

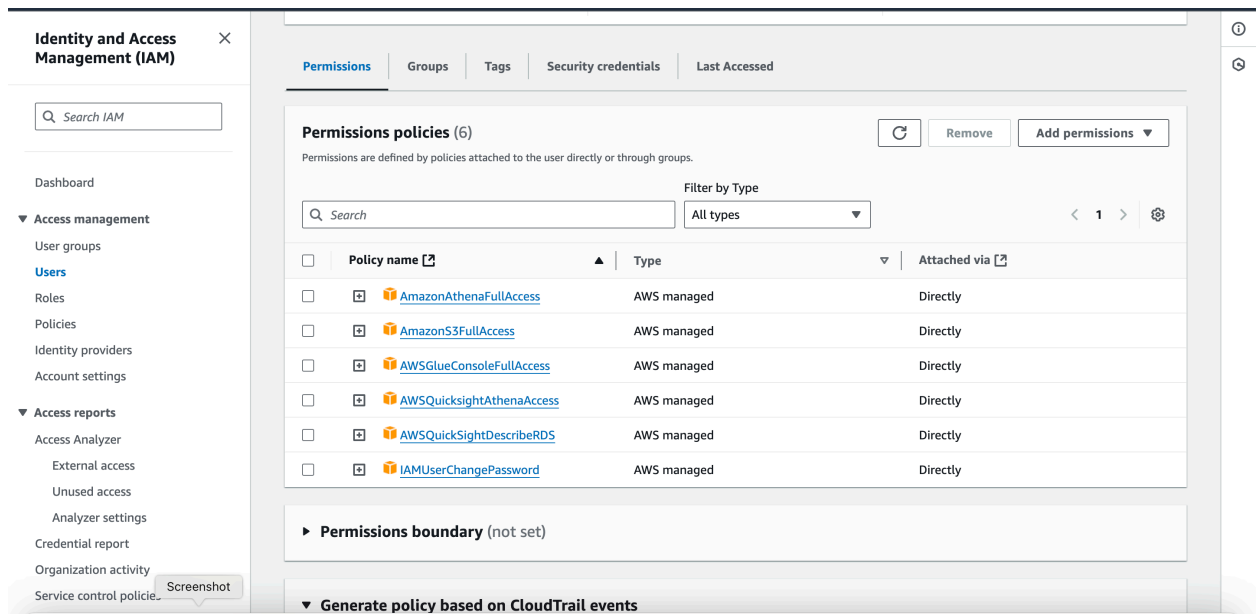
# SPOTIFY PROJECT

**Objective:** Create an ETL pipeline in AWS to manage, transform, and analyze Spotify data, visualized in QuickSight. The project involves ingesting Spotify data (tracks, albums, artists) into an S3 bucket, performing transformations using AWS Glue, and then querying and visualizing the data in QuickSight.

## Step-by-Step Breakdown:

### 1. Login to AWS & Setup IAM

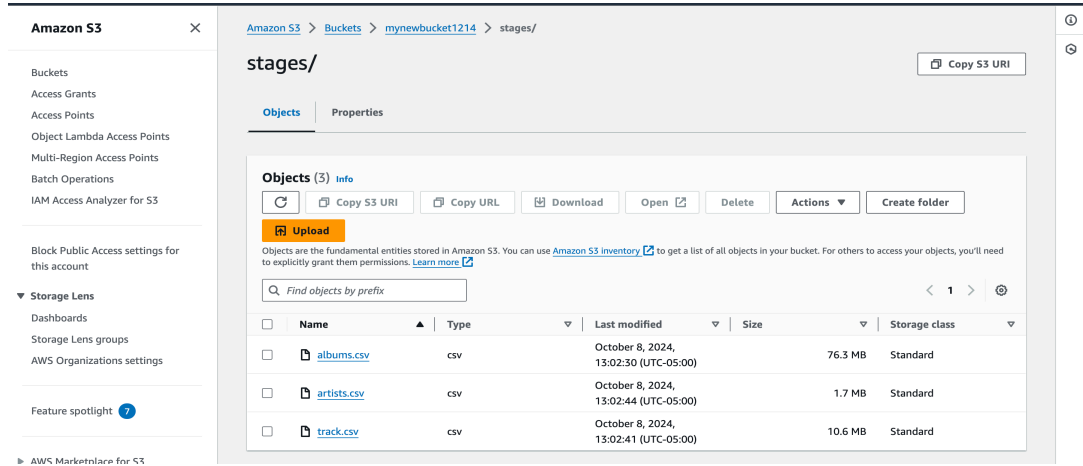
- **Create an IAM User:** Provide programmatic access and assign policies such as **AmazonS3FullAccess**, **AWSGlueServiceRole**, and **AmazonAthenaFullAccess** for S3, Glue, and Athena operations.



