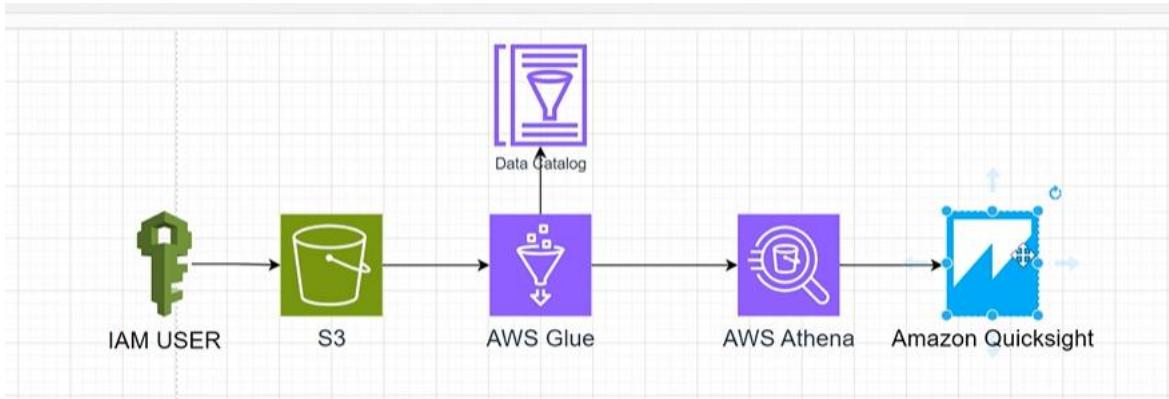


End to End AWS Data Engineering Project

Data Pipeline design:



IAM (identity and access management) where you can create users, groups, policies etc.

Creating a user:

Click on create user

The screenshot shows the 'Create user' wizard in the AWS IAM console. The left sidebar shows 'Access management' with 'Users' selected. The main area is titled 'User details'. The 'User name' field contains 'Sai'. Below it, a note says: 'The user name can have up to 64 characters. Valid characters: A-Z, a-z, 0-9, and + = , . @ _ - (hyphen)'. A checked checkbox says 'Provide user access to the AWS Management Console - optional'. In the 'Console password' section, 'Custom password' is selected, and the password '@01' is entered. A note says: 'Must be at least 8 characters long' and 'Must include at least three of the following mix of character types: uppercase letters (A-Z), lowercase letters (a-z), numbers (0-9), and symbols ! @ # \$ % ^ & * () _ + - (hyphen) = [] { } | '.

Click next

Permissions options

- Add user to group
Add user to an existing group, or create a new group. We recommend using groups to manage user permissions by job function.
- Copy permissions
Copy all group memberships, attached managed policies, and inline policies from an existing user.
- Attach policies directly
Attach a managed policy directly to a user. As a best practice, we recommend attaching policies to a group instead. Then, add the user to the appropriate group.

Permissions policies (1/1440)

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

Filter by Type		Attached entities
Policy name	Type	
<input type="checkbox"/> AccessAnalyzerServiceRolePolicy	AWS managed	0
<input type="checkbox"/> AccountManagementFromVercel	AWS managed	0
<input checked="" type="checkbox"/> AdministratorAccess	AWS managed - job function	0
<input type="checkbox"/> AdministratorAccess-Amplify	AWS managed	0

The above implies the user (sai) get all the admin access

Click on next

Review and create

Review your choices. After you create the user, you can view and download the autogenerated password, if enabled.

User details

User name	Console password type	Require password reset
sai	Custom password	No

Permissions summary

Name	Type	Used as
AdministratorAccess	AWS managed - job function	Permissions policy

Tags - optional

Tags are key-value pairs you can add to AWS resources to help identify, organize, or search for resources. Choose any tags you want to associate with this user.

No tags associated with the resource.

[Add new tag](#)

You can add up to 50 more tags.

[Cancel](#) [Previous](#) [Create user](#)

Click create user

U see the interface as shown:

User created successfully

You can view and download the user's password and email instructions for signing in to the AWS Management Console.

[View user](#)

Step 1 Specify user details **Step 2** Set permissions **Step 3** Review and create **Step 4** Retrieve password

Retrieve password

You can view and download the user's password below or email users instructions for signing in to the AWS Management Console. This is the only time you can view and download this password.

Console sign-in details

Console sign-in URL: <https://351596828353.sigin.aws.amazon.com/console>

User name:

Console password: [Show](#)

[Email sign-in instructions](#)

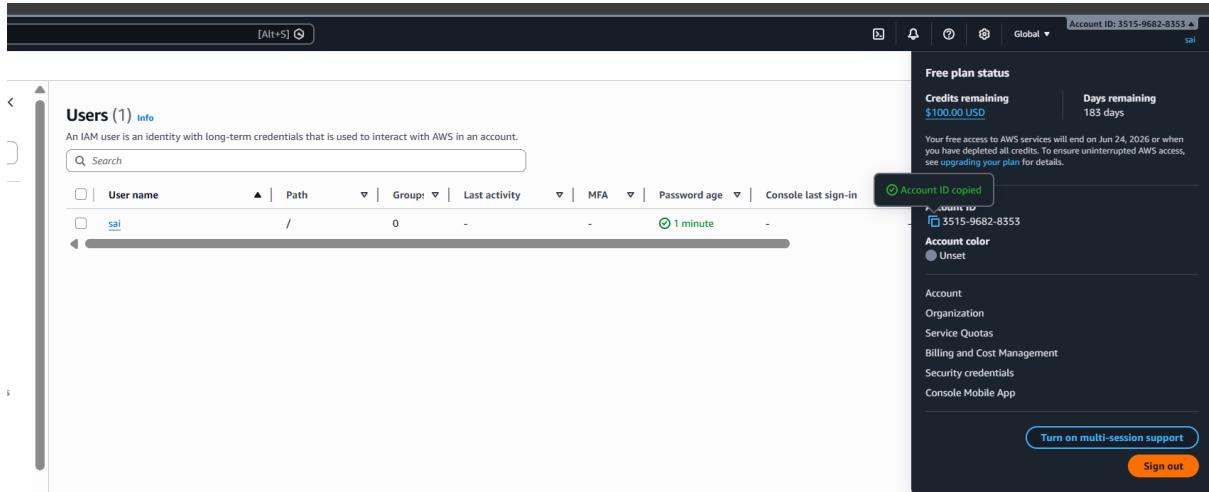
[Cancel](#) [Download .csv file](#) [Return to users list](#)

