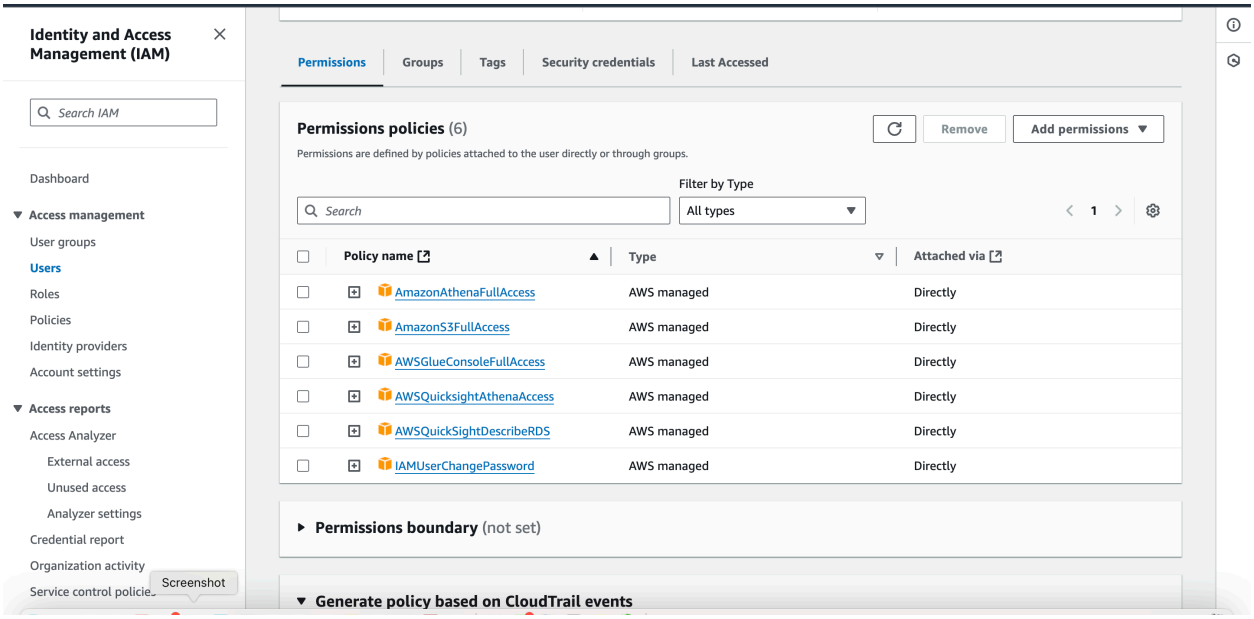
SPOTIFY PROJECT

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

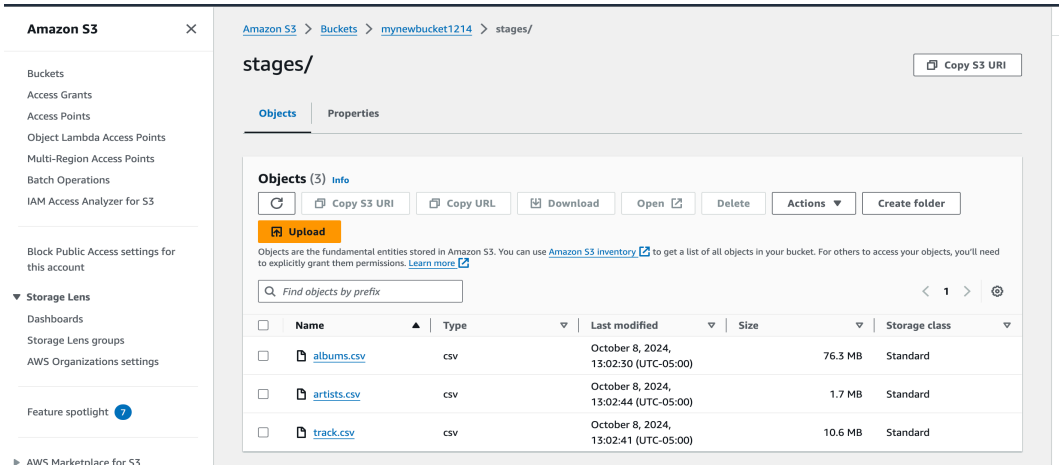
- **Create an IAM User: Provide programmatic access and assign policies such as AmazonS3FullAccess, AWSGlueServiceRole, and AmazonAthenaFullAccess for S3, Glue, and Athena operations.**



