

An ETL Data Pipeline with Storage layer, ETL (processing and datastore) layer and visualization layer:

Storage Layer – S3

- 1) Create S3 bucket and a folder in it, then upload dataset to the created folder:

[Amazon S3](#) > [Buckets](#) > [Create bucket](#)

Create bucket [Info](#)

Buckets are containers for data stored in S3. [Learn more](#)

General configuration

AWS Region

Asia Pacific (Sydney) ap-southeast-2

Bucket name [Info](#)

mod6etlbucket

Bucket name must be unique within the global namespace and follow the bucket naming rules. [See rules for](#)

► Advanced settings

After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.

Cancel

Create bucket

General purpose buckets (3) [Info](#)

Buckets are containers for data stored in S3. [Learn more](#)

	Name	AWS
<input type="radio"/>	mod6etlbucket	Asia
<input type="radio"/>	s3jzm	Asia
<input type="radio"/>	s3q2	Asia

mod6etlbucket

Info

Objects

Properties

Permissions

Metrics

Management

Access Points

Objects (0)

Info

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

<

1

>

Name

Type

Last modified

Size

Storage class

No objects

Folder

Folder name

education_data

/

Folder names can't contain "/". [See rules for naming](#)

Server-side encryption

Info

Server-side encryption protects data at rest.

The following encryption settings apply only to the folder object and not to sub-folder objects.

Server-side encryption

☒ Do not specify an encryption key

The bucket settings for default encryption are used to encrypt the folder object when storing it in Amazon S3.

☐ Specify an encryption key

The specified encryption key is used to encrypt the folder object before storing it in Amazon S3.

If your bucket policy requires objects to be encrypted with a specific encryption key, you must specify the same encryption key when you create a folder. Otherwise, folder creation will fail.

Cancel

Create folder

mod6etlbucket

Info

Objects

Properties

Permissions

Metrics

Managem

Objects (1)

Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inve](#)

Find objects by prefix

Name

▲

Type

education_data/

Folder

education_data/

Copy S3 URI

Objects

Properties

Objects (0)

Info

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

< 1 >

Name

▲

Type

▼

Last modified

▼

Size

▼

Storage class

▼

No objects

Upload [Info](#)

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 9.0 KB)

Remove

Add files

Add folder

All files and folders in this table will be uploaded.

< 1 >

<input checked="" type="checkbox"/>	Name	Folder
-------------------------------------	------	--------

<input checked="" type="checkbox"/>	education.csv	-
-------------------------------------	---------------	---

Destination [Info](#)

Destination

[s3://mod6etlbucket/education_data/](#)

► Destination details

Bucket settings that impact new objects stored in the specified destination.

► Permissions

Grant public access and access to other AWS accounts.

► Properties

Specify storage class, encryption settings, tags, and more.

Cancel

Upload

Summary

Destination
s3://mod6etlbucket/education_data/

Succeeded
1 file, 9.0 KB

Files and folders

Configuration

Files and folders (1 Total, 9.0 KB)

Find by name

Name	Folder	Type	Size	Status	Error
education.csv	-	text/csv	9.0 KB	Succeeded	-

ETL Layer – AWS Glue and Athena

2) Set up AWS Glue IAM role for S3 data read/write:

AWS Glue

Sydney

Welcome to AWS Glue
Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.

Prepare your account for AWS Glue
Admins: Grant access to AWS Glue and set a default IAM role.
Set up roles and users

Catalog and search for datasets
View your databases & tables and catalog data using Crawlers.
Go to the Data Catalog

Move and transform data
Transform data using a visual, notebook, or code interface.
Author and edit ETL jobs

Selected roles (0)
Choose the roles you want to have access to AWS Glue.

Remove Choose roles

< 1 > ⚙

Role name

Creation date (UTC)

No roles selected

Select IAM roles

All roles (9)
All roles across your organization.

Refresh Create IAM role

Find role

< 1 > ⚙

Role name	Creation date (UTC)
AWSServiceRoleForApplicationAutoScaling_DynamoDBTable	January 31, 2024 at 15:14...
AWSServiceRoleForSupport	January 13, 2024 at 05:35...

