

Group Member's Name:-

Muhammad Arsalan (2303.KHI.DEG.025)

Abdul Rehman (2303.KHI.DEG.035)

Arshad Shiwani (2303.KHI.DEG.026)

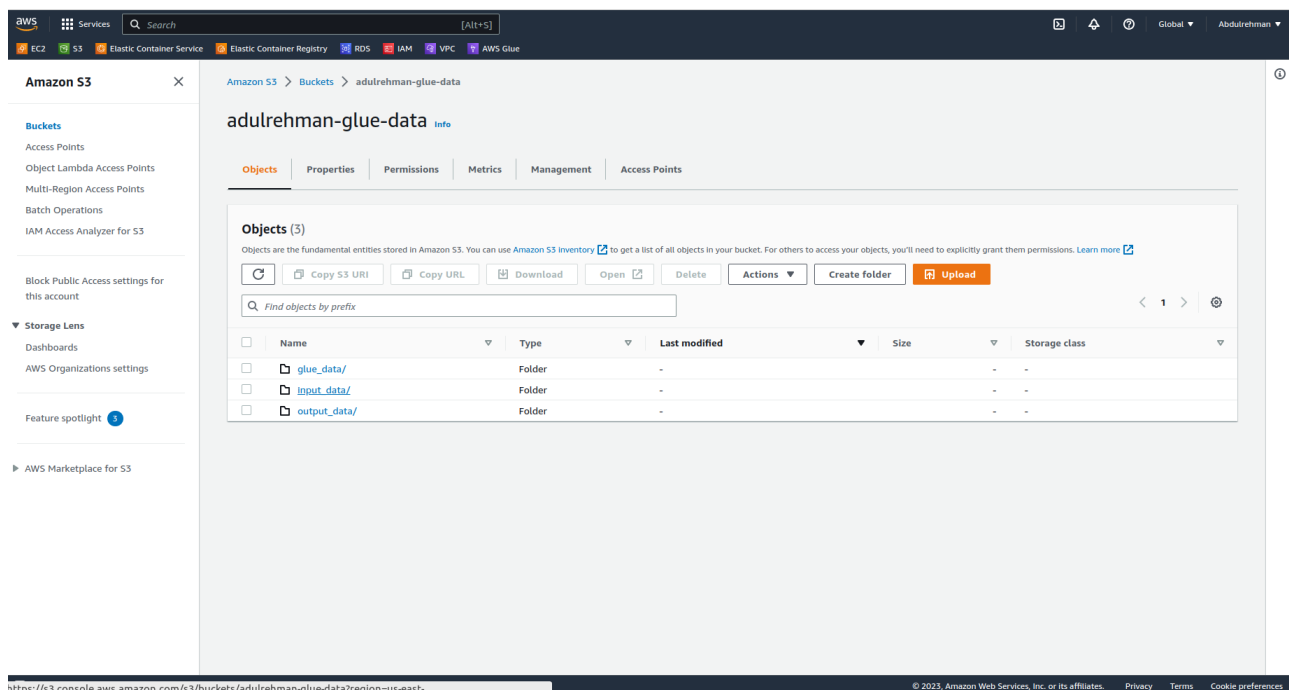
we created directories on s3 Buckets.

Q1. Why do we need to take this step?

Ans: Specifying the input and output data locations is essential for the Glue job. Typically, both the data source and the generated output data are stored in an S3 bucket, which provides a reliable and scalable storage solution.

Q2. What is this service's purpose?

Ans: S3 serves as a feature-rich storage service offering durability, availability, security, and scalability. With its redundant storage across multiple devices and facilities, data stored in S3 is highly resilient and protected against data loss. Its comprehensive set of features ensures reliable and secure storage for a variety of applications and data types.



This is the output Data.

The screenshot shows the Amazon S3 console interface. The left sidebar contains navigation options like Buckets, Access Points, and Storage Lens. The main content area displays the 'location/' bucket with 5 objects. The objects are listed in a table with columns for Name, Type, Last modified, Size, and Storage class.

