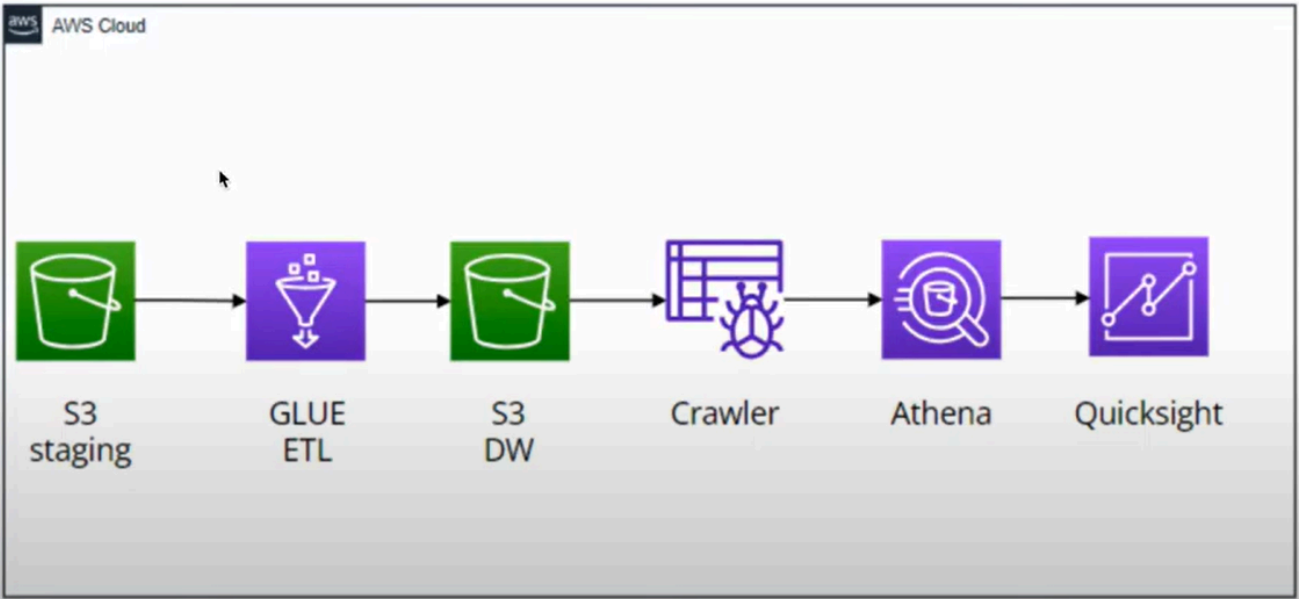


ARCHITECTURE



1. First I created IAM User. gave necessary permissions, Logged in with IAM User

The screenshot shows the AWS IAM console interface. The left sidebar contains navigation links for Identity and Access Management (IAM), Access management, Access reports, CloudShell, and Feedback. The main content area displays the 'Permissions policies (6)' for the user 'spotifyprojectrupak'. The policies are listed in a table with columns for Policy name, Type, and Attached via. The policies are:

Policy name	Type	Attached via
AmazonAthenaFullAccess	AWS managed	Directly
AmazonS3FullAccess	AWS managed	Directly
AmazonS3OutpostsFullAccess	AWS managed	Directly
AWSGlueConsoleFullAccess	AWS managed	Directly
AWSQuicksightAthenaAccess	AWS managed	Directly
IAMUserChangePassword	AWS managed	Directly

Below the table, there is a section for 'Permissions boundary (not set)' and a section for 'Generate policy based on CloudTrail events'.