Now click return to users list

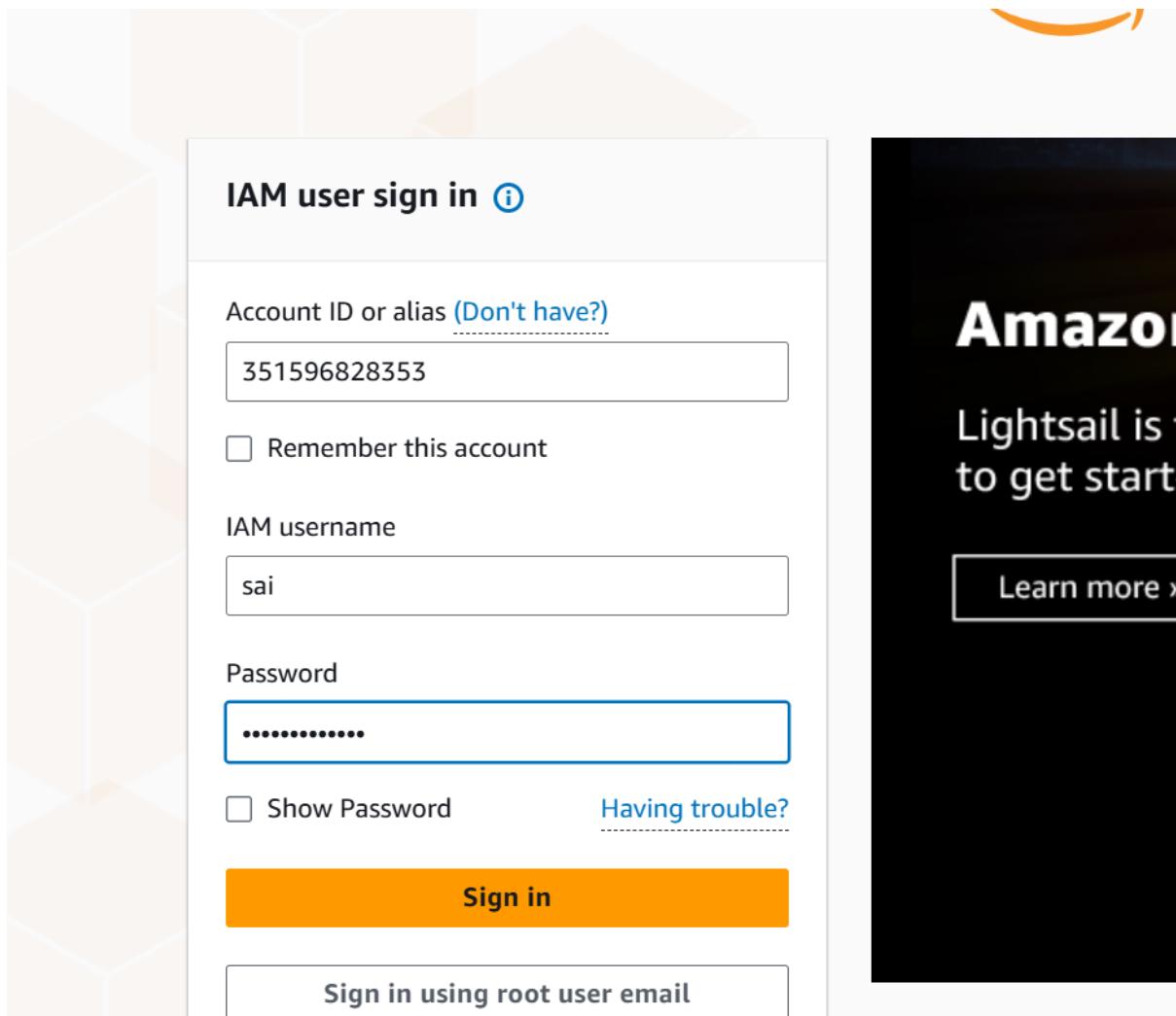
Log out and login as a IAM user

Before logging out, copy ur account ID: 351596828353

Click on sign out



Logging as IAM user



The interface looks like this and now click on s3 to create my bucket

S3 is a kind of cloud storage where u can store all kinds of files like structured, unstructured etc

Let us create a bucket

The screenshot shows the Amazon S3 console. At the top right, there is a 'Create a bucket' button. Below it, a section titled 'Pricing' states: 'With S3, there are no minimum fees. You only pay for what you use. Prices are based on the location of your S3 bucket.' On the left, there is a 'How it works' section.

Before that changed the region as follows, click on setting button below beside the name

The screenshot shows the 'Edit localisation and default region' settings page. It includes sections for 'Localisation' (Language set to 'Browser default') and 'Default region' (Default Region set to 'United States (N. Virginia) us-east-1').

If u give the bucket name as test, as it is a name globally.so give a name which is unique globally

The screenshot shows the 'Create bucket' configuration page. Under 'General configuration', the 'Bucket type' is set to 'General purpose'. The 'Bucket name' field contains 'test', which is highlighted in red with an error message: 'Bucket with the same name already exists'. There is also a note: 'Bucket names must be 3 to 63 characters and unique within the global namespace. Bucket names must also begin with a letter or number.' The 'Copy settings from existing bucket - optional' section is present at the bottom.

General configuration

AWS Region

Europe (Stockholm) eu-north-1

Bucket type | Info

General purpose

Recommended for most use cases and access patterns. General storage classes that redundantly store objects across multiple

Bucket name | Info

firstbucket

Bucket with the same name already exists

Bucket names must be 3 to 63 characters and unique within the glob

Copy settings from existing bucket - optional

Only the bucket settings in the following configuration are copied.

[Choose bucket](#)

Format: s3://bucket/prefix

Keep all the default settings, I have given bucket name as 'sai01project01'

≡ Amazon S3 > Buckets > Create bucket

Buckets are containers for data stored in S3.

General configuration

AWS Region

Europe (Stockholm) eu-north-1

Bucket type | Info

General purpose

Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of

storage classes that redundantly store objects across multiple Availability Zones.

Directory

Recommended for low-latency use cases. These buckets use only the S3 Express On processing of data within a single Availability Zone.

Bucket name | Info

sai01project01

Bucket names must be 3 to 63 characters and unique within the global namespace. Bucket names must also begin and end with a letter or number. Valid characters are a-z, 0-9, periods (.), and hyphens (-). [Learn more](#)

Copy settings from existing bucket - optional

Only the bucket settings in the following configuration are copied.

[Choose bucket](#)

Format: s3://bucket/prefix

Object Ownership [Info](#)

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

Object Ownership

ACLs disabled (recommended)

All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

ACLs enabled

Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

≡ Amazon S3 > Buckets > Create bucket

No tags associated with this bucket.

[Add new tag](#)

You can add up to 50 tags.

Default encryption [Info](#)

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type [Info](#)

Secure your objects with two separate layers of encryption. For details on pricing, see [DSSE-KMS pricing](#) on the Storage tab of the [Amazon S3 pricing page](#).

Server-side encryption with Amazon S3 managed keys (SSE-S3)

Server-side encryption with AWS Key Management Service keys (SSE-KMS)

Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)

Bucket Key

Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. [Learn more](#)

Disable

Enable

Advanced settings

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

Cancel

Create bucket

Click create bucket

Successfully created bucket "sai01project01". To upload files and folders, or to configure additional bucket settings, choose View details.

General purpose buckets All AWS Regions

Directory buckets

General purpose buckets (1) Info

Buckets are containers for data stored in S3.

Name	AWS Region	Creation date
sai01project01	Europe (Stockholm) eu-north-1	December 24, 2025, 18:55:54 (UTC+05:30)

Account snapshot Info Updated daily

Storage Lens provides visibility into storage usage and activity trends.

External access summary - new Info Updated daily

External access findings help you identify bucket permissions that allow public access or access from other AWS accounts.

Click on it

Create a folder in it

sai01project01 Info

Objects Metadata Properties Permissions Metrics Management Access Points

Actions Create folder

Objects (0)

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](#)

Create folder

Create folder as shown

Folder

Folder name /orders/

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

Don't 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.

