**Big Data Analytics Summer’24**

**Final Project : End-to-End Data Engineering Project Using AWS Cloud and PySpark**

**Project Title:** Spotify Data Engineering

**Group-4**

**Group Members:**

1. Jai Saxena (Roll No. 50)
2. Muhammad Armughan Yaqoob (Roll No. 68),
3. Jefferey Lin (Roll No. 27)

**Project Overview:**

This project aims to implement a comprehensive data engineering solution leveraging AWS cloud services and PySpark. The objective is to build a robust ETL pipeline that processes raw Spotify data into a structured format, allowing for insightful business analytics. The project will encompass data ingestion, transformation, storage, and visualization, using AWS Glue, Athena, QuickSight, and S3.

**Project Objectives:**

1. **Data Ingestion and Storage**

* Ingest raw Spotify data into AWS S3.
* Organize data into staging and data warehouse buckets.

1. **Data Transformation**

* Utilize AWS Glue to create an ETL pipeline for transforming raw data.
* Ensure accurate data transformation by handling necessary joins and data cleaning.

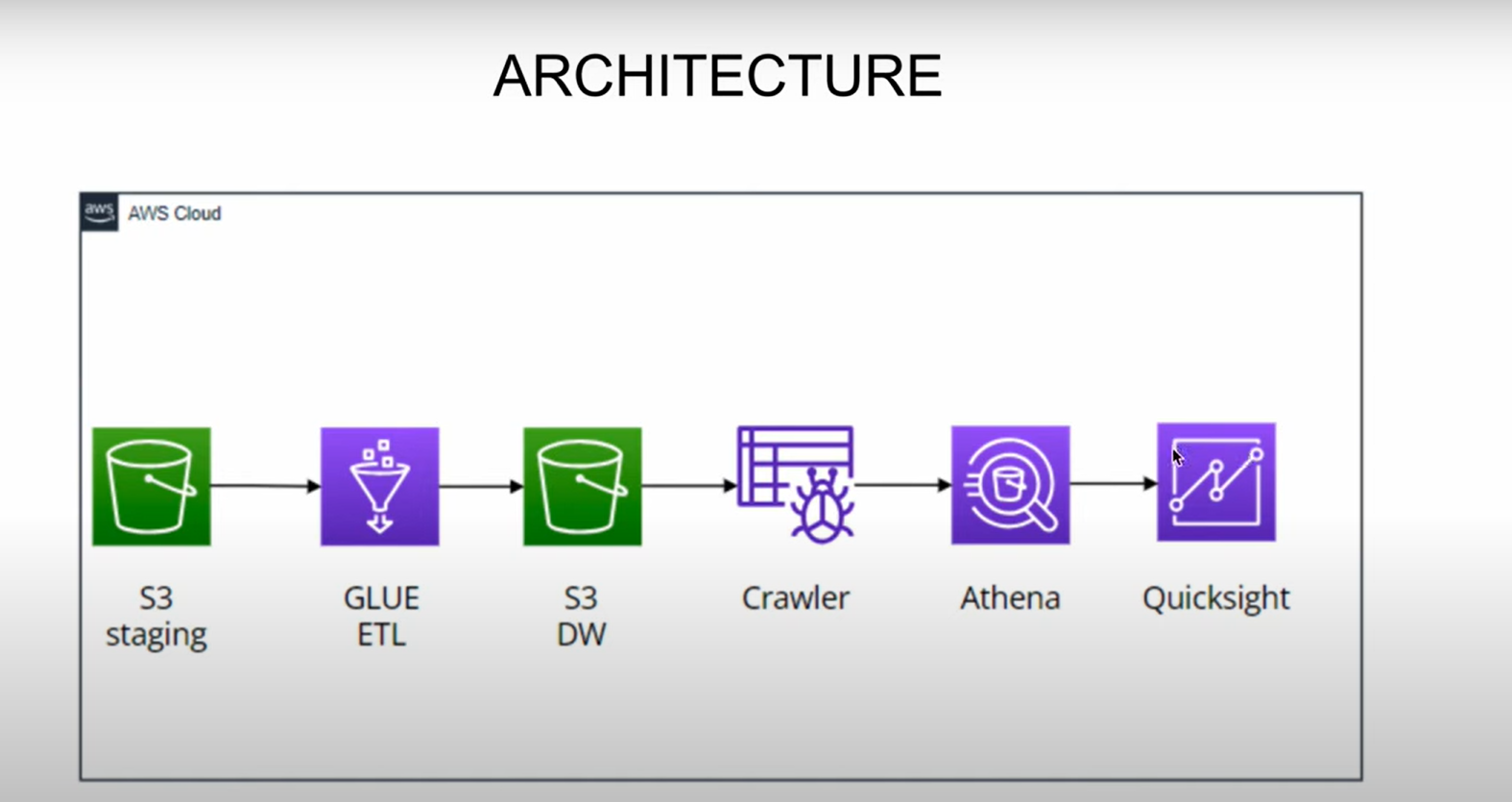
1. **Data Querying**

* Use AWS Athena to query the transformed data.

1. **Data Visualization**

* Leverage AWS QuickSight to create visualizations and gain business insights from the data.

**Architecture and Workflow:**



1. **Data Storage**

* **AWS S3**: Data will be stored in S3 buckets divided into 'staging' and 'data warehouse' folders. Raw CSV files will be uploaded into the staging folder.

1. **ETL Pipeline**

* **AWS Glue**: Create a Glue ETL job to process data from the staging bucket, apply necessary transformations, and load it into the data warehouse bucket. This includes:
* **Data Crawling**: Use AWS Glue Crawlers to scan data in S3, creating metadata tables in the AWS Glue Data Catalog.
* **Data Transformation**: Employ AWS Glue's visual ETL interface to join datasets (albums, artists, tracks) and perform data cleaning operations.
* **Data Storage**: Save transformed data in Parquet format in the data warehouse bucket for efficient querying.

1. **Data Querying**

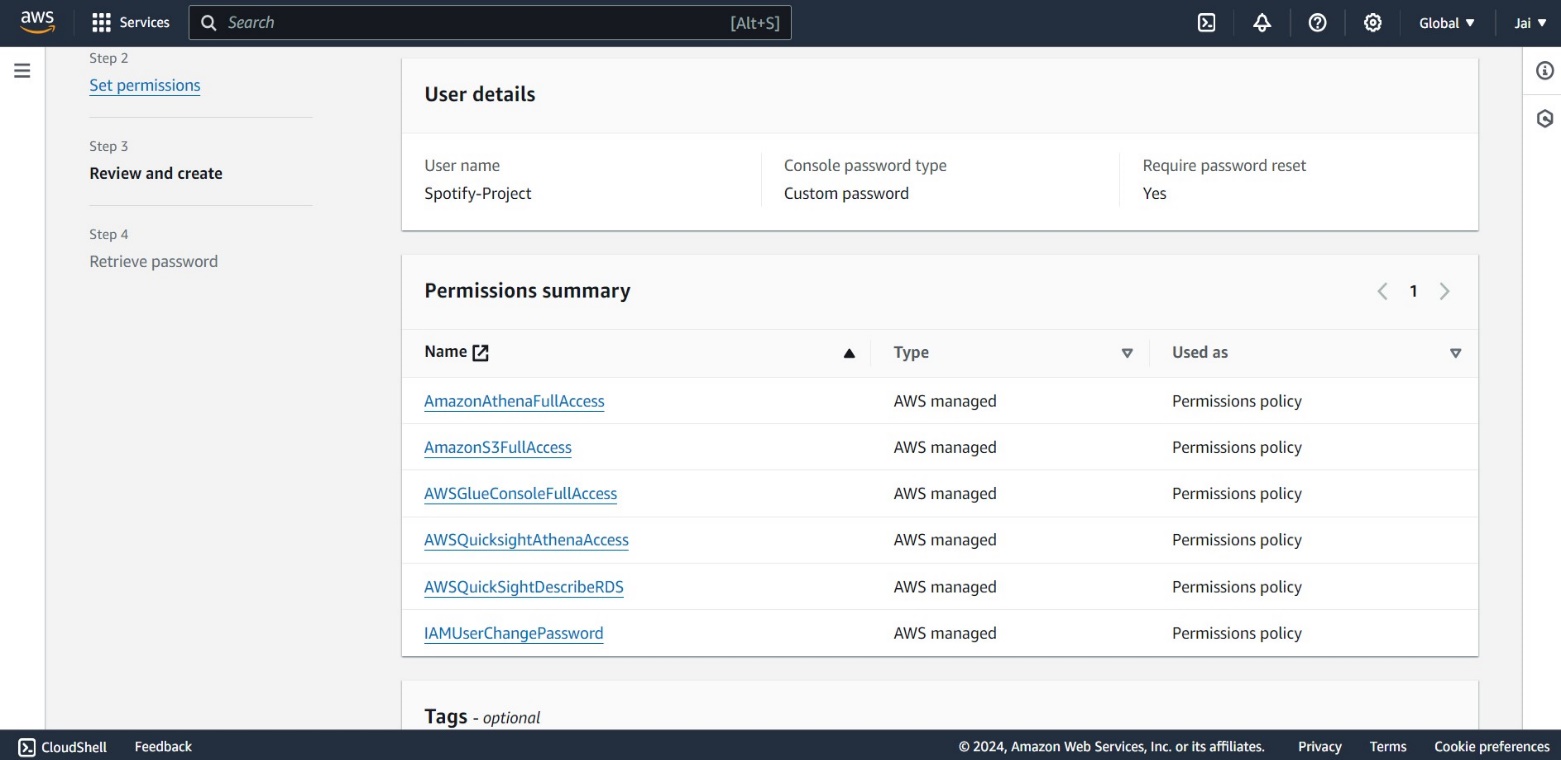
* **AWS Athena**: Query the data stored in the data warehouse bucket using SQL queries. Athena will utilize the metadata created by the Glue Crawler.