2. S3 bucket configure and add data to staging

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

Amazon S3

Buckets

spotifyprojectrupak

Amazon S3

General purpose buckets

Directory buckets

Table buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight 11

spotifyprojectrupak

Info

Objects

Metadata

Properties

Permissions

Metrics

Management

Access Points

Objects (2)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

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](#)

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
	Datawarehouse/	Folder	-	-	-
	Staging/	Folder	-	-	-

CloudShellFeedback

© 2025, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

Amazon S3

Buckets

spotifyprojectrupak

Staging/

Amazon S3

General purpose buckets

Directory buckets

Table buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Staging/

Copy S3 URI

Objects

Properties

Objects (3)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

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](#)

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
	albums.csv	csv	April 16, 2025, 20:30:57 (UTC-04:00)	76.3 MB	Standard
	artists.csv	csv	April 16, 2025, 20:30:59 (UTC-04:00)	1.7 MB	Standard
	track.csv	csv	April 16, 2025, 20:31:12 (UTC-04:00)	10.6 MB	Standard

CloudShellFeedback

© 2025, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

3. Create Glue Visual ETL transformation

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

spotify-glue-job

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Successfully updated job

Successfully updated job spotify-glue-job. To run the job choose the Run Job button.

Last modified on 4/16/2025, 9:02:33 PM

Actions

Save

Run

+

Data source - S3 bucket Albums

Data source - S3 bucket Artists

Transform - Join Join Albums and ID

Data source - S3 bucket tracks

Transform - Join Join tracks and Albums

Transform - DropFields Drop Fields

Data target - S3 bucket Destination

CloudShellFeedback

© 2025, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

Added necessary IAM role to it by creation of the role by the root user login.

☰

ⓘ

✕

☑ **Successfully updated job**
Successfully updated job spotify-glue-job. To run the job choose the Run Job button.

spotify-glue-job

Last modified on 4/16/2025, 9:02:33 PM

Actions ▾

Save

Run

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Descriptions can be up to 2048 characters long.

IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

s3_access_glue_spotify ▾

Ⓒ

Type

Always first you select in IAM, “for” what service you have to give access to and then “from” what service.

4. Running pipeline

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

spotify-glue-job

Last modified on 4/16/2025, 9:02:33 PM

Actions

Save

Run

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Job runs (1/1)

Info

Last updated (UTC)

April 17, 2025 at 01:13:19

View details

Stop job run

Troubleshoot with AI

Table View

Card View

Filter job runs by property

<

1

>

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue version
<div></div> Succeeded	0	04/16/2025 21:10:13	04/16/2025 21:12:24	2 m 2 s	3 DPUs	G.1X	4.0

Run details

Input arguments (10)

Continuous logs

Run insights

Metrics

Troubleshooting analysis - preview

Spark UI

Job name

Start time (Local)

Glue version

Last modified on (Local)

CloudShell

Feedback

© 2025, Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie preferences

5. Output is S3 Data warehousing bucket

Amazon S3

General purpose buckets

Directory buckets

Table buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight

Objects

Properties

Objects (8)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

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](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	run-1744852315079-part-block-0-r-00000-snappy.parquet	parquet	April 16, 2025, 21:12:07 (UTC-04:00)	5.8 MB	Standard
<input type="checkbox"/>	run-1744852315079-part-block-0-r-00001-snappy.parquet	parquet	April 16, 2025, 21:12:06 (UTC-04:00)	5.8 MB	Standard
<input type="checkbox"/>	run-1744852315079-part-block-0-r-00002-snappy.parquet	parquet	April 16, 2025, 21:12:07 (UTC-04:00)	5.8 MB	Standard
<input type="checkbox"/>	run-1744852315079-part-block-0-r-00003-snappy.parquet	parquet	April 16, 2025, 21:12:06 (UTC-04:00)	5.8 MB	Standard
<input type="checkbox"/>	run-1744852315079-part-block-0-r-00004-	parquet	April 16, 2025, 21:12:07 (UTC-04:00)	5.8 MB	Standard

6. Now we will run Glue crawler – which will convert our file in datawarehouse to database

Creating crawler

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

AWS Glue > Crawlers > Add crawler

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Zero-ETL integrations

ETL jobs

Visual ETL

Step 1

Set crawler properties

Step 2

Choose data sources and classifiers

Step 3

Configure security settings

Step 4

Set output and scheduling

Step 5

Review and create

Set crawler properties

Crawler details

Name

Enter a unique crawler name

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - optional

Enter a description

Descriptions can be up to 2048 characters long.

