Serverless Data Pipeline

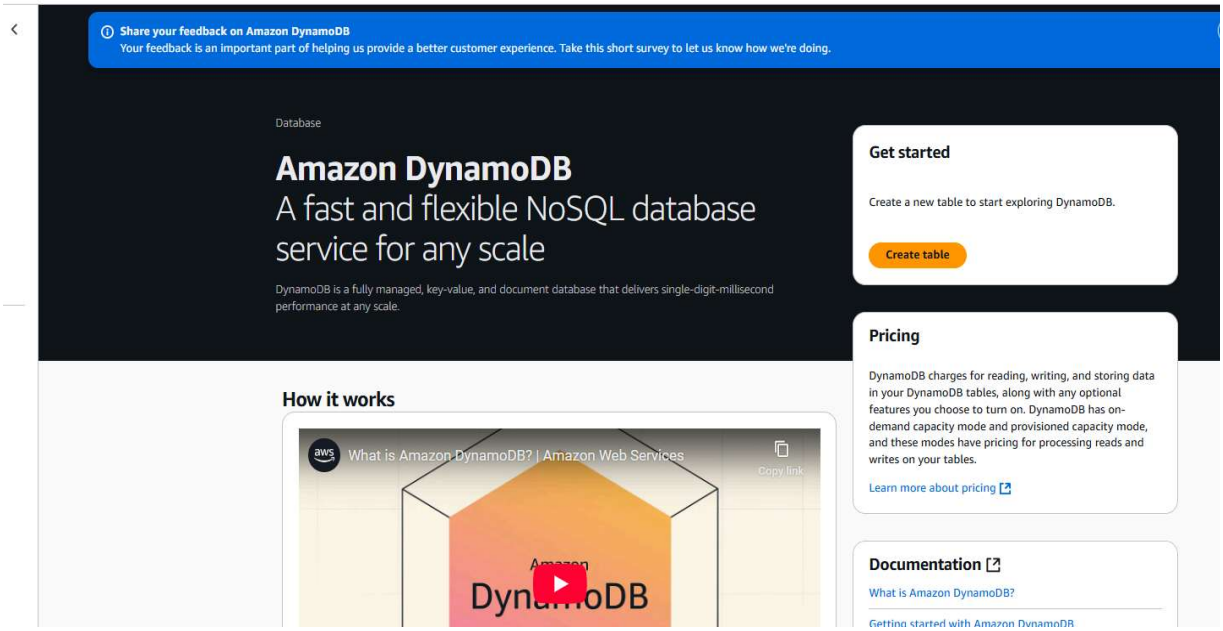
Overall Architecture

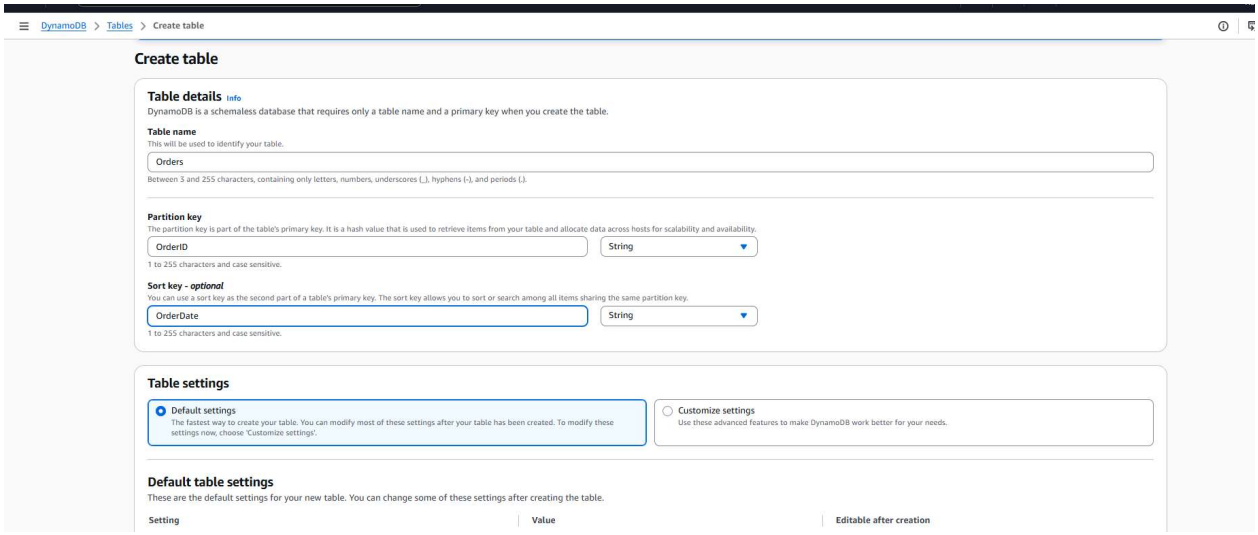
- Source: DynamoDB table (Orders)
- Transformation & Cataloging: Glue Crawlers & ETL Job
- Storage & Analytics: S3 (Parquet), Query via Athena
- Roles: One IAM role for DynamoDB crawling, another for S3 crawling and Glue ETL operational access

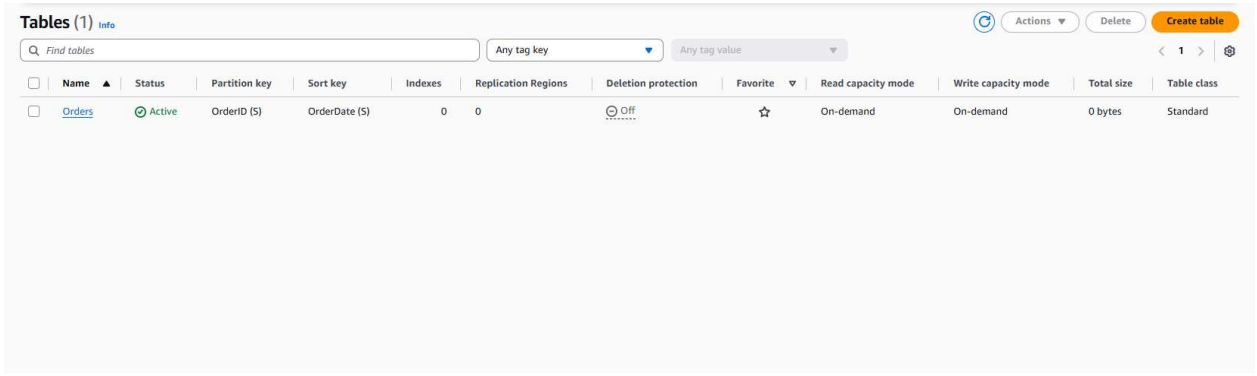
Step 1: Create DynamoDB Table

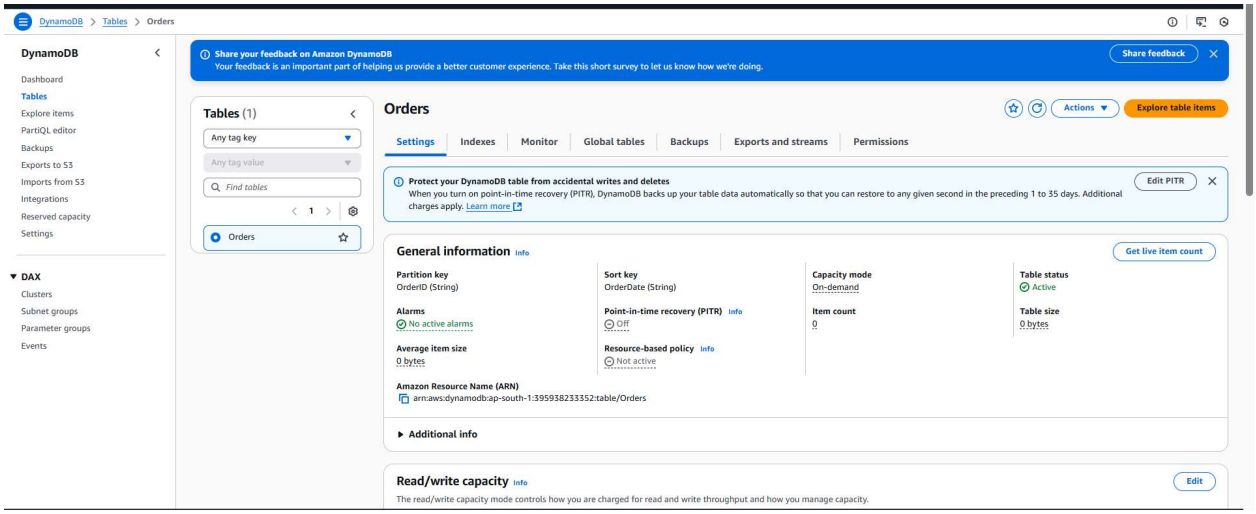
- Table Name: Orders
- Partition Key: OrderID (String)
- Sort Key: OrderDate (String)
- Capacity: On-Demand

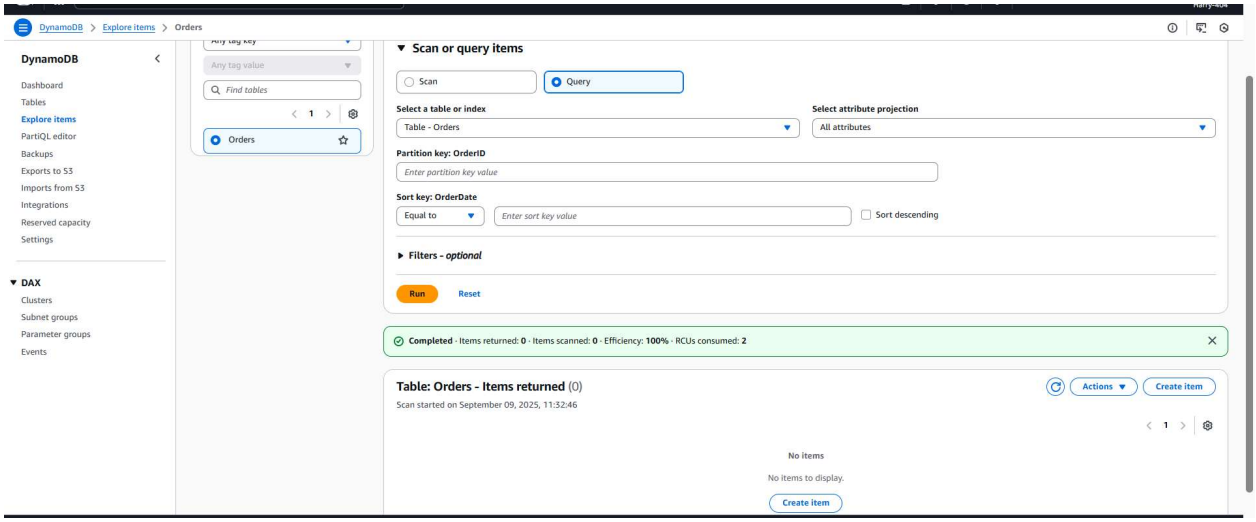
Console Path: DynamoDB → Tables → Create Table




DynamoDB > Tables > Create table
Create table
Table details Info
DynamoDB is a schemaless database that requires only a table name and a primary key when you create the table.
Table name
This will be used to identify your table.
Orders
Between 3 and 255 characters, containing only letters, numbers, underscores (_), hyphens (-), and periods (.).
Partition key
The partition key is part of the table's primary key. It is a hash value that is used to retrieve items from your table and allocate data across hosts for scalability and availability.
OrderID
String
1 to 255 characters and case sensitive.
Sort key - optional
You can use a sort key as the second part of a table's primary key. The sort key allows you to sort or search among all items sharing the same partition key.
OrderDate
String
1 to 255 characters and case sensitive.
Table settings
Default settings
The fastest way to create your table. You can modify most of these settings after your table has been created. To modify these settings now, choose 'Customize settings'.
Customize settings
Use these advanced features to make DynamoDB work better for your needs.
Default table settings
These are the default settings for your new table. You can change some of these settings after creating the table.
Setting
Value
Editable after creation


Tables (1) Info
Actions
Delete
Create table
Find tables
Any tag key
Any tag value
Name
Status
Partition key
Sort key
Indexes
Replication Regions
Deletion protection
Favorite
Read capacity mode
Write capacity mode
Total size
Table class
Orders
Active
OrderID (S)
OrderDate (S)
0
0
Off
On-demand
On-demand
0 bytes
Standard


DynamoDB > Tables > Orders
DynamoDB
Dashboard
Tables
Explore items
PartiQL editor
Backups
Exports to S3
Imports from S3
Integrations
Reserved capacity
Settings
DAX
Clusters
Subnet groups
Parameter groups
Events
Share your feedback on Amazon DynamoDB
Your feedback is an important part of helping us provide a better customer experience. Take this short survey to let us know how we're doing.
Share feedback
Tables (1)
Any tag key
Any tag value
Find tables
Orders
Orders
Actions
Explore table items
Settings
Indexes
Monitor
Global tables
Backups
Exports and streams
Permissions
Protect your DynamoDB table from accidental writes and deletes
When you turn on point-in-time recovery (PITR), DynamoDB backs up your table data automatically so that you can restore to any given second in the preceding 1 to 35 days. Additional charges apply. Learn more
Edit PITR
General information Info
Get live item count
Partition key
OrderID (String)
Sort key
OrderDate (String)
Capacity mode
On-demand
Table status
Active
Alarms
No active alarms
Point-in-time recovery (PITR) Info
Off
Item count
0
Table size
0 bytes
Average item size
0 bytes
Resource-based policy Info
Not active
Amazon Resource Name (ARN)
arn:aws:dynamodb:ap-south-1:395938233352:table/Orders
Additional info
Read/write capacity Info
Edit
The read/write capacity mode controls how you are charged for read and write throughput and how you manage capacity.

