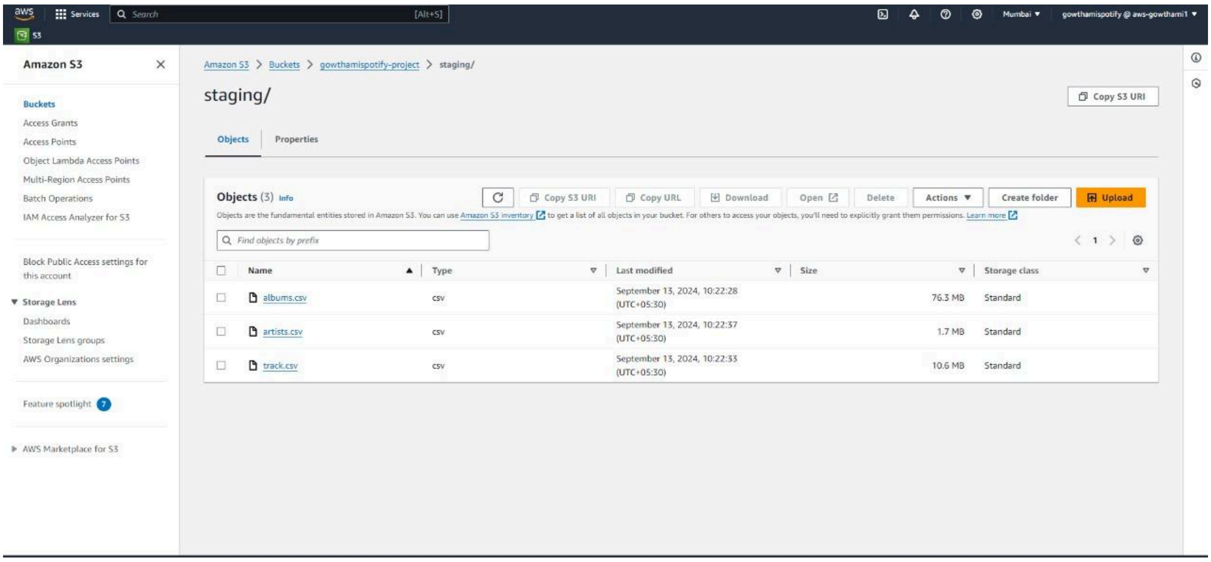
2. Create an S3 Bucket

S3 is used to store the raw Spotify data and processed data in different layers.

Steps:

- In the AWS Management Console, navigate to **S3**.
- Create a new bucket for storing data.
- Inside the bucket, create two folders (or "layers"):

    - **Staging**: Where raw data will be initially uploaded.
    - **Data Warehouse**: Where transformed and processed data will be stored after the ETL process.

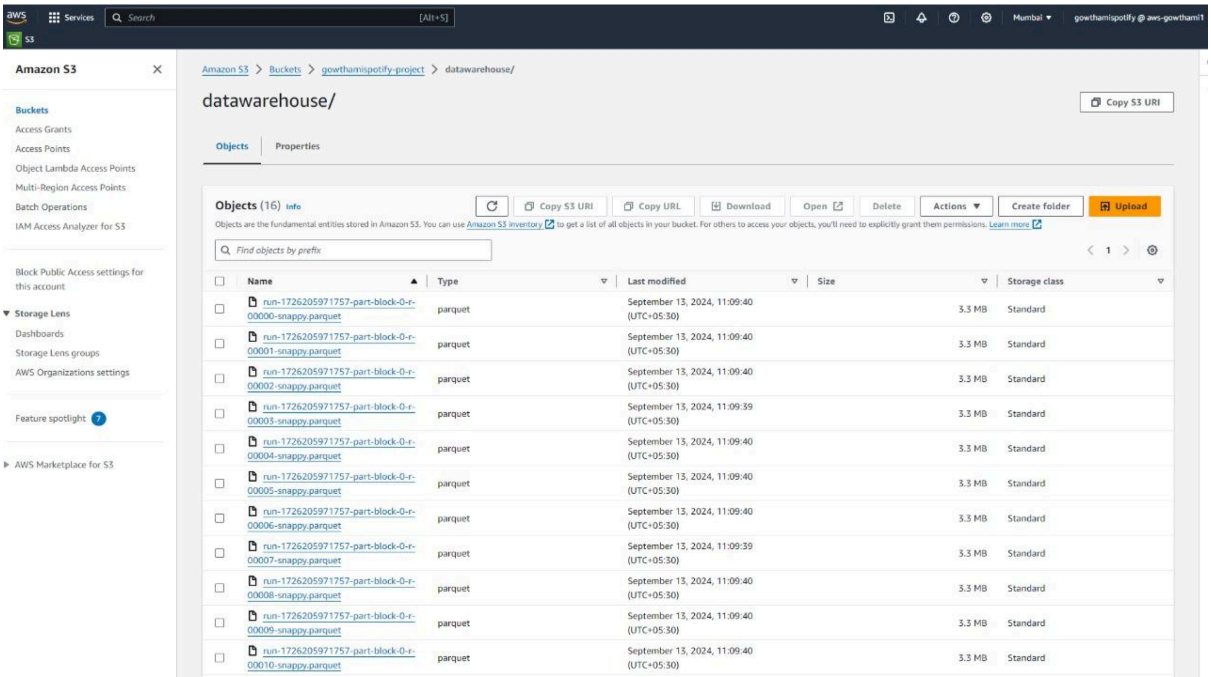


staging/

Objects (3) Info

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
□	albums.csv	csv	September 13, 2024, 10:22:28 (UTC+05:30)	76.3 MB	Standard
□	artists.csv	csv	September 13, 2024, 10:22:37 (UTC+05:30)	1.7 MB	Standard
□	track.csv	csv	September 13, 2024, 10:22:33 (UTC+05:30)	10.6 MB	Standard



datawarehouse/

Objects (16) Info

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
□	run-1726205971757-part-block-0-r-00000-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00001-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00002-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00003-snappy.parquet	parquet	September 13, 2024, 11:09:39 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00004-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00005-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00006-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00007-snappy.parquet	parquet	September 13, 2024, 11:09:39 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00008-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00009-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard
□	run-1726205971757-part-block-0-r-00010-snappy.parquet	parquet	September 13, 2024, 11:09:40 (UTC+05:30)	3.3 MB	Standard

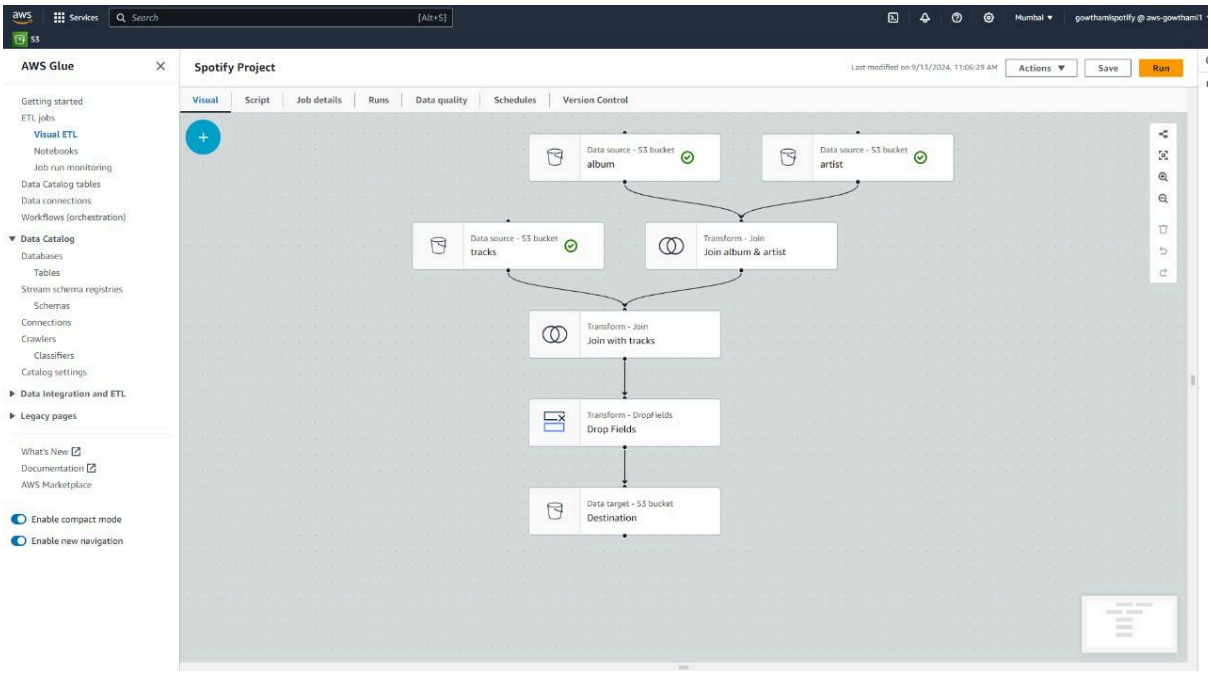
3. Create a Data Pipeline Using AWS Glue

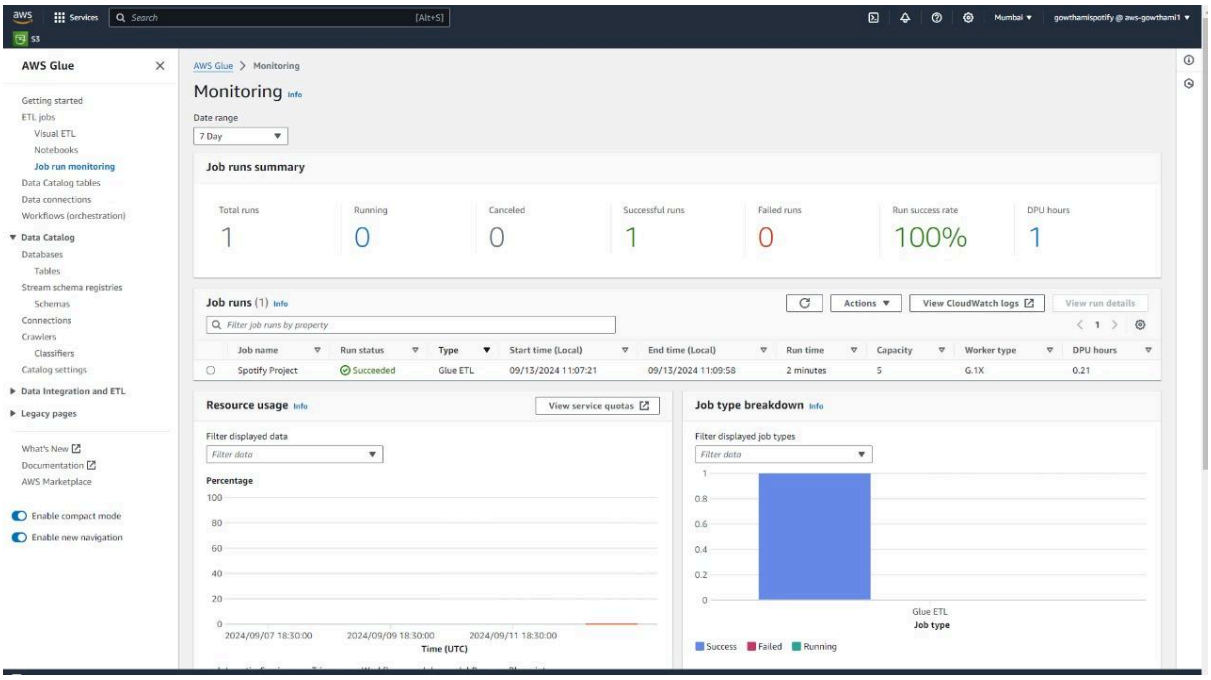
Set up an ETL pipeline that transfers and transforms data from the Staging layer to the Data Warehouse layer using AWS Glue.