Create folder

sai01project01 Info

Objects (1)

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](#)

Name	Type	Last modified	Size	Storage class
orders/	Folder	-	-	-

Lets download the Kaggle data

<https://www.kaggle.com/datasets/vivek468/superstore-dataset-final?resource=download>

Open the excel, filter 'order date' with jan 1st 2017, copy and paste in notepad, save

The data looks like this

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Catag	Product Name	Sales	Q1
850		849	CA-2017-1	01-01-2017 #####	Standard	CGA-14725 Guy Armst	Consumer United Sta Lorain	Ohio			44052	East	FUR-FU-1	Furniture	Furnishing	Linden 10"	48.896	
4011		4010	CA-2017-1	01-01-2017 #####	Standard	CS-20725 Steven Cai	Consumer United Sta Los Angeles	California			90036	West	FUR-FU-1	Furniture	Furnishing	Howard Mi	474.43	
6684		6683	CA-2017-1	01-01-2017 #####	First Class	DP-13390 Dennis Pai	Home Offi	United Sta Franklin	Wisconsin		53132	Central	OFF-BI-10	Office Sup	Binders	Wilson Jon	3.6	
8071		8070	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	OFF-ST-10	Office Sup	Storage	SACFO Bo	454.56	
8072		8071	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	FUR-FU-1	Furniture	Furnishing	Tenex Car	141.42	
8073		8072	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	FUR-CH-1	Furniture	Chairs	Office Star	310.744	
8074		8073	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	OFF-AR-10	Office Sup	Art	Fluorescei	12.736	
8075		8074	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	OFF-BI-10	Office Sup	Binders	GBC Instai	6.47	
8076		8075	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	OFF-BI-10	Office Sup	Binders	Pressboar	13.748	
8077		8076	CA-2017-1	01-01-2017 #####	Standard	CJM-15250 Janet Mart	Consumer United Sta Huntsville	Texas			77340	Central	OFF-AP-10	Office Sup	Appliance	Fellowes S	15.224	
9996																		
9997																		
9998																		

Again creating a folder in orders

Amazon S3 > Buckets > sal01project01 > orders/ > Create folder

Your bucket policy might block folder creation
If your bucket policy prevents uploading objects without specific tags, metadata, or access control list (ACL) grantees, you will not be able to create a folder using this configuration. Instead, you can use the [upload configuration](#) to upload an empty folder and specify the appropriate settings.

Folder

Folder name: snapshot_day=2017-01-01 /

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

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

Don't 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](#)

When we create AWS crawler, when we create the structure of the folder with the name like this, there is a partition with all of this, we can discuss later

Amazon S3 > Buckets > sal01project01 > orders/ > snapshot_day=2017-01-01/ > Upload

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 SDKs 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 (0)
All files and folders in this table will be uploaded.

Remove	Add files	Add folder

Destination [Info](#)
Destination: [s3://sal01project01/orders/snapshot_day=2017-01-01/](#)

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

Amazon S3 > Buckets > sai01project01 > orders/ > snapshot_day=2017-01-01/ > Upload

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

Files and folders (1 total, 2.5 KB)

All files and folders in this table will be uploaded.				
<input type="text"/> Find by name		Folder	Type	Size
<input type="checkbox"/>	Name	-	text/csv	2.5 KB
sample superstore jan01 2017.csv				

Destination [Info](#)

Destination: [s3://sai01project01/orders/snapshot_day=2017-01-01/](#)

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.

Upload succeeded

For more information, see the [Files and folders](#) table.

Upload: status

After you navigate away from this page, the following information is no longer available.

Summary

Destination	Succeeded	Failed
s3://sai01project01/orders/snapshot_day=2017-01-01/	1 file, 2.5 KB (100.00%)	0 files, 0 B (0%)

Files and folders Configuration

Files and folders (1 total, 2.5 KB)

All files and folders in this table will be uploaded.				
<input type="text"/> Find by name		Folder	Type	Size
<input type="checkbox"/>	Name	-	text/csv	2.5 KB
sample superstore jan01 2017.csv				

Now let us create AWS crawler

Search glue

Glue is an ETL tool and u can run also crawler and create a data catalog using glue

Here is the GLUE interface

eu-north-1.console.aws.amazon.com/glue/home?region=eu-north-1#/v2/getting-started

aws Search [Alt+S]

AWS Glue

- Getting started**
 - ETL jobs
 - Visual ETL
 - Notebooks
 - Job run monitoring
 - Data Catalog tables
 - Data connections
 - Workflows (orchestration)
 - Zero-ETL integrations [New](#)
- Data Catalog**
 - Databases
 - Tables
 - Stream schema registries
 - Schemas
 - Connections
 - Crawlers
 - Classifiers
 - Catalog settings
- Data Integration and ETL**
 - Legacy pages
- What's New [New](#)

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 [Updated](#)

Use Zero-ETL integrations to replicate data in near real-time, or ETL jobs to transform data in visual, notebook, or code interface.

[Go to Zero-ETL integrations](#)

[Go to ETL jobs](#)

Resources and tutorials [New](#)

Getting started with AWS Glue: [Documentation](#) [AWS Training](#)

Glue in 5 Minutes Videos: [Authoring](#), [GenAI](#), [Monitoring](#), [Orchestration](#)

[Using connectors and connections](#)

[AWS Glue Documentation home](#)

Examples: [AWS Glue blog posts](#) [AWS Glue on GitHub](#)

Data integration and management

Monitor & debug ETL jobs and track usage

[Go to job run monitoring](#)

Connect to your data stores

[Go to connections](#)

We are going into crawler and data cataloging

Databases (0)
A database is a set of associated table definitions, organized into a logical group.

Name	Description	Location URI	Source catalog	Created on (UTC)
No resources No resources to display.				

Last updated (UTC) December 24, 2025 at 17:52:53 Edit Delete Add database

Click on add data base ,give name and click create database

Create a database
Create a database in the AWS Glue Data Catalog.

Database details

Name
db_project01

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

Description - optional
Enter text

Descriptions can be up to 2048 characters long.

Database settings

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

An S3 location is required for managed tables and Zero-ETL integrations.

Cancel Create database

Click on crawlers, create crawler

Crawlers (0) info
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from L...
No resources No resources to display.						

Last updated (UTC) December 24, 2025 at 17:56:47 Action Run Create crawler

Give name and next

AWS Glue > Crawlers > Add crawler

Step 1 Set crawler properties

Crawler details info

Name
orders_project

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

Click Add datasource

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) Info

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.

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://bucket/prefix/object

[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.

Choose S3 path

S3 buckets > sail01project01

Objects (1/1)

Find object by prefix

Key

orders/

Last modified

Size

[Cancel](#) [Choose](#)

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.

[Cancel](#) [Previous](#) [Next](#) [Add an S3 data source](#)

Optional 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.
 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.

No schema changes are made or no new tables will be added to the Data Catalog after the first crawl run.

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

Click add an s3 datasource

S3 path
Browse for or enter an existing S3 path.
 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.

No schema changes are made or no new tables will be added to the Data Catalog after the first crawl run.

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

The above meaning is the folders which were crawled will not be crawled after addition of new files

Click next

Crawlers > Add crawler

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 (1) Info

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

Type	Data source	Parameters
S3	s3://sa101project01/orders/	Recrawl new only

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

Crawlers > Add crawler

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

Configure security settings

IAM role Info

Existing IAM role

Choose an IAM role

Create new 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

We have created IAM user and user logged in and using AWS services. similarly for AWS services to work they need some access . so, IAM role is used to give different services access.

As I Want to run the AWS crawler job, I have to use an IAM role for that using which role I want to run the services and to that role I have to give all the access which it needs.