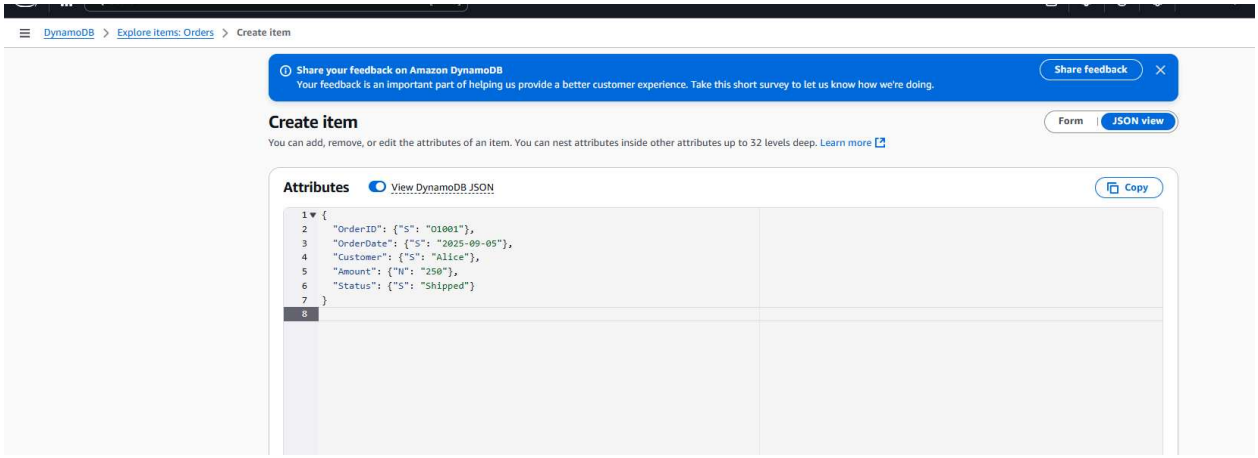


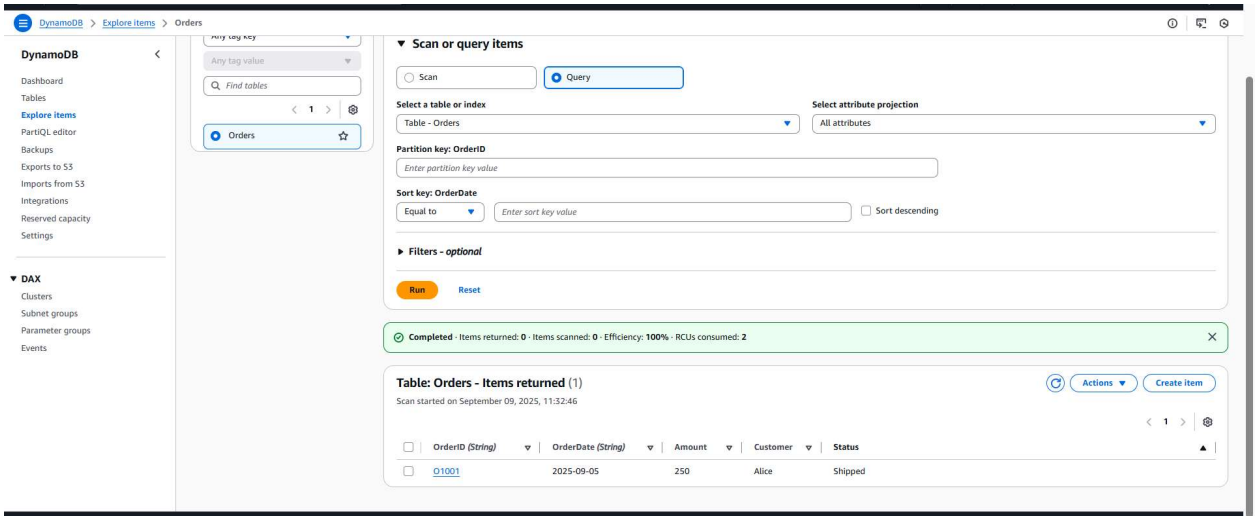
Step 2: Insert Sample Data

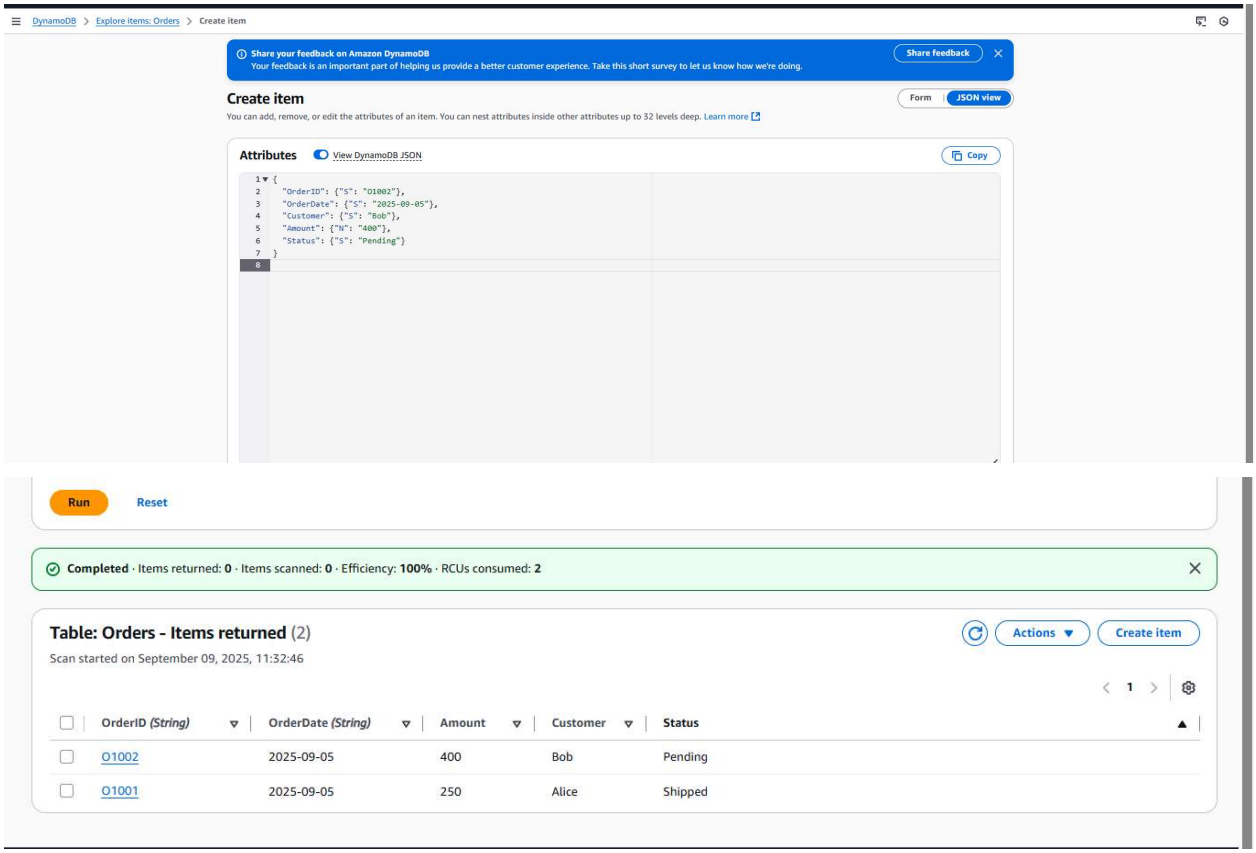
Insert multiple orders manually via the DynamoDB Console → Orders → Create item. Example (JSON view):

```
{
  "OrderID": "O1001",
  "OrderDate": "2025-09-05",
  "Customer": "Alice",
  "Amount": 250,
  "Status": "Shipped"
}
```

Repeat for other customers and statuses (like "Pending"), 10-15 records for better simulation.

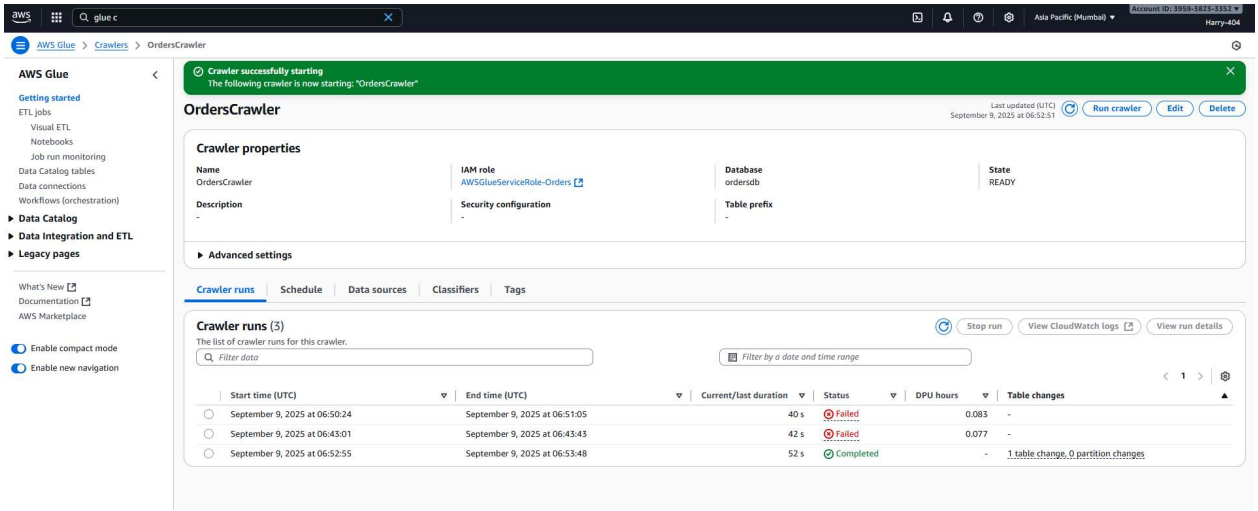

DynamoDB > Explore items: Orders > Create item
Share your feedback on Amazon DynamoDB
Your feedback is an important part of helping us provide a better customer experience. Take this short survey to let us know how we're doing.
Share feedback
Create item
You can add, remove, or edit the attributes of an item. You can nest attributes inside other attributes up to 32 levels deep. Learn more
Form
JSON view
Attributes
View DynamoDB JSON
Copy
{
"OrderID": {"S": "O1001"},
"OrderDate": {"S": "2025-09-05"},
"Customer": {"S": "Alice"},
"Amount": {"N": "250"},
"Status": {"S": "Shipped"}
}


DynamoDB > Explore items > Orders
DynamoDB
Dashboard
Tables
Explore items
PartiQL editor
Backups
Exports to S3
Imports from S3
Integrations
Reserved capacity
Settings
DAX
Clusters
Subnet groups
Parameter groups
Events
Find tables
Orders
Scan or query items
Scan
Query
Select a table or index
Table - Orders
Select attribute projection
All attributes
Partition key: OrderID
Enter partition key value
Sort key: OrderDate
Equal to
Enter sort key value
Sort descending
Filters - optional
Run
Reset
Completed · Items returned: 0 · Items scanned: 0 · Efficiency: 100% · RCUs consumed: 2
Table: Orders - Items returned (1)
Scan started on September 09, 2025, 11:32:46
Actions
Create item
OrderID (String)
OrderDate (String)
Amount
Customer
Status
O1001
2025-09-05
250
Alice
Shipped



Step 3: Glue Crawler 1 (OrdersCrawler): Catalog DynamoDB

- AWS Glue Console: Crawlers → Add crawler
- Name: OrdersCrawler
- Source: DynamoDB → Select "Orders" table

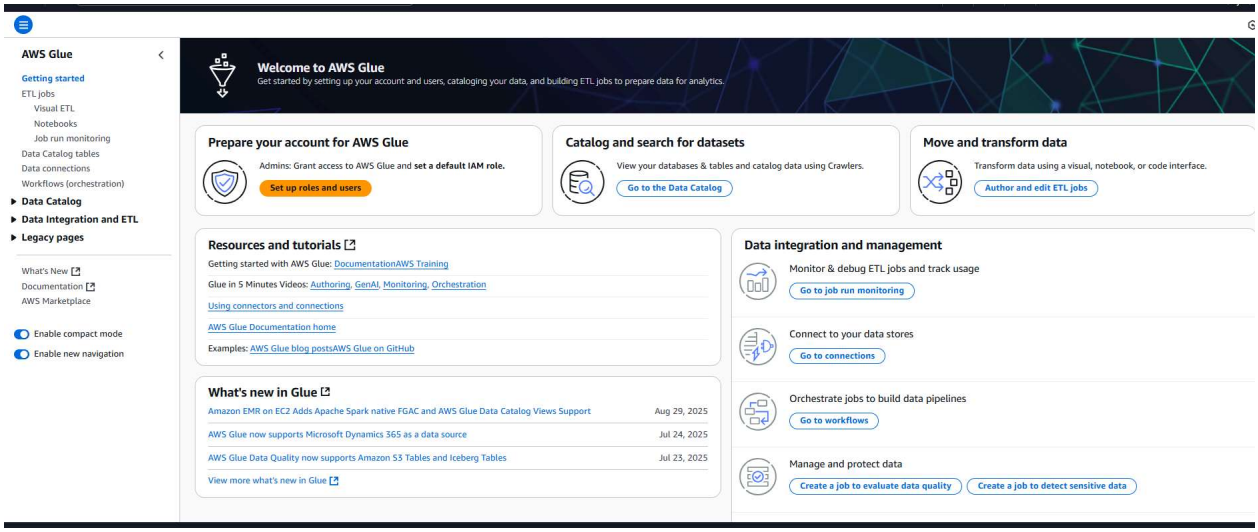


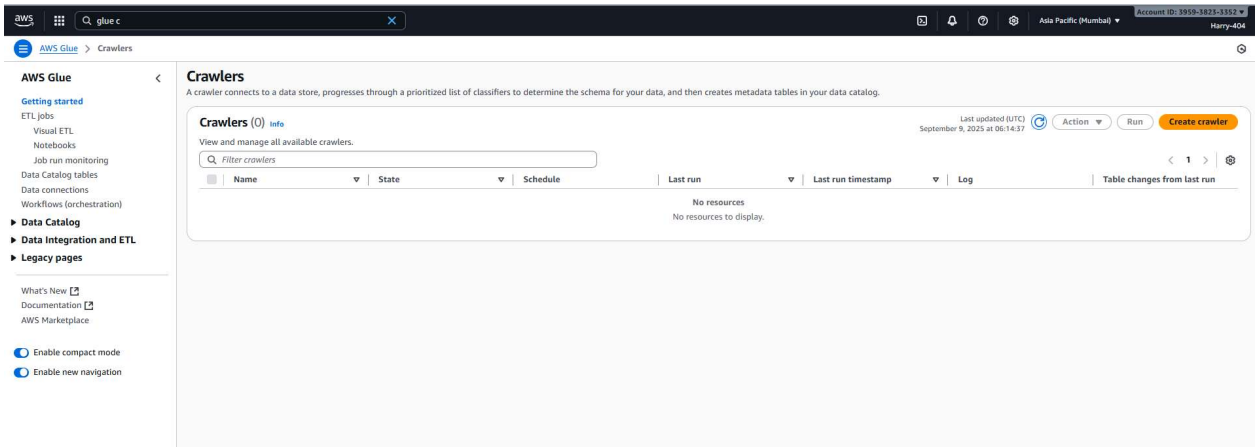
IAM Role: Create (or select) a role with DynamoDB read access and Glue permissions

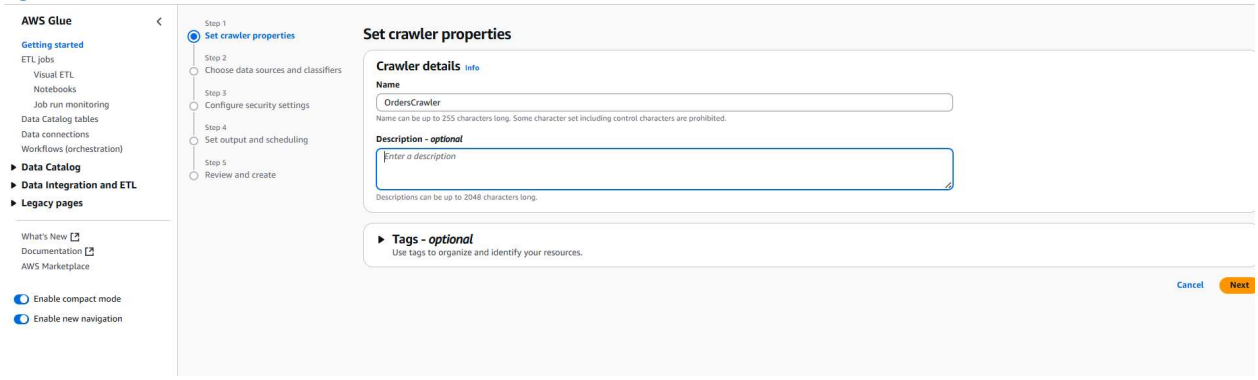
- Example: `AWSGlueServiceRole-Orders`
- Attach: `AmazonDynamoDBReadOnlyAccess`, minimum `AWSGlueServiceRole`

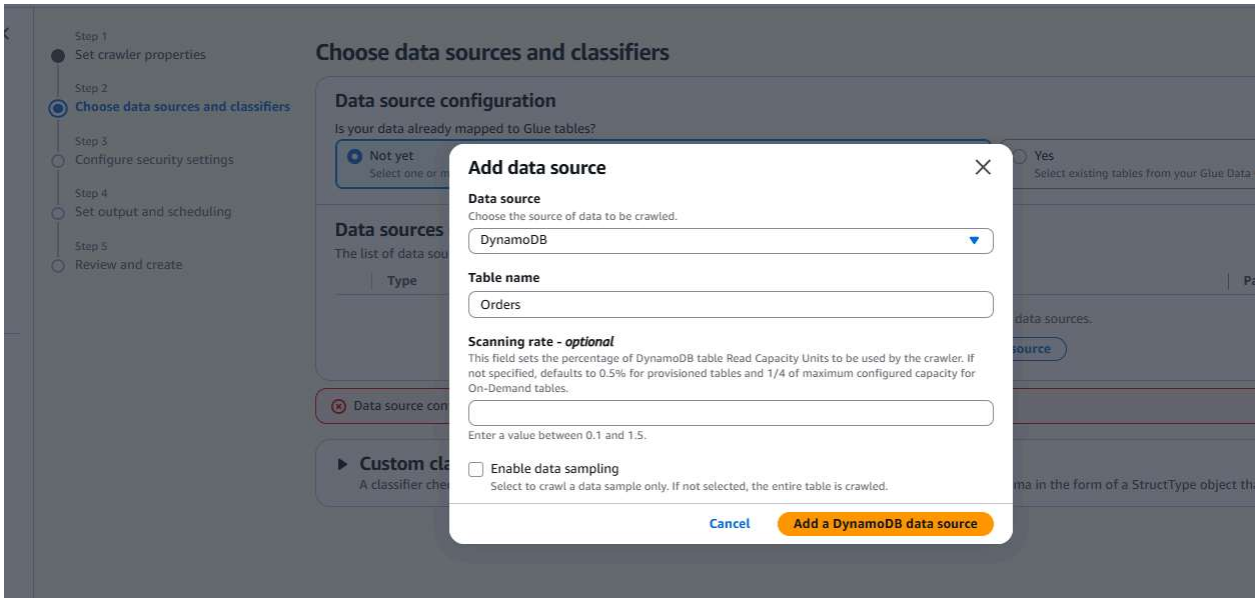
- Output: Glue Data Catalog → Database: `ordersdb`

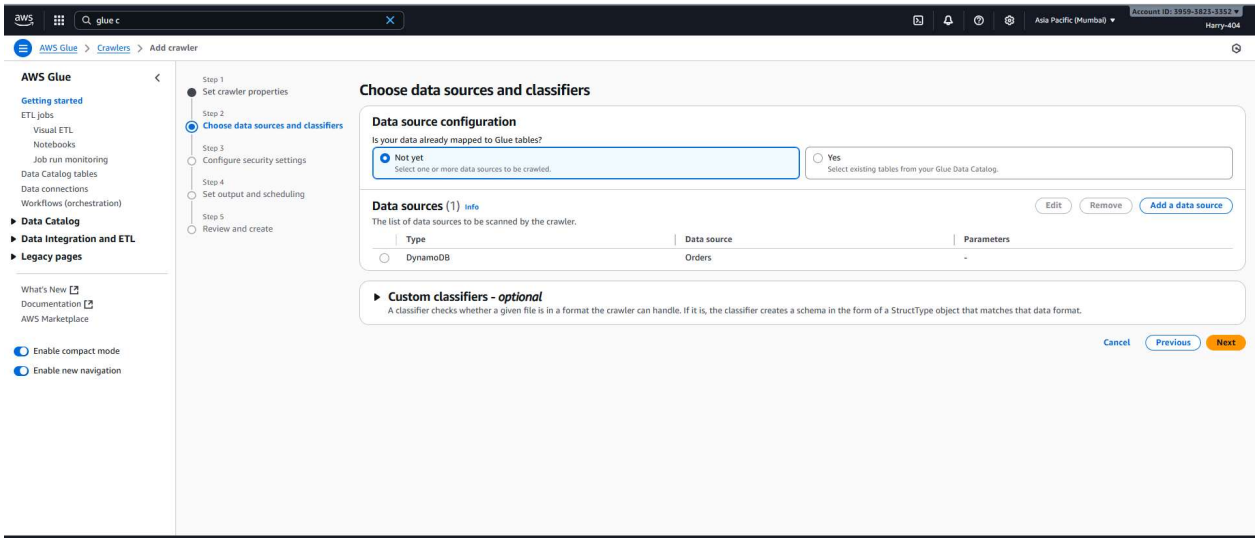
Run the crawler; it will infer the schema: OrderID, OrderDate, Customer, Amount, Status.
IAM role trust and permissions are handled as per wizard prompts and AWS docs.