Steps:

- Go to **AWS Glue** in the console.
- Create a new **Glue Job** for the ETL process.

- Configure the job to read from the **Staging** S3 folder, apply transformations (if any), and write the processed data to the **Data Warehouse** folder in S3.





Date range: 7 Day

Job runs summary

Total runs	Running	Canceled	Successful runs	Failed runs	Run success rate	DPU hours
1	0	0	1	0	100%	1

Job runs (1)

Job name	Run status	Type	Start time (Local)	End time (Local)	Run time	Capacity	Worker type	DPU hours
Spotify Project	Succeeded	Glue ETL	09/13/2024 11:07:21	09/13/2024 11:09:58	2 minutes	5	G.1X	0.21

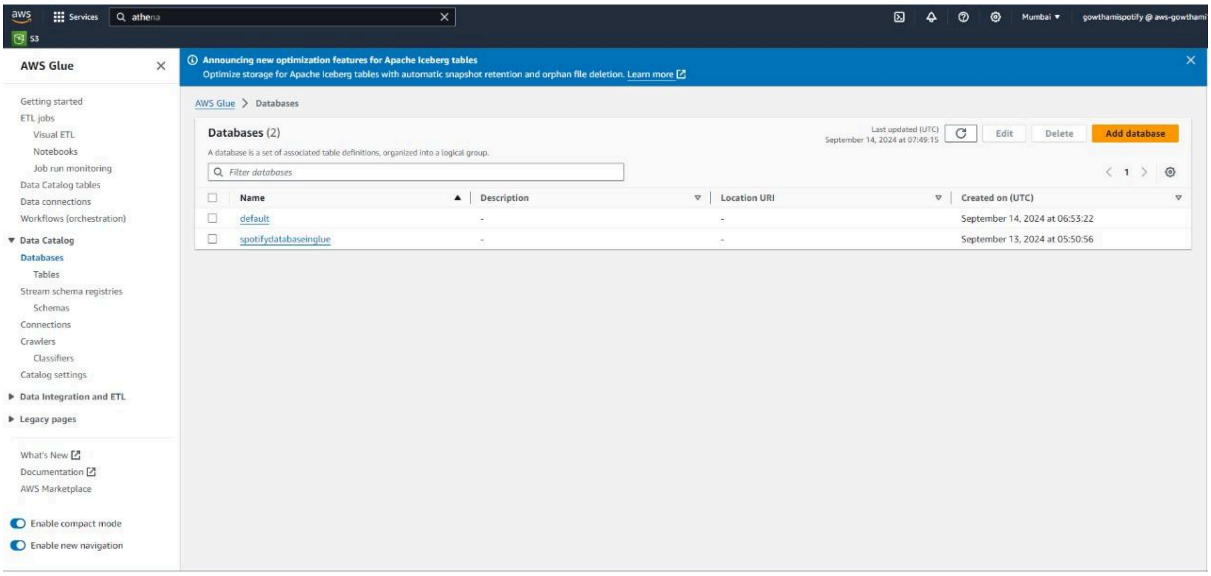
4. Set Up a Glue Crawler

AWS Glue Crawler is used to automatically catalog the data in the S3 Data Warehouse layer.

Steps:

- In the AWS Glue Console, create a new **Crawler**.

- Specify the **S3 Data Warehouse** folder as the data source.
- Run the crawler, which will create a **Data Catalog** and a **Database** in AWS Glue.



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

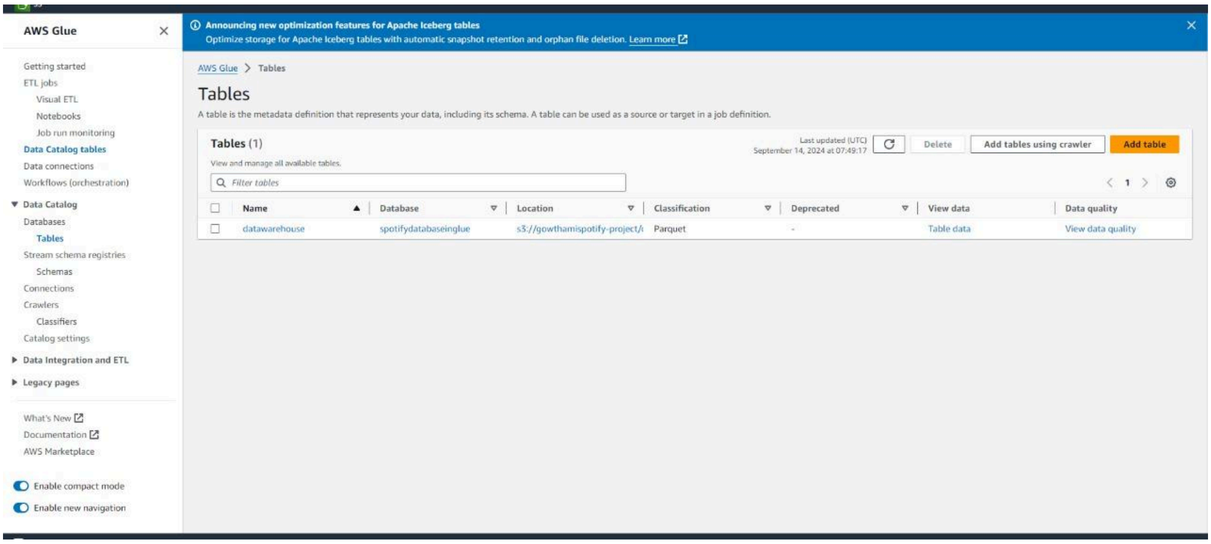
AWS Glue > Databases

Databases (2)