Add crawler

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

Configure security settings

IAM role Info

Existing IAM role

Choose an IAM role

Create new 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

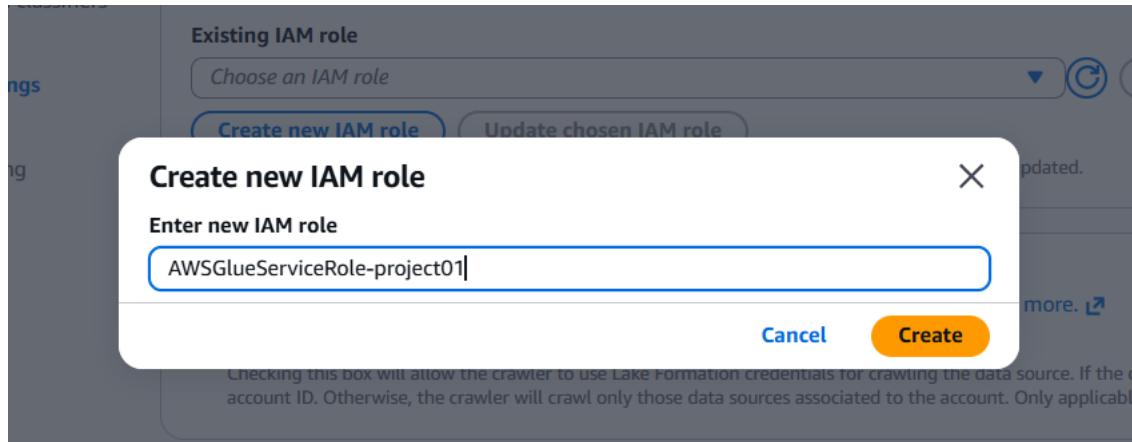
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable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

Security configuration - optional

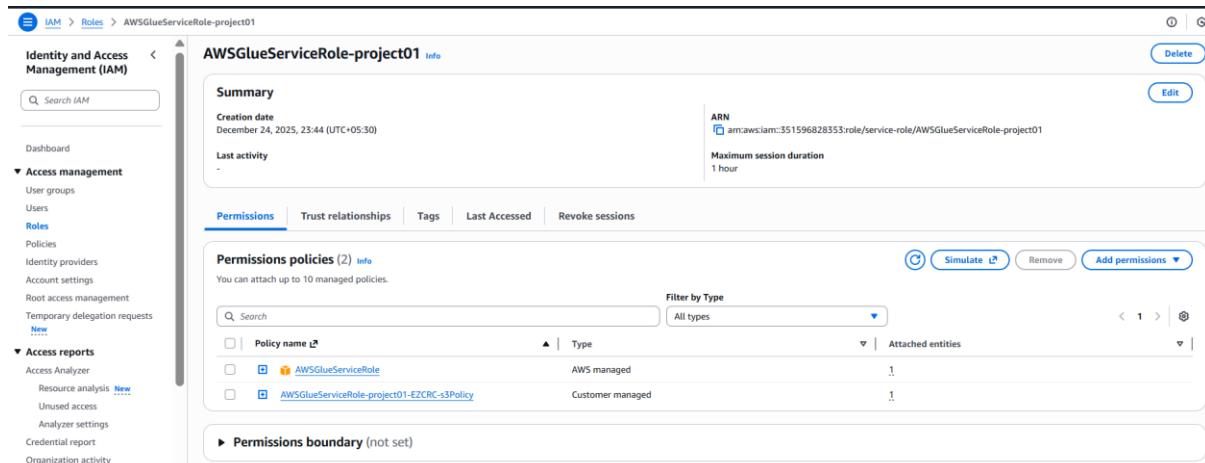
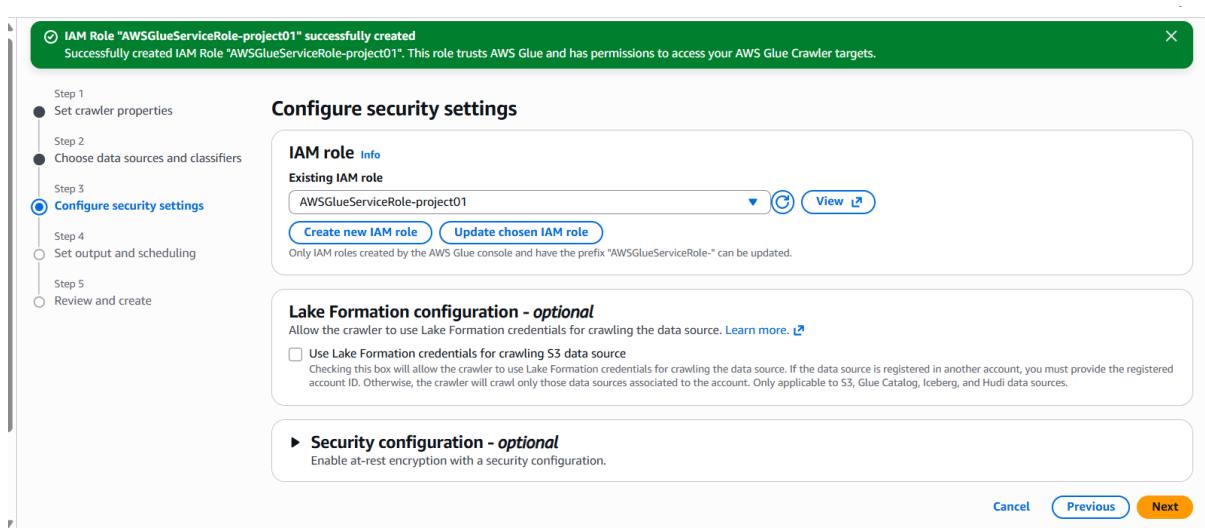
Enable at-rest encryption with a security configuration.

Cancel Previous Next

Project01 is the name and AWSGlueServiceRole is by default available



We had created IAM role as shown, click on view to see



Identity and Access Management (IAM)

AWS managed

Customer managed

AWSGlueServiceRole-project01-EZCRC-s3Policy

```

1- [ {
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": [
        "arn:aws:s3:::sa101project01/orders/*"
      ],
      "Condition": {
        "StringEquals": {
          "aws:ResourceAccount": "351596828353"
        }
      }
    }
  ]
}
  
```

Two roles, one is for glue service role and other is which gives access to s3 click on + button and see for which service that u have access

One role is for glue services and other is it created a s3 policy where it can access everything from this path: sa101project01/orders/

As we have given data source as S3 path, it has automatically taken the access to that s3 path by creating a policy

Click next

IAM Role "AWSGlueServiceRole-project01" successfully created

Successfully created IAM Role "AWSGlueServiceRole-project01". This role trusts AWS Glue and has permissions to access your AWS Glue Crawler targets.

Configure security settings

IAM role [Info](#)

Existing IAM role: **AWSGlueServiceRole-project01**

[Create new IAM role](#) [Update chosen IAM role](#)

Lake Formation configuration - optional

Use Lake Formation credentials for crawling S3 data source

Security configuration - optional

Enable at-rest encryption with a security configuration.

Set output and scheduling

Output configuration [Info](#)

Target database: **db_project01**

[Clear selection](#) [Add database](#)

Table name prefix - optional:

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.

On demand

Hourly

Daily

Weekly

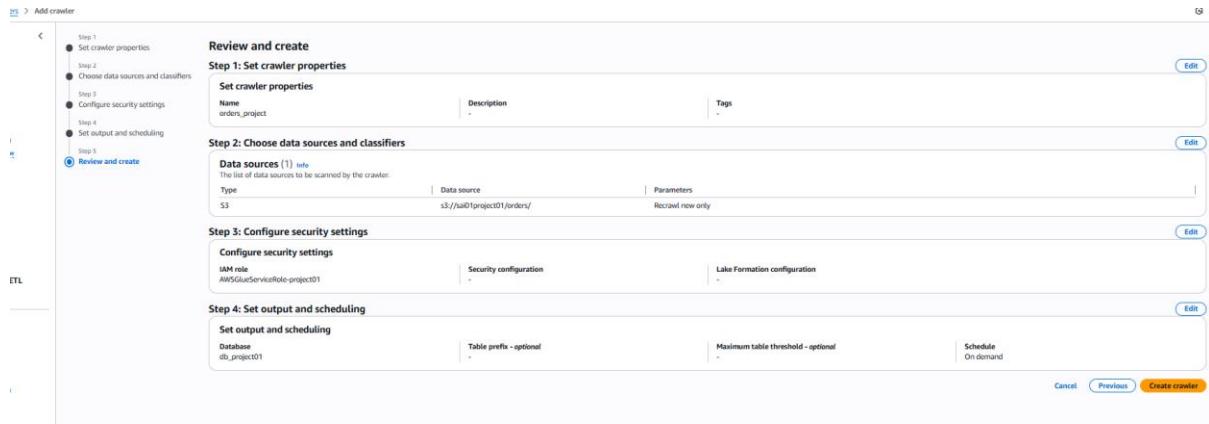
Monthly

Custom (cron expression)

On demand

Click on next

Create crawler



Here is the interface

Click on run crawler

Working of crawler...