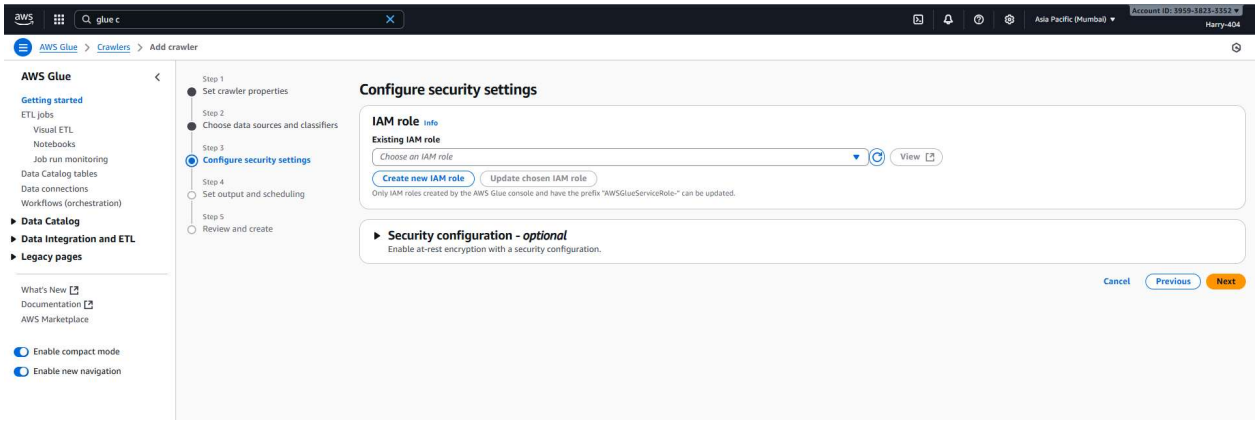

AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Welcome to AWS Glue
Get started 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
Resources and tutorials
Getting started with AWS Glue: DocumentationAWS Training
Glue in 5 Minutes Videos: Authoring, GenAI, Monitoring, Orchestration
Using connectors and connections
AWS Glue Documentation home
Examples: AWS Glue blog postsAWS Glue on GitHub
What's new in Glue
Amazon EMR on EC2 Adds Apache Spark native FGAC and AWS Glue Data Catalog Views Support
Aug 29, 2025
AWS Glue now supports Microsoft Dynamics 365 as a data source
Jul 24, 2025
AWS Glue Data Quality now supports Amazon S3 Tables and Iceberg Tables
Jul 23, 2025
View more what's new in Glue
Data integration and management
Monitor & debug ETL jobs and track usage
Go to job run monitoring
Connect to your data stores
Go to connections
Orchestrate jobs to build data pipelines
Go to workflows
Manage and protect data
Create a job to evaluate data quality
Create a job to detect sensitive data

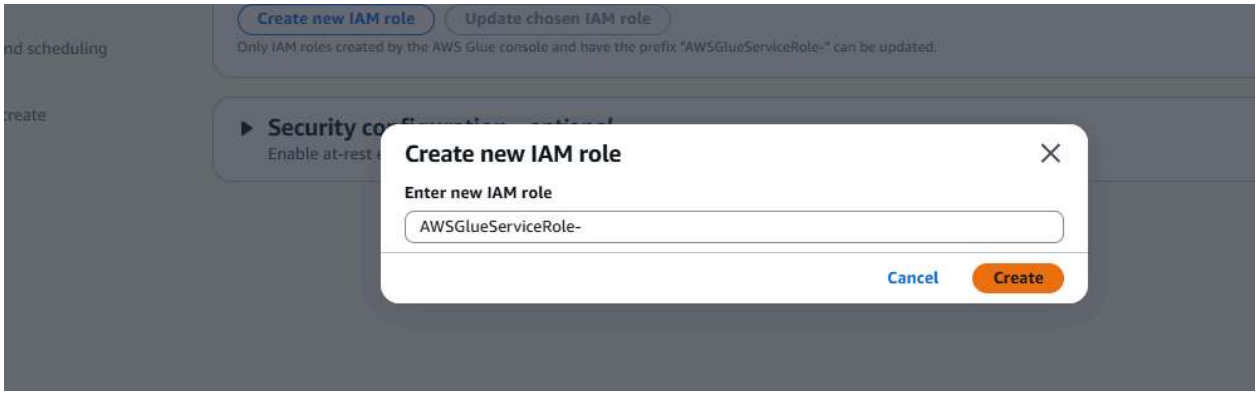

aws
glue c
Asia Pacific (Mumbai)
Harry-404
AWS Glue > Crawlers
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Crawlers
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.
Crawlers (0) Info
View and manage all available crawlers.
Last updated (UTC)
September 9, 2025 at 06:14:37
Action
Run
Create crawler
Filter crawlers
Name
State
Schedule
Last run
Last run timestamp
Log
Table changes from last run
No resources
No resources to display.

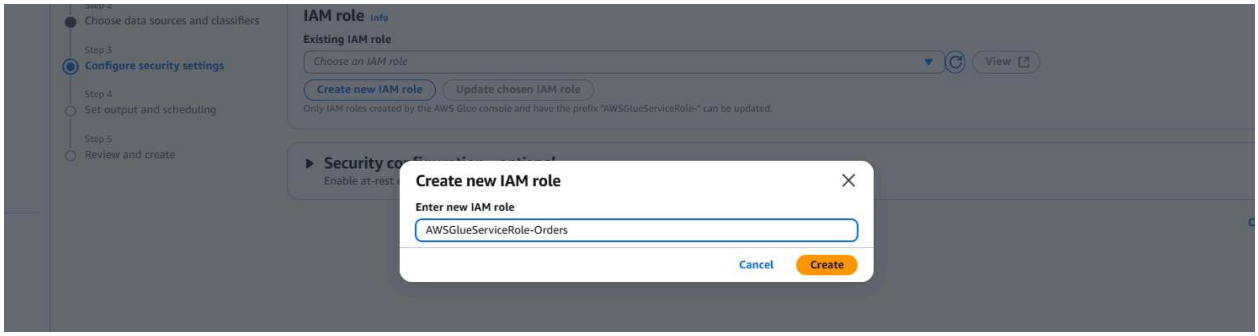

AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Step 1
Set crawler properties
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 Info
Name
OrdersCrawler
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

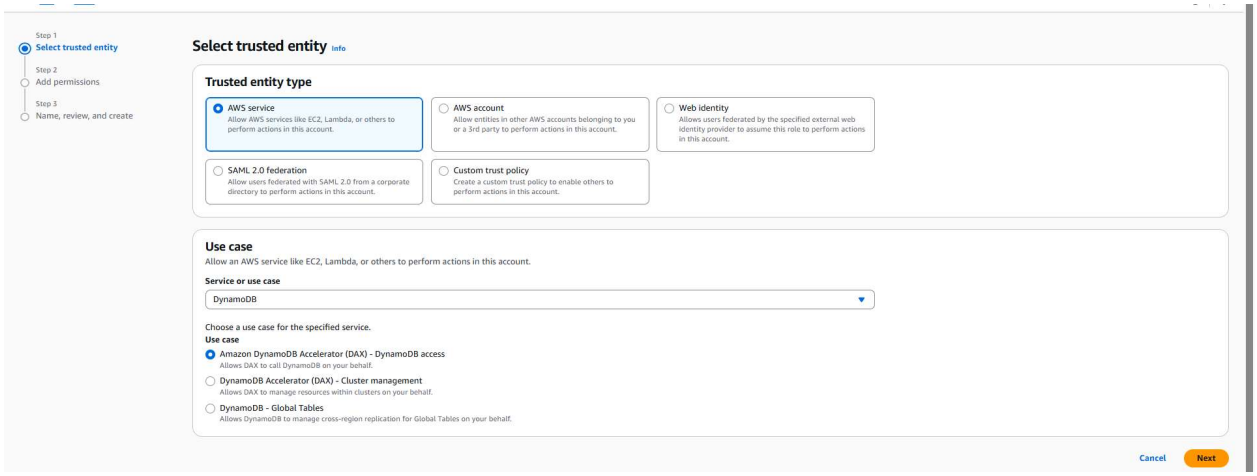

Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Yes
Data sources
Type
Add data source
Data source
Choose the source of data to be crawled.
DynamoDB
Table name
Orders
Scanning rate - optional
This field sets the percentage of DynamoDB table Read Capacity Units to be used by the crawler. If not specified, defaults to 0.5% for provisioned tables and 1/4 of maximum configured capacity for On-Demand tables.
Enter a value between 0.1 and 1.5.
Enable data sampling
Select to crawl a data sample only. If not selected, the entire table is crawled.
Cancel
Add a DynamoDB data source

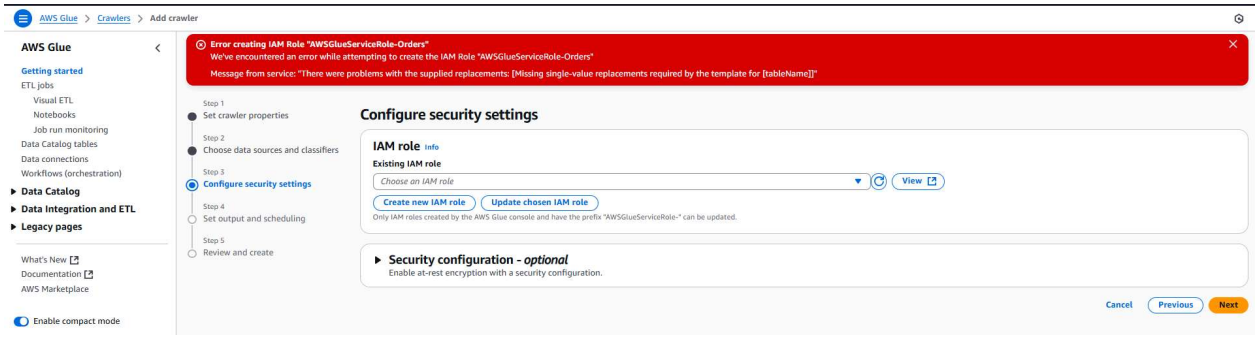

glue c
Asia Pacific (Mumbai)
Harry-404
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
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.
Edit
Remove
Add a data source
Type
Data source
Parameters
DynamoDB
Orders
-
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

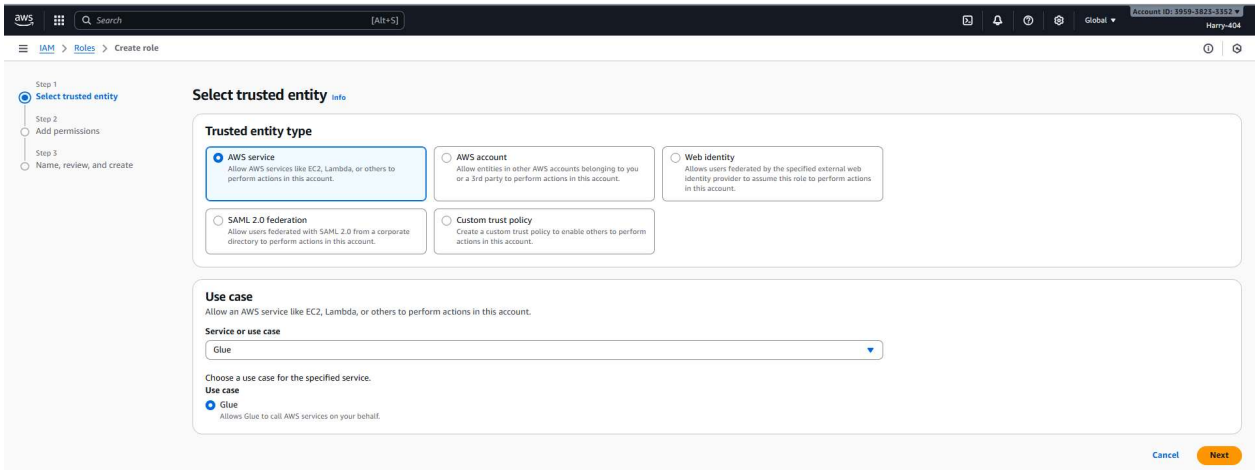

glue c
Asia Pacific (Mumbai)
Harry-404
AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
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
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.
Security configuration - optional
Enable at-rest encryption with a security configuration.
Cancel
Previous
Next

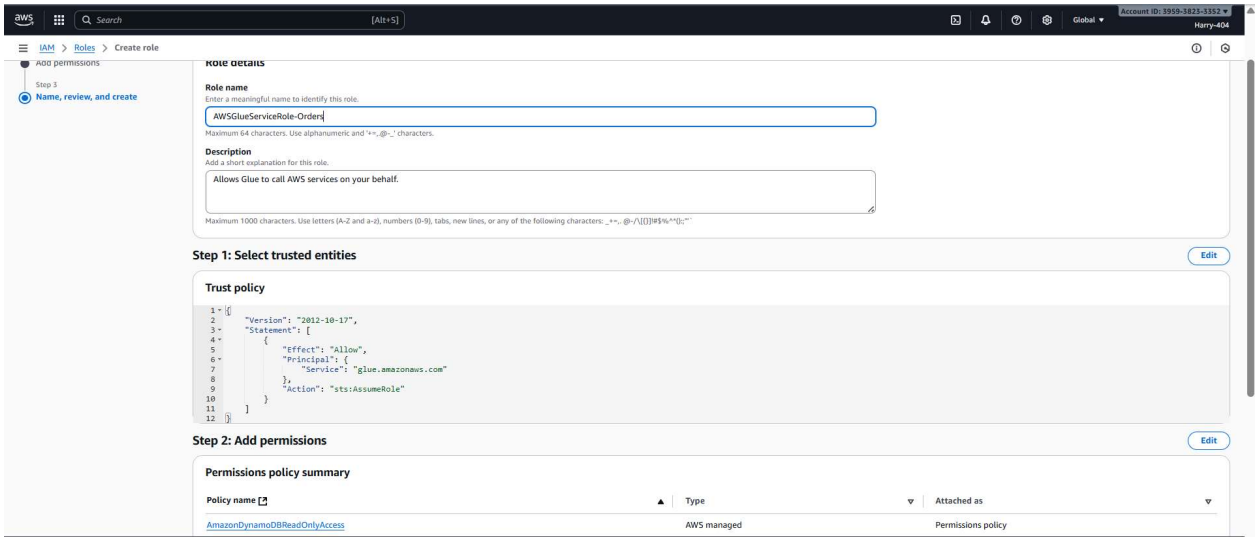

Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.
Create new IAM role
Enter new IAM role
AWSGlueServiceRole-
Cancel
Create