1. **Data Visualization**

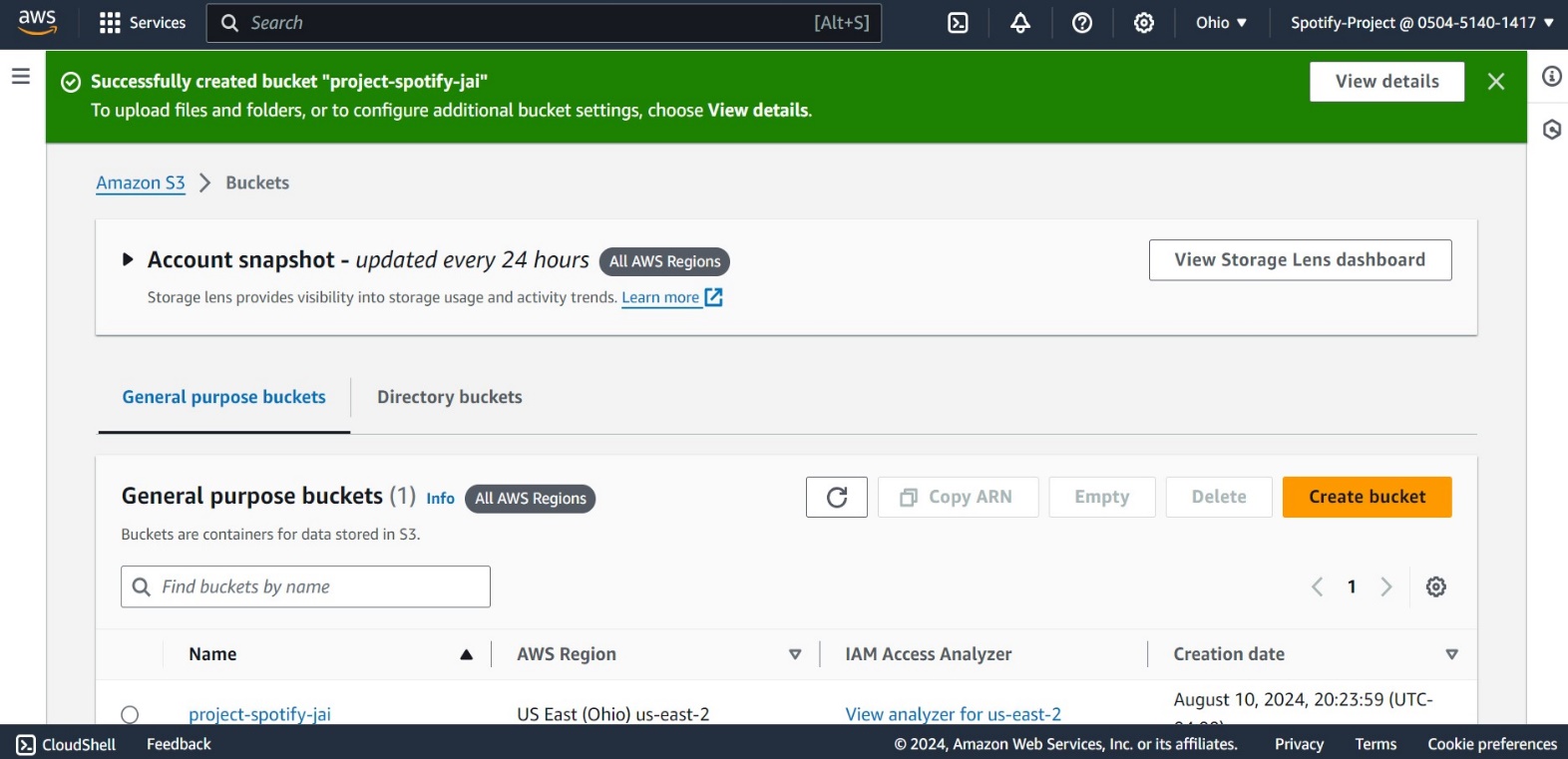
* **AWS QuickSight**: Connect QuickSight to Athena to build dashboards and reports, providing insights into Spotify data such as album popularity, artist metrics, and track features.

**Detailed Steps**

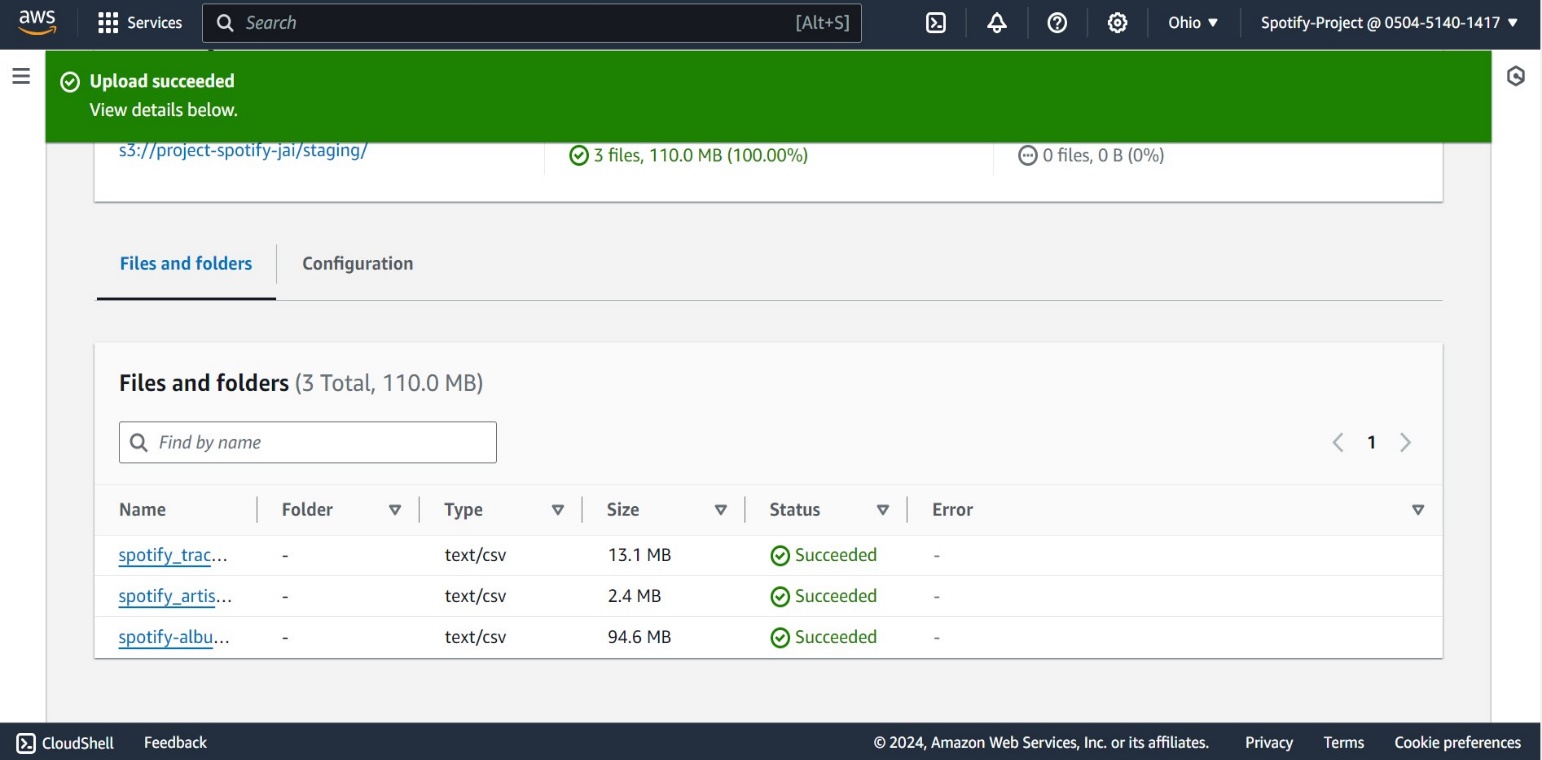
* **Setting Up AWS Environment**
* Create IAM users with appropriate permissions for S3, Glue, Athena, and QuickSight.



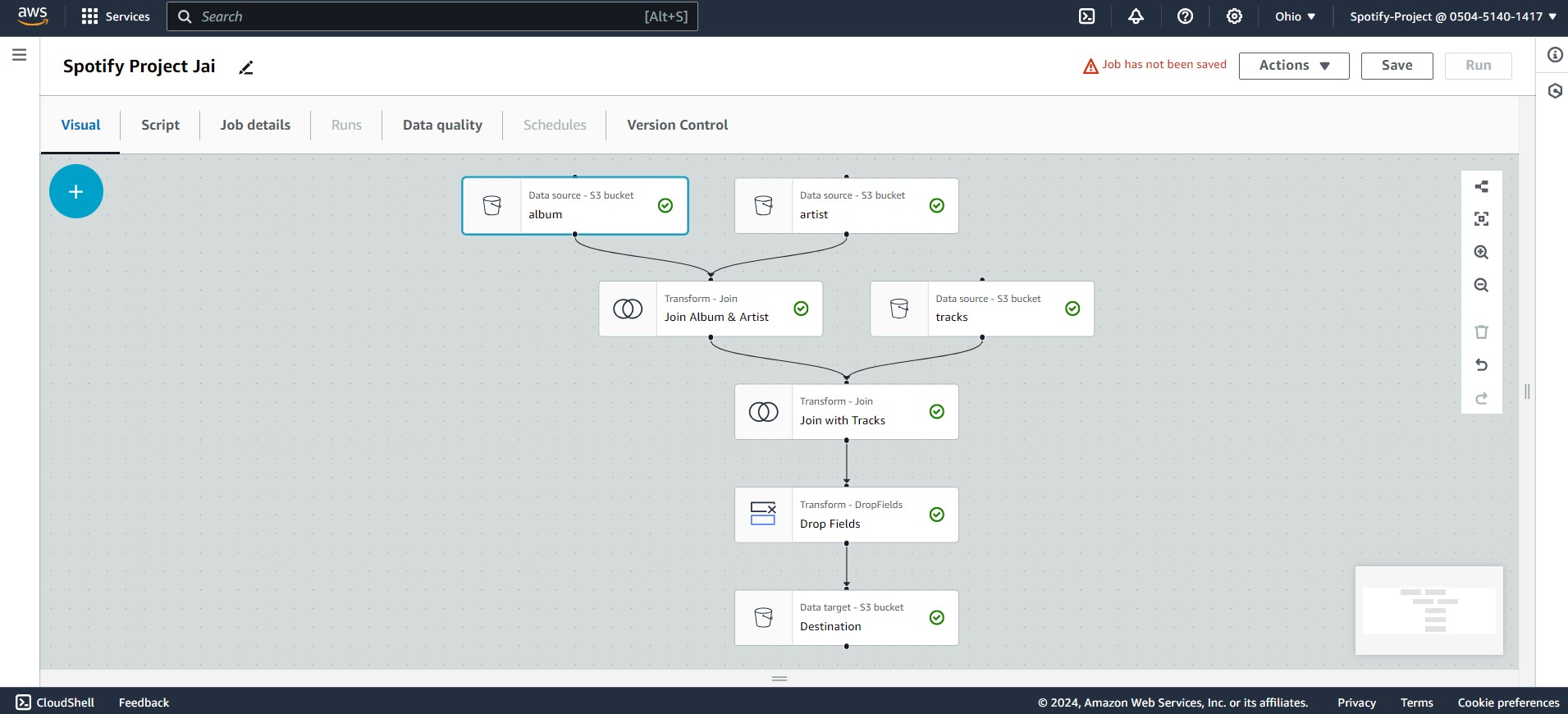
* Set up S3 buckets with 'staging' and 'data warehouse' folders.