Name	Type	Last modified	Size	Storage class
run-1684323485565-part-block-0-r-00014-snappy.parquet	parquet	May 17, 2023, 16:38:14 (UTC+05:00)	599.0 B	Standard
run-1684323485565-part-block-0-r-00002-snappy.parquet	parquet	May 17, 2023, 16:38:13 (UTC+05:00)	599.0 B	Standard
run-1684323485565-part-block-0-r-00021-snappy.parquet	parquet	May 17, 2023, 16:38:13 (UTC+05:00)	599.0 B	Standard
run-1684323485565-part-block-0-r-00025-snappy.parquet	parquet	May 17, 2023, 16:38:13 (UTC+05:00)	599.0 B	Standard
run-1684323485565-part-block-0-r-00031-snappy.parquet	parquet	May 17, 2023, 16:38:13 (UTC+05:00)	599.0 B	Standard

The screenshot shows the Amazon S3 console interface. The left sidebar contains navigation options like Buckets, Access Points, and Storage Lens. The main content area displays the 'input_data/' bucket with 2 objects. The objects are listed in a table with columns for Name, Type, Last modified, Size, and Storage class.

Name	Type	Last modified	Size	Storage class
earnings/	Folder	-	-	-
location/	Folder	-	-	-

Here we save csv file.

aws

Services

Search

[Alt+S]

EC2

S3

Elastic Container Service

Elastic Container Registry

RDS

IAM

VPC

AWS Glue

Global

Abdulrehman

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > adulrehman-glue-data > Input_data/ > location/

location/

Copy S3 URI

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 permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
	locations.csv	csv	May 17, 2023, 15:52:37 (UTC+05:00)	916.0 B	Standard

CloudShell

Feedback

Language

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

aws

Services

Search

[Alt+S]

EC2

S3

Elastic Container Service

Elastic Container Registry

RDS

IAM

VPC

AWS Glue

Global

Abdulrehman

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > adulrehman-glue-data > Input_data/ > earnings/

earnings/

Copy S3 URI

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 permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
	date=2023-05-16/	Folder	-	-	-

CloudShell

Feedback

Language

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Here we created IAM Role

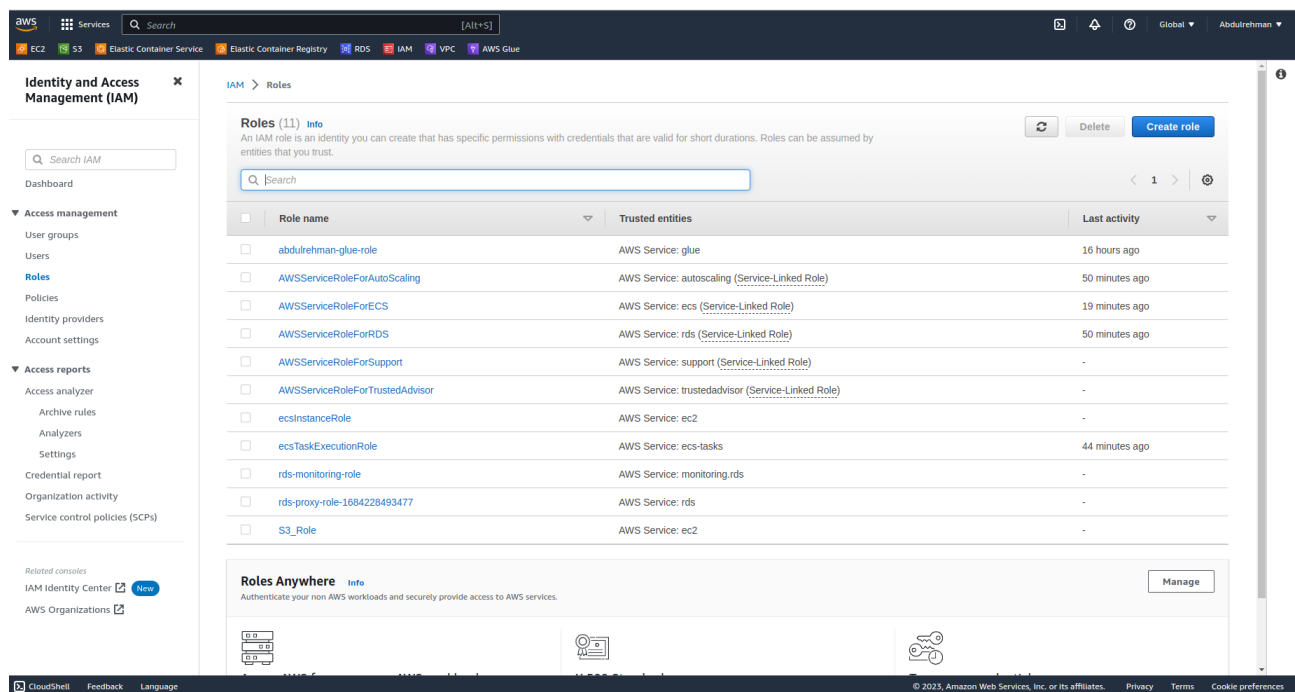
Q1. Why do we need to take this step?