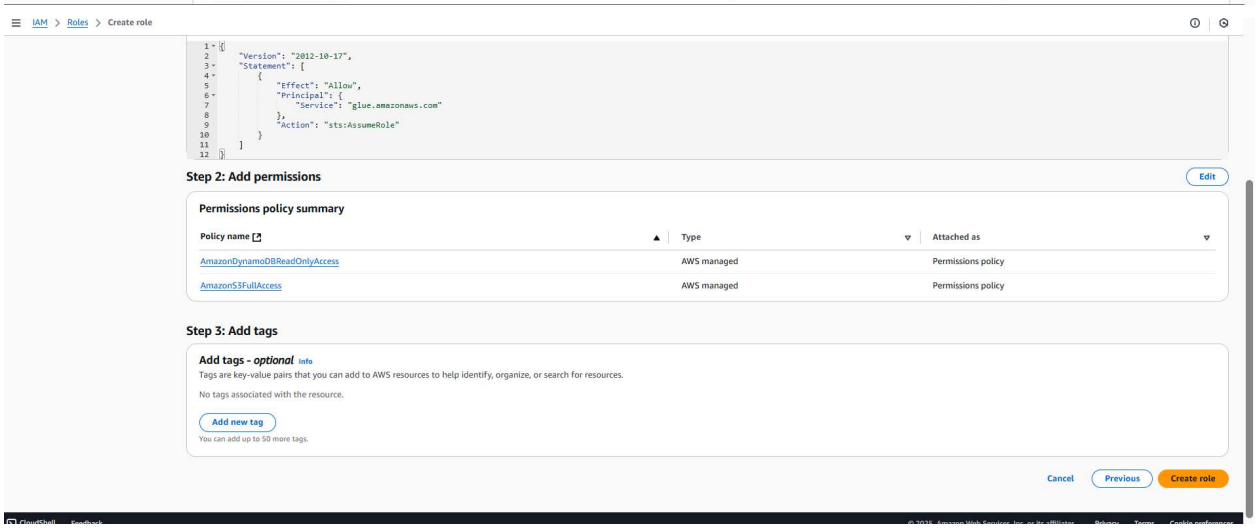

Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
IAM role Info
Existing IAM role
Choose an IAM role
View
Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.
Create new IAM role
Enter new IAM role
AWSGlueServiceRole-Orders
Cancel
Create

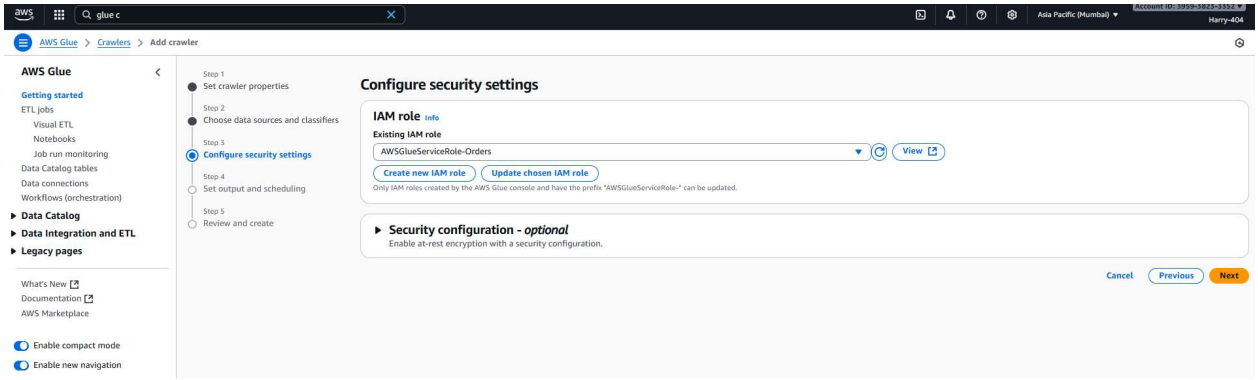

Step 1
Select trusted entity
Step 2
Add permissions
Step 3
Name, review, and create
Select trusted entity Info
Trusted entity type
AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.
AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.
Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.
SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.
Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.
Use case
Allow an AWS service like EC2, Lambda, or others to perform actions in this account.
Service or use case
DynamoDB
Choose a use case for the specified service.
Use case
Amazon DynamoDB Accelerator (DAX) - DynamoDB access
Allows DAX to call DynamoDB on your behalf.
DynamoDB Accelerator (DAX) - Cluster management
Allows DAX to manage resources within clusters on your behalf.
DynamoDB - Global Tables
Allows DynamoDB to manage cross-region replication for Global Tables on your behalf.
Cancel
Next

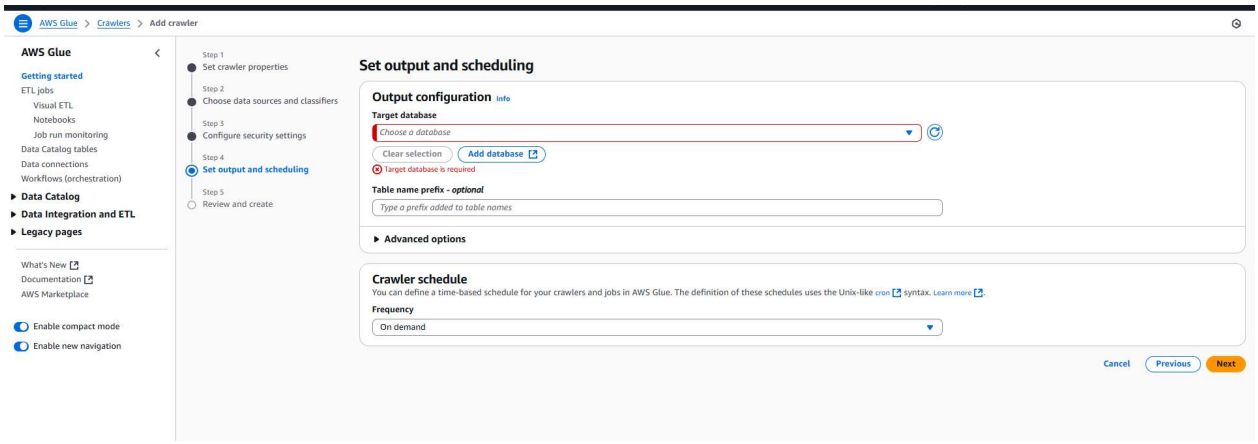

AWS Glue > Crawlers > Add crawler
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Error creating IAM Role "AWSGlueServiceRole-Orders"
We've encountered an error while attempting to create the IAM Role "AWSGlueServiceRole-Orders"
Message from service: "There were problems with the supplied replacements: [Missing single-value replacements required by the template for [tableName]]"
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
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.
Security configuration - optional
Enable at-rest encryption with a security configuration.
Cancel
Previous
Next


Search
[Alt+S]
Global
Account ID: 3959-3823-3352
Harry-404
IAM > Roles > Create role
Step 1
Select trusted entity
Step 2
Add permissions
Step 3
Name, review, and create
Select trusted entity Info
Trusted entity type
AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.
AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.
Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.
SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.
Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.
Use case
Allow an AWS service like EC2, Lambda, or others to perform actions in this account.
Service or use case
Glue
Choose a use case for the specified service.
Use case
Glue
Allows Glue to call AWS services on your behalf.
Cancel
Next


Search
[Alt+S]
Global
Account ID: 3959-3823-3352
Harry-404
IAM > Roles > Create role
Add permissions
Step 3
Name, review, and create
Role details
Role name
Enter a meaningful name to identify this role.
AWSGlueServiceRole-Orders
Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.
Description
Add a short explanation for this role.
Allows Glue to call AWS services on your behalf.
Maximum 1000 characters. Use letters (A-Z and a-z), numbers (0-9), tabs, new lines, or any of the following characters: _+=, .@-/\[{}]!#$%^*();:"`
Step 1: Select trusted entities
Edit
Trust policy
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": {
"Service": "glue.amazonaws.com"
},
"Action": "sts:AssumeRole"
}
]
}
Step 2: Add permissions
Edit
Permissions policy summary
Policy name
Type
Attached as
AmazonDynamoDBReadOnlyAccess
AWS managed
Permissions policy

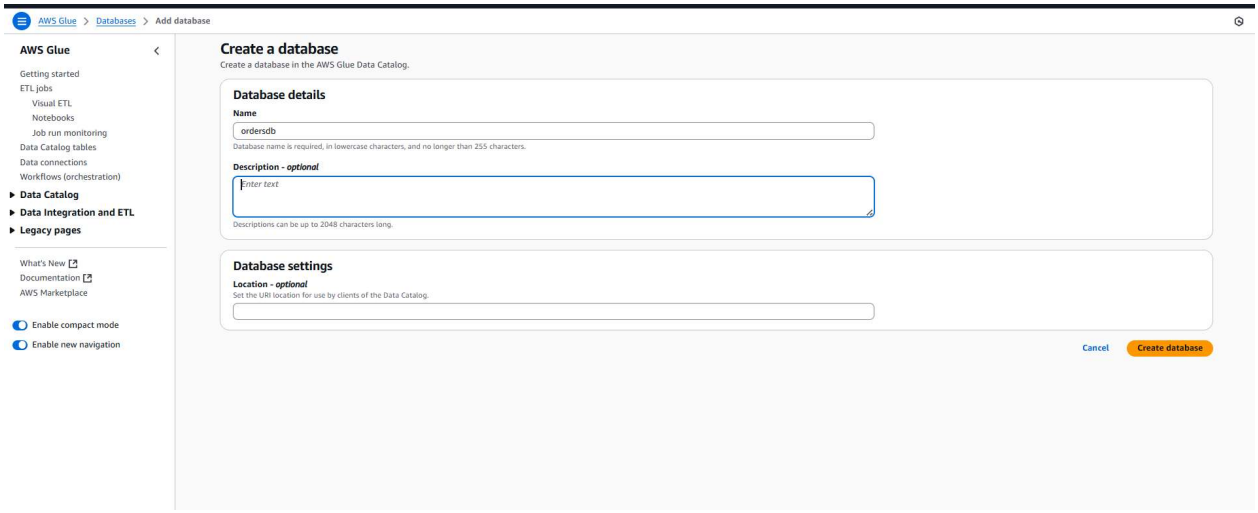

IAM > Roles > Create role
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": {
"Service": "glue.amazonaws.com"
},
"Action": "sts:AssumeRole"
}
]
}
Step 2: Add permissions
Edit
Permissions policy summary
Policy name
Type
Attached as
AmazonDynamoDBReadOnlyAccess
AWS managed
Permissions policy
AmazonS3FullAccess
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

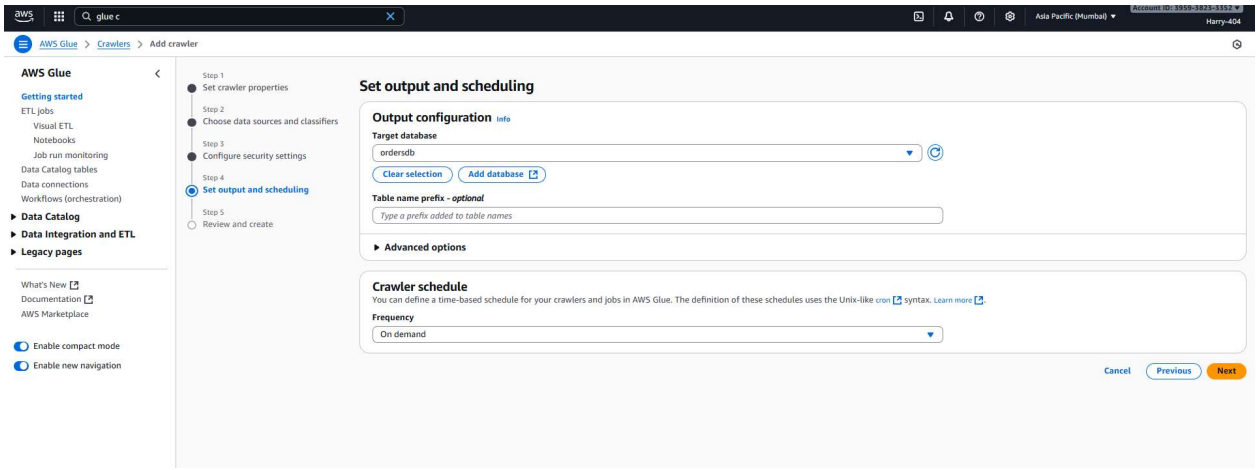


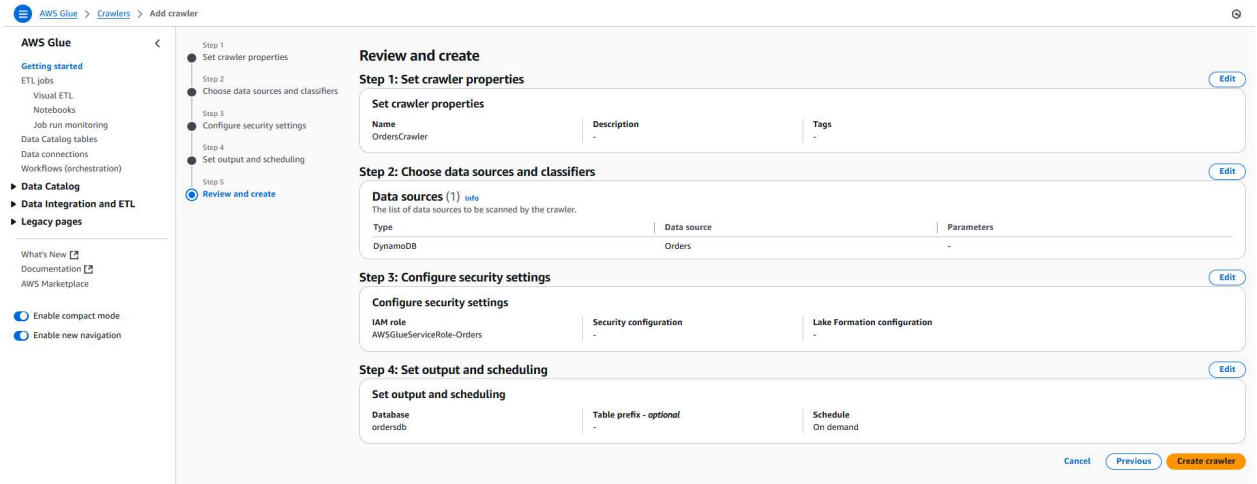


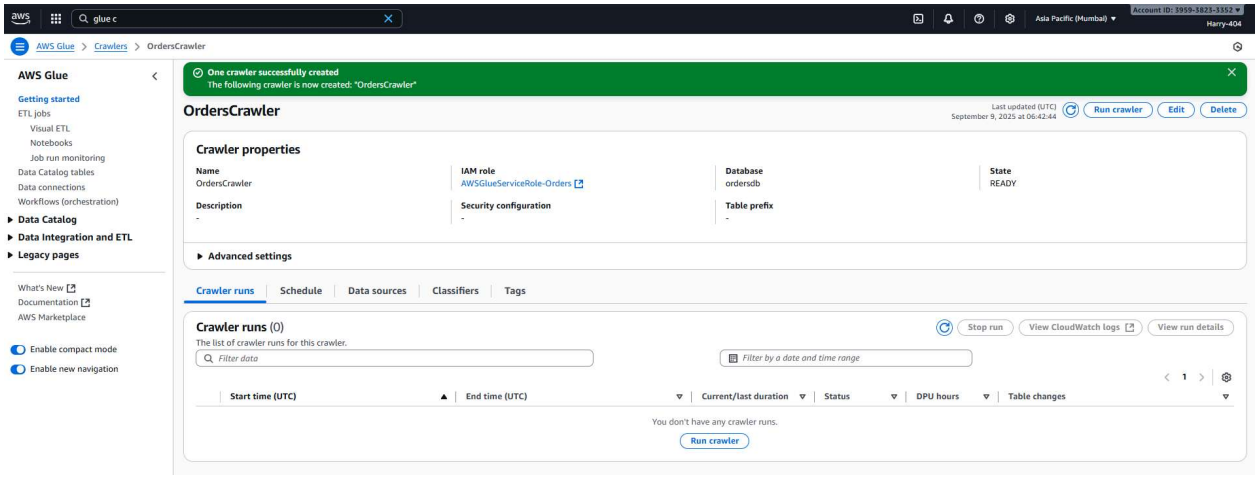
Create Glue Database:

- In AWS Glue Console, go to Databases → Add database
- Database name: ordersdb (must be all lowercase)
- Click Create database


AWS Glue > Databases > Add database
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Create a database
Create a database in the AWS Glue Data Catalog.
Database details
Name
ordersdb
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.
Cancel
Create database


glue c
Asia Pacific (Mumbai)
Harry-404
AWS Glue > 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
Set output and scheduling
Output configuration Info
Target database
ordersdb
Clear selection
Add database
Table name prefix - optional
Type a prefix added to table names
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