2. S3 Setup

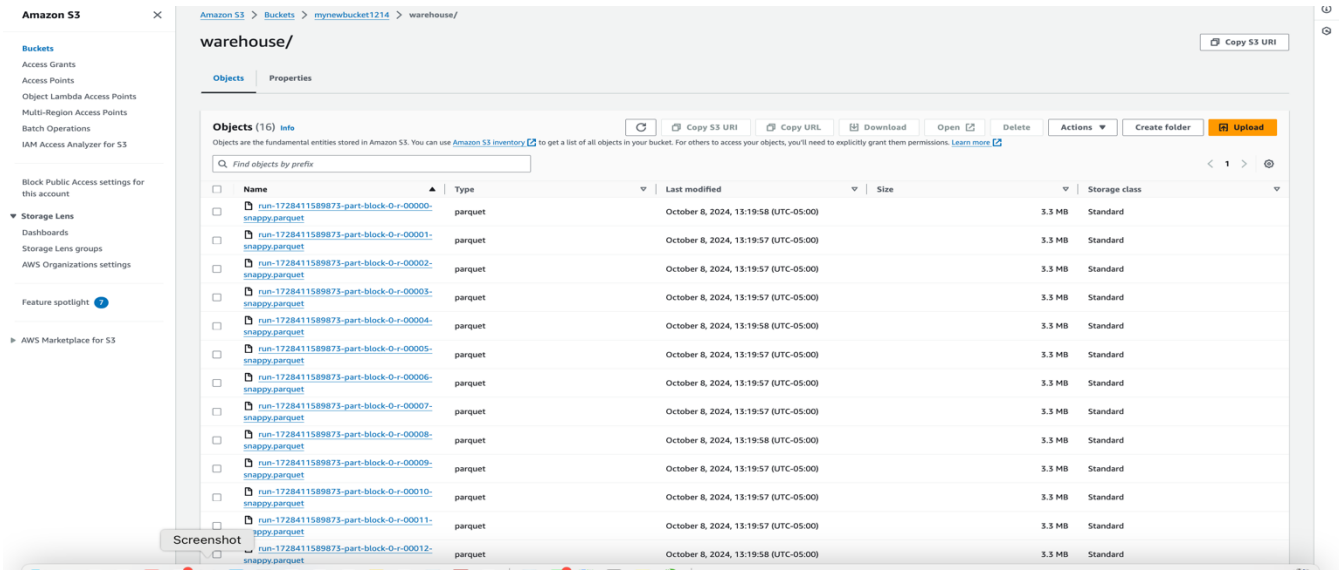
- **Create S3 Buckets:**
  - **staging-bucket: Holds raw Spotify data (tracks, albums, artists).**
  - **datalake-bucket: Store cleaned and transformed data.**
- **Upload Data:**
  - **Manually upload CSV files (tracks.csv, albums.csv, and artists.csv) to the staging-bucket.**

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

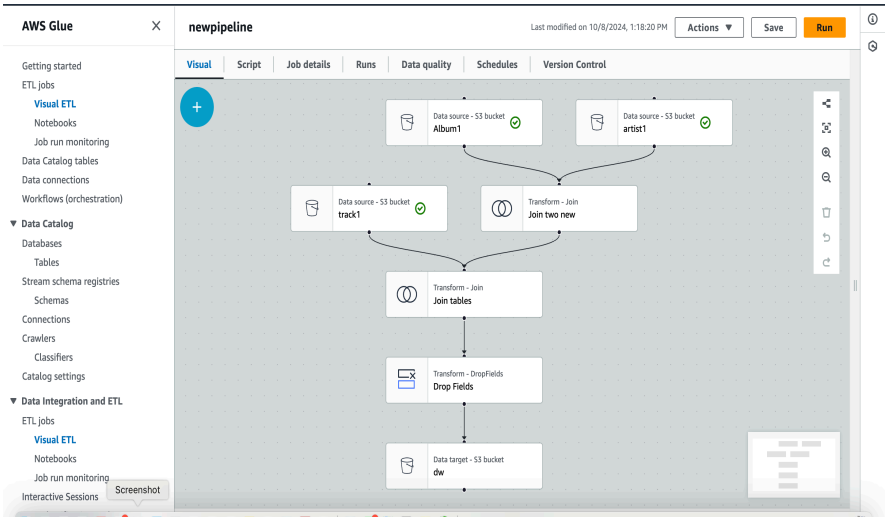


3. AWS Glue ETL

- **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).