Ans: Security, Least Privilege, Separation of Responsibilities, Ease of Management:

we creating IAM roles for AWS Glue ensures secure and controlled access to data and resources while maintaining efficient management practices.

Q2. What is this service's purpose?

Ans: Certainly! In short, the purpose of this process is to create an IAM role specifically for AWS Glue, which is an ETL service provided by AWS. This role grants necessary permissions to Glue for accessing data and performing data transformation tasks. By creating this role and associating it with Glue, you enable the service to interact with your data stored in Amazon S3 and perform automated data preparation and transformation tasks for analysis and other purposes.



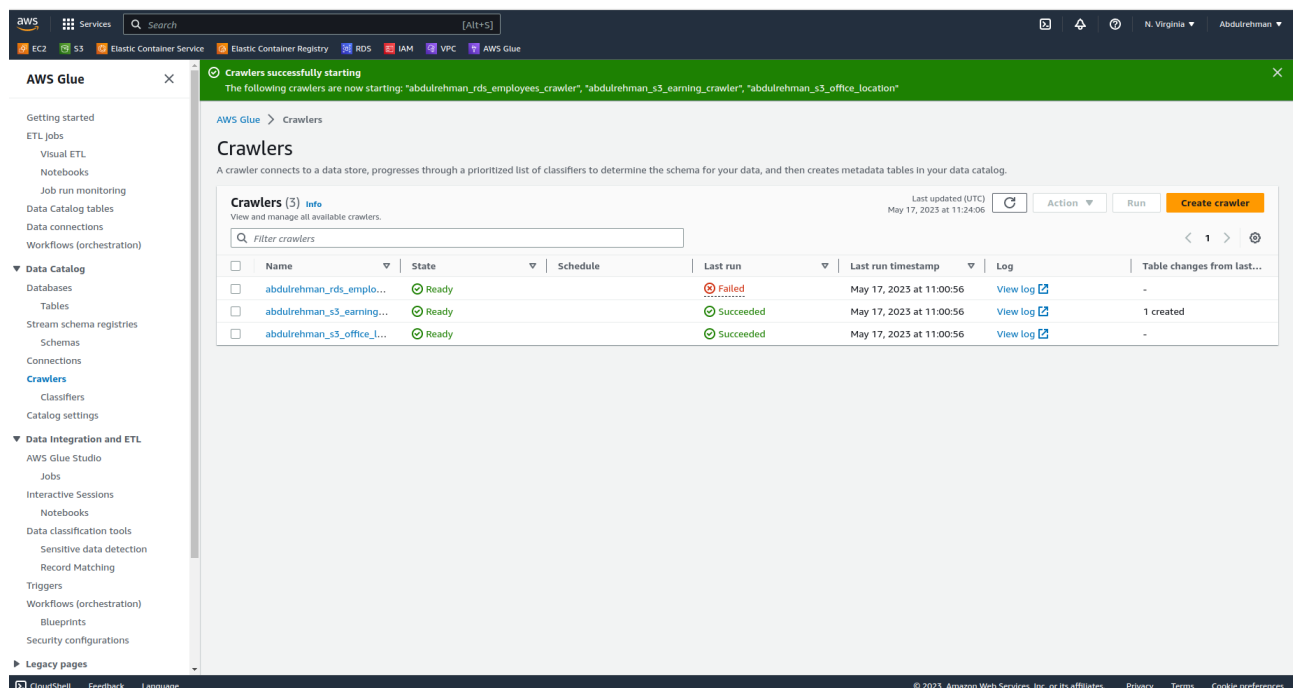
we created Glue Crawlers:

Q1. Why do we need to take this step?

Ans: Taking the step of creating Glue Crawlers in AWS Glue is crucial for efficiently preparing our data sources for ETL (Extract, Transform, Load) processing. By leveraging Glue Crawlers, we can simplify and automate the process of discovering and cataloging metadata about our data sources. This automated discovery saves us valuable time, minimizes the chances of errors, and enhances the overall quality of our data.

Q2. What is this service's purpose?

Ans: The purpose of AWS Glue, particularly its Glue Crawlers, extends beyond simple data processing. Glue Crawlers play a vital role in inferring the schema of our data sources. This inference capability significantly streamlines and automates the data preparation process required for ETL processing. By automatically discovering and cataloging metadata about our data sources, Glue Crawlers not only accelerate the ETL pipeline but also contribute to error reduction and improved data quality. This service empowers organizations to effectively handle and derive insights from their data assets in a more efficient and reliable manner.



The screenshot displays the AWS Glue console interface. At the top, a green banner indicates "Crawlers successfully starting" with the message: "The following crawlers are now starting: 'abdulrehman_rds_employees_crawler', 'abdulrehman_s3_earning_crawler', 'abdulrehman_s3_office_location'". Below this, the "Crawlers" section is visible, showing a list of three crawlers. The first crawler, "abdulrehman_rds_emplo...", is in a "Failed" state. The other two, "abdulrehman_s3_earning..." and "abdulrehman_s3_office_I...", are in a "Ready" state. The console also features a left-hand navigation menu with options like "Getting started", "Data Catalog", and "Data Integration and ETL".

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last...
abdulrehman_rds_emplo...	Failed		Failed	May 17, 2023 at 11:00:56	View log	-
abdulrehman_s3_earning...	Ready		Succeeded	May 17, 2023 at 11:00:56	View log	1 created
abdulrehman_s3_office_I...	Ready		Succeeded	May 17, 2023 at 11:00:56	View log	-