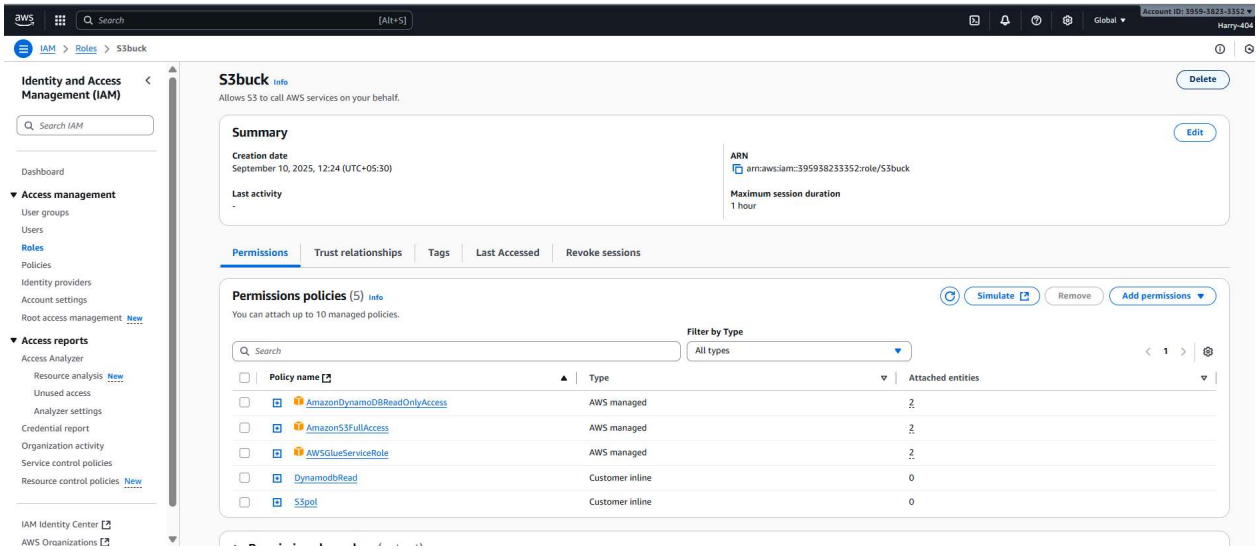

AWS Glue > Crawlers > Add crawler
Review and create
Step 1: Set crawler properties
Edit
Set crawler properties
Name
OrdersCrawler
Description
-
Tags
-
Step 2: Choose data sources and classifiers
Edit
Data sources (1) Info
The list of data sources to be scanned by the crawler.
Type
Data source
Parameters
DynamoDB
Orders
-
Step 3: Configure security settings
Edit
Configure security settings
IAM role
AWSGlueServiceRole-Orders
Security configuration
-
Lake Formation configuration
-
Step 4: Set output and scheduling
Edit
Set output and scheduling
Database
ordersdb
Table prefix - optional
-
Schedule
On demand
Cancel
Previous
Create crawler

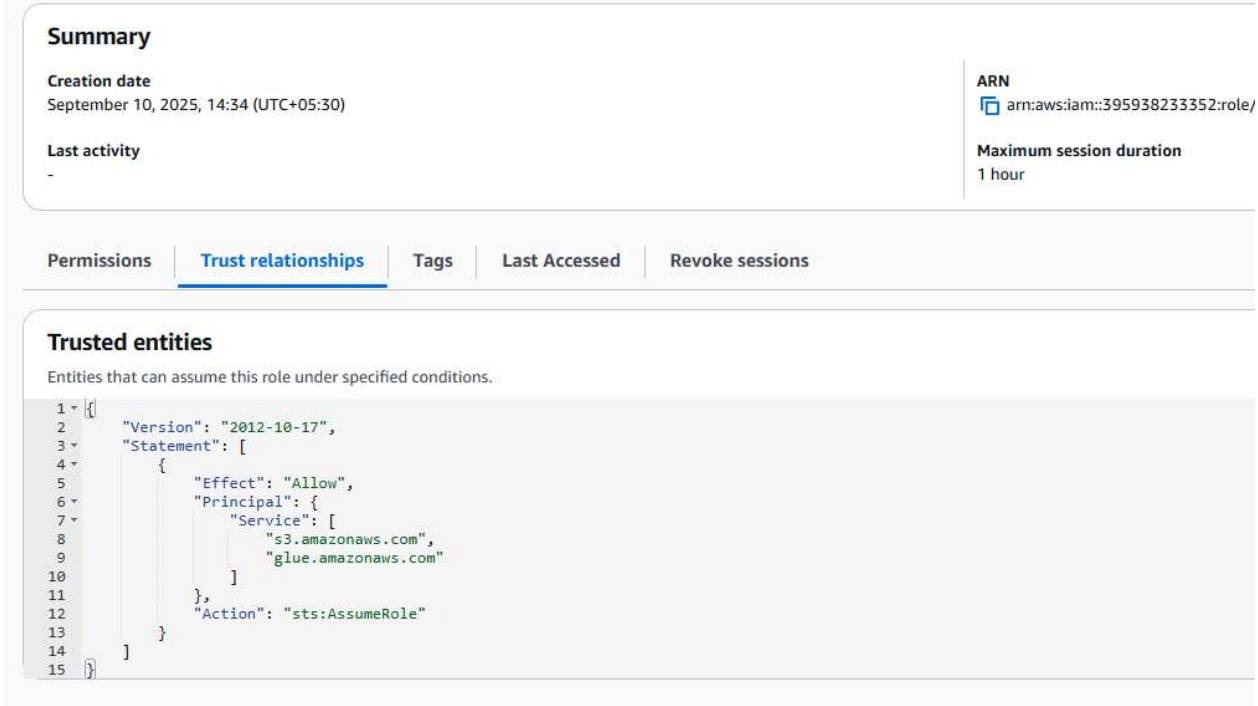


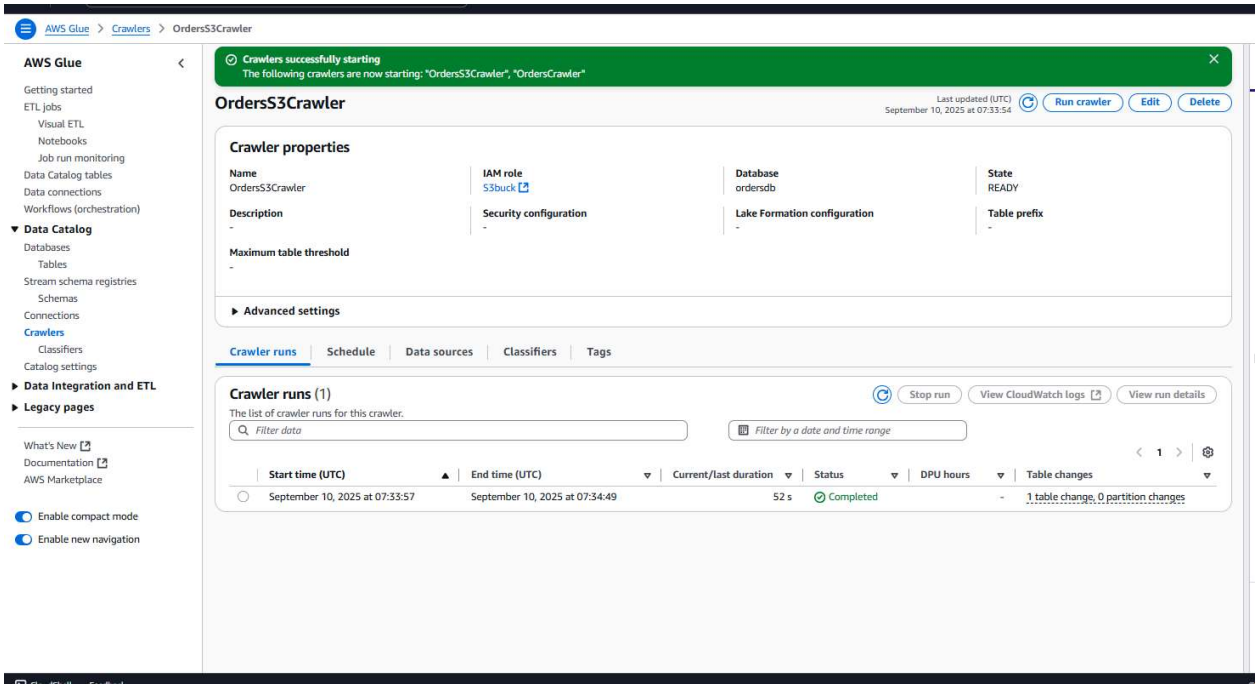
Step 4: Glue Crawler 2 (OrdersS3Crawler): Catalog Cleaned S3 Data

- Purpose: After Glue ETL job writes transformed data to S3, this crawler catalogs the S3 Parquet files.
- Name: OrdersS3Crawler
- Source: S3 → Bucket or prefix (e.g., `s3://my-orders-analytics/shipped/`)
- IAM Role: Create/select role (e.g., `S3buck`) with S3 read, Glue, and DynamoDB read-only permissions
    - Attach: `AmazonS3FullAccess`, `AmazonDynamoDBReadOnlyAccess`, `AWSGlueServiceRole`
    - Inline Permission: `DynamDbRead, S3pol`
- Output: Glue Data Catalog → Database: `ordersdb` (or a separate database if preferred)

Run the crawler after ETL job has created Parquet files in S3 to update the schema in Glue Catalog as per S3 data structure.


Search
[Alt+S]
Global
Account ID: 3959-3823-3352
Harry-404
IAM > Roles > S3buck
Identity and Access Management (IAM)
Search IAM
Dashboard
Access management
User groups
Users
Roles
Policies
Identity providers
Account settings
Root access management New
Access reports
Access Analyzer
Resource analysis New
Unused access
Analyzer settings
Credential report
Organization activity
Service control policies
Resource control policies New
IAM Identity Center
AWS Organizations
S3buck Info
Allows S3 to call AWS services on your behalf.
Delete
Summary
Edit
Creation date
September 10, 2025, 12:24 (UTC+05:30)
ARN
arn:aws:iam::395938233352:role/S3buck
Last activity
-
Maximum session duration
1 hour
Permissions
Trust relationships
Tags
Last Accessed
Revoke sessions
Permissions policies (5) Info
You can attach up to 10 managed policies.
Simulate
Remove
Add permissions
Filter by Type
Search
All types
Policy name
Type
Attached entities
AmazonDynamoDBReadOnlyAccess
AWS managed
2
AmazonS3FullAccess
AWS managed
2
AWSGlueServiceRole
AWS managed
2
DynamodbRead
Customer inline
0
S3pol
Customer inline
0


Summary
Creation date
September 10, 2025, 14:34 (UTC+05:30)
ARN
arn:aws:iam::395938233352:role/
Last activity
-
Maximum session duration
1 hour
Permissions
Trust relationships
Tags
Last Accessed
Revoke sessions
Trusted entities
Entities that can assume this role under specified conditions.
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": {
"Service": [
"s3.amazonaws.com",
"glue.amazonaws.com"
]
},
"Action": "sts:AssumeRole"
}
]
}



Step 5: Glue ETL Job (Visual Editor/Studio, PySpark Script)

- Source: Glue Data Catalog Table: ordersdb.orders (from DynamoDB, cataloged by OrdersCrawler)
- Target: S3 → `s3://my-orders-analytics/shipped/`
    - Format: Parquet (compression: Snappy, ideal for analytics)
- Transformation: Filter for orders where Status = "Shipped"
- Role: Use the same IAM role as for S3 Crawler, with access to S3, Glue, and DynamoDB.
- Method: Either visual mapping or PySpark code as below:

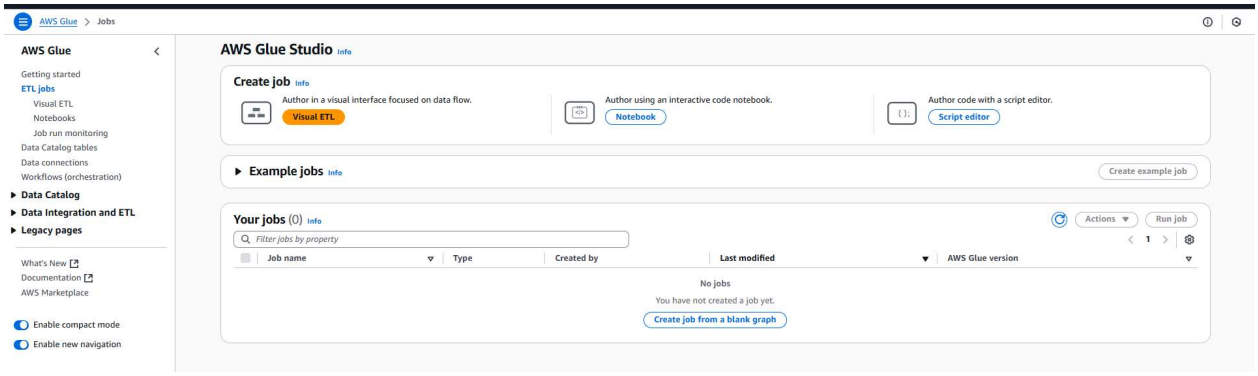
Sample Glue PySpark Script

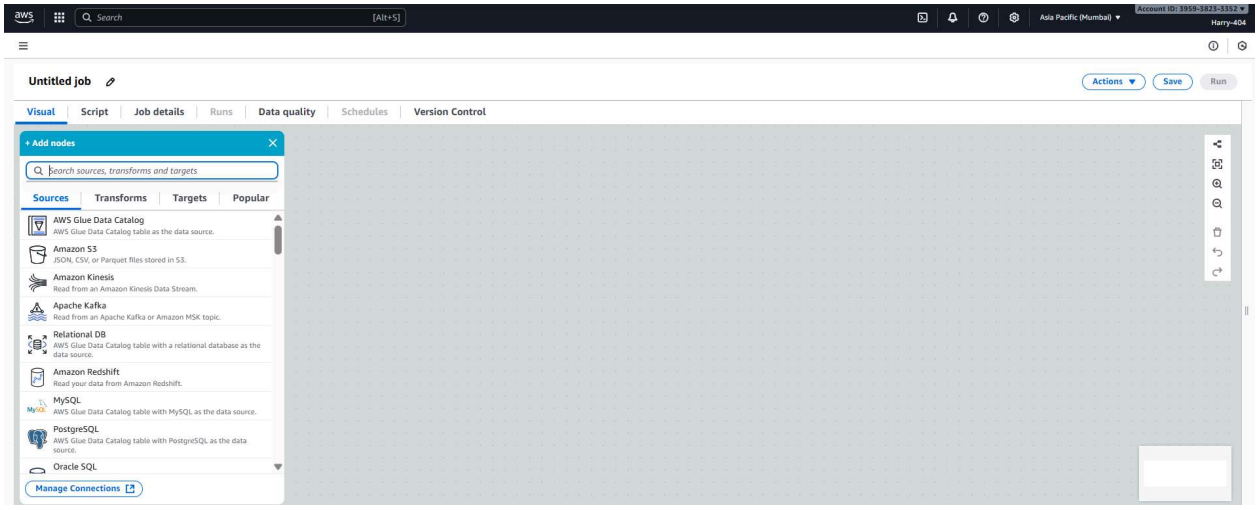
```
python

import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
```

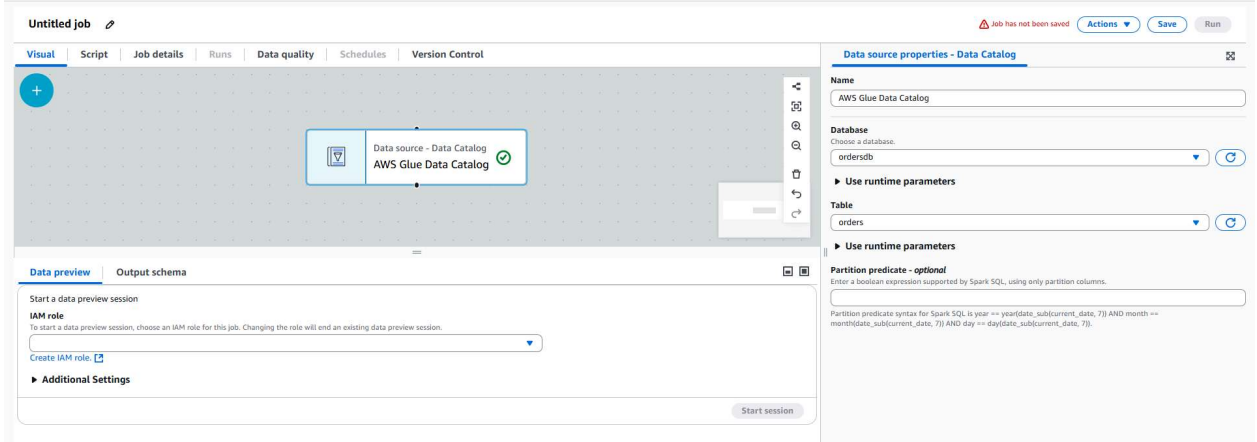
```
args = getResolvedOptions(sys.argv, ['JOB_NAME'])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)

# Read DynamoDB (via Glue Catalog)
datasource = glueContext.create_dynamic_frame.from_catalog(
    database="ordersdb",
    table_name="orders"
)

# Filter only shipped orders
filtered = Filter.apply(datasource, lambda x: x["status"] ==
"Shipped")

# Write to S3 as Parquet
glueContext.write_dynamic_frame.from_options(
    frame=filtered,
    connection_type="s3",
    connection_options={"path":
"s3://my-orders-analytics/shipped/"},
    format="parquet"
)
```