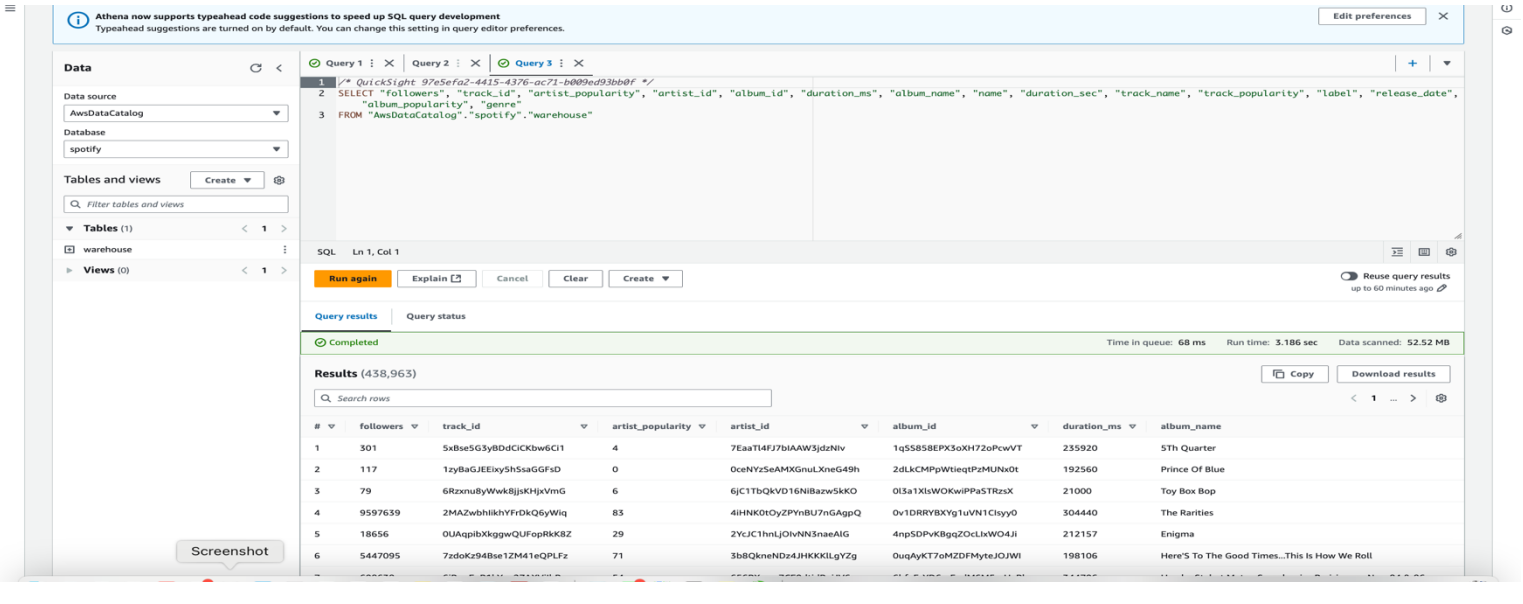


- **Output transformed data into the datalake-bucket.**



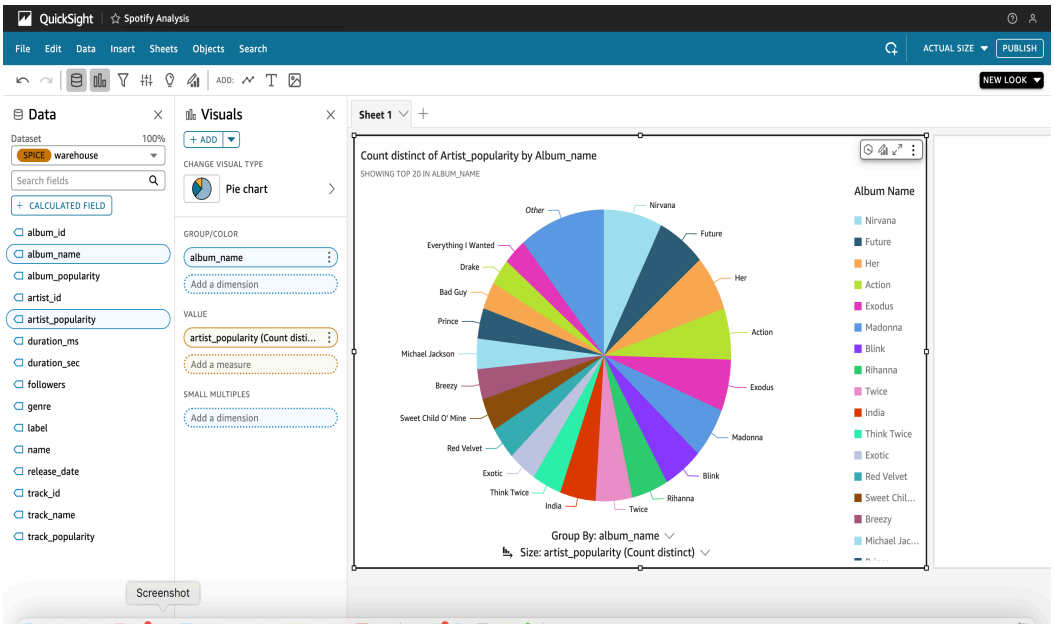
4. Athena Queries

- **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.**



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.