Identity and Access Management (IAM)

Search IAM

Dashboard

Access management

User groups

Users

Roles

Policies

IAM > Dashboard

IAM Dashboard

Security recommendations **2**

⚠️ Add MFA for root user

Add MFA for root user - Enable multi-factor authentication for this account.

⚠️ Deactivate or delete access keys for root user

Deactivate or delete the access keys for the root user to an IAM user to improve security.

IAM > Roles

Roles (9) Info

Refresh

Delete

Create role

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Search

< 1 > ⚙️

<input type="checkbox"/>	Role name	Trusted entities	Last activity
<input type="checkbox"/>	AWSServiceRoleForApplicationAutoScaling_DynamoDBTable	AWS Service: dynamodb.application-	12 hours ago
<input type="checkbox"/>	AWSServiceRoleForSupport	AWS Service: support/Service-Linker	-

glue

Other services

DataBrew

Glue

Choose a service or type




Cancel

Next

Permissions policies (2/912) [Info](#)

Choose one or more policies to attach to your new role.




× Filter by Type All types

<input type="checkbox"/>	<input type="checkbox"/>	Policy name ↗	Type
<input type="checkbox"/>	<input type="checkbox"/>	 AmazonDMSRedshiftS3Role	AWS managed
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 AmazonS3FullAccess	AWS managed
<input type="checkbox"/>	<input type="checkbox"/>	 AmazonS3ObjectLambdaExecutionRol...	AWS managed

Permissions policies (2/912) [Info](#)

Choose one or more policies to attach to your new role.

× Filter by Type All types 3 matches < 1 > ⚙

<input type="checkbox"/>	<input type="checkbox"/>	Policy name ↗	Type	Description
<input type="checkbox"/>	<input type="checkbox"/>	 AmazonSageMakerServiceCatalogProd...	AWS managed	Service role policy used by the AWS Glue...
<input type="checkbox"/>	<input type="checkbox"/>	 AWSGlueServiceNotebookRole	AWS managed	Policy for AWS Glue service role which all...
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 AWSGlueServiceRole	AWS managed	Policy for AWS Glue service role which all...

▶ Set permissions boundary - optional

Cancel Previous Next

Name, review, and create

Role details

Role name

Enter a meaningful name to identify this role.

Maximum 64 characters. Use alphanumeric and '+=,.,@-_' characters.

Step 2: Add permissions

Edit

Permissions policy summary

Policy name	Type	Attached as
AmazonS3FullAccess	AWS managed	Permissions policy
AWSGlueServiceRole	AWS managed	Permissions policy

Step 3: Add tags

Add tags - optional [Info](#)

Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.

No tags associated with the resource.

Add new tag

You can add up to 50 more tags.

Cancel

Previous

Create role

Role Mod6_Glue-Visual-ETL_role created.

[IAM](#) > Roles

Roles (10) [Info](#)

An IAM role is an identity you can create that has spec

mod6

☐

Role name

☐

[Mod6_Glue-Visual-ETL_role](#)

Back to AWS Glue console, apply the created role:

Select IAM roles

All roles (10)
All roles across your organization.

Find role

< 1 >

<input type="checkbox"/>	Role name	Creation date (UTC)
<input type="checkbox"/>	AWSServiceRoleForApplicationAutoScaling_DynamoDBTable	January 31, 2024 at 15:14...
<input type="checkbox"/>	AWSServiceRoleForSupport	January 13, 2024 at 05:35...
<input type="checkbox"/>	AWSServiceRoleForTrustedAdvisor	January 13, 2024 at 05:35...
<input type="checkbox"/>	EC2_Q1-Kinesis_role	January 28, 2024 at 06:00...
<input type="checkbox"/>	Ec2Kinesisagentrole	January 14, 2024 at 02:35...
<input type="checkbox"/>	KinesisFirehoseServiceRole-KDS-S3-S-ap-southeast-2-1706439300041	January 28, 2024 at 11:03...
<input type="checkbox"/>	KinesisFirehoseServiceRole-KDS-S3-Y-ap-southeast-2-1705209477854	January 14, 2024 at 05:20...
<input type="checkbox"/>	Mod5_SQS-Lambda_Role	January 31, 2024 at 13:55...
<input checked="" type="checkbox"/>	Mod6_Glue-Visual-ETL_role	February 1, 2024 at 05:44:...
<input type="checkbox"/>	PutEventFunction-role-fa9c1y3j	January 29, 2024 at 13:00...