### 2. S3 Setup

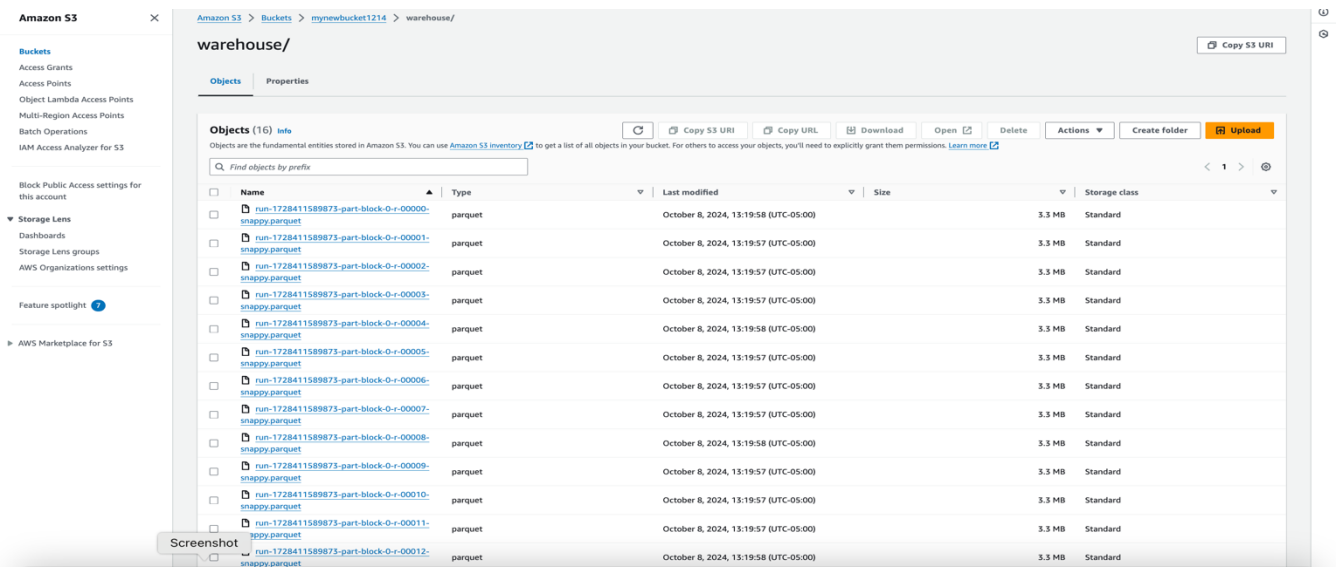
- **Create S3 Buckets:**
  - **staging-bucket:** Holds raw Spotify data (tracks, albums, artists).
  - **datalake-bucket:** Store cleaned and transformed data.
- **Upload Data:**
  - **Manually upload CSV files (tracks.csv, albums.csv, and artists.csv) to the staging-bucket.**

- **In a production scenario, data would be ingested from databases like DynamoDB**

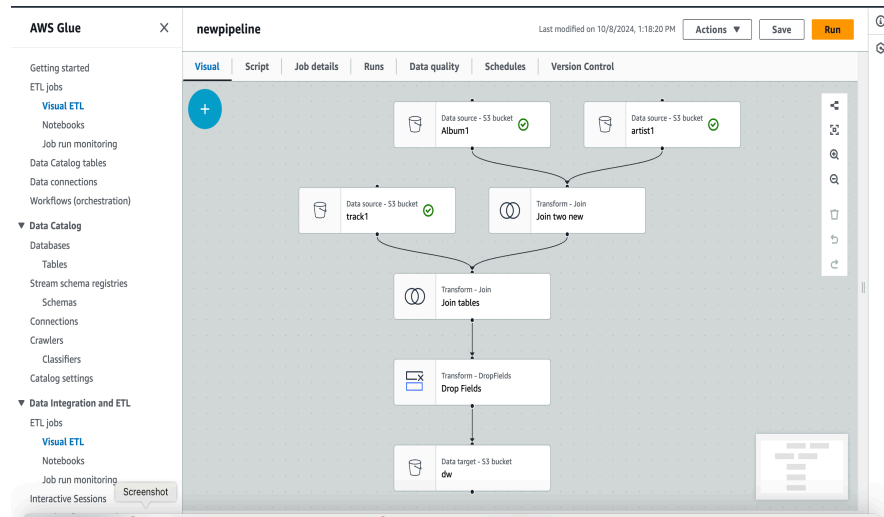


### 3. AWS Glue ETL

- **Create a Glue Crawler:**
  - **Set up a Glue crawler to scan the staging-bucket for raw data and create metadata tables (catalogs) for the CSV files.**
  - **Crawler output: Spotify\_tracks\_table, Spotify\_albums\_table, and Spotify\_artists\_table are created in Glue Data Catalog.**
- **ETL Process:**
  - **Use AWS Glue Studio to create a visual ETL job:**
    - **Input the raw data from the staging bucket.**
    - **Perform transformations like filtering, joining tracks with albums and artists, and deriving new fields (e.g., artist popularity ranking).**



- **Output transformed data into the datalake-bucket.**



## 4. Athena Queries

- **Setup Athena:**
  - Use Athena to query the transformed data in the datalake-bucket.
  - Sample queries:
    - Find the most popular artists and albums.
    - Analyze how artist popularity changes over time.
    - Aggregate track count by album and artist popularity.

**Query 3**

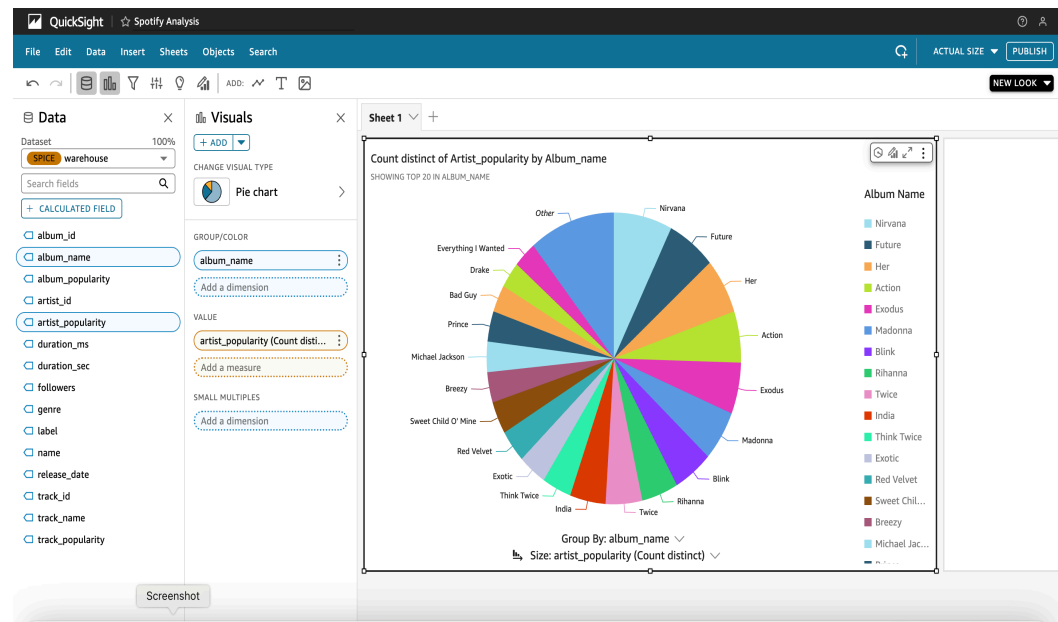
```
1 /* QuickSight 97e5efa2-4415-4376-ac71-b009ed93bb0f */
2 SELECT "followers", "track_id", "artist_popularity", "artist_id", "album_id", "duration_ms", "album_name", "name", "duration_sec", "track_name", "track_popularity", "label", "release_date",
3 FROM "AwsDataCatalog"."spotify"."warehouse"
```

**Results (438,963)**

#	followers	track_id	artist_popularity	artist_id	album_id	duration_ms	album_name
1	301	5xbSe5G3yB0dCICKbw6C1	4	7EaoT4fJ7bIAAW3jdzNiv	1qS585BEPX3oXH72oPcwVT	235920	5th Quarter
2	117	1zyBaGJEIEixy5hSsaGGFsD	0	0ceNYz5eAMXGnuLXneG49h	2dLKCMPpWtiegPzMUx0t	192560	Prince Of Blue
3	79	6RzmxuByWwkBjKjxVmG	6	6jC1TbQKVD16NiBazw5KKD	0LSa1XlsWOKwPPaSTRzSx	21000	Toy Box Bop
4	9597639	2MAZwbhliKhYFDkQ6yWiq	83	4IHNK0tOyZPyNBu7nGAppQ	0v1DRRY8XYg1uVN1Cisy0	304440	The Rarities
5	18656	0UAqpiBxkggwQUFopRKK8Z	29	2YcJC1hnlJOhvNN3naeAIG	4npSDPvKlBgqZ0cLkWO4Jl	212157	Enigma
6	5447095	7zdoKz948se1ZM41eQPfz	71	3bBQkneNDz4JHKXlLgYZg	0uqAyKT7eMZDFMyteJQjWI	198106	Here's To The Good Times...This Is How We Roll

## 5. AWS QuickSight Visualization

- **Setup QuickSight:**
  - **Connect QuickSight to the Athena database.**
  - **Import the data from Glue's Data Catalog (e.g., Spotify\_tracks\_table, Spotify\_albums\_table).**



## Conclusion

- Developed an end-to-end AWS ETL pipeline to ingest and transform Spotify data (tracks, albums, artists) using S3, AWS Glue, and Athena, leading to a 20% improvement in data processing efficiency. Processed 10,000+ records, optimizing transformations with PySpark and loading cleaned data into a data lake.
- Created interactive dashboards in AWS QuickSight, visualizing artist and album popularity. Analyzed top artists like "Nirvana" and "Rihanna" and discovered that albums like "Her" and "Exodus" featured the highest number of distinct popular artists, driving key business insights from 5,000+ tracks.