As our data is in s3 which is a file, if u want to query this data u need to have information about data which is meta data like what are the columns, datatypes etc. Then only I can query the data using sql.

So this crawler crawls over the s3 file, identify what are the columns, delimiters etc in the file internally by itself and u will be able to see table but it doesn't store any data in it, the data is in s3 only, it is just creating meta data

Crawler properties

- Name: orders_project
- IAM role: AWSServiceRole-project01
- Description: -
- Maximum table threshold: -
- State: READY
- Table prefix: -

Crawler runs (1)

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours
December 25, 2025 at 05:02:08	December 25, 2025 at 05:03:18	01 min 09 s	Completed	0.043

Table changes
1 table change, 1 partition change

AWS Glue > Tables

Tables

Name	Database	Location	Classification	Deprecated	View data	Data quality
orders	db_project01	s3://sa101project01/orders/	CSV	-	Table data	View data quality

Table name is orders becoz we gave folder name 'orders' in s3

DB that we created

Click on the orders table, u see the meta data as shown

AWS Glue

- Getting started
- ETL jobs
- Visual ETL
- Notebooks
- Job run monitoring
- Data Catalog tables**
- Data connections
- Workflows (orchestration)
- Zero-ETL integrations [New](#)
- Data Catalog**
- Databases
- Tables**
- Stream schema registries
- Schemas
- Connections
- Crawlers
- Classifiers
- Catalog settings
- Data Integration and ETL**
- Legacy pages**

What's New [New](#)
Documentation [New](#)
AWS Marketplace

Enable compact mode

Schema (22) View and manage the table schema.

#	Column name	Data type	Partition key
1	row id	bigint	-
2	order id	string	-
3	order date	string	-
4	ship date	string	-
5	ship mode	string	-
6	customer id	string	-
7	customer name	string	-
8	segment	string	-
9	country	string	-
10	city	string	-
11	state	string	-
12	postal code	bigint	-
13	region	string	-

AWS Glue

- Getting started
- ETL jobs
- Visual ETL
- Notebooks
- Job run monitoring
- Data Catalog tables**
- Data connections
- Workflows (orchestration)
- Zero-ETL integrations [New](#)
- Data Catalog**
- Databases
- Tables**
- Stream schema registries
- Schemas
- Connections
- Crawlers
- Classifiers
- Catalog settings
- Data Integration and ETL**
- Legacy pages**

What's New [New](#)
Documentation [New](#)
AWS Marketplace

Table overview Data quality - new

Table details

Name: orders	Classification: CSV	Deprecated
Database: db_project01	Location: s3://sai01project01/orders/	Column statistics: No statistics
Description:	Connection:	

Last updated: December 25, 2025 at 05:03:17

Advanced properties

Partitions (1) The list of partitions for this table.

snapshot_day
2017-01-01

Properties View files View Properties

The crawler created a partition already, if I created one more folder and run the crawler again it will create another partition, these partitions will help in querying in athena

So now, if u click on view files, u see one file

Amazon S3

Buckets General purpose buckets Directory buckets Table buckets Vector buckets [New](#)

Access management and security Access Points Access Points for FSx Access Grants IAM Access Analyzer

Storage management and insights Storage Lens Batch Operations

Account and organization settings

snapshot_day=2017-01-01/

Objects (1) 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](#)

Name	Type	Last modified	Size	Storage class
sample_superstore_jan01_2017.csv	csv	December 24, 2025, 21:24:30 (UTC+05:30)	2.5 KB	Standard

So u can click on every partition and see which s3 files are available

Let us create another folder in s3

Filtered to 4th jan, copy in notepad, save it

	A	B	C	D	E	F	G	H	I	J	K	L
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer Segment	Customer	Country	City	State	Postal Code	
2403	2402	CA-2017-145877	04-01-2017 #####	Second Class	AS-10090	Adam Shill	Consumer	United States	Springfield	Missouri	65807	
2404	2403	CA-2017-145877	04-01-2017 #####	Second Class	AS-10090	Adam Shill	Consumer	United States	Springfield	Missouri	65807	
2618	2617	CA-2017-147942	04-01-2017 #####	Standard	CMS-17365	Maribeth S.	Consumer	United States	San Francisco	California	94110	
2619	2618	CA-2017-147942	04-01-2017 #####	Standard	CMS-17365	Maribeth S.	Consumer	United States	San Francisco	California	94110	
4219	4218	CA-2017-149881	04-01-2017 #####	First Class	NC-18535	Nick Creb	Corporate	United States	San Francisco	California	94110	
4220	4219	CA-2017-149881	04-01-2017 #####	First Class	NC-18535	Nick Creb	Corporate	United States	San Francisco	California	94110	
4330	4329	CA-2017-118360	04-01-2017 #####	Standard	CJC-15775	John Cast	Consumer	United States	New York City	New York	10011	
5489	5488	CA-2017-134495	04-01-2017 #####	Second Class	BF-11020	Barry Fran.	Corporate	United States	Jacksonville	Florida	32216	
5490	5489	CA-2017-134495	04-01-2017 #####	Second Class	BF-11020	Barry Fran.	Corporate	United States	Jacksonville	Florida	32216	
5491	5490	CA-2017-134495	04-01-2017 #####	Second Class	BF-11020	Barry Fran.	Corporate	United States	Jacksonville	Florida	32216	
5492	5491	CA-2017-134495	04-01-2017 #####	Second Class	BF-11020	Barry Fran.	Corporate	United States	Jacksonville	Florida	32216	
6009	6008	CA-2017-129028	04-01-2017 #####	First Class	GB-14530	George Be	Corporate	United States	Florence	South Carolina	29501	
9996												

Created a folder in s3 with snapshot_day=2017-01-04

The screenshot shows the Amazon S3 console interface. On the left, there's a sidebar with navigation links like 'Amazon S3', 'Buckets', 'Access management and security', and 'Storage management and insights'. The main area shows a folder named 'orders/'. Inside 'orders/' are two sub-folders: 'snapshot_day=2017-01-01/' and 'snapshot_day=2017-01-04/'. The 'snapshot_day=2017-01-04/' folder is currently selected. A green banner at the top of the main area says 'Successfully created folder "snapshot_day=2017-01-04".' Below the banner, there are buttons for 'Copy S3 URI', 'Copy URL', 'Download', 'Open in browser', 'Delete', 'Actions', 'Create folder', and 'Upload'.

I uploaded my csv file in the folder

The screenshot shows the AWS Lambda 'Upload: status' page. At the top, it says 'Upload succeeded' with a link to 'Files and folders table'. Below that is a summary table with columns for Destination (s3://sai01/project01/orders/snapshot_day=2017-01-04/), Status (Succeeded), and Failed (0 files, 0 B (0%)). At the bottom, there are tabs for 'Files and folders' and 'Configuration'. The 'Files and folders' tab shows a table with one item: 'sample superstore jan04 2017.csv' (Type: text/csv, Size: 3.0 KB, Status: Succeeded).