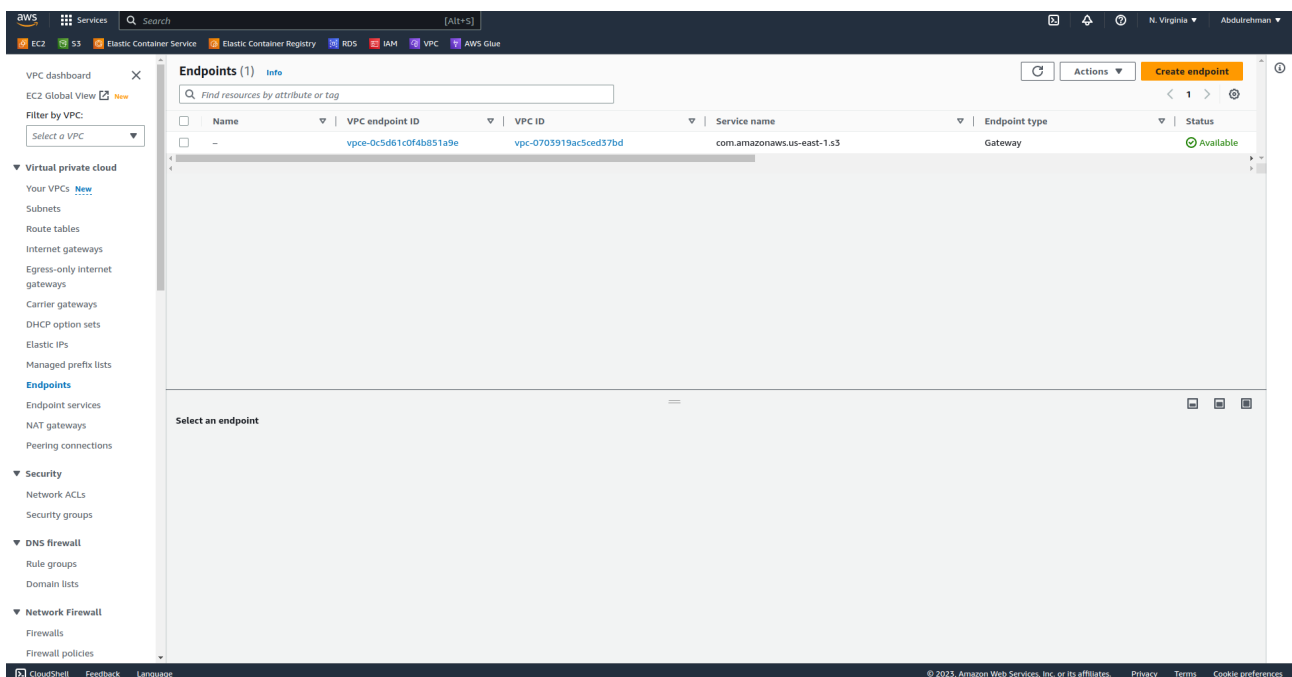
This is the Endpoint

Q1. Why do we need to take this step?

Ans: Creating a VPC endpoint is necessary to securely enable the Glue job to access S3. Since the Glue job interacts with data stored in S3, establishing a VPC endpoint ensures that the data transfers occur within the VPC, enhancing security and protecting sensitive information.

Q2. What is this service's purpose?

Ans: The purpose of creating a VPC endpoint is to control and manage access to AWS services within a VPC. It allows for granular control over which services can be accessed, specifies allowed VPCs and subnets, supports encryption with AWS PrivateLink, and enables the use of VPC security groups and IAM policies to enforce access controls. This enhances security, network isolation, and compliance adherence.



Q1. Why do we need to take this step?

Ans: Defining inbound and outbound traffic rules in security groups is essential for controlling network traffic to and from our AWS resources. This proactive measure helps mitigate security risks, unauthorized access attempts, potential data breaches, and other security threats by allowing us to precisely manage and restrict network communication.

Q2. What is this service's purpose?

Ans: The purpose of creating security group rules is to establish granular control over inbound and outbound network traffic for our AWS resources. By defining these rules, we can safeguard our resources from unauthorized access, potential data breaches, and various security threats. This capability empowers organizations to enforce strict network security policies and protect their assets effectively.

The screenshot displays the AWS Management Console's 'Security Groups' page. The left-hand navigation pane shows the 'Virtual private cloud' section expanded, with 'Security groups' selected. The main content area, titled 'Security Groups (7) info', contains a table listing seven security groups. Each row includes a checkbox, the group name, ID, VPC ID, description, owner, and counts for inbound and outbound rules. The groups include 'flask_-5916', 'launch-wizard-1', 'redis-3322', 'default', 'default VPC security gr...', 'EC2ContainerService-a...', and 'launch-wizard-2'.