Last updated (UTC) September 14, 2024 at 07:49:15

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

□	Name	Description	Location URI	Created on (UTC)
□	default	-	-	September 14, 2024 at 06:53:22
□	spotifydatabaseinglue	-	-	September 13, 2024 at 05:50:56



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

AWS Glue > Tables

Tables

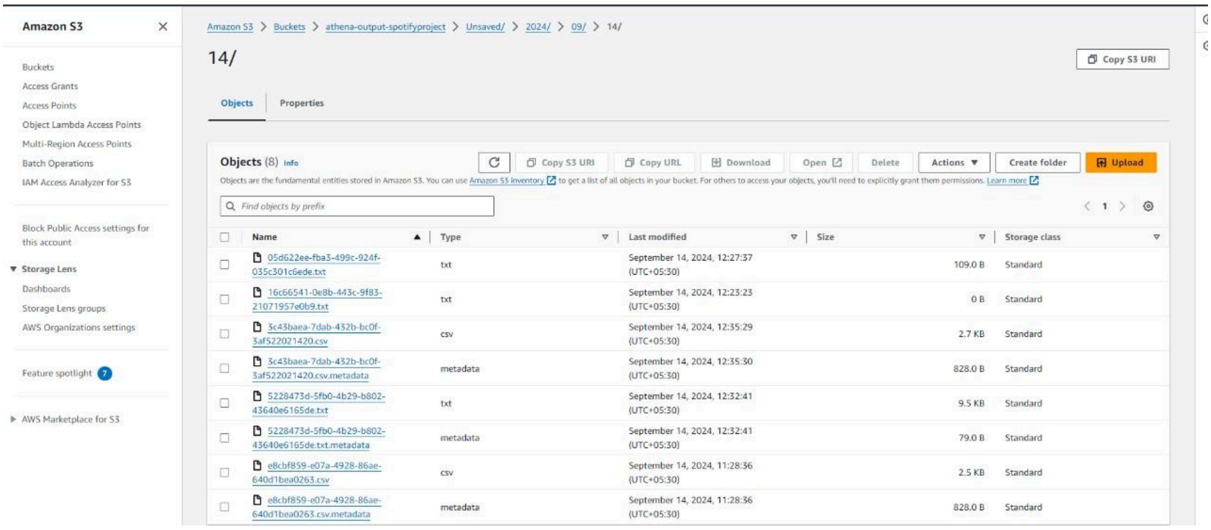
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (1)

Last updated (UTC) September 14, 2024 at 07:49:17

View and manage all available tables.

□	Name	Database	Location	Classification	Deprecated	View data	Data quality
□	datawarehouse	spotifydatabaseinglue	s3://gowthamispotify-project/	Parquet	-	Table data	View data quality



Amazon S3 > Buckets > athena-output-spotifyproject > Unsaved/ > 2024/ > 09/ > 14/

14/

Objects (8)

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

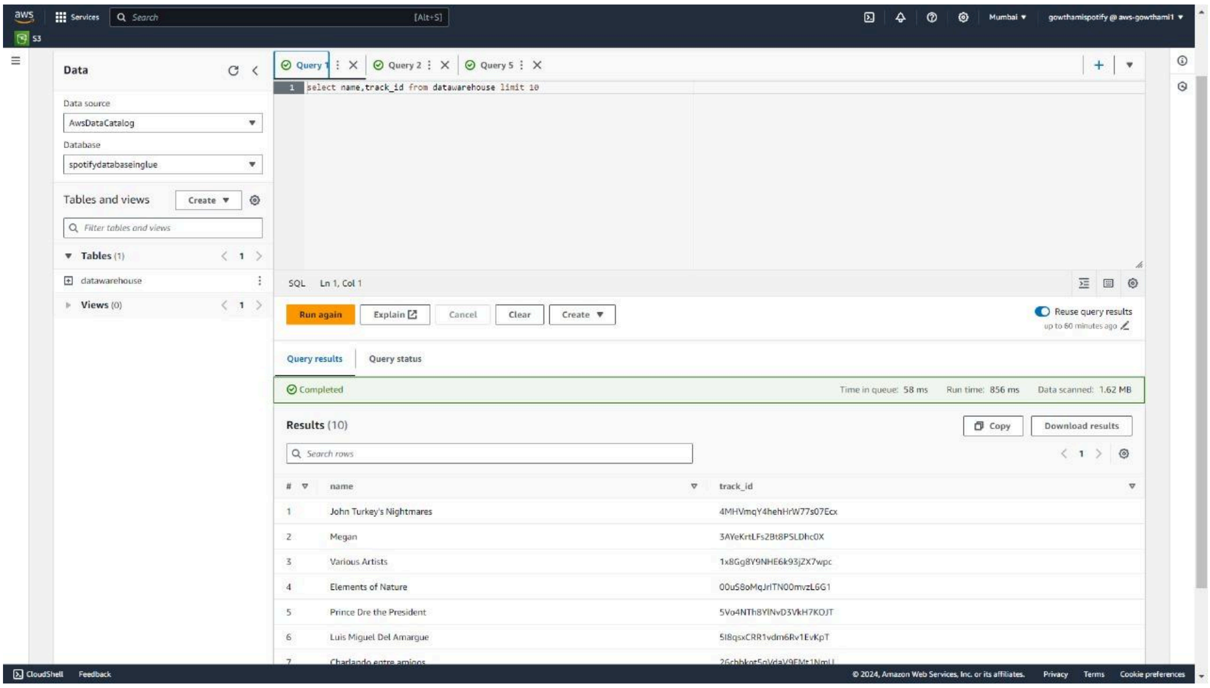
□	Name	Type	Last modified	Size	Storage class
□	05d622ee-fba3-499c-924f-035c301c6ede.txt	txt	September 14, 2024, 12:27:37 (UTC+05:30)	109.0 B	Standard
□	16c66541-0e8b-443c-9f83-21071957e0b9.txt	txt	September 14, 2024, 12:23:23 (UTC+05:30)	0 B	Standard
□	3c43baea-7dab-432b-bc0f-3af522021420.csv	csv	September 14, 2024, 12:35:29 (UTC+05:30)	2.7 KB	Standard
□	3c43baea-7dab-432b-bc0f-3af522021420.csv.metadata	metadata	September 14, 2024, 12:35:30 (UTC+05:30)	828.0 B	Standard
□	5228473d-5fb0-4b29-b802-43640e6165de.txt	txt	September 14, 2024, 12:32:41 (UTC+05:30)	9.5 KB	Standard
□	5228473d-5fb0-4b29-b802-43640e6165de.txt.metadata	metadata	September 14, 2024, 12:32:41 (UTC+05:30)	79.0 B	Standard
□	e8cbf859-e07a-4928-86ae-640d1bea0263.csv	csv	September 14, 2024, 11:28:36 (UTC+05:30)	2.5 KB	Standard
□	e8cbf859-e07a-4928-86ae-640d1bea0263.csv.metadata	metadata	September 14, 2024, 11:28:36 (UTC+05:30)	828.0 B	Standard