Note: u can add multiple files in it, np

Now let us run the crawler in glue as shown

Click run

The screenshot shows the AWS Glue interface under the 'Crawlers' section. A banner at the top announces optimization features for Apache Iceberg tables. The crawler list table has one entry:

Name	State	Last run	Last run timestamp	Log	Table changes from last r...
orders_project	Ready	Succeeded	December 25, 2025 at 05:...	View log	1 created

If u click on the crawler, u see the status and shows u all the run

The screenshot shows the AWS Glue interface under the 'orders_project' crawler details. The 'Crawler runs' tab is selected, displaying two completed runs:

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
December 25, 2025 at 06:02:13	December 25, 2025 at 06:03:13	01 min	Completed	-	0 table changes, 1 partition change
December 25, 2025 at 05:02:08	December 25, 2025 at 05:03:18	01 min 09 s	Completed	0.043	1 table change, 1 partition change

Now we are running on demand but in production u set a schedule, as ur files are coming daily overnight. Next day if u schedule at 5 am what ever the new folders had created it will crawl

The screenshot shows the AWS Glue interface under the 'orders_project' crawler details. The 'Schedule' tab is selected, showing the current frequency as 'On demand'.

While creating crawler, we had chosen crawl only new subfolders instead of crawl all folders

Now let us go to tables, u see another partition added as shown

AWS Glue > Tables > orders

Table details

- Name: orders
- Database: db_project01
- Description: -
- Last updated: December 25, 2025 at 05:03:17

Advanced properties

Partitions (2)

snapshot.day	Files
2017-01-01	View files
2017-01-04	View files

Lets say if my s3 folder is not given in the format snapshot like may be with only date, now u see it creates partition name with some random name eg: partition key

So its better to follow the format but u can follow anything as per wish

Now these partition will help us in running the queries faster as below

Conclusion: the data catalog has all the information of all the folders in the s3 location (orders)

Lets go to athena, click on query editor

eu-north-1.console.aws.amazon.com/athena/home?region=eu-north-1#/landing-page

Amazon Athena

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.

Begin querying your data

- Query your data in Amazon SageMaker Unified Studio**
Run and schedule queries in a single data and AI development environment.
- Query your data in Athena console**
Discover the query editor and start querying right away.

Pricing
Region: EU (Stockholm)

The screenshot shows the Amazon Athena Query Editor. On the left, the 'Data' sidebar is open, showing the 'Data source' set to 'AwsDataCatalog', 'Catalog' set to 'None', and 'Database' set to 'db_project01'. Under 'Tables and views', there is a table named 'orders' which is listed as 'Partitioned'. The main area is titled 'Query 1' and contains a single row with the number '1'. Below the query editor are buttons for 'Run', 'Explain', 'Cancel', 'Clear', and 'Create'. At the bottom, there are tabs for 'Query results' and 'Query stats'.

Data source:

U see AwsDataCatalog means datacatalog that I created using glue

U know ur database which u created

It is also showing orders table that we created, it shows it is a partitioned table

It s all the data types

The screenshot shows the 'Tables and views' section of the Amazon Athena Query Editor. On the left, the 'Catalog' is set to 'None' and the 'Database' is set to 'db_project01'. The 'Tables and views' section shows one table named 'orders'. The table details are as follows:

Column	Type
row id	bigint
order id	string
order date	string
ship date	string
ship mode	string
customer id	string
customer name	string
segment	string
country	string
city	string
state	string
postal code	bigint

The 'Views (0)' section is empty.

It has snapshot day column as well

Tables (1)			
city	string	⋮	▲
state	string	⋮	
postal code	bigint	⋮	
region	string	⋮	
product id	string	⋮	
category	string	⋮	
sub-category	string	⋮	
product name	string	⋮	
sales	double	⋮	
quantity	bigint	⋮	
discount	double	⋮	
profit	double	⋮	
snapshot_day	string (Partitioned)	⋮	▼

Click on preview table

Amazon Athena > Query editor

Data source: AwsDataCatalog

Catalog: None

Database: db_project01

Tables and views: Create ▾

Tables (1): orders Partitioned

Views (0)

Run Query

- Preview Table
- Generate table DDL
- Load partitions
- Insert
- Insert into editor
- Manage
- Delete table
- Generate statistics - new
- View properties
- View in Glue ↗

Run Explain Cancel Clear Create ▾

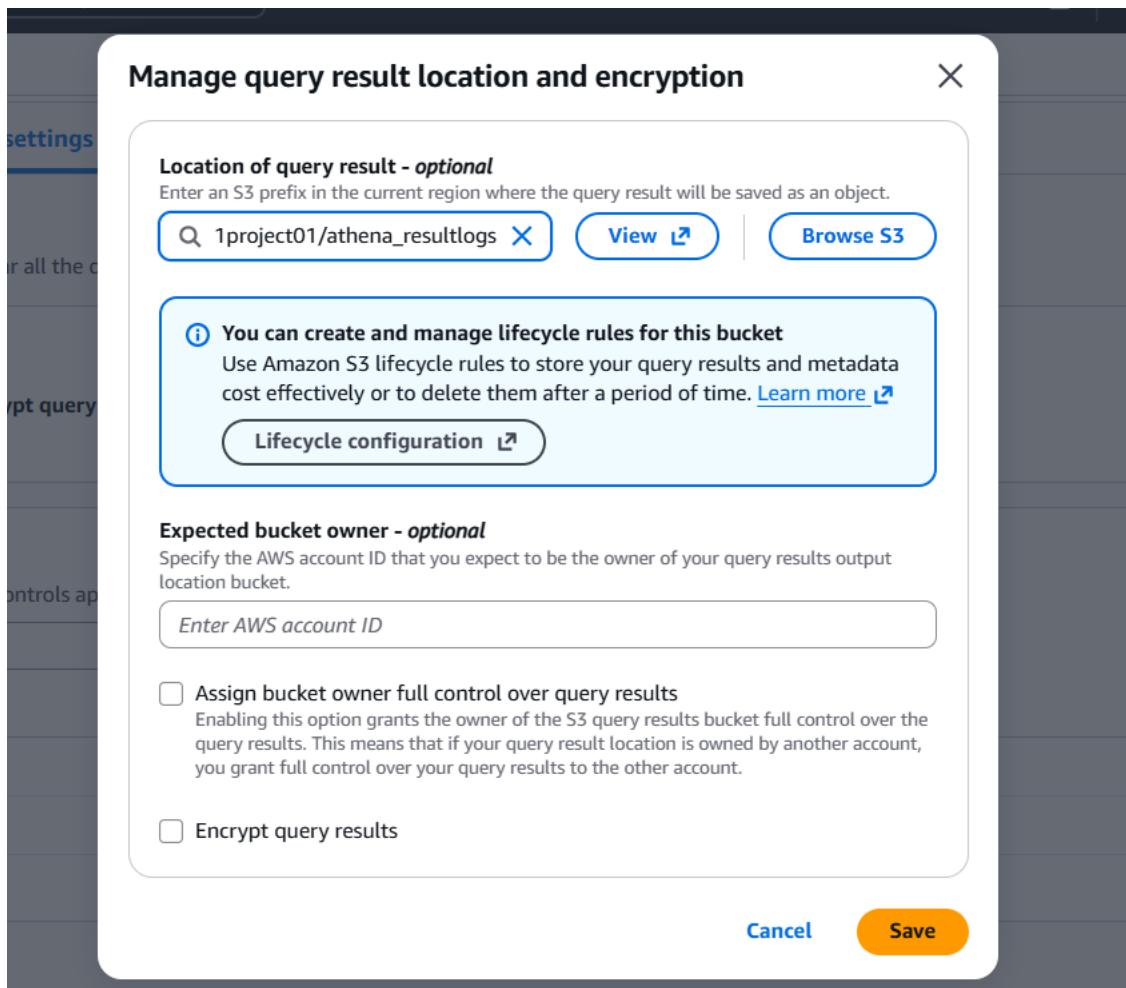
Query results | Query stats

When athena runs, it stores results and that location u need to provide by above

Click on query settings beside saved queries, click on manage

Before that we created a folder as shown

Now specify this location



Save

Lets go back to editor and run query

Amazon Athena > Query editor

Editor Recent queries Saved queries Query settings Workgroup primary

Data source: AmazonDataCatalog Catalog: None Database: db_project01 Tables and views: orders Views: None