Cancel

Confirm

Selected roles (1)
Choose the roles you want to have access to AWS Glue.

Remove

Choose roles

< 1 >

<input type="checkbox"/>	Role name	Creation date (UTC)
<input type="checkbox"/>	Mod6_Glue-Visual-ETL_role	February 1, 2024 at 05:44:47

Selected users (0)
Choose the users you want to have access to AWS Glue.

Remove

Choose users

< 1 >

<input type="checkbox"/>	User name	Console last activity	Creation date (UTC)
No users selected			

Choose users

Cancel

Next

Choose S3 locations

Targeted users and roles
0 users and 1 roles

Choose access to Amazon S3

☐ No additional access
Do not change permissions.

☐ Add access to specific Amazon S3 locations
Choose specific S3 paths that you want to grant access to.

☒ Grant full access to Amazon S3
Grant access to all S3 resources in your AWS account.

Data access permissions

Set the type of data access for the Glue users and roles.

Data access permissions

☐ Read only (recommended)

☒ Read and write

Cancel Previous **Next**

Choose a default AWS Glue service role

IAM role for AWS Glue

☒ Create the standard AWS Glue service role and set it as the default (recommended)
AWS will create an IAM role with the IAM policies needed to run AWS Glue jobs, then set it as the default.

☐ Set an existing IAM role as the default
Select an IAM role that you've configured to use as an AWS Glue service role. Glue will set this role as the default, but won't add any permissions to it. [Learn more](#)

☒ AWS will create and configure this IAM role for you:

- AWSGlueServiceRole

Cancel Previous **Next**

Step 3: Choose a default service role

[Edit](#)

Choose a default AWS Glue service role

Selected service role
AWSGlueServiceRole

Cancel Previous **Apply changes**

3) Configure query output location in Athena:

Athena

Sydney

Analytics

Amazon Athena

Start querying data instantly.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 and other federated data sources using standard SQL.

Get started

☒ Query your data with Trino SQL
Use Query editor to analyze data on S3, on premises, or on other clouds.

☐ Analyze your data using PySpark and Spark SQL
Use notebooks to build interactive Spark applications.

Launch query editor

Amazon Athena

>

Query editor

Editor

Recent queries

Saved queries

Settings

Workgroup primary

Query result and encryption settings

Manage

Query result location and encryption

Query result location

Encrypt query results

Expected bucket owner

Assign bucket owner full control over query results

-

-

-

Turned off

Amazon Athena

>

Query editor

>

Manage settings

Manage settings

Query result location and encryption

Location of query result - optional

Enter an S3 prefix in the current region where the query result will be saved as an object.

Q

s3://mod6etlbucket

X

View

Browse S3

You can create and manage lifecycle rules for this bucket

Use Amazon S3 lifecycle rules to store your query results and metadata cost effectively or to delete them after a period of time.

[Learn more](#)

Lifecycle configuration

Expected bucket owner - optional

Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

Enter AWS account ID

☐ Assign bucket owner full control over query results

Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

☐ Encrypt query results

Cancel

Save

4) At Athena, create table catalogue for the dataset:

Editor Recent queries Saved queries Settings

Data Query 2

Data source: AwsDataCatalog

Database: Choose a database

Tables and views: Create ▲

Filter tables and views

Tables (0)

Views (0)

Create a table from data source

S3 bucket data

AWS Glue Crawler

Create with SQL

Set crawler properties

Crawler details info

Name: education_crawler

Description - optional: Enter a description

Tags - optional: Use tags to organize and identify your resources.

Cancel Next

Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

Not yet (selected) Yes

Data sources (0) info

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

Add a data source

You don't have any data sources.

Add a data source

Data source configuration cannot be empty.

Add data source

Data source

Choose the source of data to be crawled.

S3

Network connection - optional

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Clear selection

Add new connection

Location of S3 data

In this account

In a different account

S3 path

Browse for or enter an existing S3 path.

s3://mod6etlbucket/education_data/

View

Browse S3

All folders and files contained in the S3 path are crawled. For example, type s3://mybucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs

This field is a global field that affects all S3 data sources.

Crawl all sub-folders

Crawl all folders again with every subsequent crawl.