5. Set Up Athena and Query Editor

AWS Athena allows you to run SQL queries on the data stored in S3 using the Glue Data Catalog.

Steps:

- In the AWS Management Console, navigate to **Athena**.
- Set up the Query Editor by specifying the S3 bucket where query results will be stored.
- Use SQL to query the data stored in the Glue Catalog and S3 Data Warehouse.



Data source: AwsDataCatalog

Database: spotifydatabaseinglue

Tables (1): datawarehouse

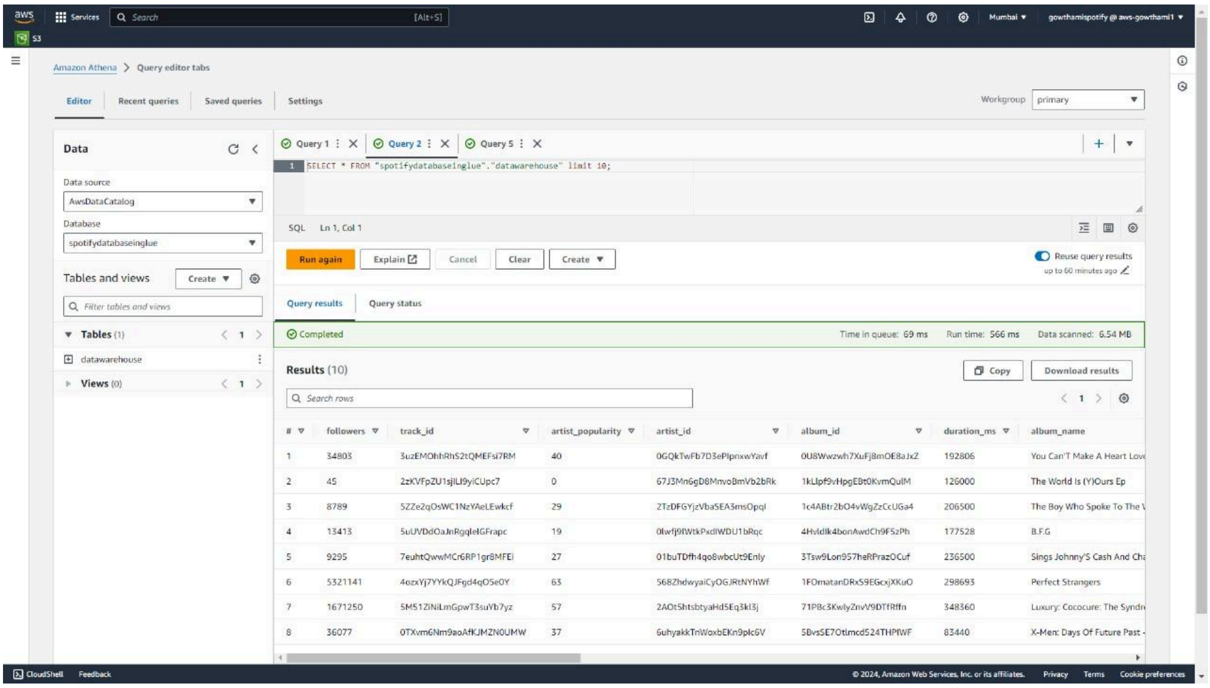
Views (0)

```
select name,track_id from datawarehouse limit 10
```

Completed   Time in queue: 58 ms   Run time: 856 ms   Data scanned: 1.62 MB

Results (10)

#	name	track_id
1	John Turkey's Nightmares	4MHVmqY4hehHrW77s07Ecx
2	Megan	3AYeKrtLFs2Bt8PSLDhc0X
3	Various Artists	1x8Gg8Y9NHE6k93jZX7wpc
4	Elements of Nature	00uS8oMqJrlTN00mvzL6G1
5	Prince Dre the President	5Vo4NTh8YlNvD3VkH7KOJT
6	Luis Miguel Del Amargue	5I8qsxCRR1vdm6Rv1EvKpT
7	Charlando entre amigos	2GcbbkotSoVdaV9EMt1NmU