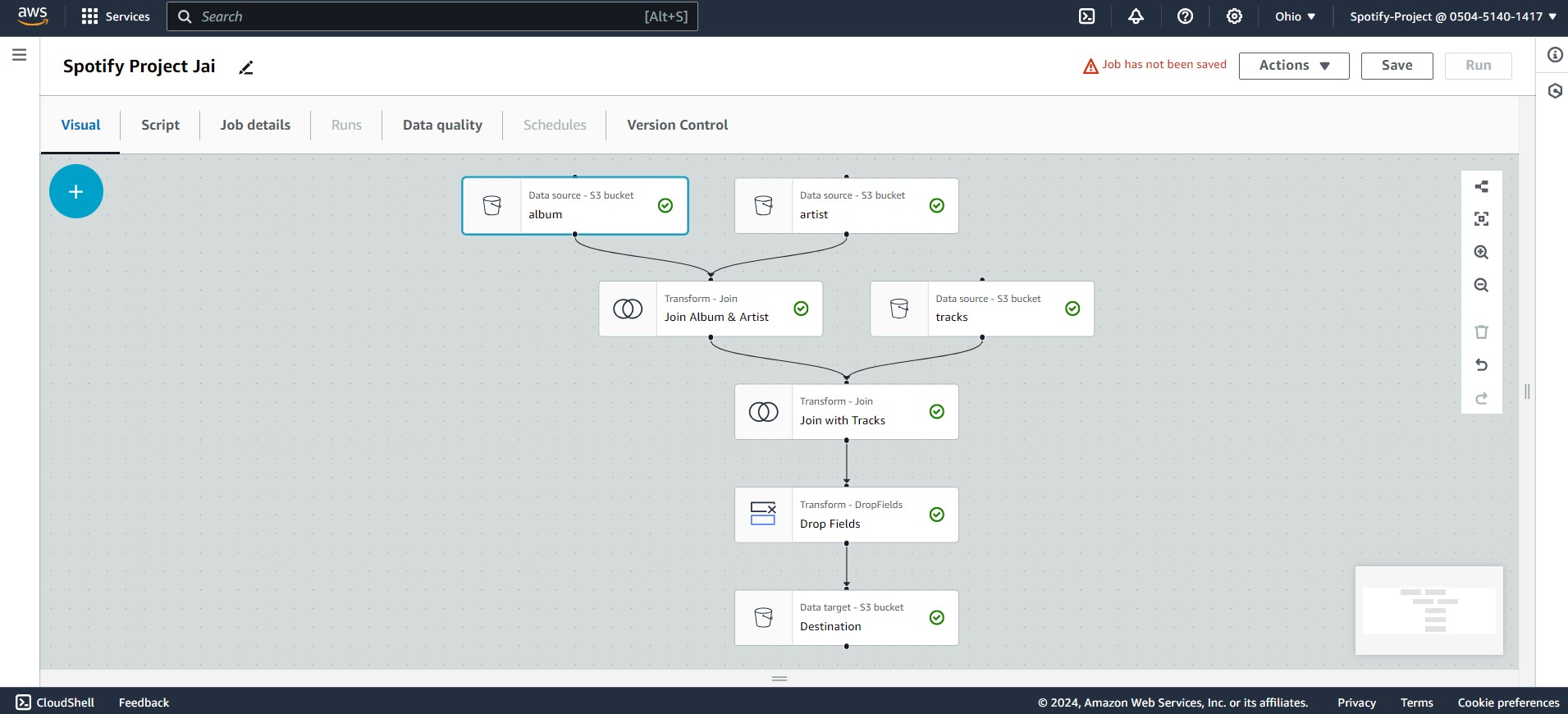
* **Data Preparation**
* Upload the Spotify dataset CSV files (albums, artists, tracks) to the S3 staging folder.



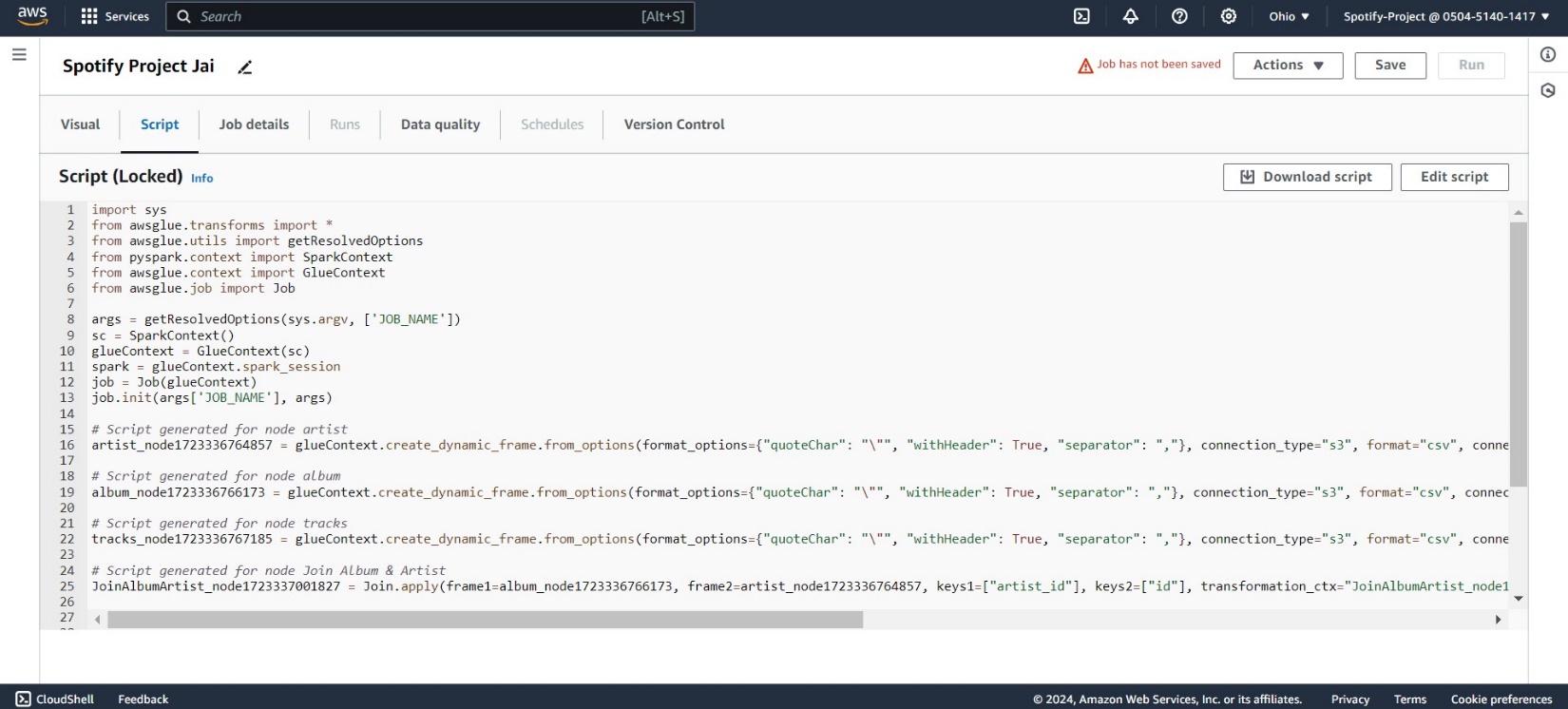
* Preprocess the data as necessary, including cleaning and structuring it.



* **Building the ETL Pipeline**
* **Create Glue Crawler**:
  + Set up a Glue Crawler to create a metadata catalog of the data stored in S3.
* **Develop Glue ETL Job**:
  + Use AWS Glue's visual ETL to drag and drop data sources, define transformation logic, and configure the target output format (Parquet).

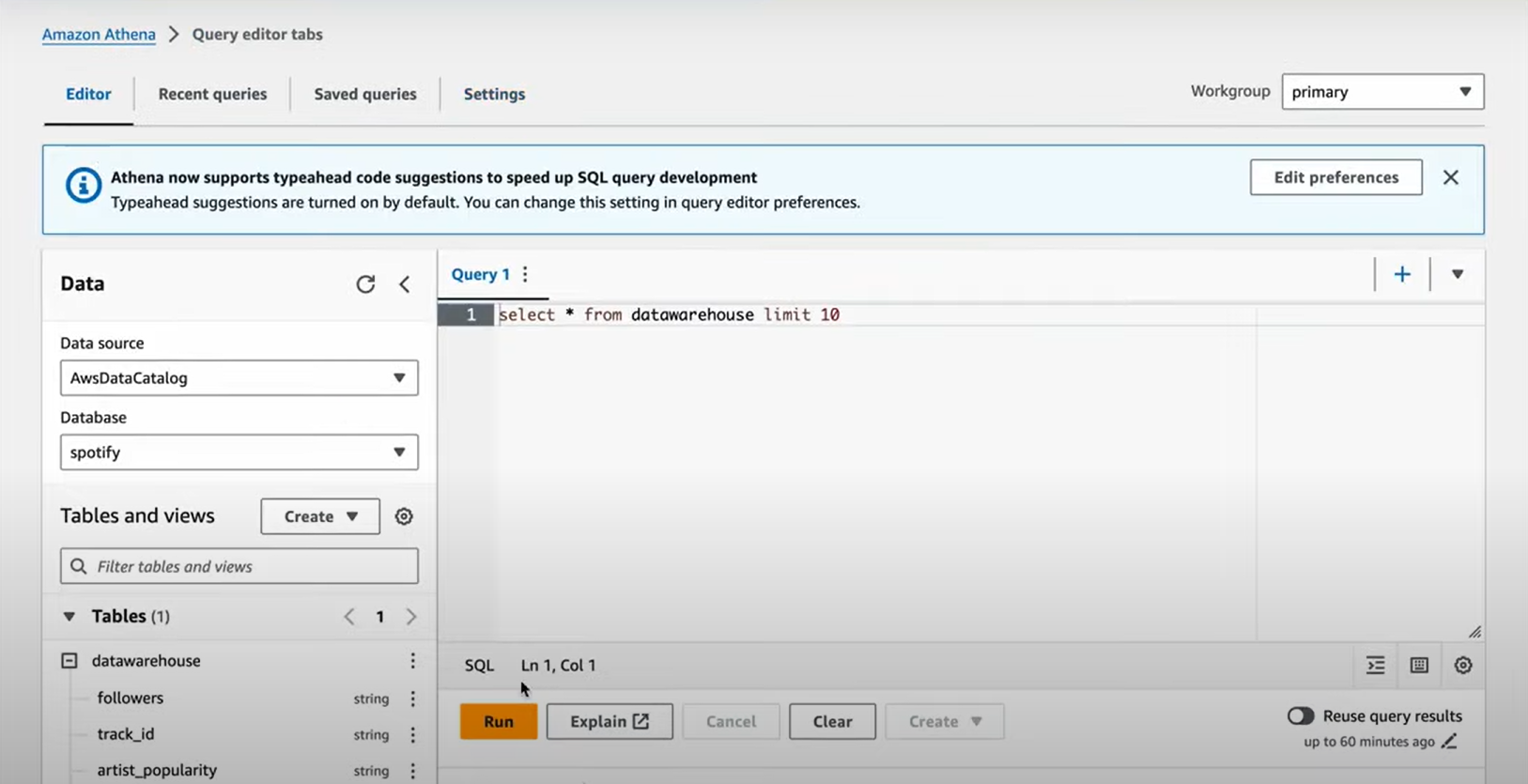


* **Run and Monitor Glue Job**:
  + Execute the Glue job and monitor its performance and success. Troubleshoot any issues related to data access or transformation.