	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules count
<input type="checkbox"/>	-	sg-0f13991ba55c1d8c0	flask_-5916	vpc-09357c0948d2fe501	2023-05-17T17:00:39....	806413649470	1 Permission entry	1 Permission entry
<input type="checkbox"/>	-	sg-02342fc17b5f51cdb	launch-wizard-1	vpc-0703919ac5ced37bd	launch-wizard-1 create...	806413649470	2 Permission entries	1 Permission entry
<input type="checkbox"/>	-	sg-0c5d3ea41dc8edddd	redis-3322	vpc-09357c0948d2fe501	2023-05-17T17:01:45....	806413649470	1 Permission entry	1 Permission entry
<input type="checkbox"/>	-	sg-040275c5b262d2cb4	default	vpc-0703919ac5ced37bd	default VPC security gr...	806413649470	2 Permission entries	1 Permission entry
<input type="checkbox"/>	-	sg-0307e72cdc0f7d72	default	vpc-09357c0948d2fe501	default VPC security gr...	806413649470	1 Permission entry	1 Permission entry
<input type="checkbox"/>	-	sg-05e7acd235ea5a2c	EC2ContainerService-a...	vpc-09357c0948d2fe501	ECS Allowed Ports	806413649470	1 Permission entry	1 Permission entry
<input type="checkbox"/>	-	sg-0c0056481be7747a8	launch-wizard-2	vpc-0703919ac5ced37bd	launch-wizard-2 create...	806413649470	2 Permission entries	1 Permission entry

The screenshot shows the 'Details' page for the security group 'sg-040275c5b262d2cb4 - default'. The breadcrumb trail indicates the path: 'VPC > Security Groups > sg-040275c5b262d2cb4 - default'. The 'Details' section provides key information: the name is 'default', the ID is 'sg-040275c5b262d2cb4', the description is 'default VPC security group', and the VPC ID is 'vpc-0703919ac5ced37bd'. It also lists the owner as '806413649470', 2 inbound rules, and 1 outbound rule. Below this, the 'Inbound rules' tab is active, showing two rules. A 'Run Reachability Analyzer' banner is visible above the rules table.

Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description
-	sgr-0bd7522b016581...	IPv4	PostgreSQL	TCP	5432	0.0.0.0/0	Rule for external access
-	sgr-0c48c77b508e293a2	-	All TCP	TCP	0 - 65535	sg-040275c5b262d2c...	Rule for Glue Crawler

aws Services Search [Alt+S]

EC2 S3 Elastic Container Service Elastic Container Registry RDS IAM VPC AWS Glue

VPC dashboard X

EC2 Global View **New**

Filter by VPC: Select a VPC

Virtual private cloud

Your VPCs **New**

Subnets

Route tables

Internet gateways

Egress-only internet gateways

Carrier gateways

DHCP option sets

Elastic IPs

Managed prefix lists

Endpoints

Endpoint services

NAT gateways

Peering connections

Security

Network ACLs

Security groups

DNS firewall

Rule groups

Domain lists

Network Firewall

Firewalls

Firewall policies

CloudShell Feedback Language

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

VPC > Security Groups > sg-040275c5b262d2cb4 - default

sg-040275c5b262d2cb4 - default Actions

Details

Security group name default Security group ID sg-040275c5b262d2cb4 Description default VPC security group VPC ID vpc-0703919ac5ced37bd

Owner 806413649470 Inbound rules count 2 Permission entries Outbound rules count 1 Permission entry

Inbound rules Outbound rules Tags

You can now check network connectivity with Reachability Analyzer Run Reachability Analyzer

Outbound rules (1/1) Manage tags Edit outbound rules

Filter security group rules

<input checked="" type="checkbox"/>	Name	Security group rule...	IP version	Type	Protocol	Port range	Destination	Description
<input checked="" type="checkbox"/>	-	sgr-085703539c0929...	IPv4	All traffic	All	All	0.0.0.0/0	-

Here we created Buckets.

aws Services Search [Alt+S]

EC2 S3 Elastic Container Service Elastic Container Registry RDS IAM VPC AWS Glue

Amazon S3 X

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets

Account snapshot View Storage Lens dashboard

Buckets (2) Info Copy ARN Empty Delete Create bucket

Find buckets by name

	Name	AWS Region	Access	Creation date
<input type="radio"/>	adulrehman-glue-data	US East (N. Virginia) us-east-1	Bucket and objects not public	May 16, 2023, 12:27:52 (UTC+05:00)
<input type="radio"/>	a-aws-handson	US East (N. Virginia) us-east-1	Bucket and objects not public	May 15, 2023, 15:49:01 (UTC+05:00)

Q1. Why do we need to take this step?