Workgroup: primary

Data source: AwsDataCatalog

Database: spotifydatabaseinglue

Tables: datawarehouse

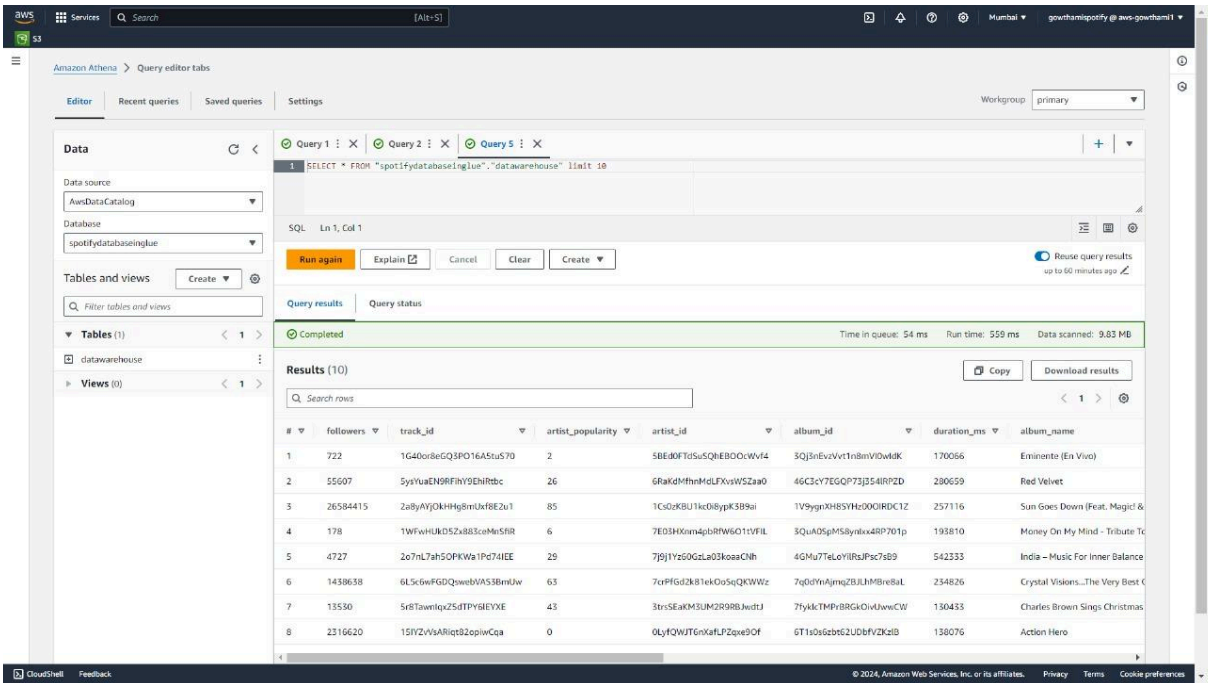
```
SELECT * FROM "spotifydatabaseinglue"."datawarehouse" limit 10;
```

SQL Ln 1, Col 1

Completed — Time in queue: 69 ms Run time: 566 ms Data scanned: 6.54 MB

Results (10)

#	followers	track_id	artist_popularity	artist_id	album_id	duration_ms	album_name
1	34803	3uzEMOhhRhS2tQMEFsi7RM	40	0GQkTwFb7D3ePIpnxwYavf	0U8Wwzwh7XuFj8mOE8aJxZ	192806	You Can'T Make A Heart Love
2	45	2zKVFpZU1sjlLI9yiCUpc7	0	67J3Mn6gD8MnvoBmVb2bRk	1kLlpf9vHpgEBt0KvmQulM	126000	The World Is (Y)Ours Ep
3	8789	5ZZe2qOsWC1NzYAeLEwkcf	29	2TzDFGYjzVbaSEA3msOpql	1c4ABtr2bO4vWgZzCcUGa4	206500	The Boy Who Spoke To The V
4	13413	5uUVDdOaJnRgqleIGFrapc	19	0lwfj9IWtkPxdIWDU1bRqc	4Hvldlk4bonAwdCh9F5zPh	177528	B.F.G
5	9295	7euhtQwwMCr6RP1gr8MFEi	27	01buTDfh4qo8wbcUt9Enly	3Tsw9Lon957heRPrazOCuf	236500	Sings Johnny'S Cash And Cha
6	5321141	4ozxYj7YYkQJFgd4qO5e0Y	63	568ZhdwyaiCyOGJRtNYhWf	1FOmatanDRx59EGcxjXKuO	298693	Perfect Strangers
7	1671250	5M51ZiNiLmGpwT3suYb7yz	57	2AOt5htsbtyaHd5Eq3kl3j	71PBc3KwlyZnvV9DTfRffn	348360	Luxury: Cococure: The Syndr
8	36077	0TXvm6Nm9aoAfKJMZN0UMW	37	6uhyakkTnWoxbEKn9plc6V	5BvsSE7Otlmcd524THPIWF	83440	X-Men: Days Of Future Past -



Workgroup: primary

Data source: AwsDataCatalog

Database: spotifydatabaseinglue

Tables: datawarehouse

```
SELECT * FROM "spotifydatabaseinglue"."datawarehouse" limit 10
```

SQL Ln 1, Col 1

Completed — Time in queue: 54 ms Run time: 559 ms Data scanned: 9.83 MB

Results (10)