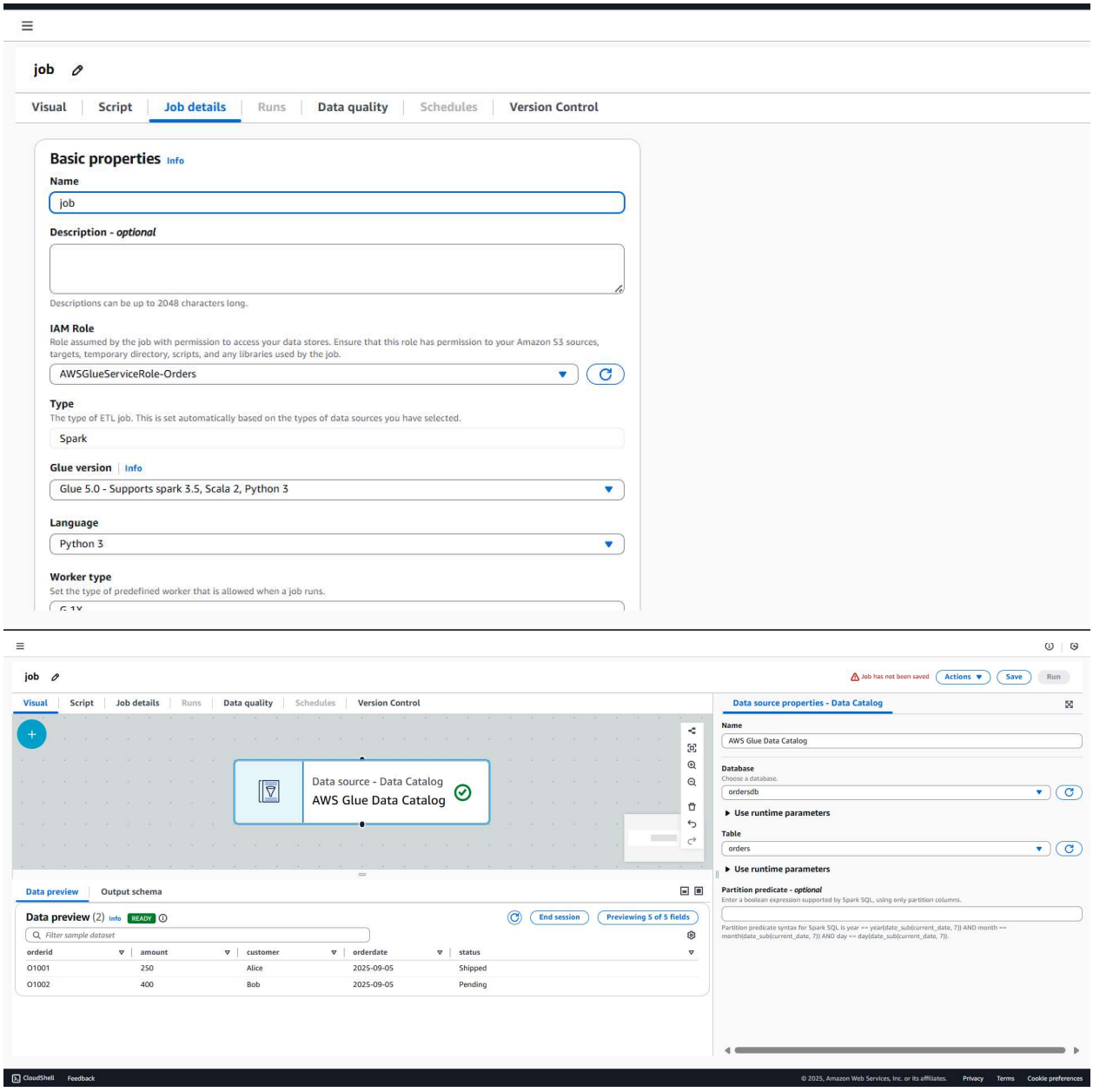
```
job.commit()
```

- In Glue Studio Visual Mode: Use drag-and-drop to set up the same source, filter (status = shipped), S3 as Parquet target. Map columns as shown in your screenshots.

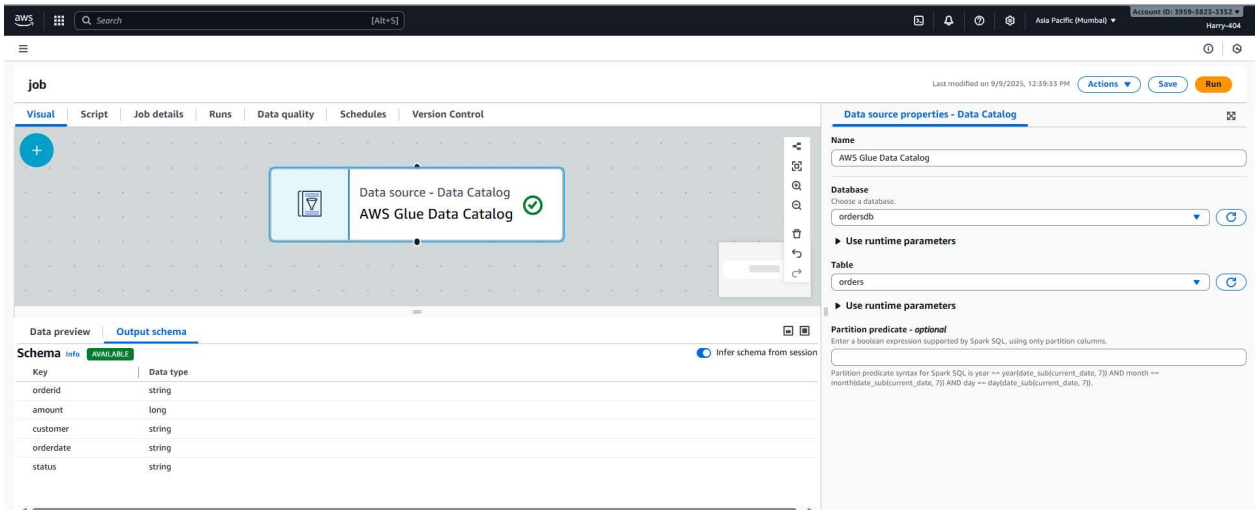


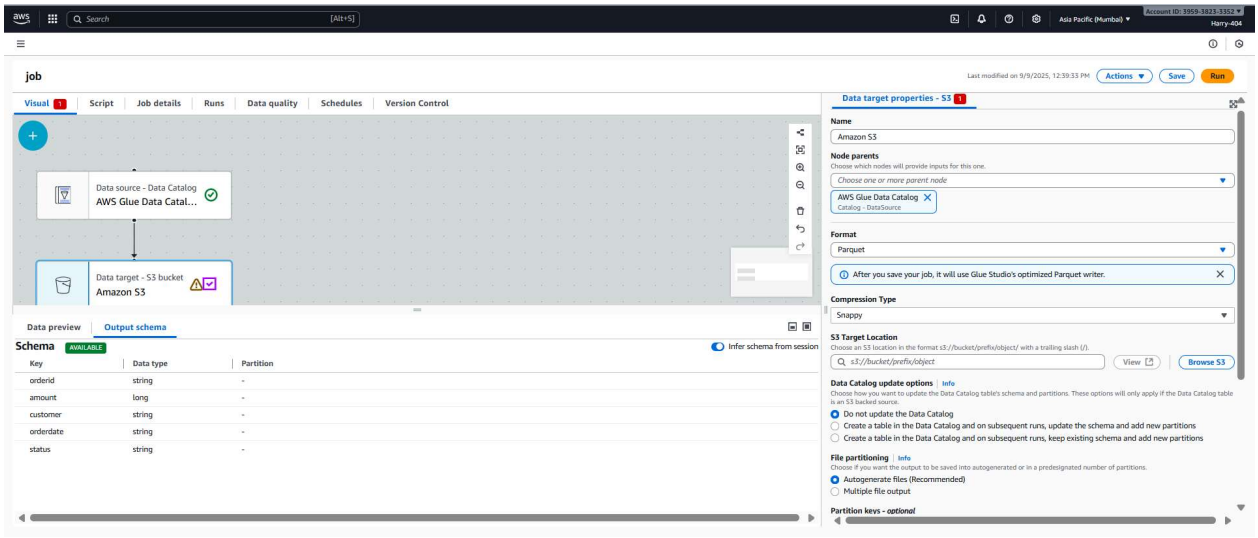


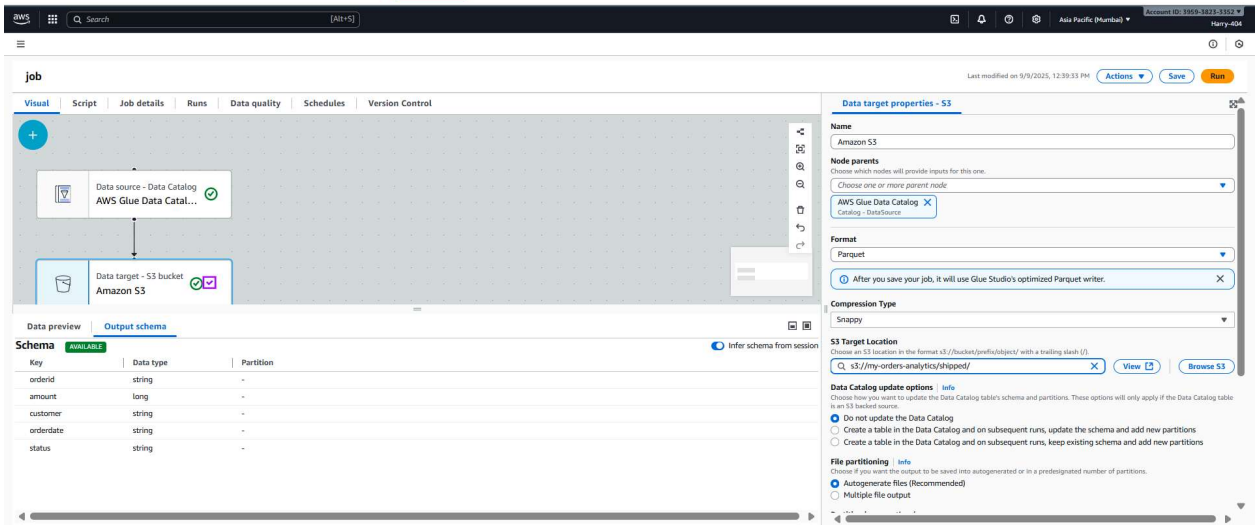


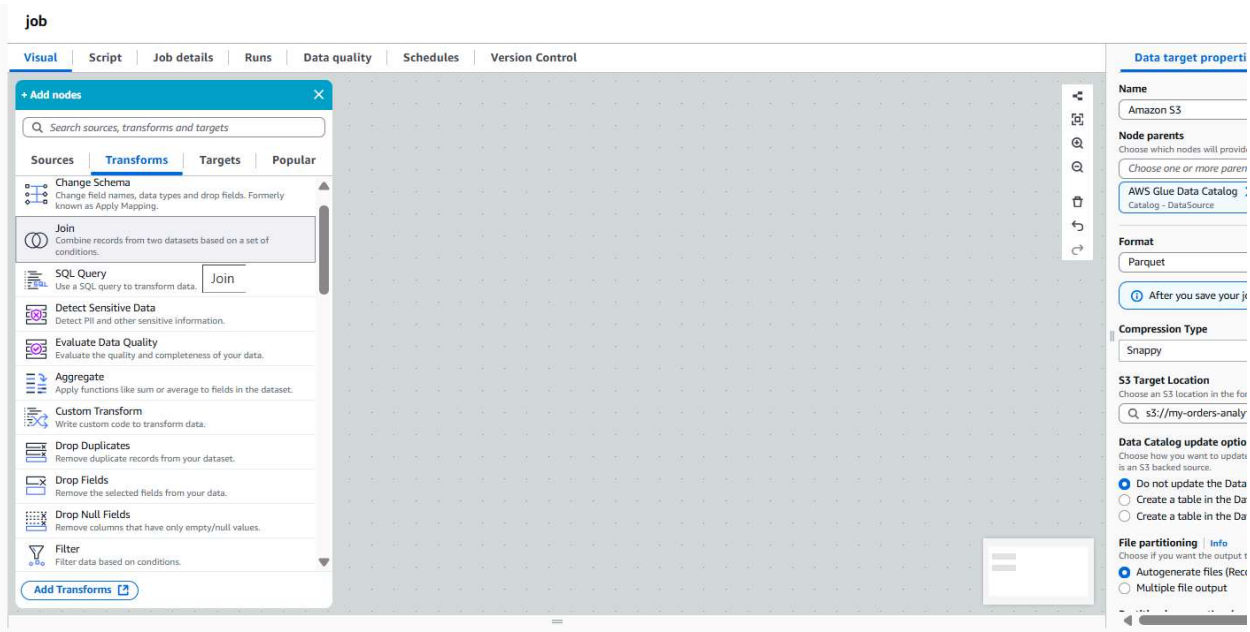

job
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Basic properties Info
Name
job
Description - optional
Descriptions can be up to 2048 characters long.
IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
AWSGlueServiceRole-Orders
Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
Spark
Glue version Info
Glue 5.0 - Supports spark 3.5, Scala 2, Python 3
Language
Python 3
Worker type
Set the type of predefined worker that is allowed when a job runs.

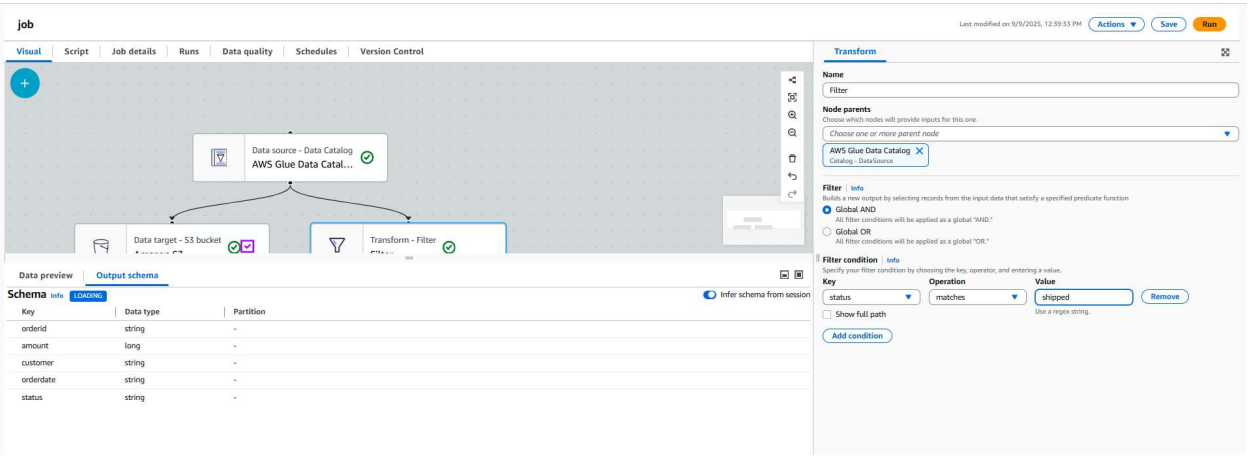
job
Job has not been saved
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catalog
Data source properties - Data Catalog
Name
AWS Glue Data Catalog
Database
Choose a database.
ordersdb
Use runtime parameters
Table
orders
Use runtime parameters
Partition predicate - optional
Enter a boolean expression supported by Spark SQL, using only partition columns.
Data preview
Output schema
Data preview (2) Info
READY
End session
Previewing 5 of 5 fields
Filter sample dataset
orderid
amount
customer
orderdate
status
O1001
250
Alice
2025-09-05
Shipped
O1002
400
Bob
2025-09-05
Pending
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

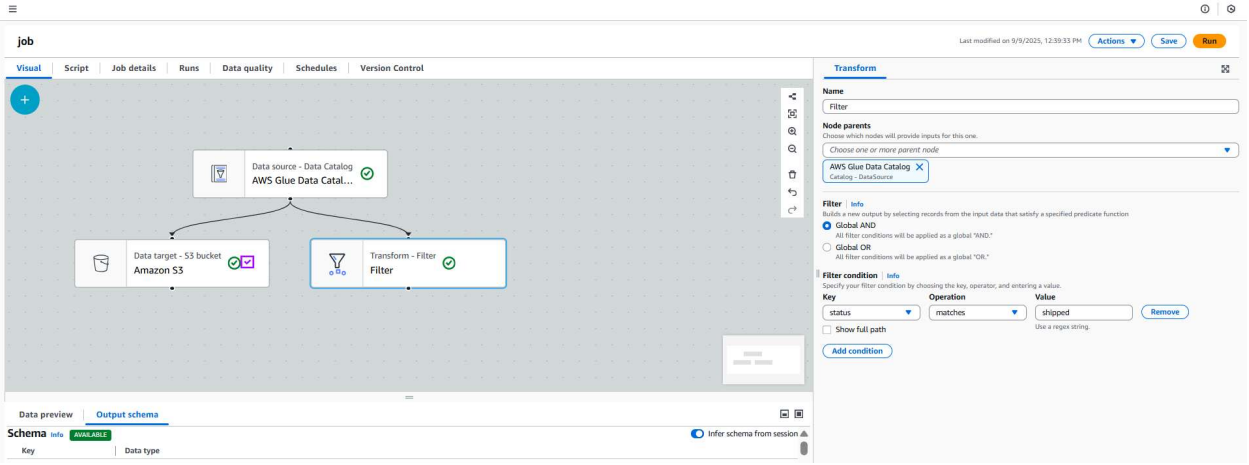

job
Last modified on 9/9/2025, 12:39:33 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catalog
Data source properties - Data Catalog
Name
AWS Glue Data Catalog
Database
Choose a database.
ordersdb
Use runtime parameters
Table
orders
Use runtime parameters
Partition predicate - optional
Enter a boolean expression supported by Spark SQL, using only partition columns.
Data preview
Output schema
Schema
AVAILABLE
Infer schema from session
Key
Data type
orderid string
amount long
customer string
orderdate string
status string