Ans: Creating a database in AWS Glue is essential for effective data preparation in ETL processing. It organizes data, improves query performance, aids data discovery, and enables access control.

Q2. What is this service's purpose?

Ans: The purpose of creating a database in AWS Glue is to organize and manage data sources for ETL processing, facilitating efficient data cataloging, query optimization, data discovery, and access control.

The screenshot shows the AWS Glue console interface. The left sidebar contains navigation options like 'Getting started', 'Data Catalog', and 'Data Integration and ETL'. The main content area is titled 'abdulrehman-glue-database' and shows the 'Database properties' section with fields for Name, Description, Location, and Created on (UTC). Below this, the 'Tables (3)' section lists three tables: 'average_earning', 'earnings', and 'location'. Each table entry includes its Name, Database, Location, Classification, and a 'View data' link.

Name	Database	Location	Classification	Deprecated	View data
average_earning	abdulrehman-glue-database	s3://adulrehman-glue-data/output_d	parquet	-	Table data
earnings	abdulrehman-glue-database	s3://adulrehman-glue-data/input_dat	csv	-	Table data
location	abdulrehman-glue-database	s3://adulrehman-glue-data/input_dat	csv	-	Table data

AWS Glue Untitled-job Last modified on 5/17/2023, 4:36:36 PM Try new UI End session Actions Save Run

Successfully started job
Successfully started job Untitled-job. Navigate to **Run details** for more details.

Visual Script Job details Runs Data quality **New** Schedules Version Control

Source Action Target Undo Redo Remove

Data source - S3 bucket Amazon S3 Data source - S3 bucket Amazon S3

Transform - Join Join

Transform - SQL Query SQL Query

Data target - S3 bucket Amazon S3

Data target properties - S3 Output schema Data preview

Format Parquet

Compression Type Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://adulrehman-glue-data/output_data/location/ View Browse S3

Data Catalog update options Info
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.
☐ 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

Database
Choose the database from the AWS Glue Data Catalog.
adulrehman-glue-database

Use runtime parameters

Table name
Enter a table name for the AWS Glue Data Catalog.
average_earning

Partition keys - optional

CloudShell Feedback Language © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

AWS Glue Untitled-job Last modified on 5/17/2023, 4:36:36 PM Try new UI End session Actions Save Run

Successfully started job
Successfully started job Untitled-job. Navigate to **Run details** for more details.

Visual Script Job details Runs Data quality **New** Schedules Version Control

Source Action Target Undo Redo Remove

Data source - S3 bucket Amazon S3 Data source - S3 bucket Amazon S3

Transform - Join Join

Transform - SQL Query SQL Query

Data target - S3 bucket Amazon S3

Data source properties - S3 Output schema Data preview

Name Amazon S3

S3 source type Info
☐ S3 location
Choose a file or folder in an S3 bucket.
☒ Data Catalog table

Database
Choose a database.
adulrehman-glue-database

Use runtime parameters

Table
location

Use runtime parameters

Partition predicate - optional
Enter a boolean expression supported by Spark SQL, using only partition columns.
Partition predicate syntax for Spark SQL is year == year(date_sub(current_date, 7)) AND month == month(date_sub(current_date, 7)) AND day == day(date_sub(current_date, 7)).

CloudShell Feedback Language © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

aws

Services

Search

[Alt+S]

EC2S3Elastic Container ServiceElastic Container RegistryRDSIAMVPCAWS Glue

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

AWS Glue Studio

Jobs

Interactive Sessions

Notebooks

Data classification tools

Sensitive data detection

Record Matching

Triggers

Workflows (orchestration)

Blueprints

Security configurations

Legacy pages

Untitled-job

Last modified on 5/17/2023, 4:36:36 PM

Try new UI

End session

Actions

Save

Run

Successfully started job

Successfully started job Untitled-job. Navigate to [Run details](#) for more details.