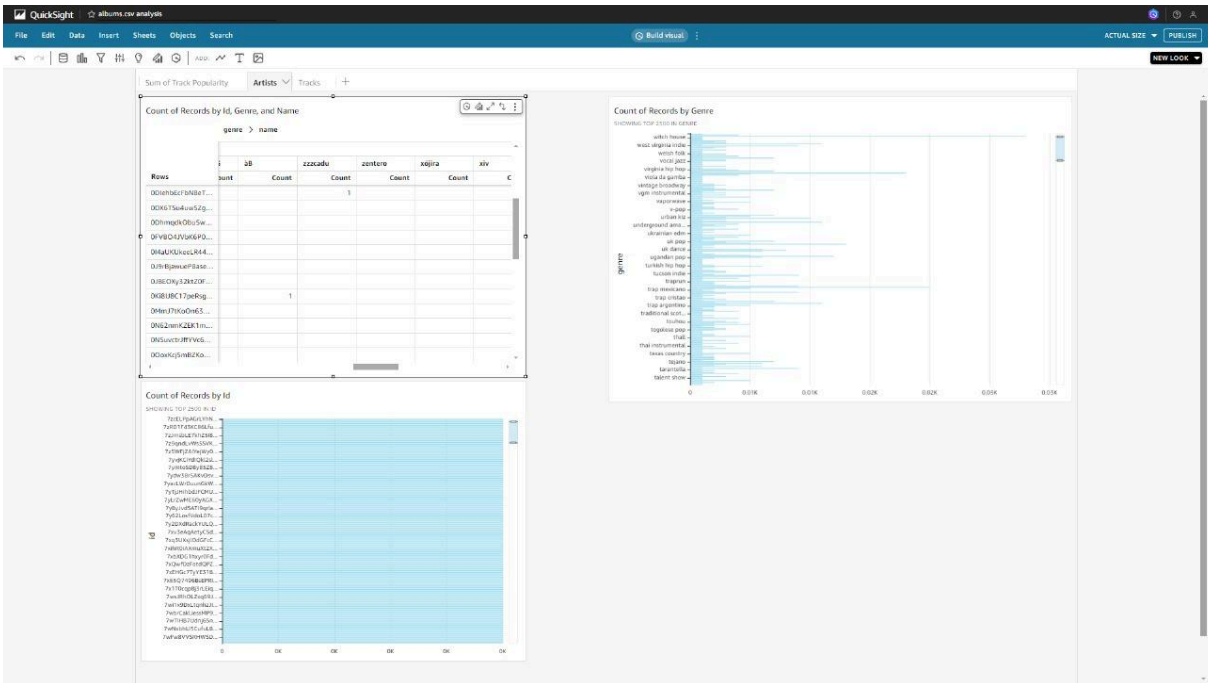
#	followers	track_id	artist_popularity	artist_id	album_id	duration_ms	album_name
1	722	1G40or8eGQ3PO16A5tuS70	2	5BEd0FTdSuSQhEBOOcWvf4	3Qj3nEvzVvt1n8mVI0wldK	170066	Eminente (En Vivo)
2	55607	5ysYuaEN9RFihY9EhiRtbc	26	6RaKdMfhnMdLFXvsWSZaa0	46C3cY7EGQP73j354IRPZD	280659	Red Velvet
3	26584415	2a8yAYjOkHHg8mUxf8E2u1	85	1CsOzKBU1kc0i8ypK3B9ai	1V9ygnXH8SYHz00OIRDC1Z	257116	Sun Goes Down (Feat. Magic! &
4	178	1WFwHUkD5Zx883ceMnSfiR	6	7E03HXnm4pbRfW6O1tVFIL	3QuA0SpMS8ynlxx4RP701p	193810	Money On My Mind - Tribute To
5	4727	2o7nL7ah5OPKWa1Pd74IEE	29	7j9j1Yz60GzLa03koaaCNh	4GMu7TeLoYilRsJPsc7sB9	542333	India – Music For Inner Balance
6	1438638	6L5c6wFGDQswebVAS3BmUw	63	7crPfGd2k81ekOoSqQKWWz	7q0dYnAjmqZBJLhMBre8aL	234826	Crystal Visions...The Very Best (
7	13530	5r8TawnIqxZ5dTPY6IEYXE	43	3trsSEaKM3UM2R9RBJwdtJ	7fyklcTMPrBRGkOivUwwCW	130433	Charles Brown Sings Christmas
8	2316620	15IYZvVsARiqt82opiwCqa	0	0LyfQWJT6nXafLPZqxe9Of	6T1s0s6zbt62UDbfVZKzIB	138076	Action Hero