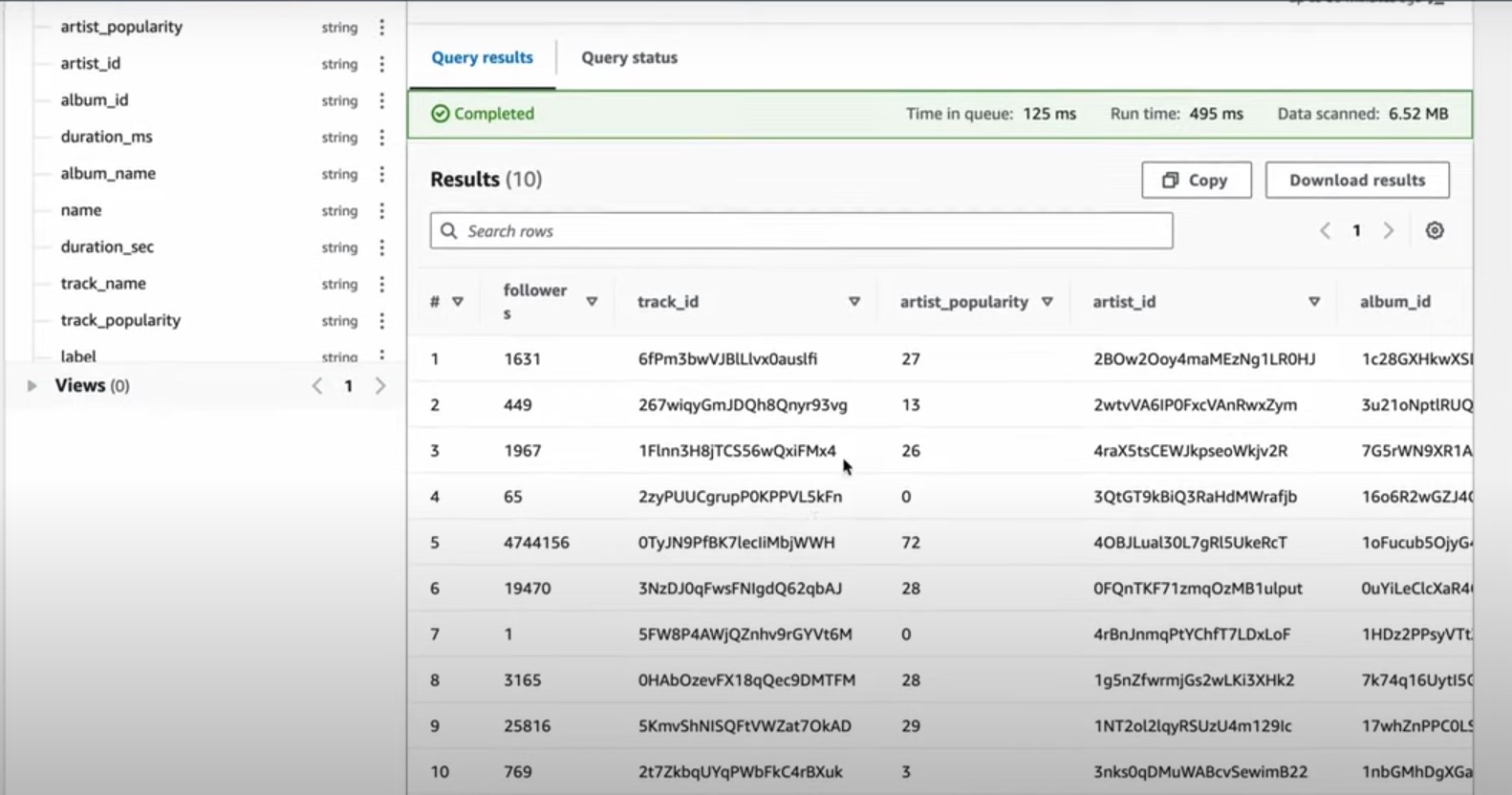


* **Query and Analysis**
* **Set Up AWS Athena**:
  + Configure Athena to query the data catalog created by the Glue Crawler.
* **Perform SQL Queries**:
  + Write and execute SQL queries to extract meaningful insights from the transformed data.

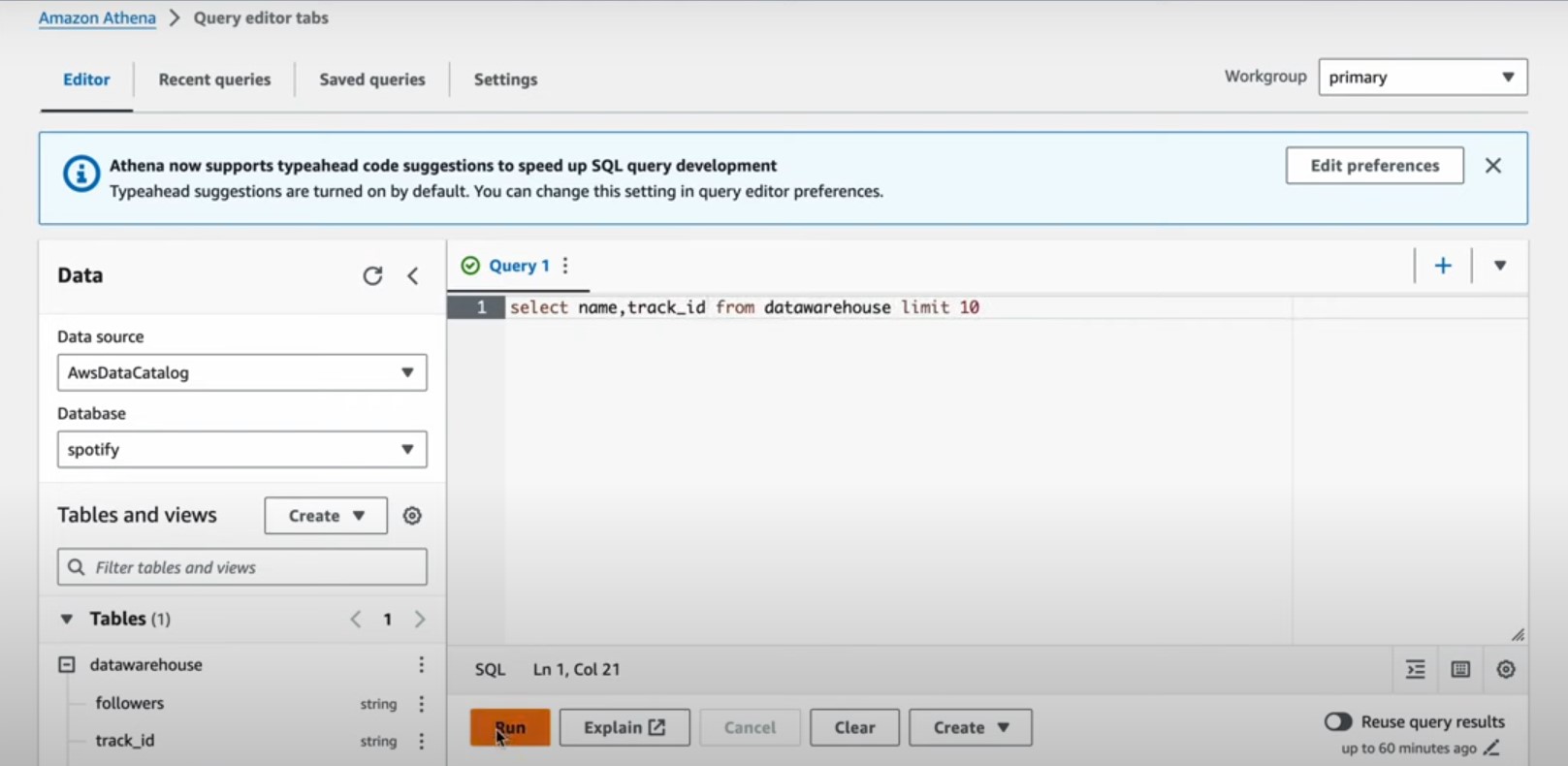
**Query-1**



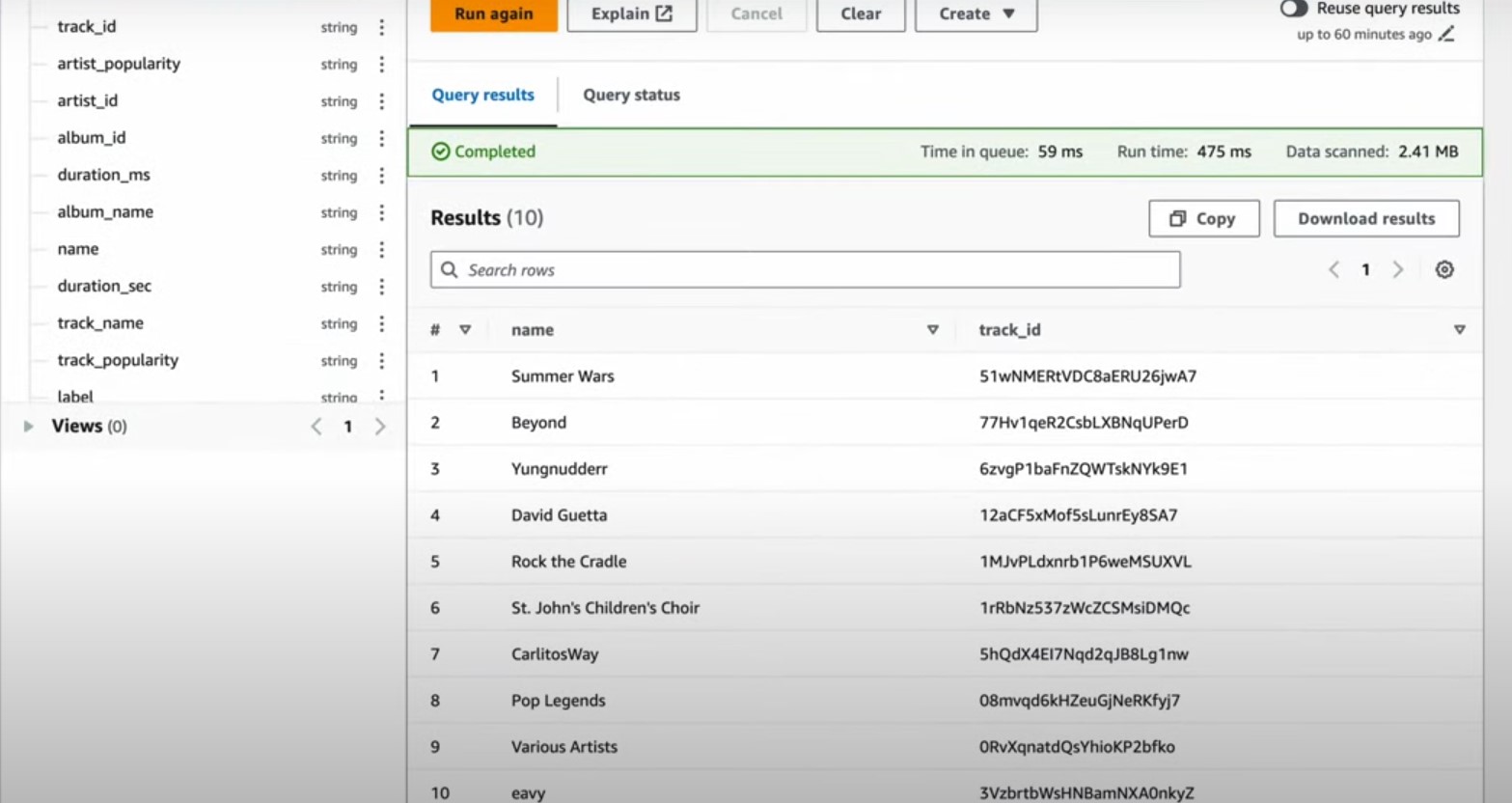
**Output**:



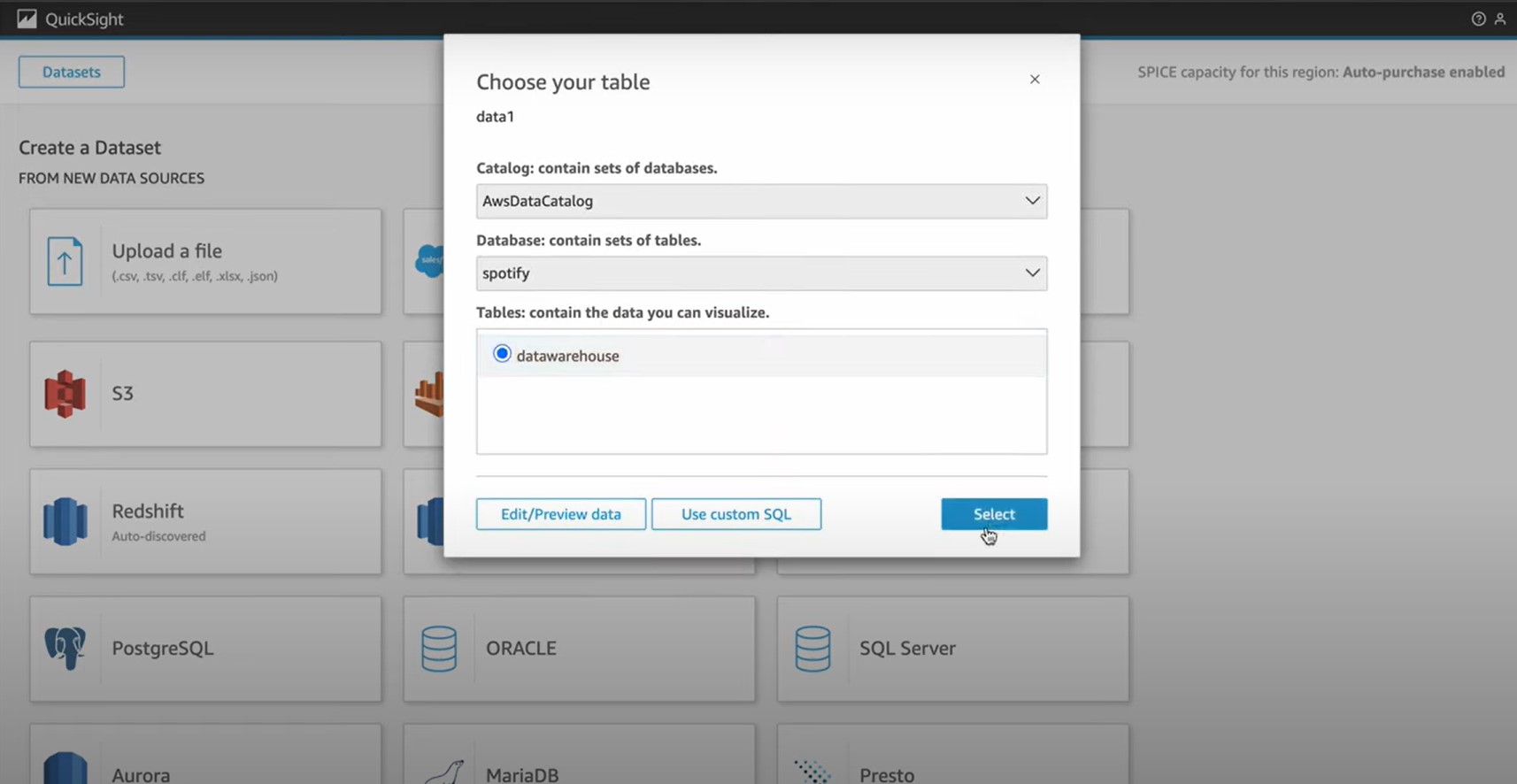
**Query-2**

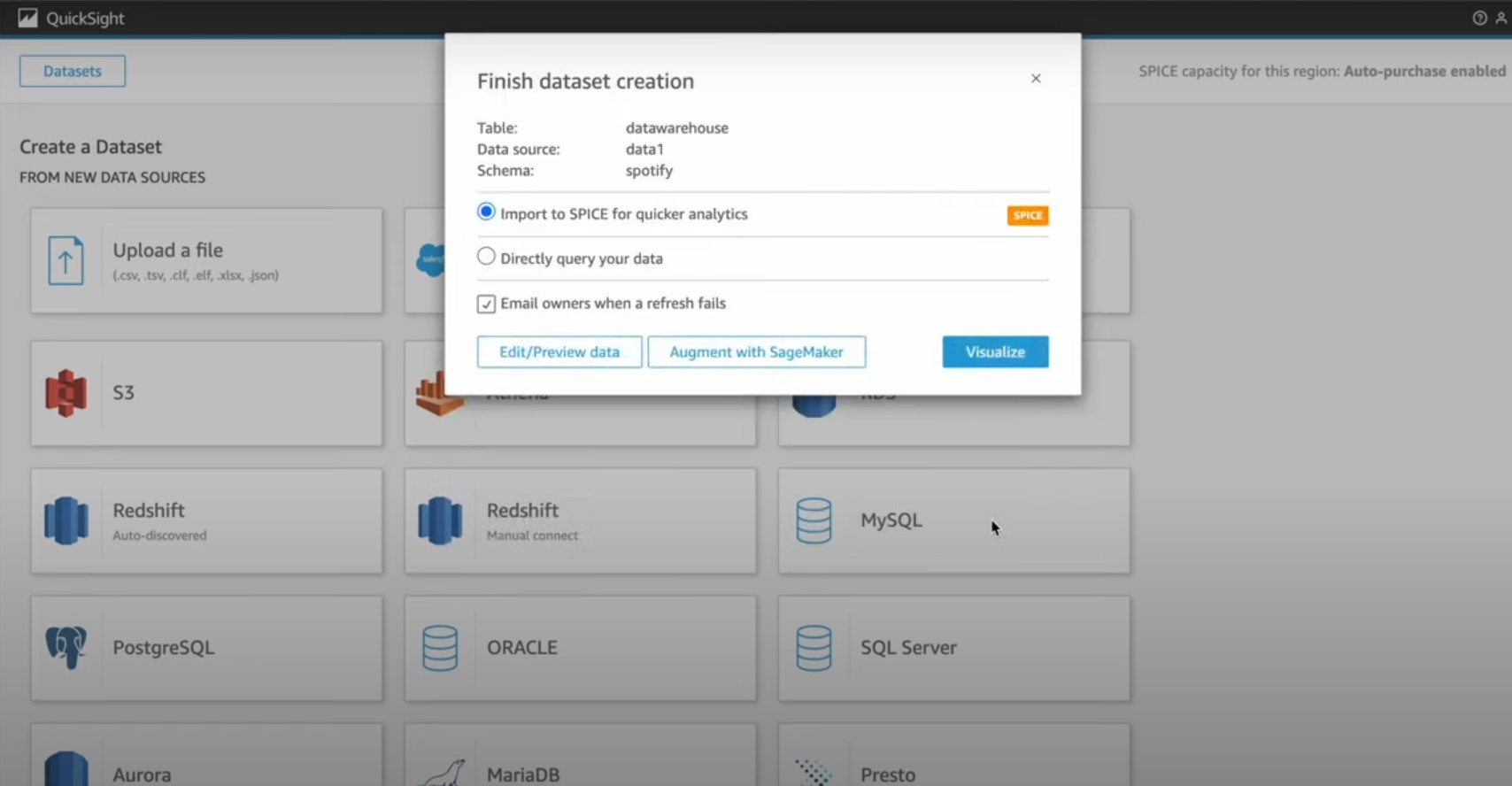


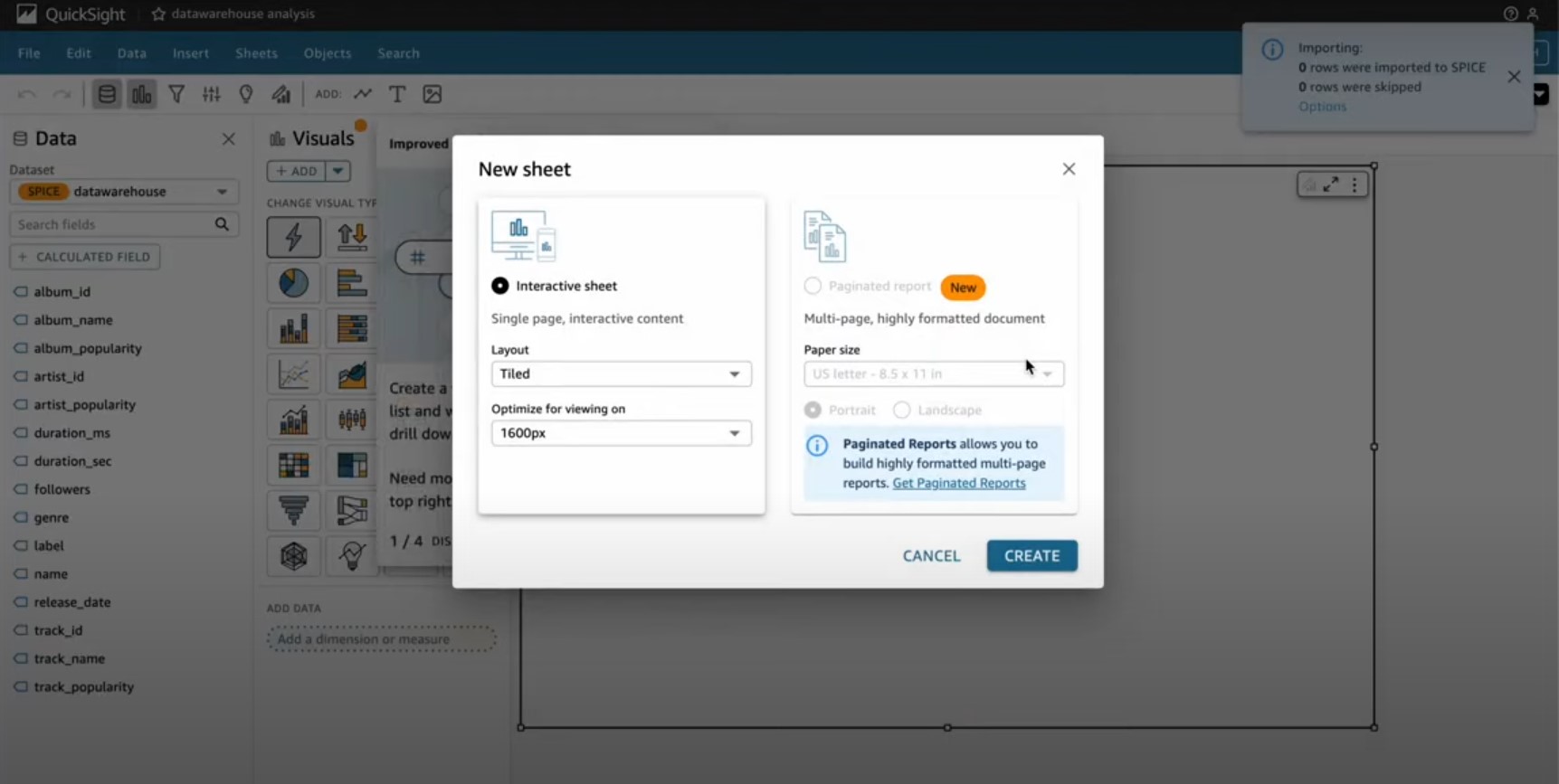
**Output:**



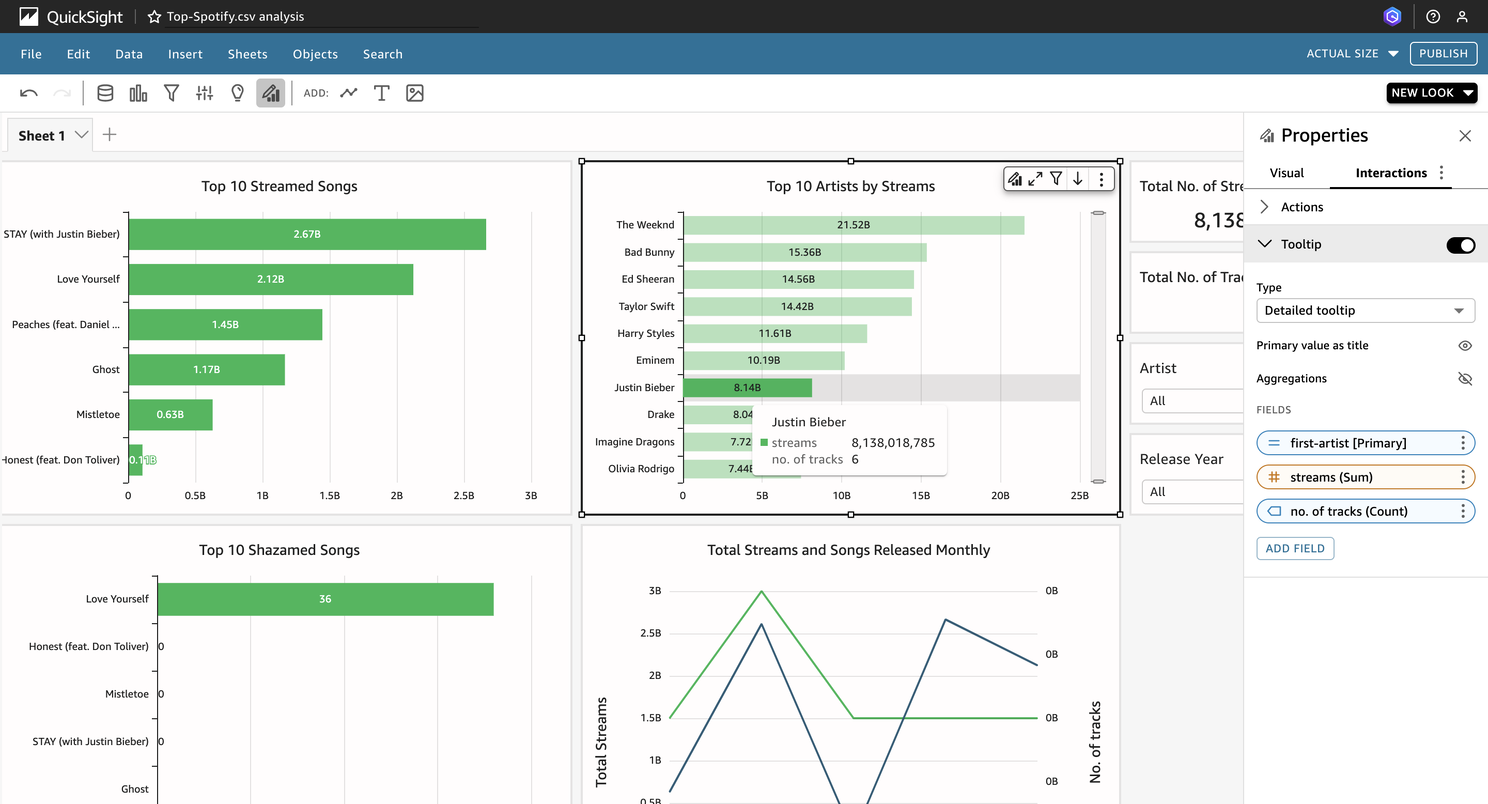
* **Visualization**
  + **Configure AWS QuickSight**:
    - Connect QuickSight to Athena and create visualizations.