VisualScriptJob detailsRunsData qualityNewSchedulesVersion Control

Source

Action

Target

Undo

Redo

Remove

Transform

Output schema

Data preview

Data source - S3 bucket

Amazon S3

Data source - S3 bucket

Amazon S3

Transform - Join

Join

Transform - SQL Query

SQL Query

Data target - S3 bucket

Amazon S3

Name

Join

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent nodes

Amazon S3

Amazon S3

S3 - DataSource

S3 - DataSource

The parents of this node have overlapping field names. AWS Glue Studio can add an Apply Mapping node to rename them and avoid downstream issues.

Custom prefix

Add a prefix to the field names of the parent node on the right

right

Resolve it

Join type

Select the type of join to perform.

Inner join

Select all rows from both datasets that meet the join condition.

Join conditions

Select a field from each parent node for the join condition.

Amazon S3

emp_id

=

Amazon S3

emp_id

CloudShell

Feedback

Language

© 2023, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

aws

Services

Search

[Alt+S]

EC2S3Elastic Container ServiceElastic Container RegistryRDSIAMVPCAWS Glue

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

AWS Glue Studio

Jobs

Interactive Sessions

Notebooks

Data classification tools

Sensitive data detection

Record Matching

Triggers

Workflows (orchestration)

Blueprints

Security configurations

Legacy pages

Untitled-job

Last modified on 5/17/2023, 4:36:36 PM

Try new UI

End session

Actions

Save

Run

Successfully started job

Successfully started job Untitled-job. Navigate to [Run details](#) for more details.

VisualScriptJob detailsRunsData qualityNewSchedulesVersion Control

Source

Action

Target

Undo

Redo

Remove

Data source properties - S3

Output schema

Data preview

Data source - S3 bucket

Amazon S3

Data source - S3 bucket

Amazon S3

Transform - Join

Join

Transform - SQL Query

SQL Query

Data target - S3 bucket

Amazon S3

Name

Amazon S3

S3 source type

info

☐ S3 location

Choose a file or folder in an S3 bucket.

☒ Data Catalog table

Database

Choose a database.

abdulrehman-glue-database

Use runtime parameters

Table

earnings

Use runtime parameters

Partition predicate - optional

Enter a boolean expression supported by Spark SQL, using only partition columns.

Partition predicate syntax for Spark SQL is year == year(date_sub(current_date, 7)) AND month == month(date_sub(current_date, 7)) AND day == day(date_sub(current_date, 7)).

CloudShell

Feedback

Language

© 2023, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

AWS Glue console showing a successfully started job named "Untitled-job". The job is configured with two data sources (Amazon S3 buckets) feeding into a "Transform - Join" node, which then feeds into a "Transform - SQL Query" node. The SQL query is displayed on the right, showing a select statement with aliases and a group by clause.

Visual

Source

Action

Target

Undo

Redo

Remove

Transform

Output schema

Data preview

Name

SQL Query

Node parents

Choose one or more parent nodes

Join

Join - Transform

Associate an alias with each input source

Input sources

SQL aliases

myDataSource

SQL query

```
1
2 select
3   location,
4   avg(earnings) as average_earnings,
5   (avg(earnings) - min(earnings)) / min(earnings) * 100 AS raise_percentage
6 from
7   myDataSource
8 group by
9   location;
```

This is the Final Result

AWS Glue console showing the same job, but with a warning message indicating that the job has not been saved. The job is still in the "Visual" tab, and the SQL query is displayed on the right. The "Data preview" tab is also visible, showing a table with 5 rows and 3 columns: location, average_earnings, and raise_percentage.

Visual

Script

Job details

Runs

Data quality

Schedules

Version Control

Source

Action

Target

Undo

Redo

Remove

Transform

Output schema

Data preview

Data preview (5)

Filter sample dataset

location	average_earnings	raise_percentage
B	6286.75	155.14407467532467
C	5576.95	129.780387309433
A	5926.05	191.49286768322676
D	5889.7	185.07744435688285
E	5599.2	158.74306839186693