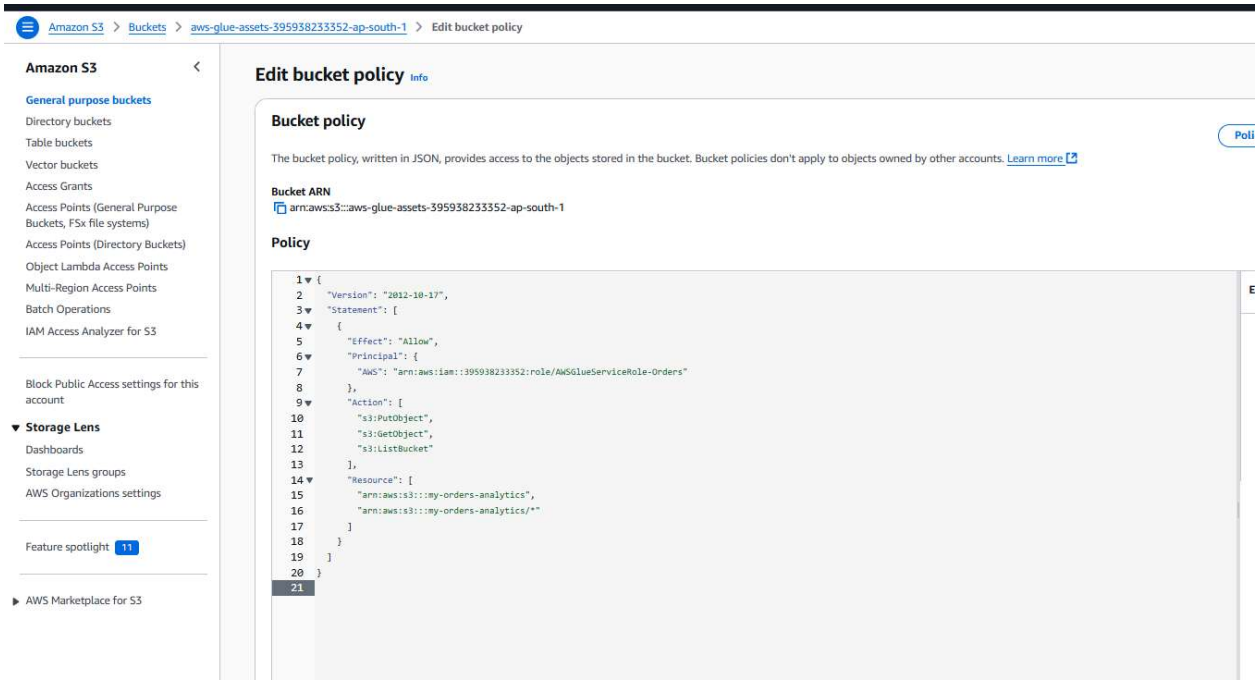

job
Last modified on 9/9/2025, 12:39:33 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catal...
Data target - S3 bucket
Amazon S3
Data target properties - S3
Name
Amazon S3
Node parents
Choose one or more parent node
AWS Glue Data Catalog
Format
Parquet
After you save your job, it will use Glue Studio's optimized Parquet writer.
Compression Type
Snappy
S3 Target Location
s3://bucket/prefix/object
View
Browse S3
Data Catalog update options
Do not update the Data Catalog
Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions
File partitioning
Autogenerate files (Recommended)
Multiple file output
Partition keys - optional
Data preview
Output schema
Schema
AVAILABLE
Infer schema from session
Key
Data type
Partition
orderid string
amount long
customer string
orderdate string
status string

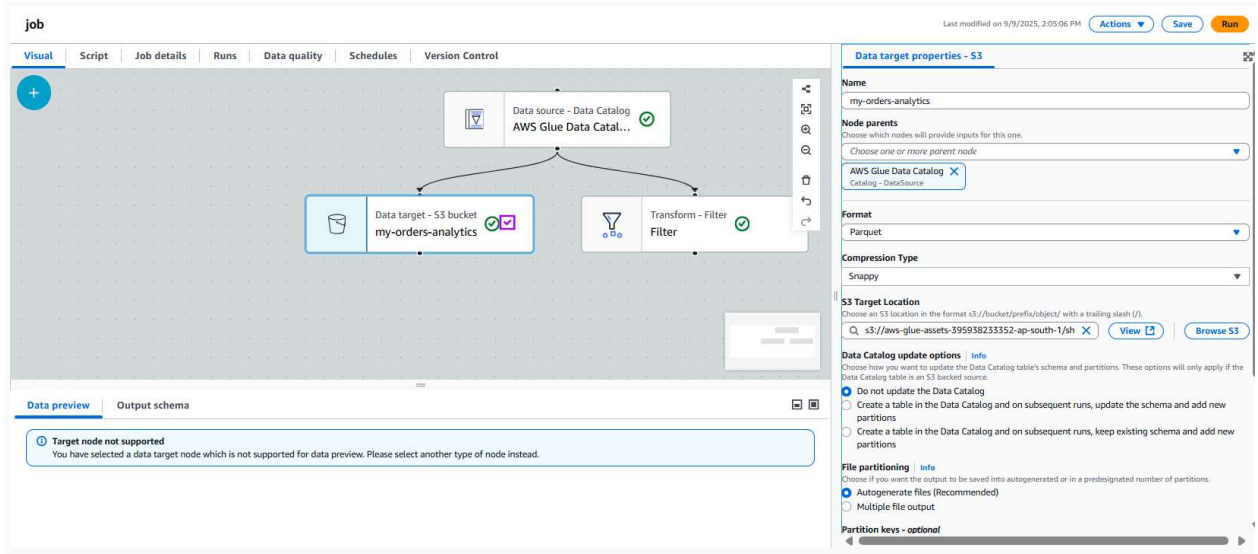

job
Last modified on 9/9/2025, 12:39:33 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catal...
Data target - S3 bucket
Amazon S3
Data target properties - S3
Name
Amazon S3
Node parents
Choose one or more parent node
AWS Glue Data Catalog
Format
Parquet
After you save your job, it will use Glue Studio's optimized Parquet writer.
Compression Type
Snappy
S3 Target Location
s3://my-orders-analytics/shipped/
View
Browse S3
Data Catalog update options
Do not update the Data Catalog
Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions
File partitioning
Autogenerate files (Recommended)
Multiple file output
Data preview
Output schema
Schema
AVAILABLE
Infer schema from session
Key
Data type
Partition
orderid string
amount long
customer string
orderdate string
status string

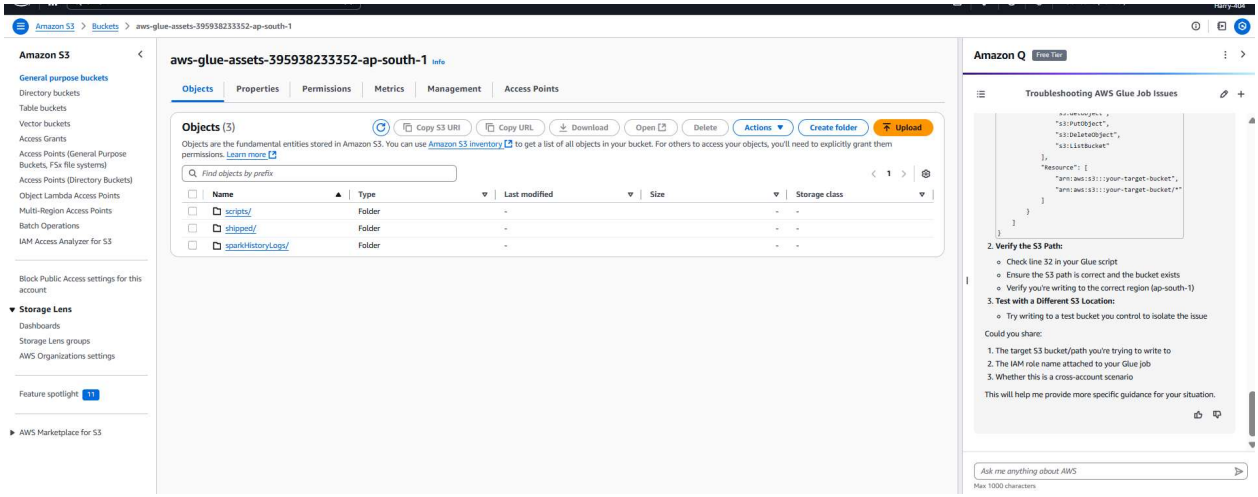

job
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
+ Add nodes
Search sources, transforms and targets
Sources
Transforms
Targets
Popular
Change Schema
Change field names, data types and drop fields. Formerly known as Apply Mapping.
Join
Combine records from two datasets based on a set of conditions.
SQL Query
Use a SQL query to transform data.
Join
Detect Sensitive Data
Detect PII and other sensitive information.
Evaluate Data Quality
Evaluate the quality and completeness of your data.
Aggregate
Apply functions like sum or average to fields in the dataset.
Custom Transform
Write custom code to transform data.
Drop Duplicates
Remove duplicate records from your dataset.
Drop Fields
Remove the selected fields from your data.
Drop Null Fields
Remove columns that have only empty/null values.
Filter
Filter data based on conditions.
Add Transforms
Data target properti
Name
Amazon S3
Node parents
Choose one or more paren
AWS Glue Data Catalog
Catalog - DataSource
Format
Parquet
After you save your j
Compression Type
Snappy
S3 Target Location
s3://my-orders-analy
Data Catalog update optio
Do not update the Data
Create a table in the Da
Create a table in the Da
File partitioning
Info
Autogenerate files (Rec
Multiple file output

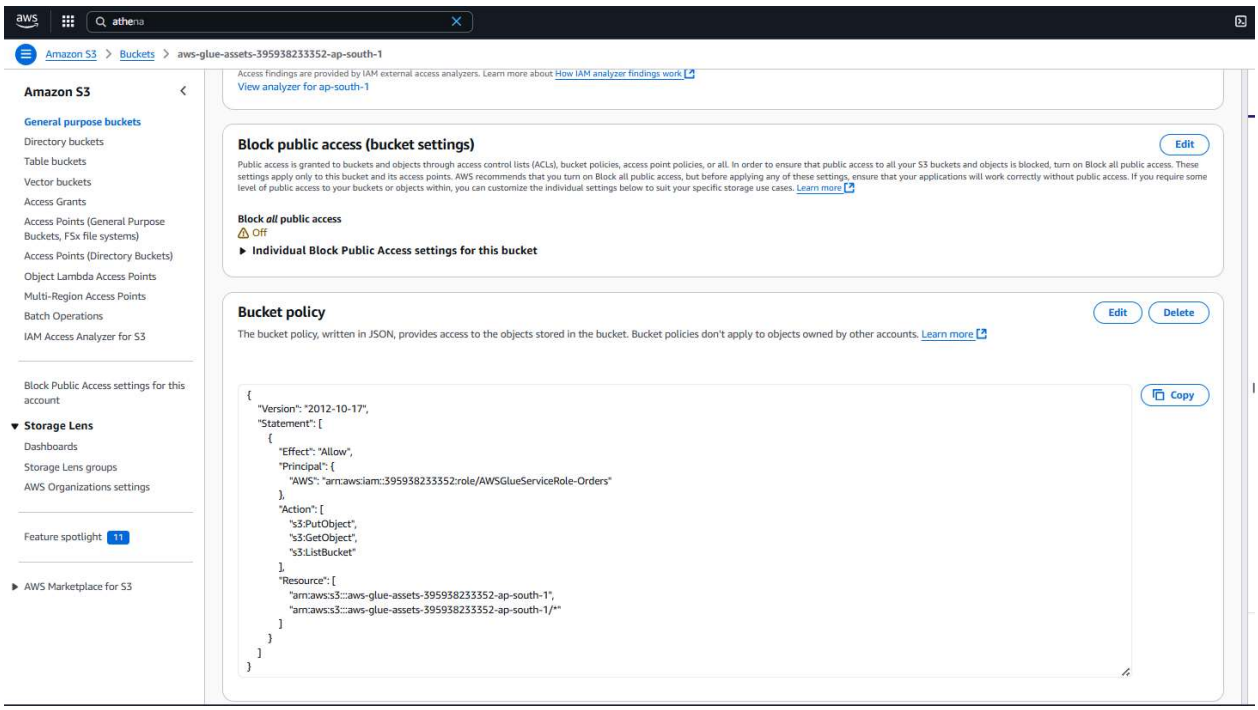

job
Last modified on 9/9/2025, 12:39:33 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catal...
Data target - S3 bucket
Transform - Filter
Data preview
Output schema
Schema
Info
LOADING
Infer schema from session
Key
Data type
Partition
orderid
string
amount
long
customer
string
orderdate
string
status
string
Transform
Name
Filter
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
AWS Glue Data Catalog
Catalog - DataSource
Filter
Info
Builds a new output by selecting records from the input data that satisfy a specified predicate function
Global AND
All filter conditions will be applied as a global "AND."
Global OR
All filter conditions will be applied as a global "OR."
Filter condition
Info
Specify your filter condition by choosing the key, operator, and entering a value.
Key
Operation
Value
status
matches
shipped
Remove
Use a regex string.
Show full path
Add condition

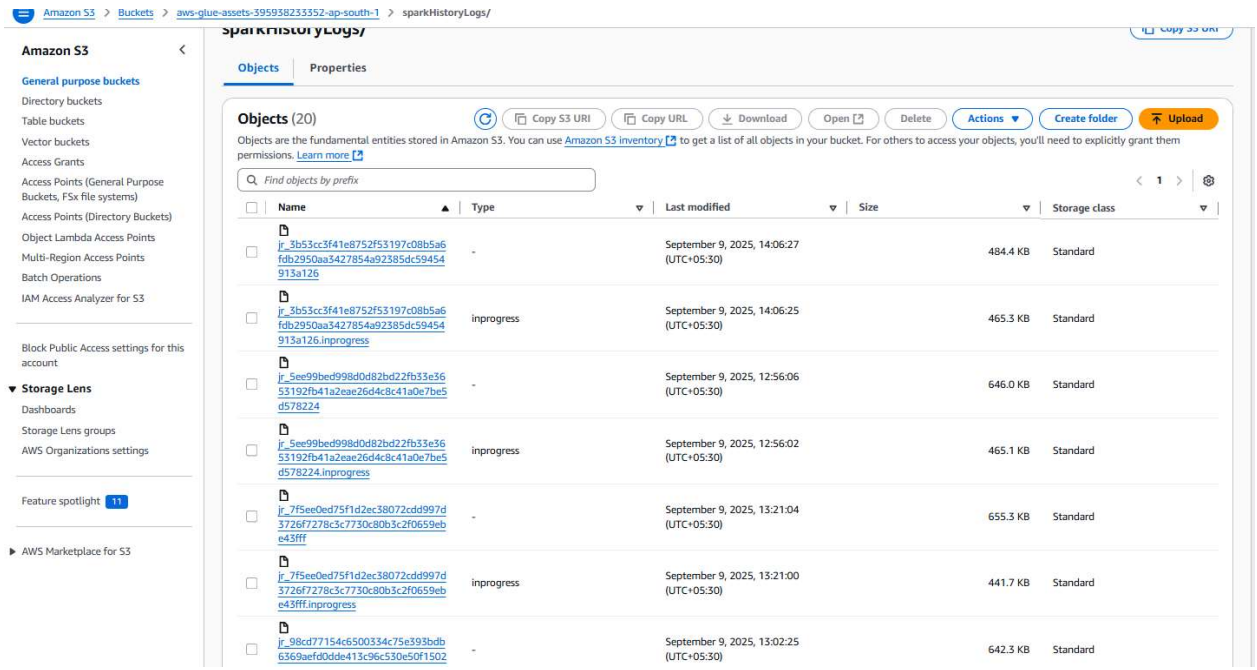

job
Last modified on 9/9/2025, 12:39:33 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catal...
Data target - S3 bucket
Amazon S3
Transform - Filter
Filter
Data preview
Output schema
Schema
Info
AVAILABLE
Infer schema from session
Key
Data type
Transform
Name
Filter
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
AWS Glue Data Catalog
Catalog - DataSource
Filter
Info
Builds a new output by selecting records from the input data that satisfy a specified predicate function
Global AND
All filter conditions will be applied as a global "AND."
Global OR
All filter conditions will be applied as a global "OR."
Filter condition
Info
Specify your filter condition by choosing the key, operator, and entering a value.
Key
Operation
Value
status
matches
shipped
Remove
Use a regex string.
Show full path
Add condition