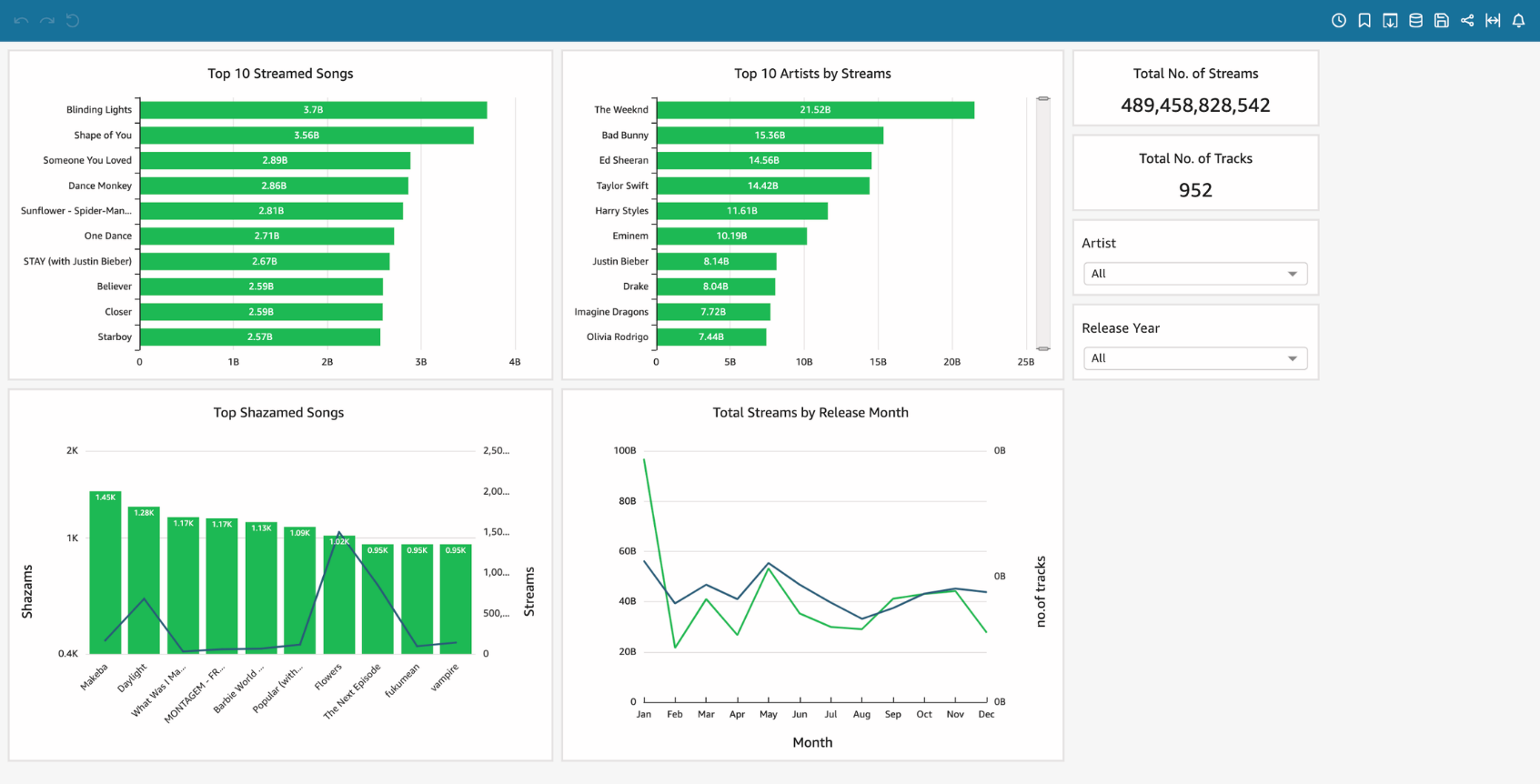


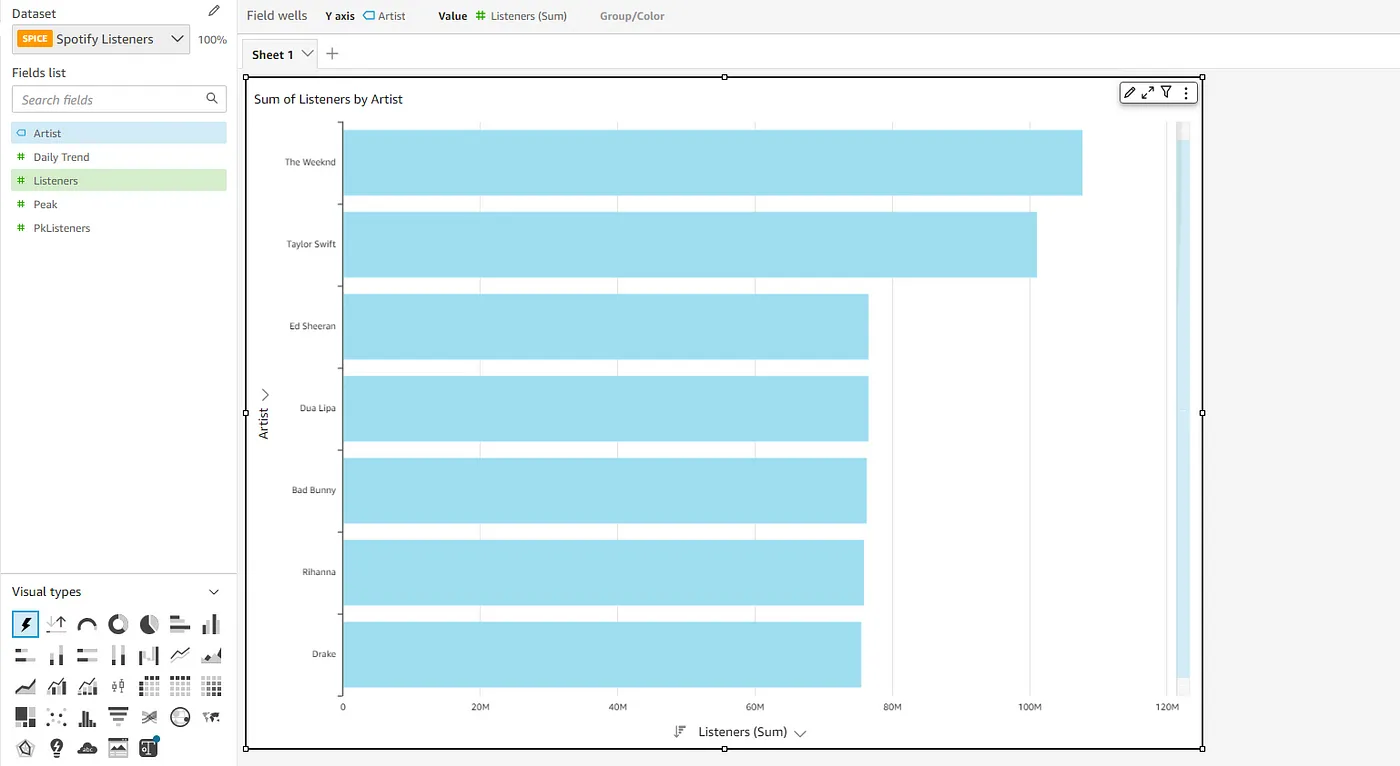


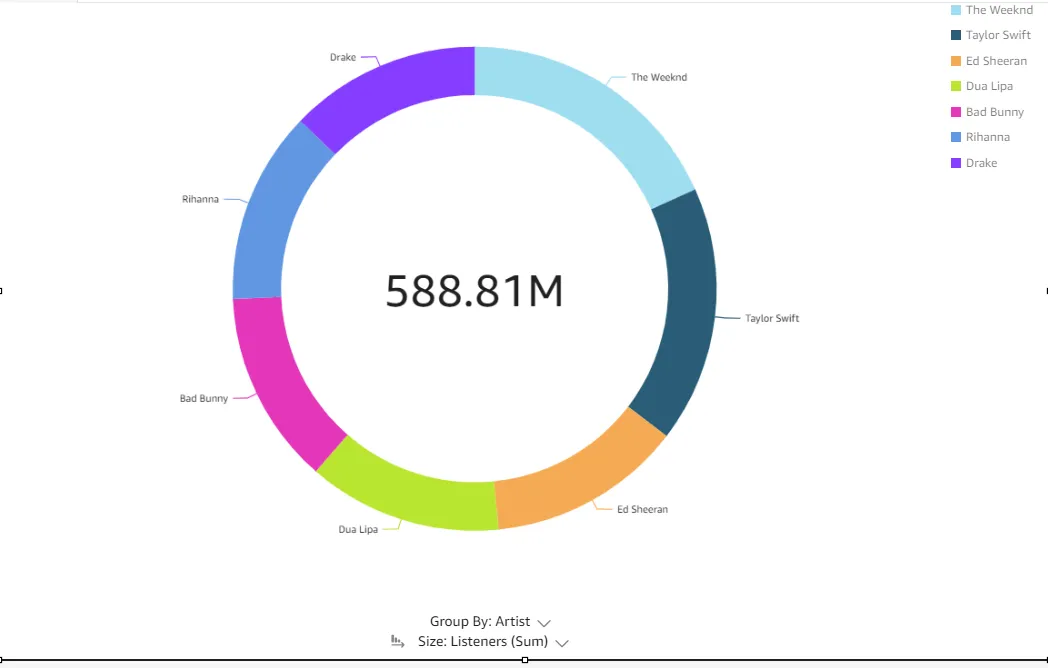


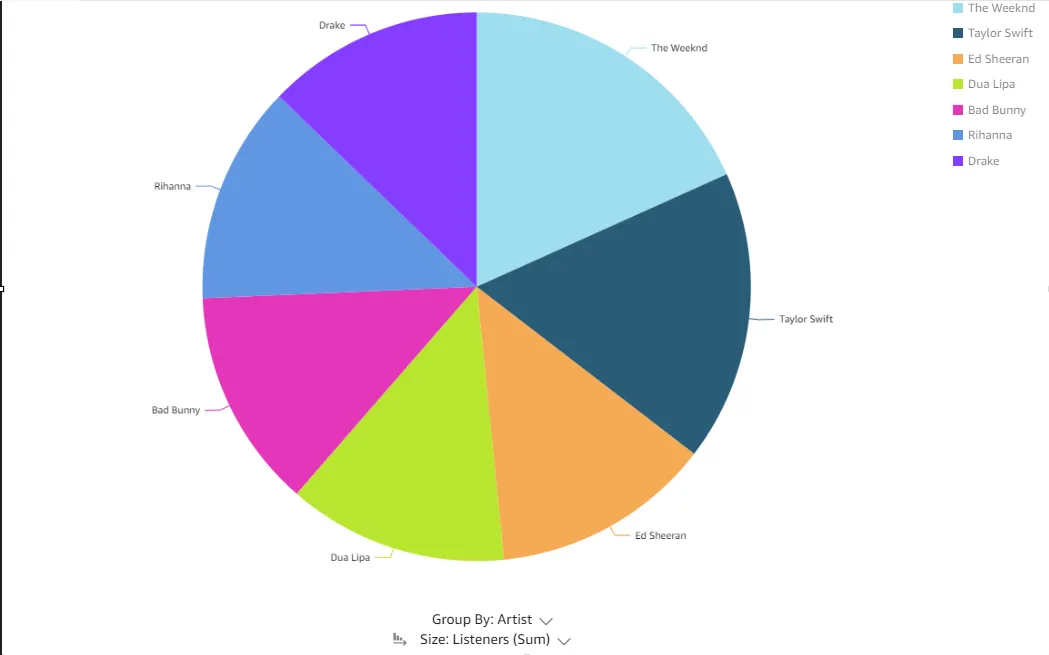
* **Develop Dashboards**:
  + Build interactive dashboards that visualize trends and insights from the Spotify dataset.