Tags - optional

Use tags to organize and identify your resources.

Cancel

Next

Adding source

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

AWS Glue > Crawlers > Add crawler

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Zero-ETL integrations

ETL jobs

Visual ETL

Step 1

Set crawler properties

Step 2

Choose data sources and classifiers

Step 3

Configure security settings

Step 4

Set output and scheduling

Step 5

Review and create

Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

Not yet

Select one or more data sources to be crawled.

Yes

Select existing tables from your Glue Data Catalog.

Data sources (0)

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
You don't have any data sources.		

Add a data source

Custom classifiers - optional

A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Cancel

Previous

Next

Now we need to create a target database after crawling is done hence creating that.

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

AWS Glue

Crawlers

Add crawler

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Zero-ETL integrations

ETL jobs

Visual ETL

Step 4

Set output and scheduling

Step 5

Review and create

Step 2: Choose data sources and classifiers

Edit

Data sources (1) Info

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://spotifyprojectrupak/Datawarehouse...	Recrawl all

Step 3: Configure security settings

Edit

Configure security settings

IAM role	Security configuration	Lake Formation configuration
s3_access_glue_spotify	-	-

Step 4: Set output and scheduling

Edit

Set output and scheduling

Database	Table prefix - optional	Maximum table threshold - optional	Schedule
crawleddatabaserupak	-	-	On demand

Cancel

Previous

Create crawler

Now, I will run the crawler

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

AWS Glue

Crawlers

crawl-s3

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Zero-ETL integrations

ETL jobs

Visual ETL

crawl-s3

Last updated (UTC)

April 17, 2025 at 03:49:35

Run crawler

Edit

Delete

Crawler properties

Name	crawl-s3	IAM role	s3_access_glue_spotify	Database	crawleddatabaserupak	State	READY
Description	-	Security configuration	-	Lake Formation configuration	-	Table prefix	-
Maximum table threshold	-						

Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
<input type="radio"/>	April 17, 2025 at 01:43:51	April 17, 2025 at 01:44:35	43 s	Completed	0.131	1 table change, 0 partition changes

Now lets test the crawler, go in data catalog tables: you will see:

aws

Search

[Alt+S]

United States (Ohio)

rupakkulkarni17

AWS Glue

Tables

datawarehouse

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Zero-ETL integrations

ETL jobs

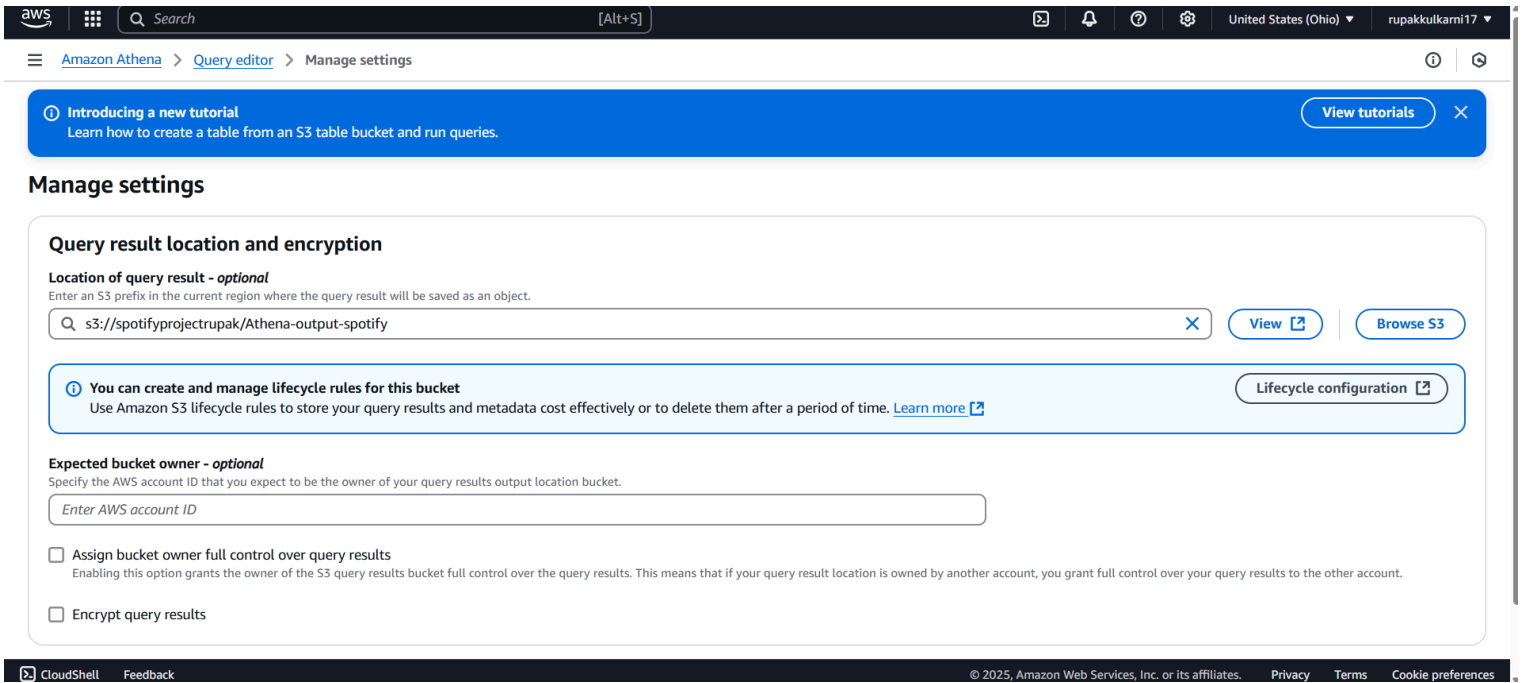
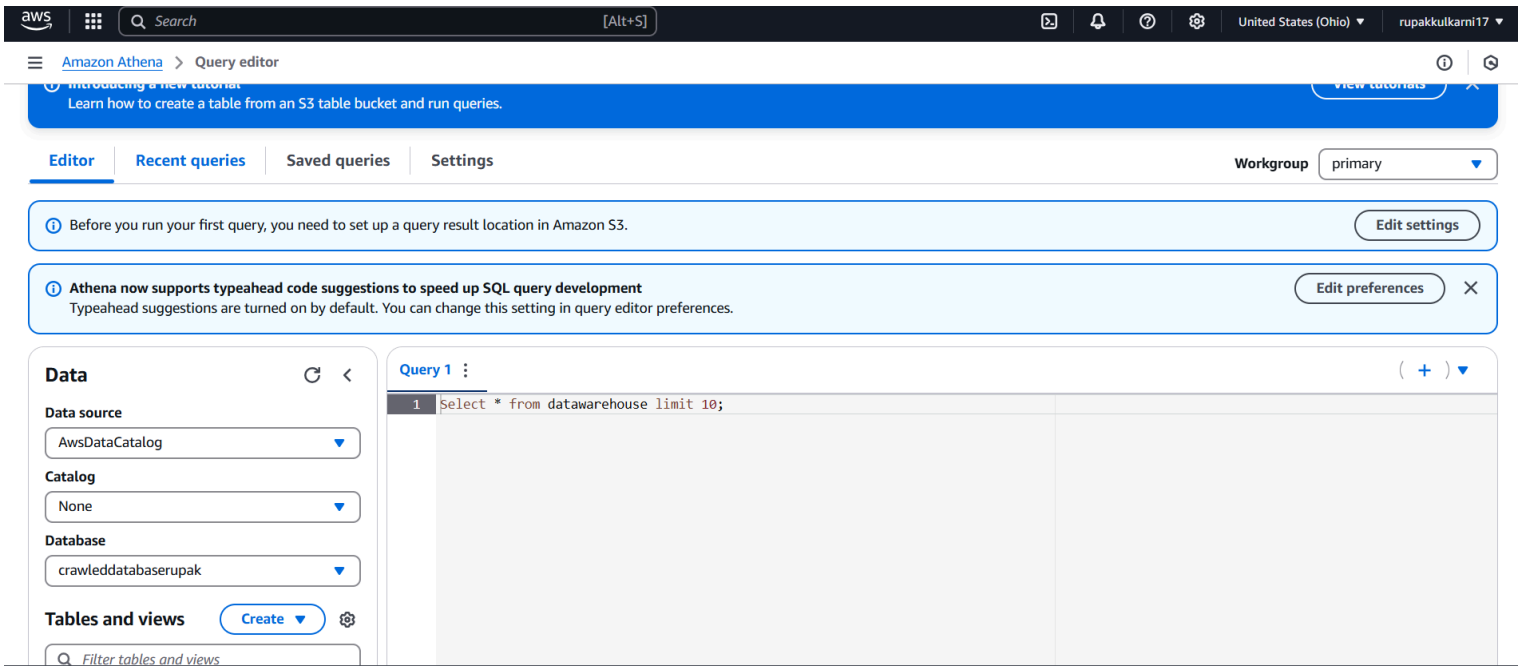
Visual ETL

Filter schemas

#	Column name	Data type	Partition key	Comment
1	followers	string	-	-
2	track_id	string	-	-
3	artist_popularity	string	-	-
4	artist_id	string	-	-
5	album_id	string	-	-
6	duration_ms	string	-	-
7	album_name	string	-	-
8	name	string	-	-
9	duration_sec	string	-	-
10	track_name	string	-	-
11	track_popularity	string	-	-
12	label	string	-	-
13	release_date	string	-	-
14	album_popularity	string	-	-
15	genre	string	-	-

Finally your transformed paraquet data is in the tabular format.

Now will will connect these these to Athena, for Querying (Database querying) – Analyze the data



For storing our query results we need to make 1 new folder in our s3 bucket. Click on settings.

Now lets connect it with Amazon Quicksight, for visualizations.

Issue faced here::: My quicksight region was different. It should be same as athena.

Also i gave access to Quicksight of Athena and all s3 buckets related to this project

QuickSight

Datasets

Create a Dataset

FROM NEW DATA SOURCES

Upload a file

(.csv, .tsv, .clf, .elf, .xlsx, .json)

Athena

MySQL

Aurora

PostgreSQL

MariaDB

Oracle

Presto

Redshift

Manual connect

SQL Server

Spark

Choose your table

data1

Catalog: contain sets of databases.

AwsDataCatalog

Database: contain sets of tables.

crawleddatabaserupak

Tables: contain the data you can visualize.

☐ datawarehouse

Edit/Preview data

Use custom SQL

Select

SPICE capacity for this region: 0 bytes of 0 bytes

S3

Redshift

Manual connect

SQL Server

Spark

QuickSight

datawarehouse analysis

File Edit Data Insert Sheets Objects Search

Dataset

SPICE datawarehouse

Search fields

+ CALCULATED FIELD

album_id

album_name

album_popularity

artist_id

artist_popularity

duration_ms

duration_sec

followers

genre

label

name

release_date

track_id

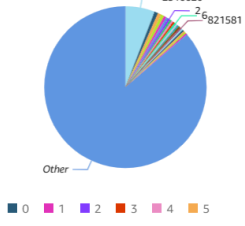
track_name

track_popularity

Sheet 1

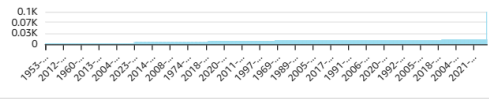
Count of Genre by Followers

SHOWING TOP 20 IN FOLLOWERS



Count of Genre by Release_date

SHOWING BOTTOM 2500 IN RELEASE_DATE



Count of Genre by Album_popularity