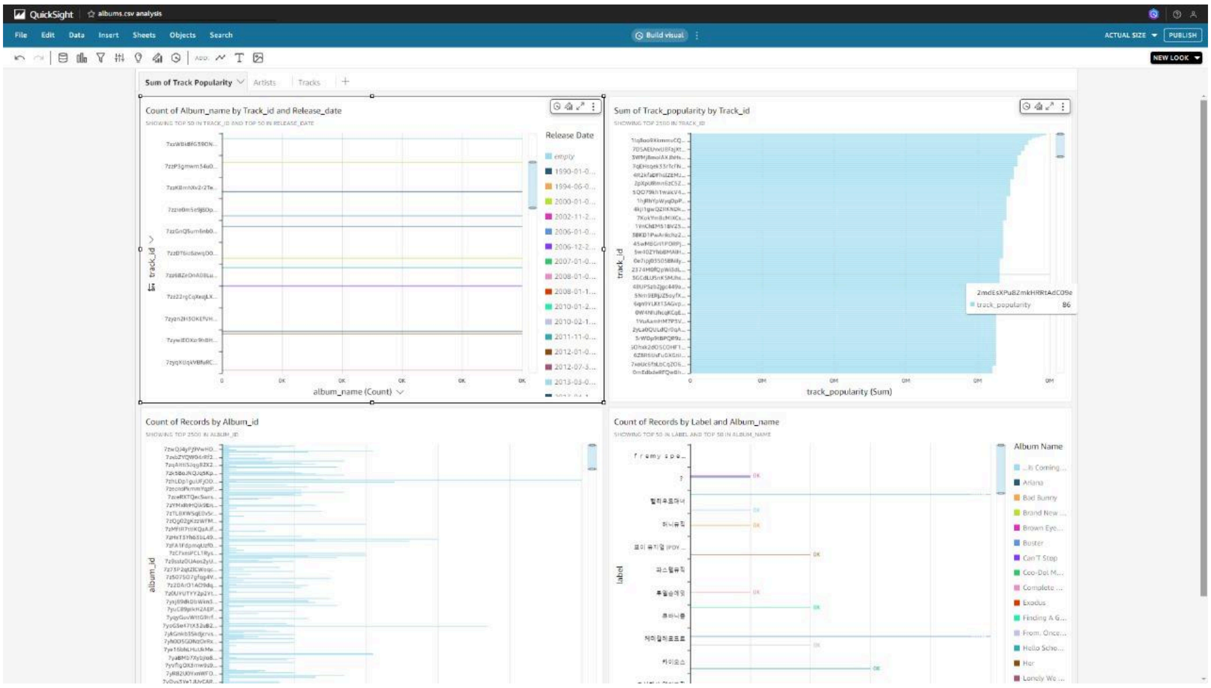
6. Visualize Data in Amazon QuickSight

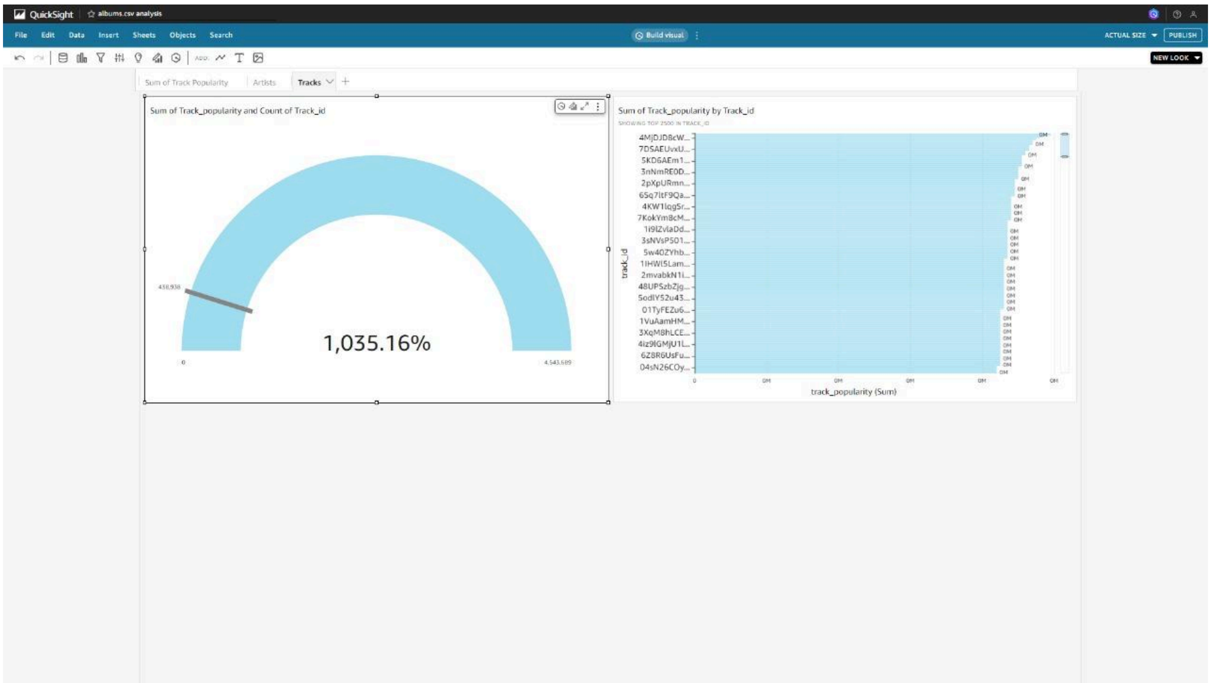
Amazon QuickSight is used to create visualizations and dashboards based on the processed data.

Steps:

- In **Amazon QuickSight**, connect to your data source by selecting Athena as the database.
- Create a new database from the Athena queries.
- Build visualizations and dashboards to display insights from the Spotify data.


QuickSight
albums.csv analysis
File Edit Data Insert Sheets Objects Search
Build visual
ACTUAL SIZE
PUBLISH
NEW LOOK
Sum of Track Popularity
Artists
Tracks
Count of Records by Id, Genre, and Name
genre > name
Rows
Count
zzzcadu
zentero
xojira
xiv
Count of Records by Genre
SHOWING TOP 2500 IN GENRE
genre
Count of Records by Id
SHOWING TOP 2500 IN ID
id


QuickSight
albums.csv analysis
File Edit Data Insert Sheets Objects Search
Build visual
ACTUAL SIZE
PUBLISH
NEW LOOK
Sum of Track Popularity
Artists
Tracks
Count of Album_name by Track_id and Release_date
SHOWING TOP 50 IN TRACK_ID AND TOP 50 IN RELEASE_DATE
track_id
album_name (Count)
Release Date
empty
1990-01-0...
1994-06-0...
2000-01-0...
2002-11-2...
2006-01-0...
2006-12-2...
2007-01-0...
2008-01-0...
2008-01-1...
2010-01-2...
2010-02-1...
2011-11-0...
2012-01-0...
2012-07-3...
2013-03-0...
Sum of Track_popularity by Track_id
SHOWING TOP 2500 IN TRACK_ID
track_id
track_popularity (Sum)
track_popularity
86
Count of Records by Album_id
SHOWING TOP 2500 IN ALBUM_ID
album_id
Count of Records by Label and Album_name
SHOWING TOP 50 IN LABEL AND TOP 50 IN ALBUM_NAME
label
Album Name
...is Coming...
Ariana
Bad Bunny
Brand New ...
Brown Eye...
Buster
Can'T Stop
Coo-Dei M...
Complete ...
Exodus
Finding A G...
From. Once...
Hello Scho...
Her
Lonely We ...


QuickSight
albums.csv analysis
File
Edit
Data
Insert
Sheets
Objects
Search
Build visual
ACTUAL SIZE
PUBLISH
ADD
NEW LOOK
Sum of Track Popularity
Artists
Tracks
Sum of Track_popularity and Count of Track_id
438,938
1,035.16%
0
4,543,689
Sum of Track_popularity by Track_id
SHOWING TOP 2500 IN TRACK_ID
track_id
4MjDJD8cW...
7DSAEUvxU...
5KD6AEm1...
3nNmREOD...
2pXpURmn...
6Sq7ltF9Qa...
4KW1lqgSr...
7KokYm8cM...
1i9lZvlaDd...
3sNVsP5O1...
5w4OZYhb...
1IHWl5Lam...
2mvabkN1i...
48UPSzbZjg...
5odlY52u43...
01TyFEZu6...
1VuAamHM...
3XqM8hLCE...
4iz9lGMjU1L...
6Z8R6UsFu...
04sN26COy...
0M
0
0M
0M
0M
0M
0M
track_popularity (Sum)