Tables (1) < 1 >

SQL: Line 1, Col 76

Run again Explain Cancel Clear Create

Completed

#	category	category
1	Furniture	1804456000000000001
2	Office Supplies	8264000000000000001
3	Technology	4823.96

Lets see how partitions help us

If u see above data scanned is 5.47kb

Lets run whole query

The screenshot shows a database query editor interface. At the top, it says "Query 2 :". Below that is a code editor window containing the SQL query: "SELECT * FROM "db_project01"."orders"". The status bar at the bottom of the editor window indicates "SQL Ln 1, Col 40". Below the editor are several buttons: "Run again", "Explain", "Cancel", "Clear", and "Create". To the right of these buttons is a checkbox labeled "Reuse query results up to 60 minutes ago". Further down, there are tabs for "Query results" and "Query stats". The "Query results" tab is selected and shows a green bar indicating "Completed". Below this, it says "Results (22)". At the bottom right of the results area are "Copy" and "Download results CSV" buttons. The status bar at the very bottom of the screen shows "Time in queue: 103 ms", "Run time: 466 ms", and "Data scanned: 5.47 KB".

U see the data scanned is 5.47kb, it is scanning all the folders

snapshot_day=2017-01-01/

Objects Properties

Objects (1)

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 permission.

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	sample superstore jan01 2017.csv	csv	December 24, 2025, 21:24:30 (UTC+05:30)	2.5 KB	Standard

sai01project01 > orders/ > snapshot_day=2017-01-04/

snapshot_day=2017-01-04/

Objects Properties

Objects (1)

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 permission.

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	sample superstore jan04 2017.csv	csv	December 25, 2025, 11:29:30 (UTC+05:30)	3.0 KB	Standard

So total $2.5 + 3 = 5.5\text{kb}$

When u query with filter as shown, u see it scanned only particular folder which is 2.52kb as per filter

In this way u can reduce scan so that it reduces ur cost and time

The screenshot shows a database query editor interface. At the top, there is a query pane with the title "Query 2" and the following SQL code:

```
1 SELECT * FROM "db_project01"."orders"
2 where snapshot_day = '2017-01-01'
```

Below the query pane, the status bar indicates "SQL Ln 2, Col 33". Underneath the status bar are several buttons: "Run again", "Explain", "Cancel", "Clear", and "Create". To the right of these buttons is a link "Reuse query results up to 60 minutes ago".

The main area of the interface is titled "Query results" and shows a green progress bar with the status "Completed". Below this, a table header "Results (10)" is visible. At the bottom right of the results area are "Copy" and "Download results CSV" buttons.

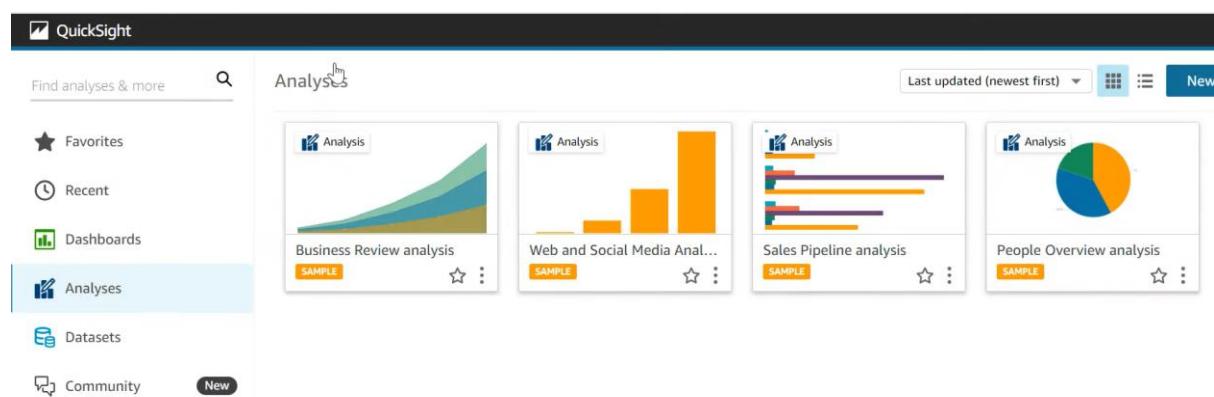
So once u created partitions, it knows where to go exactly

So try to create snapshot folders in s3 like what type of analysis u r going to do

And hence while querying u can save time and cost

Lets go to quicksight

1 month free access by logging with ur email id



Click on dataset, click new dataset

Datasets

Name	Owner	Last Modified
Sales Pipeline	Me	a few seconds ago
Business Review	Me	a few seconds ago
People Overview	Me	a few seconds ago
Web and Social Media Analytics	Me	a few seconds ago

New dataset

Click on athena

New Athena data source

Data source name

Athena workgroup

[primary]

Create data source

Query in Amazon SageMaker Unified Studio

Editor Recent queries Saved queries Query settings

Workgroup primary

Data

Query 2 :

```
1 SELECT * FROM "db_project01"."orders"
2 where snapshot_day = '2017-01-01'
```

Set up

New Athena data source

Data source name

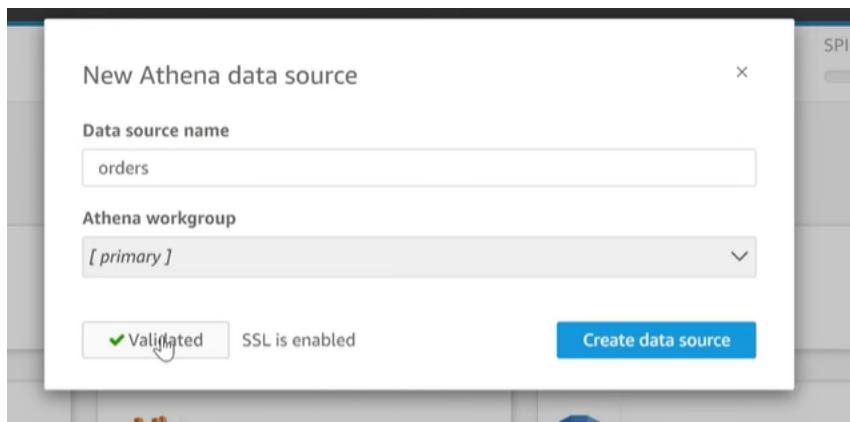
orders

Athena workgroup

[primary]

Validate connection SSL is enabled Create data source

Click validate connection



Click create data source

The screenshot shows the 'Choose your table' dialog box. It lists the 'orders' table under 'Tables: contain the data you can visualize.' Below it are sections for 'Catalog' (set to 'AwsDataCatalog') and 'Database' (set to 'DB_project01'). At the bottom are buttons for 'Edit/Preview data', 'Use custom SQL', and a large blue 'Select' button with a hand cursor icon.