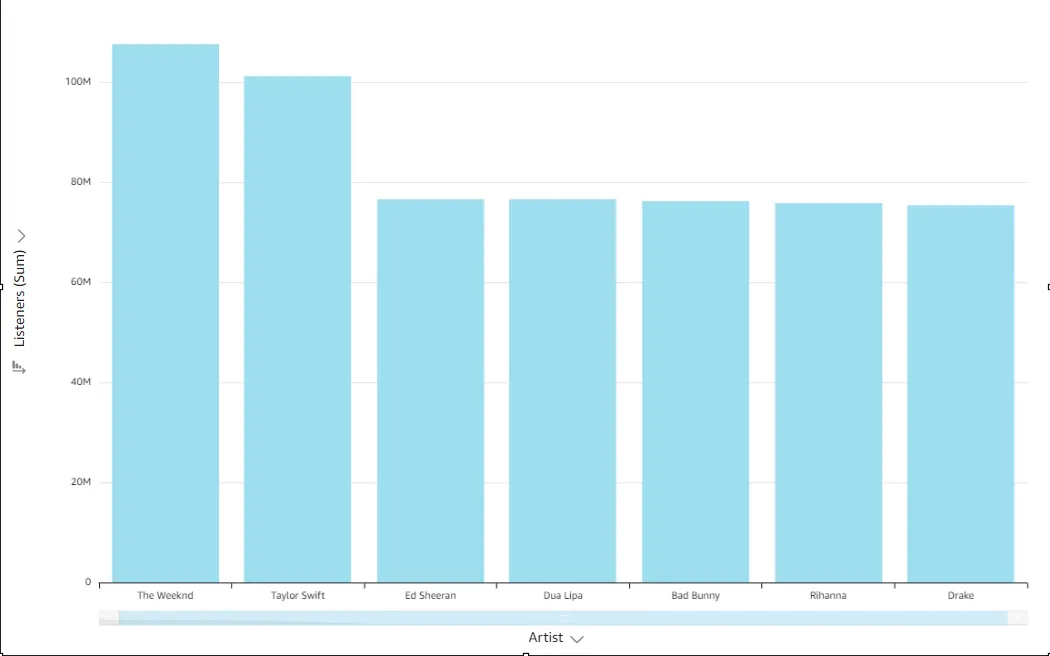


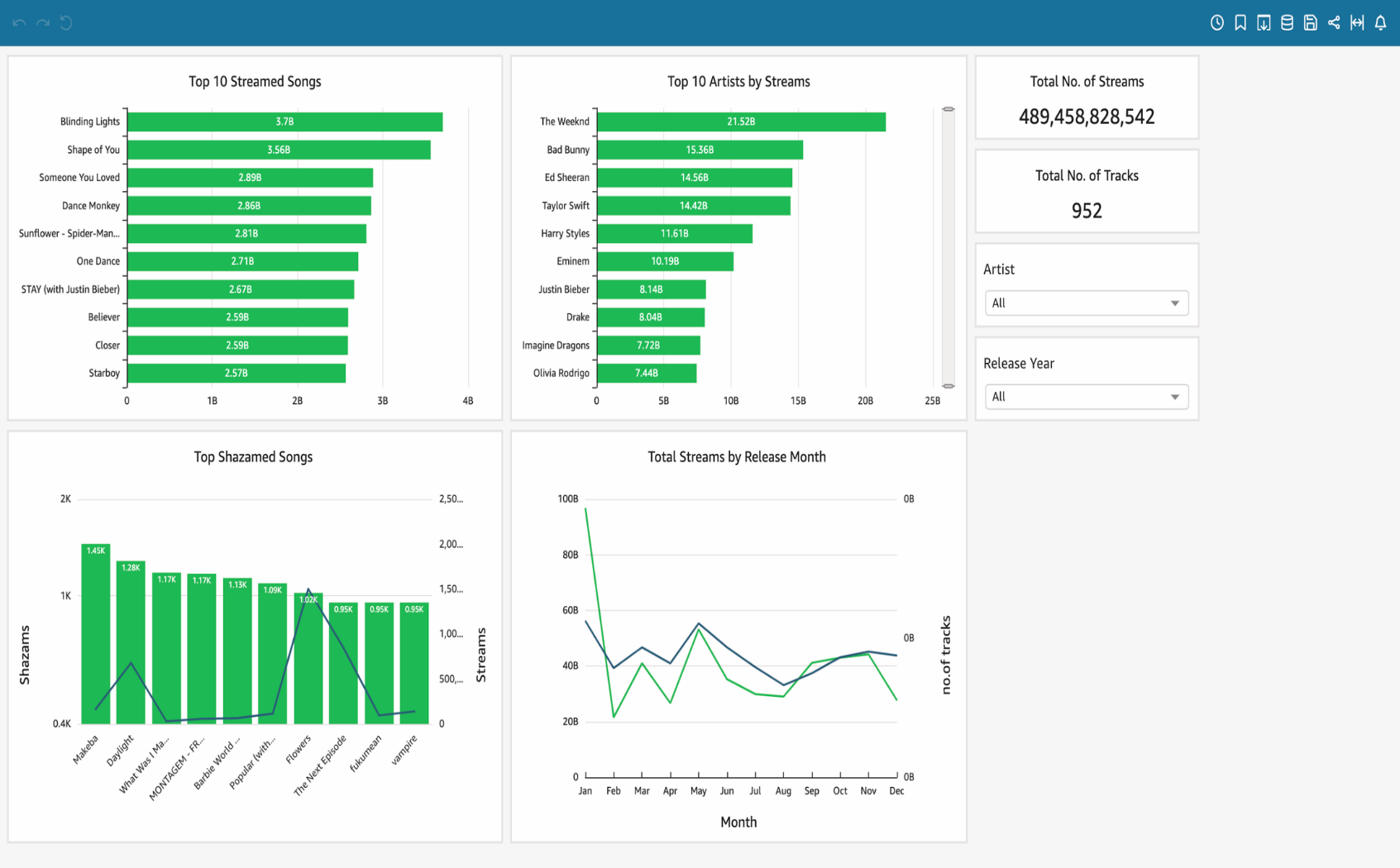


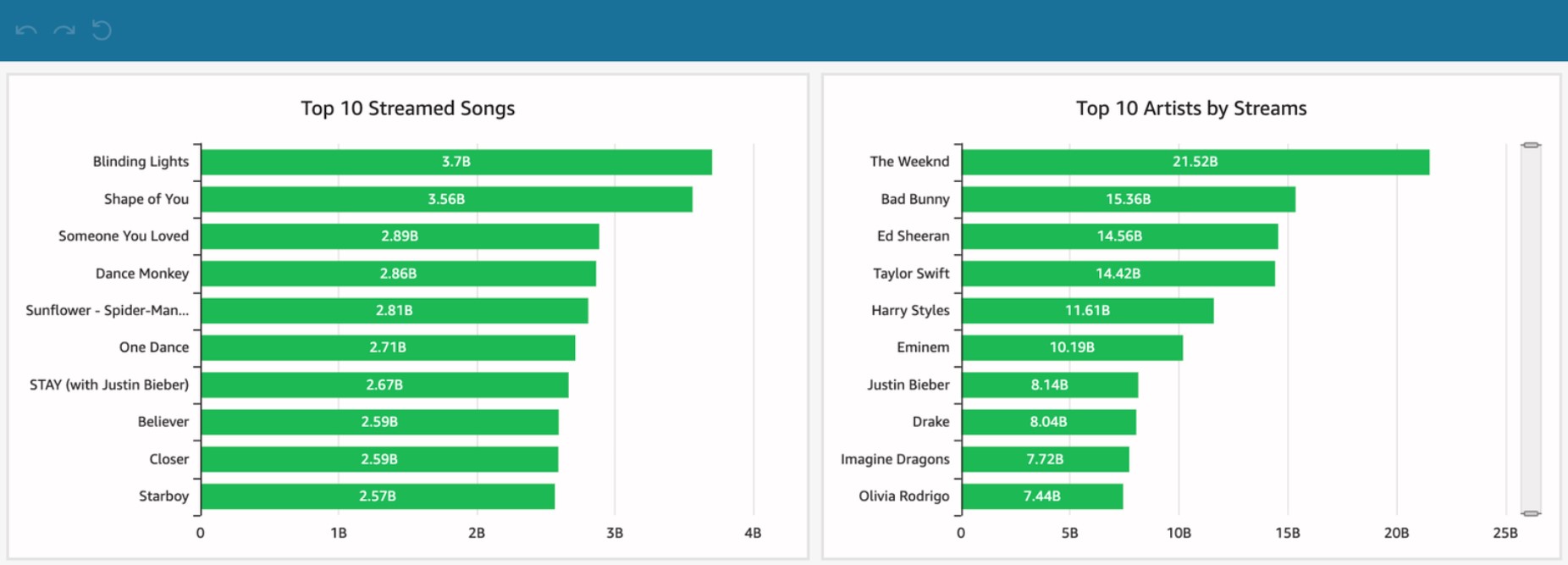


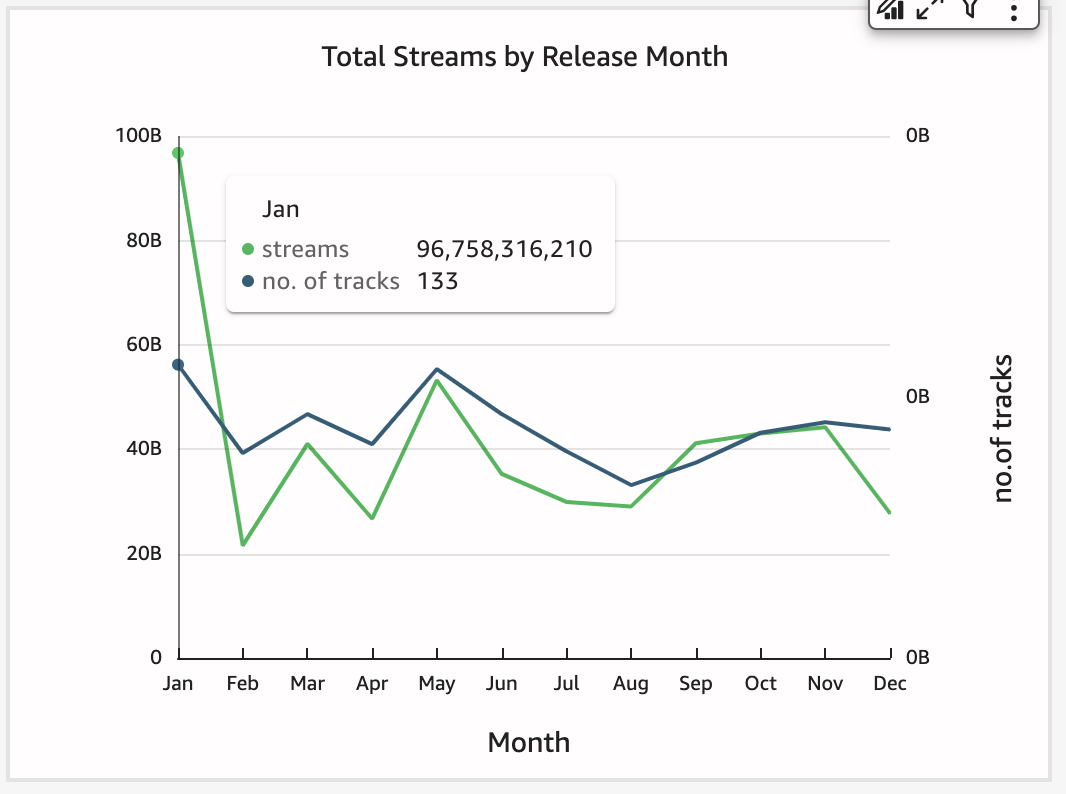








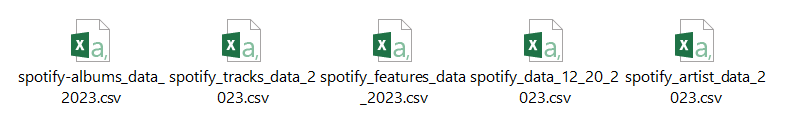




**Dataset Description**

The Spotify dataset consists of five CSV files:

* **Albums**: Details of albums, including release dates and track information.
* **Artists**: Information on artists, including names, number of followers, and genres.
* **Spotify Data**: Metadata on album popularity and related attributes.
* **Spotify Features**: Metrics such as danceability, energy, and loudness.
* **Tracks**: Track IDs, popularity, and explicit content.



**Expected Outcomes:**

1. **Successful Data Ingestion and Storage**: Raw data successfully uploaded to and organized within AWS S3.
2. **Effective ETL Pipeline**: Data transformation from raw to structured format with accurate joins and cleaning.
3. **Queryable Data**: Data available for querying through AWS Athena with correct metadata.
4. **Insightful Visualizations**: Interactive dashboards in AWS QuickSight providing actionable business insights.

**CONCLUSION SUMMARY:**

This project demonstrates a comprehensive end-to-end data engineering solution using AWS tools, providing hands-on experience with data ingestion, transformation, querying, and visualization. The skills developed in this project are essential for modern data engineering roles, emphasizing the importance of cloud-based data solutions and ETL processes.

This project offers a robust and comprehensive solution for data engineering using AWS cloud services and PySpark, aimed at transforming raw Spotify data into actionable business insights. Through meticulous planning and execution, the project achieves its primary objectives of data ingestion, transformation, querying, and visualization, providing a hands-on experience in building a scalable and efficient ETL pipeline.

The project begins by setting up a secure and organized data storage solution in AWS S3, with a clear separation between staging and data warehouse buckets. This structured approach ensures that the raw data is ingested and stored in a way that facilitates efficient processing and management. The use of AWS Glue, a powerful ETL service, allows for seamless data transformation, where raw CSV files are meticulously cleaned, joined, and converted into the Parquet format. This format not only optimizes storage but also enhances query performance, making the data readily available for further analysis.

A significant highlight of this project is the implementation of AWS Glue Crawlers, which automatically scan the data in S3 and create a metadata catalog in AWS Glue Data Catalog. This automation simplifies the management of data schemas and ensures that the data is accurately represented in the catalog, ready for querying. The transformation process is further enhanced by AWS Glue’s visual ETL interface, enabling the drag-and-drop configuration of data sources and transformation logic, which underscores the importance of user-friendly tools in modern data engineering.

Following the transformation, AWS Athena is employed to query the structured data stored in the data warehouse bucket. Athena’s integration with the Glue Data Catalog ensures that the queries are efficient and accurate, providing a streamlined process for extracting meaningful insights. The project demonstrates the power of SQL in querying large datasets, highlighting the importance of having well-structured and clean data to achieve precise analytical outcomes.

The final phase of the project involves data visualization using AWS QuickSight. By connecting QuickSight to Athena, interactive dashboards and reports are created, offering deep insights into various aspects of the Spotify dataset, such as album popularity, artist metrics, and track features. These visualizations not only serve as a tool for data exploration but also provide valuable business insights that can drive decision-making processes.

Overall, this project exemplifies the critical role of data engineering in transforming raw data into valuable business intelligence. By leveraging AWS cloud services, the project showcases the efficiency, scalability, and flexibility of cloud-based data solutions, which are increasingly becoming the standard in the industry. The hands-on experience gained through this project equips individuals with essential skills in data ingestion, transformation, querying, and visualization, all of which are fundamental to modern data engineering roles.

In summary, the project successfully delivers an end-to-end data engineering solution that not only meets the specified objectives but also provides a practical understanding of the tools and techniques required to build and manage complex data pipelines in a cloud environment. The ability to ingest, transform, and visualize data within the AWS ecosystem demonstrates the power of cloud computing in handling large-scale data engineering tasks, making this project a valuable learning experience for anyone pursuing a career in this field.