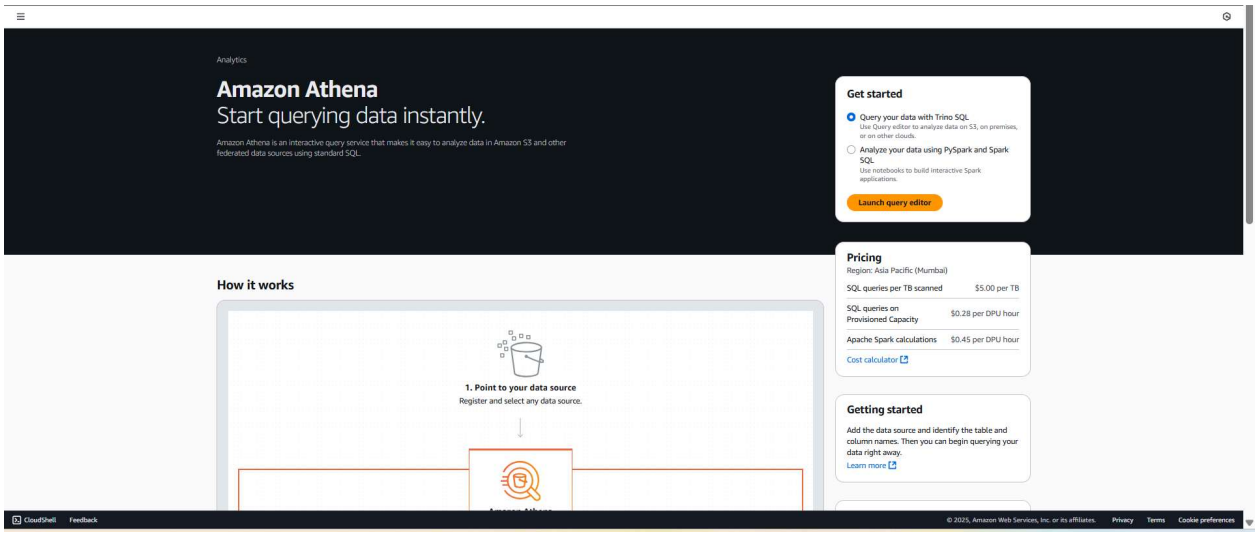

Amazon S3 > Buckets > aws-glue-assets-395938233352-ap-south-1 > Edit bucket policy
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Vector buckets
Access Grants
Access Points (General Purpose Buckets, FSx file systems)
Access Points (Directory Buckets)
Object Lambda Access Points
Multi-Region 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
AWS Marketplace for S3
Edit bucket policy Info
Bucket policy
The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts. Learn more
Bucket ARN
arn:aws:s3:::aws-glue-assets-395938233352-ap-south-1
Policy
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": {
"AWS": "arn:aws:iam::395938233352:role/AWSGlueServiceRole-Orders"
},
"Action": [
"s3:PutObject",
"s3:GetObject",
"s3:ListBucket"
],
"Resource": [
"arn:aws:s3:::my-orders-analytics",
"arn:aws:s3:::my-orders-analytics/*"
]
}
]
}


job
Last modified on 9/9/2025, 2:05:06 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Data source - Data Catalog
AWS Glue Data Catal...
Data target - S3 bucket
my-orders-analytics
Transform - Filter
Filter
Data preview
Output schema
Target node not supported
You have selected a data target node which is not supported for data preview. Please select another type of node instead.
Data target properties - S3
Name
my-orders-analytics
Node parents
Choose one or more parent node
AWS Glue Data Catalog
Catalog - DataSource
Format
Parquet
Compression Type
Snappy
S3 Target Location
s3://aws-glue-assets-395938233352-ap-south-1/sh
View
Browse S3
Data Catalog update options
Do not update the Data Catalog
Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions
File partitioning
Autogenerate files (Recommended)
Multiple file output
Partition keys - optional


Amazon S3 > Buckets > aws-glue-assets-395938233352-ap-south-1
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Vector buckets
Access Grants
Access Points (General Purpose Buckets, FSx file systems)
Access Points (Directory Buckets)
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
AWS Marketplace for S3
aws-glue-assets-395938233352-ap-south-1 Info
Objects
Properties
Permissions
Metrics
Management
Access Points
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
scripts/
Folder
shipped/
Folder
sparkHistoryLogs/
Folder
Amazon Q
Free Tier
Troubleshooting AWS Glue Job Issues
"s3:PutObject",
"s3:DeleteObject",
"s3:ListBucket"
],
"Resource": [
"arn:aws:s3:::your-target-bucket",
"arn:aws:s3:::your-target-bucket/*"
]
}
]
}
2. Verify the S3 Path:
Check line 32 in your Glue script
Ensure the S3 path is correct and the bucket exists
Verify you're writing to the correct region (ap-south-1)
3. Test with a Different S3 Location:
Try writing to a test bucket you control to isolate the issue
Could you share:
1. The target S3 bucket/path you're trying to write to
2. The IAM role name attached to your Glue job
3. Whether this is a cross-account scenario
This will help me provide more specific guidance for your situation.
Ask me anything about AWS
Max 1000 characters

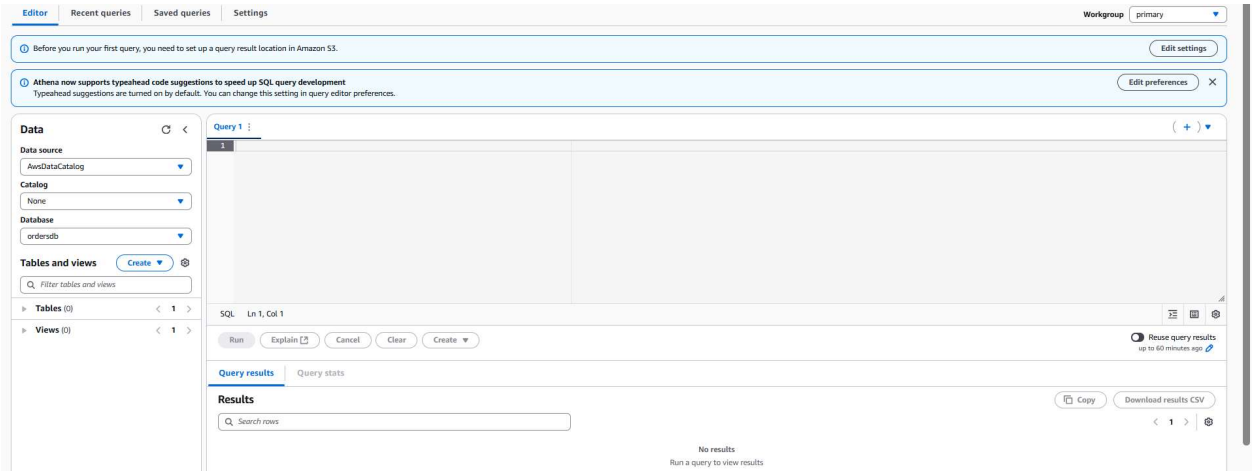

aws
athena
Amazon S3 > Buckets > aws-glue-assets-395938233352-ap-south-1
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Vector buckets
Access Grants
Access Points (General Purpose Buckets, FSx file systems)
Access Points (Directory Buckets)
Object Lambda Access Points
Multi-Region 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
AWS Marketplace for S3
Access findings are provided by IAM external access analyzers. Learn more about How IAM analyzer findings work
View analyzer for ap-south-1
Block public access (bucket settings)
Edit
Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more
Block all public access
Off
Individual Block Public Access settings for this bucket
Bucket policy
Edit
Delete
The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts. Learn more
Copy
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": {
"AWS": "arn:aws:iam::395938233352:role/AWSGlueServiceRole-Orders"
},
"Action": [
"s3:PutObject",
"s3:GetObject",
"s3:ListBucket"
],
"Resource": [
"arn:aws:s3:::aws-glue-assets-395938233352-ap-south-1",
"arn:aws:s3:::aws-glue-assets-395938233352-ap-south-1/*"
]
}
]
}

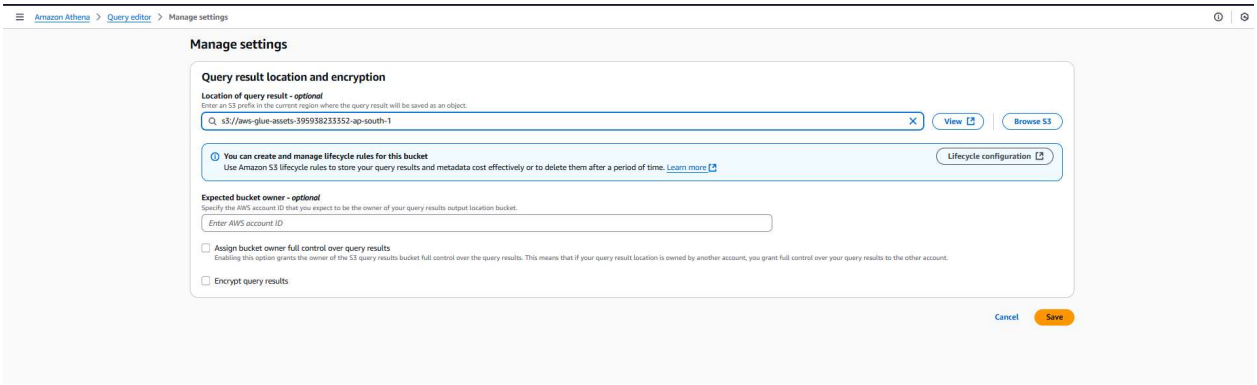


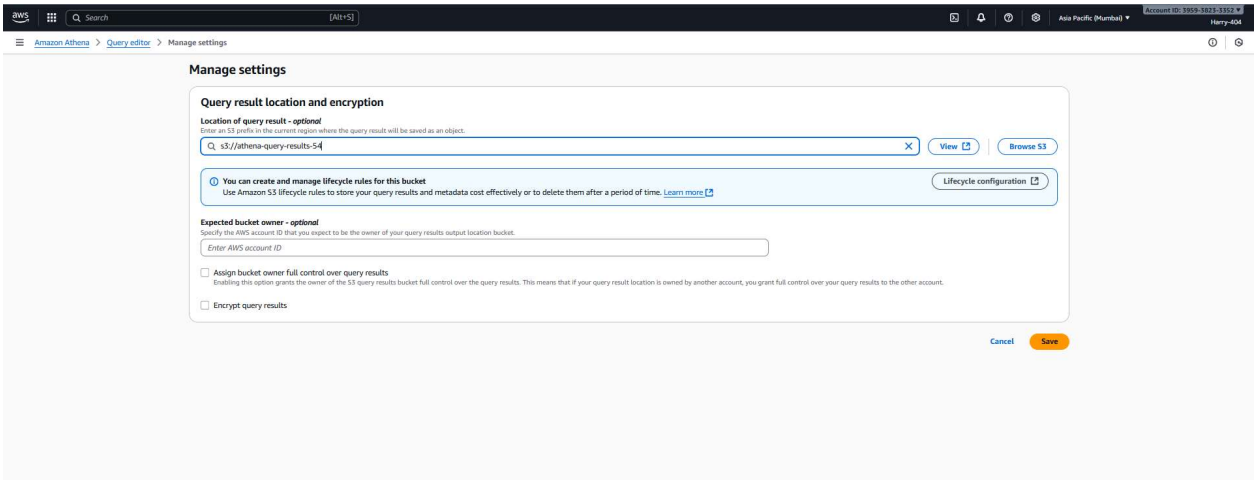


Step 6: Athena Setup and Queries

1. Set S3 query output location:
   Example: `s3://athena-query-results-54/`
2. In Athena Console:
   - Data source: `AwsDataCatalog`
   - Database: `ordersdb`
   - Table: `shipped` (or whatever table is created by the S3 crawler)







Example SQL Queries

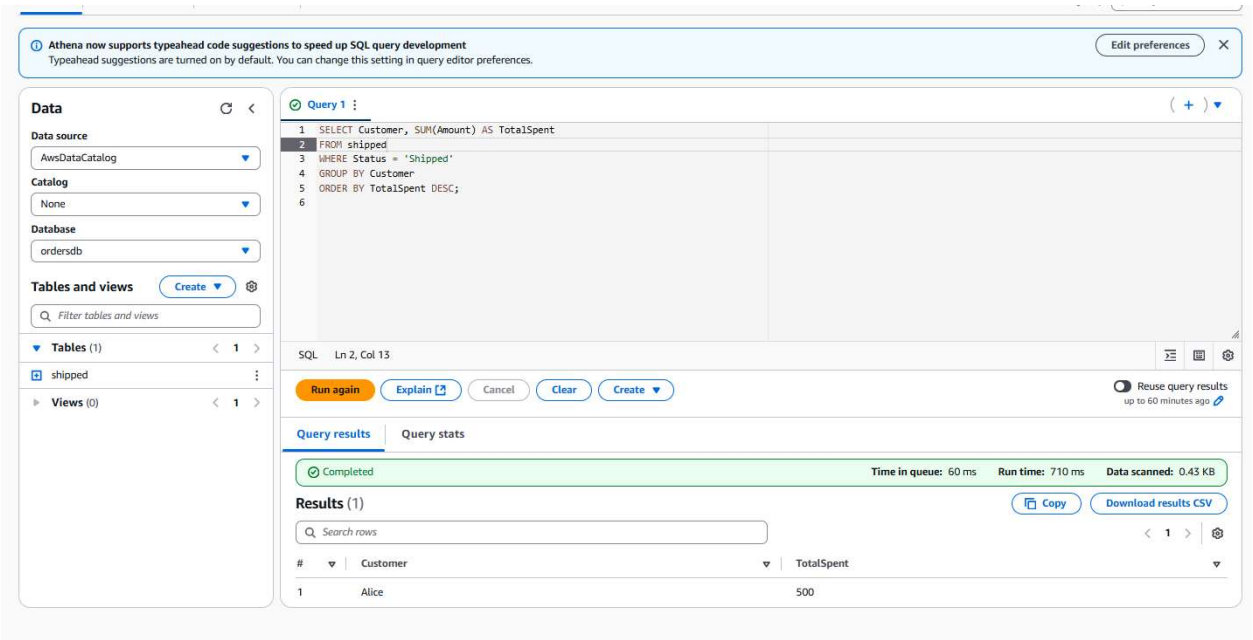
a) Top Customers by Spend

```sql
SELECT Customer, SUM(Amount) AS TotalSpent
FROM shipped
```

```sql
WHERE Status = 'Shipped'
GROUP BY Customer
ORDER BY TotalSpent DESC;
```

b) Monthly Revenue

```sql
SELECT OrderDate, SUM(Amount) AS MonthlyRevenue
FROM shipped
WHERE Status = 'Shipped'
GROUP BY OrderDate;
```



Step 7: QuickSight Visualization

- Connect QuickSight to Athena (using the same database and tables)
- Example dashboards:
    - Bar: Customer vs. TotalSpent
    - Line: Monthly revenue trend

Step 8: Cleanup

1. Delete DynamoDB Table

- Go to DynamoDB Console.
- Select the Orders table.
- Click Delete table.
- Confirm deletion.

2. Delete AWS Glue Components

a) Delete Glue Jobs

- Go to AWS Glue Console → Jobs.
- Select all the jobs you created (related to this lab).
- Click Delete.

b) Delete Glue Crawlers

- Go to AWS Glue Console → Crawlers.
- Select all crawlers related to this lab (e.g., OrdersCrawler).
- Click Delete.

c) Delete Glue Databases and Tables (Optional)

- Go to AWS Glue Console → Databases.
- Select the databases (e.g., ordersdb).
- Delete related tables.
- Delete the database if desired.

3. Empty and Delete S3 Buckets

a) `my-orders-analytics` Bucket

- Navigate to S3 Console.
- Open the bucket.
- Select all folders/files (e.g., `shipped/`).
- Click Delete to empty bucket.
- Then delete the bucket.

b) Athena Query Result Bucket (e.g., `athena-query-results`)

- Similarly, empty all query result files.
- Delete the bucket.

c) `aws-glue-assets-*` Bucket

- Do not delete this bucket as it's managed by AWS for Glue internal use.

4. Delete AWS Athena Resources

Drop Athena Tables

- In Athena Console, connect to your database (ordersdb).
- Run DROP TABLE commands to remove created tables, e.g.,
- `sql`
- `DROP TABLE IF EXISTS orders;`