Finish dataset creation

Table: orders
Data source: orders
Schema: db_namastesql

Import to SPICE for quicker analytics ✓ 20GB available SPICE

Directly query your data

Email owners when a refresh fails

QuickSight | orders analysis

File Edit Data Insert Sheets Objects Search

ADD: + T

Data

Dataset: SPICE orders

Search fields

+ CALCULATED FIELD

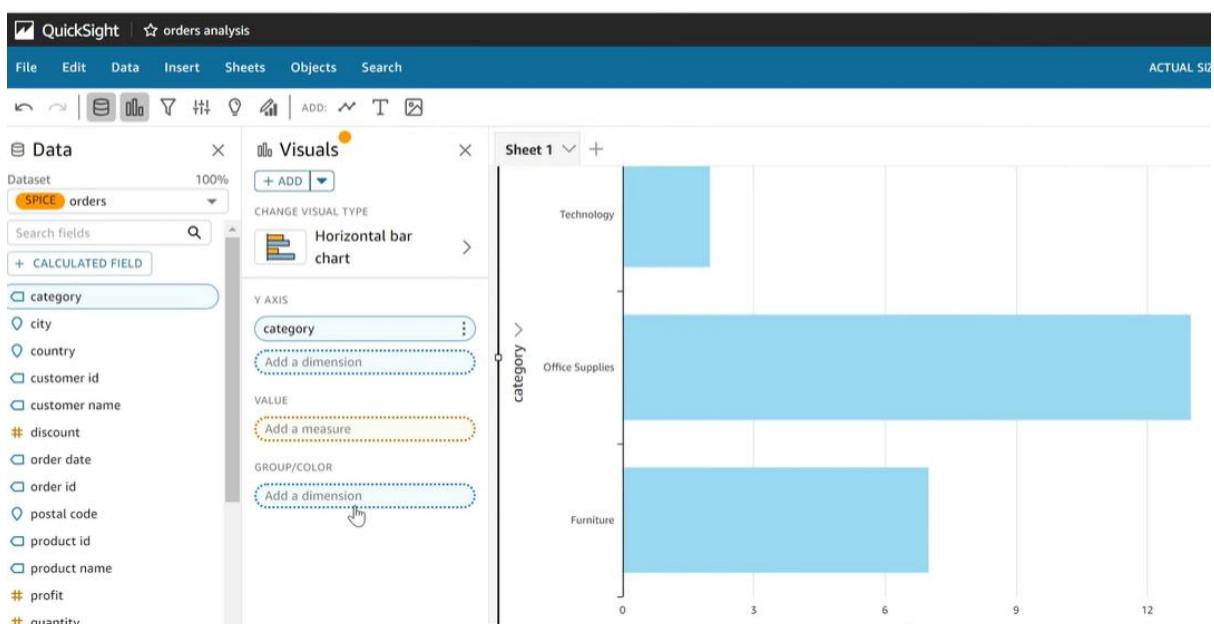
category

- city
- country
- customer id
- customer name
- discount
- order date
- order id
- postal code
- product id

Sheet 1 +

AutoGraph

Add 1 or more fields to build a visual.



Click on horizontal bar chart,

Y axis: drag category on to y axis

Sales into value

You see the graph as above

If you're unable to do the above, give access as shown

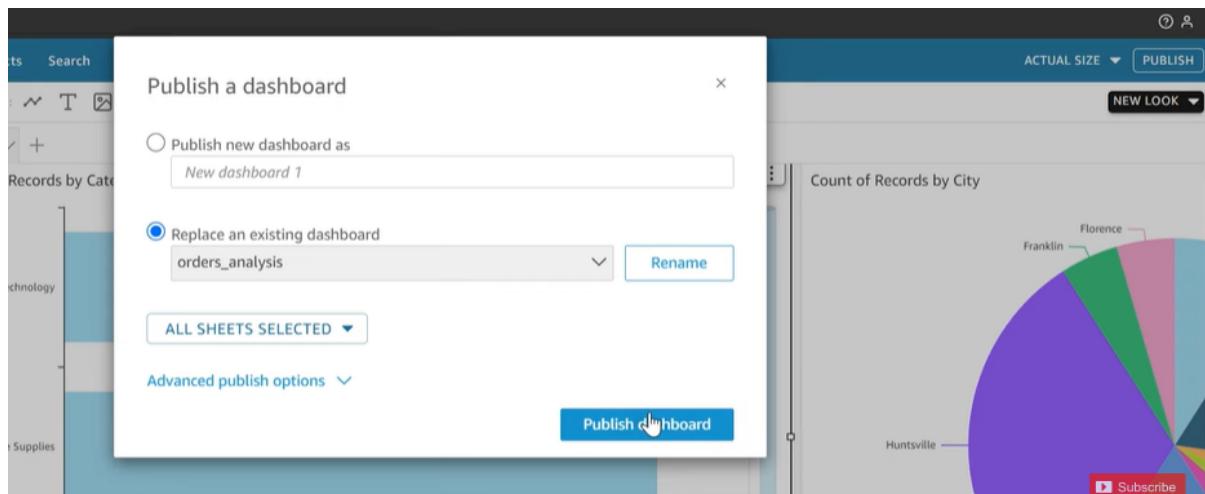
QuickSight access to AWS services

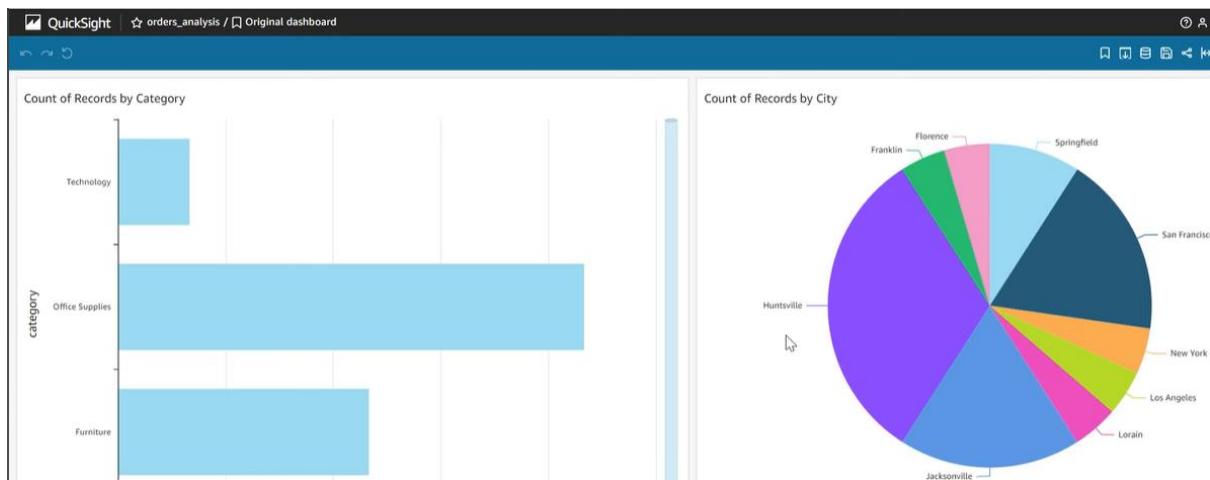
Make your existing AWS data and users available in QuickSight. [Learn more](#)

Allow access and autodiscovery for these resources

-  Amazon Redshift
-  Amazon RDS
-  IAM
-  Amazon S3 (3 buckets selected)
[Select S3 buckets](#)
-  Amazon Athena
Make sure you've chosen the right Amazon S3 buckets for QuickSight access
-  Amazon S3 Storage Analytics
-  AWS IoT Analytics
-  Amazon OpenSearch Service
-  Amazon Kinesis

Now publish





Finally, I don't need to store data in some db and for quick analysis I uploaded files in s3, run a crawler, run the queries on it in athena and create charts