Crawl new sub-folders only

Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

Crawl based on events

Rely on Amazon S3 events to control what folders to crawl.

Sample only a subset of files

Exclude files matching pattern

Cancel

Add an S3 data source

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 (1)

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

Type	Data source	Parameters
S3	s3://mod6etlbucket/education_data/	Recrawl all

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

Configure security settings

IAM role [Info](#)

Existing IAM role

Mod6_Glue-Visual-ETL_role

View

Create new IAM role

Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - optional
Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more.](#)

☐ Use Lake Formation credentials for crawling S3 data source
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

Security configuration - optional
Enable at-rest encryption with a security configuration.

Cancel

Previous

Next

Set output and scheduling

Output configuration [Info](#)

Target database

Choose a database

Clear selection

Add database

Table name prefix - optional

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name
education_database

Database name is required, in lowercase characters, and no longer than 255 characters.

Location - optional
Set the URI location for use by clients of the Data Catalog.

Description - optional
Enter text

Descriptions can be up to 2048 characters long.

Cancel

Create database

Set output and scheduling

Output configuration [Info](#)

Target database

education_database

▼

↺

Clear selection

Add database [↗](#)

Table name prefix - optional

Type a prefix added to table names

Maximum table threshold - optional

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

▶ Advanced options

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like [cron](#) syntax. [Learn more](#)

Frequency

On demand

Cancel

Previous

Next

Review and create

Step 1: Set crawler properties

Edit

Set crawler properties

Name

education_crawler

Description

-

Tags

-

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://mod6etlbucket/education_data/	Recrawl all

Step 3: Configure security settings

Edit

Configure security settings

IAM role

Mod6_Glue-Visual-ETL_role

Security configuration

-

Lake Formation configuration

-

Step 4: Set output and scheduling

Edit

Set output and scheduling

Database

education_database

Table prefix - optional

-

Maximum table threshold - optional

-

Schedule

On demand

Cancel

Previous

Create crawler

education_crawler

Last updated (UTC)
February 1, 2024 at 11:44:00

↺

Run crawler

Crawler properties

Name

education_crawler

IAM role

Mod6_Glue-Visual-ETL_role [↗](#)

Database

education_database

State

READY

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
February 2, 2024 at 13:29:28	February 2, 2024 at 13:30:23	55 s	Completed	-	1 table change, 0 partition changes

5) Back to Athena to query data:

The screenshot shows the AWS Athena Editor interface. The 'Editor' tab is selected and highlighted with a red box. The 'Data' section is active, showing the 'Database' dropdown set to 'education_database' (highlighted with a red box). Below this, the 'Tables and views' section shows a list of tables. The 'education_data' table is selected and highlighted with a red box. The table schema is displayed below it, with columns 'datasrno', 'workex', and 'gmat' all of type 'bigint' (highlighted with a green box). The 'Query 2' editor is visible on the right, showing the SQL editor with 'Ln 1, Col 1' and buttons for 'Run' and 'Expla'.

Editor | Recent queries | Saved queries | Settings

Data  < **Query 2** ⋮

1

Data source
AwsDataCatalog ▼

Database
education_database ▼

Tables and views **Create** ▼ ⚙️

🔍 Filter tables and views

▼ **Tables** (1) < 1 >

education_data ⋮

- datasrno bigint ⋮
- workex bigint ⋮
- gmat bigint ⋮

► **Views** (0) < 1 >

SQL Ln 1, Col 1

Run **Expla**

Query results

Editor Recent queries Saved queries Settings

Data

Data source
AwsDataCatalog

Database
education_database

Tables and views Create ▼

Filter tables and views

▼ Tables (1) < 1

education_data

- datasrno bigint
- workex bigint
- gmat bigint

► Views (0) < 1 >

Run Query

- Preview Table
- Generate table DDL

Insert

- Insert into editor

Manage

- Delete table
- Generate statistics - new
- View properties
- View in Glue

SQL Ln 1, Col 1

Run Explain

Query results Query sta

Query command:

Query 2 : X Query 3 : X

```
SELECT * FROM "education_database"."education_data" limit 10;
```

Query output: Success, returned with 10 rows of records as a preview.

Query results

Query stats

Completed

Time in queue: 96 ms

Results (10)

Search rows

#	datasrno	workex	gmat
1	1	21	720
2	2	107	640
3	3	57	740
4	4	99	690
5	5	208	710
6	6	136	660
7	7	70	660
8	8	103	710
9	9	79	700
10	10	22	730

Visualization Layer - AWS QuickSight

6) Data visualization with AWS QuickSight:

(i) Load dataset to QuickSight:

QuickSight

Search results for 'QuickSight'

Services (1)

Features (1)

Resources **New**

Documentation (17,096)

Services

QuickSight

Fast, easy to use business analytics

Analyses

Last updated (newest first)

New analysis

Analysis

People Overview analysis

SAMPLE

☆

⋮

Analysis

Web and Social Media Anal...

SAMPLE

☆

⋮

Analysis

Sales Pipeline analysis

SAMPLE

☆

⋮

Analysis

Business Review analysis


SAMPLE

☆

⋮


New dataset


Your Datasets


 Business Review
SPICE


Create a Dataset

FROM NEW DATA SOURCES

 Upload a file
(.csv, .tsv, .clf, .elf, .xlsx, .json)

 Salesforce
Connect to Salesforce

 Athena

 RDS

New Athena data source ×

Data source name

query3

Athena workgroup

[primary] ▼

✔ Validated

SSL is enabled

Create data source

Choose your table

query3

Catalog: contain sets of databases.

AwsDataCatalog

Database: contain sets of tables.

education_database

Tables: contain the data you can visualize.

education_data

Edit/Preview data

Use custom SQL

Select

Finish dataset creation

Table: education_data

Data source: query3

Schema: education_database

Import to SPICE for quicker analytics

SPICE

Directly query your data


☒ Email owners when a refresh fails

Edit/Preview data

Augment with SageMaker

Visualize

New sheet



☒ Interactive sheet


Single page, interactive content

Layout

Tiled

Optimize for viewing on

1600px



☐ Paginated report

New


Multi-page, highly formatted document

Paper size

US letter - 8.5 x 11 in

☒ Portrait

☐ Landscape



Paginated Reports allows you to build highly formatted multi-page reports. [Get Paginated Reports](#)

CANCEL

CREATE


(ii) Visualize data in QuickSight:

Histogram of 'gmat' column:

Visuals

+ ADD

CHANGE VISUAL TYPE

 Histogram

VALUE

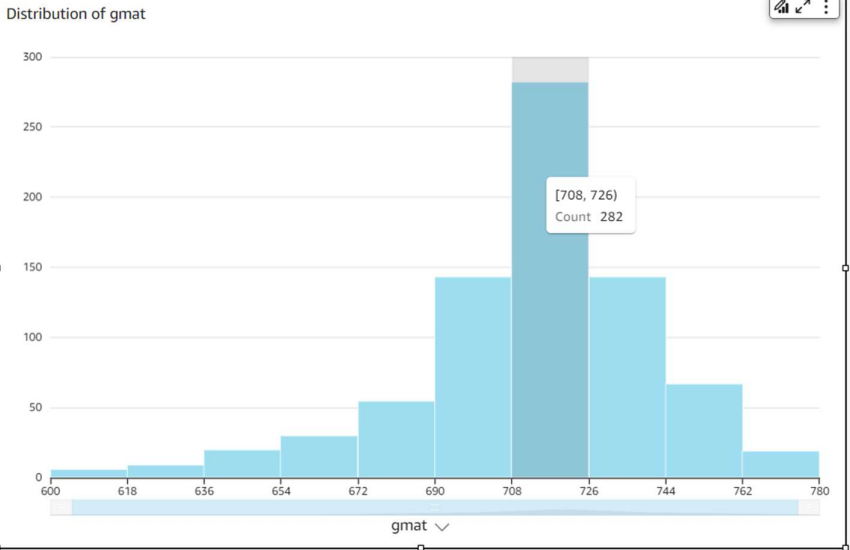
gmat

Add a measure

Sheet 1

Sheet 2

Distribution of gmat



[708, 726]
Count 282

gmat

Score Range	Count
600-618	10
618-636	15
636-654	25
654-672	35
672-690	55
690-708	145
708-726	282
726-744	145
744-762	65
762-780	25

Histogram of 'workex' column:

