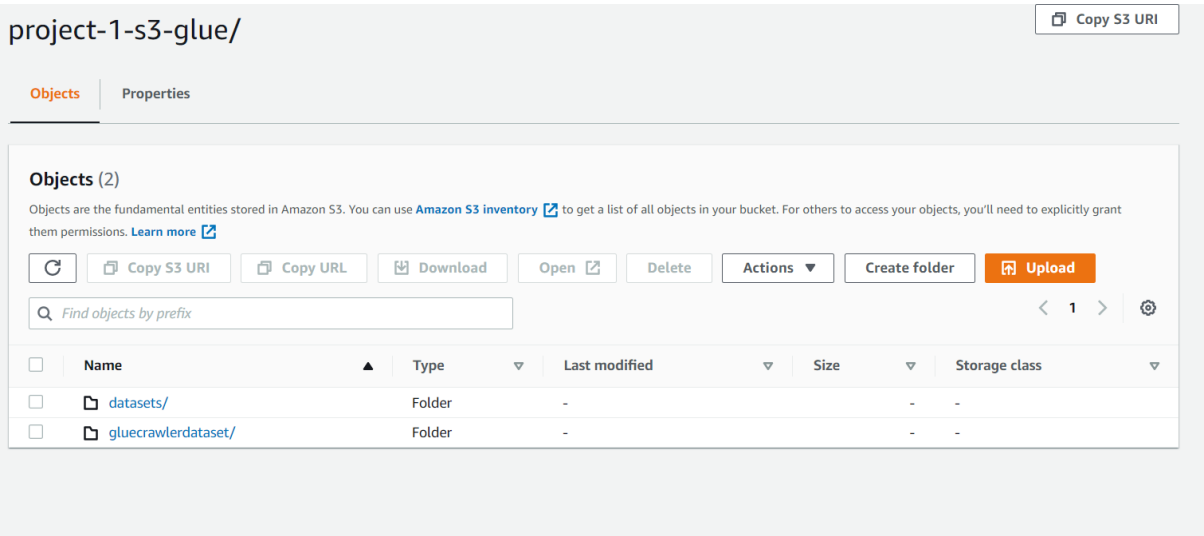
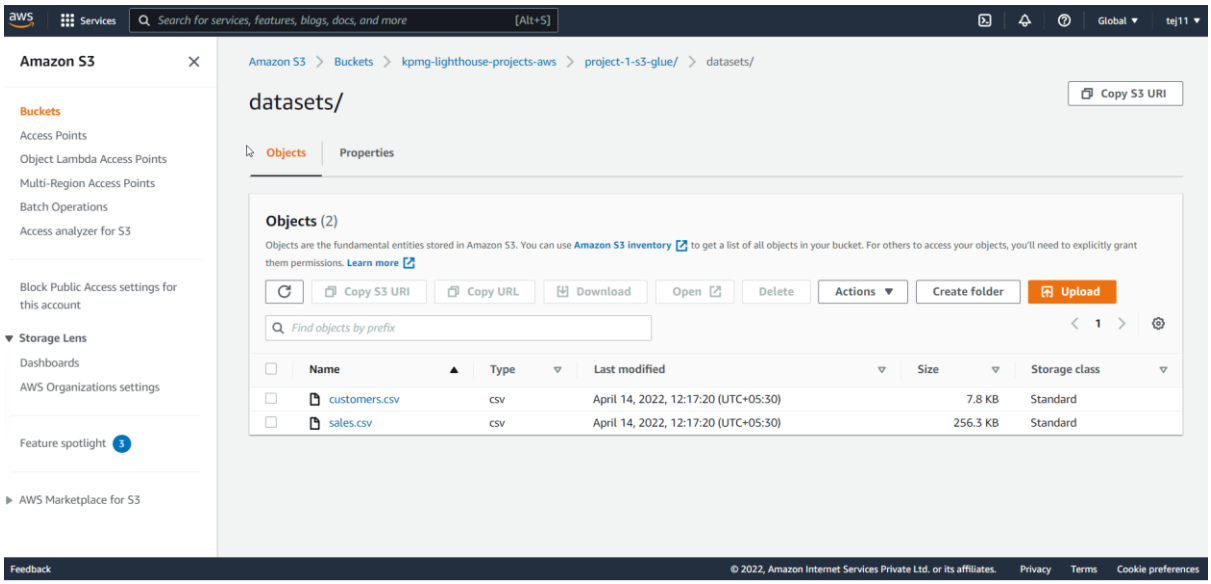
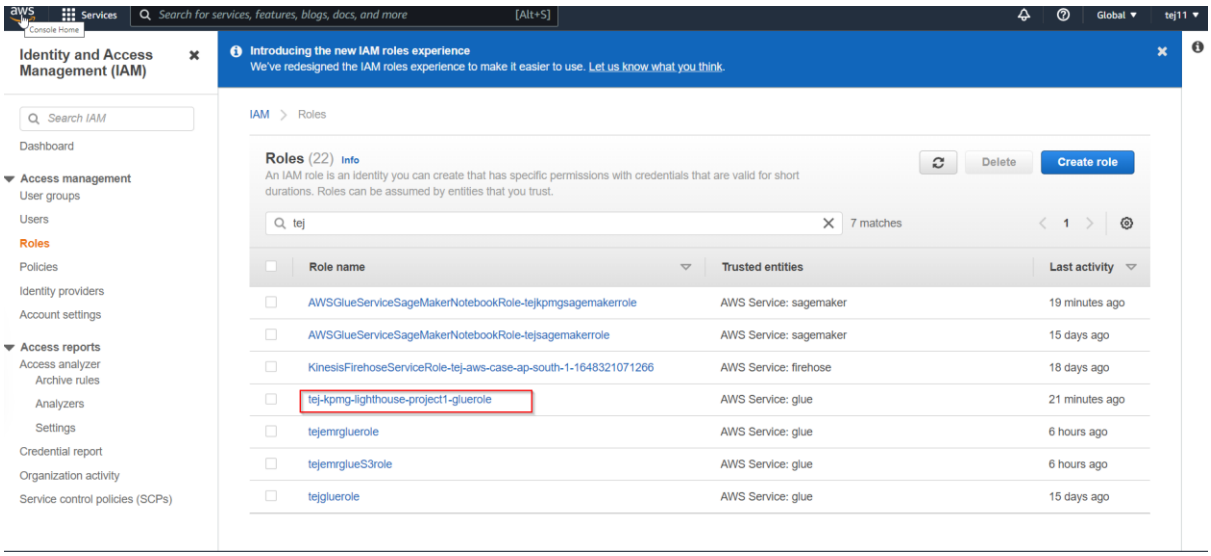
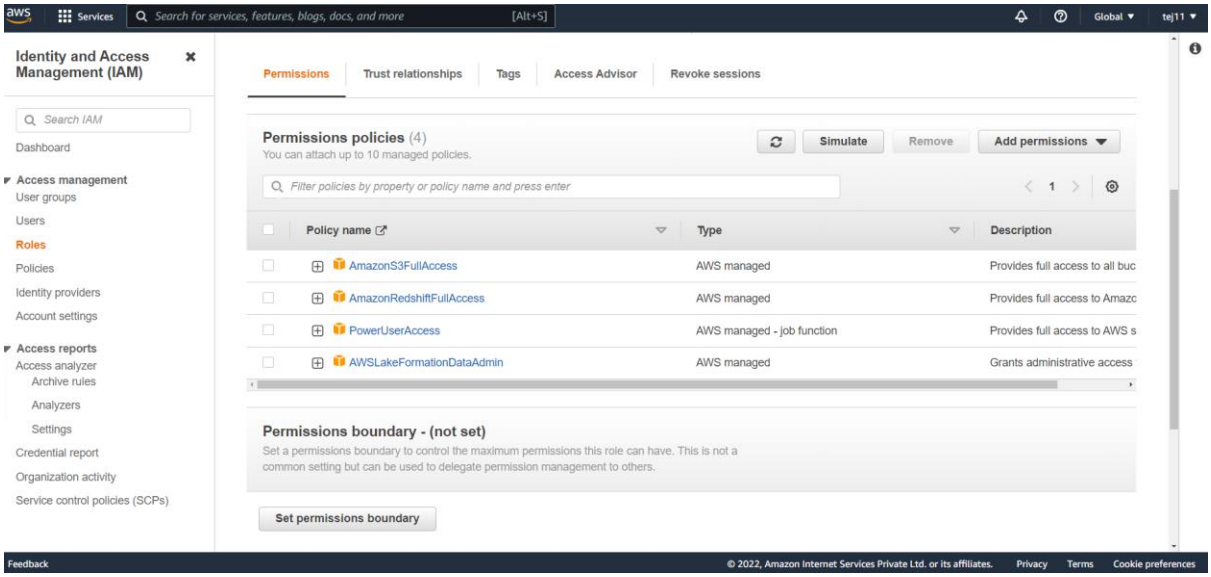
Naming conventions

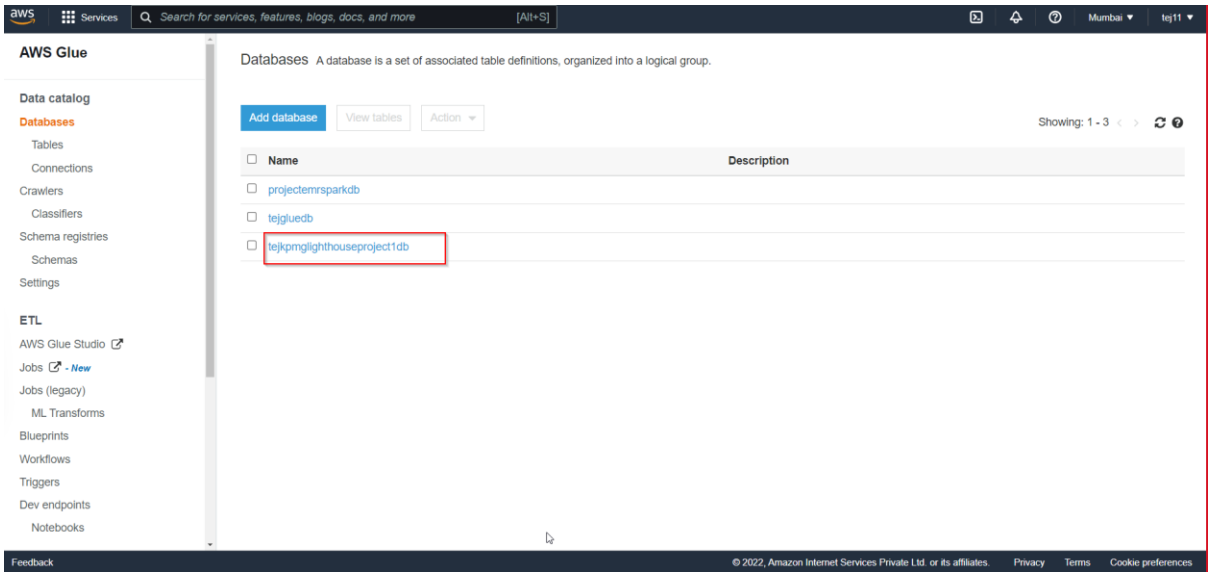


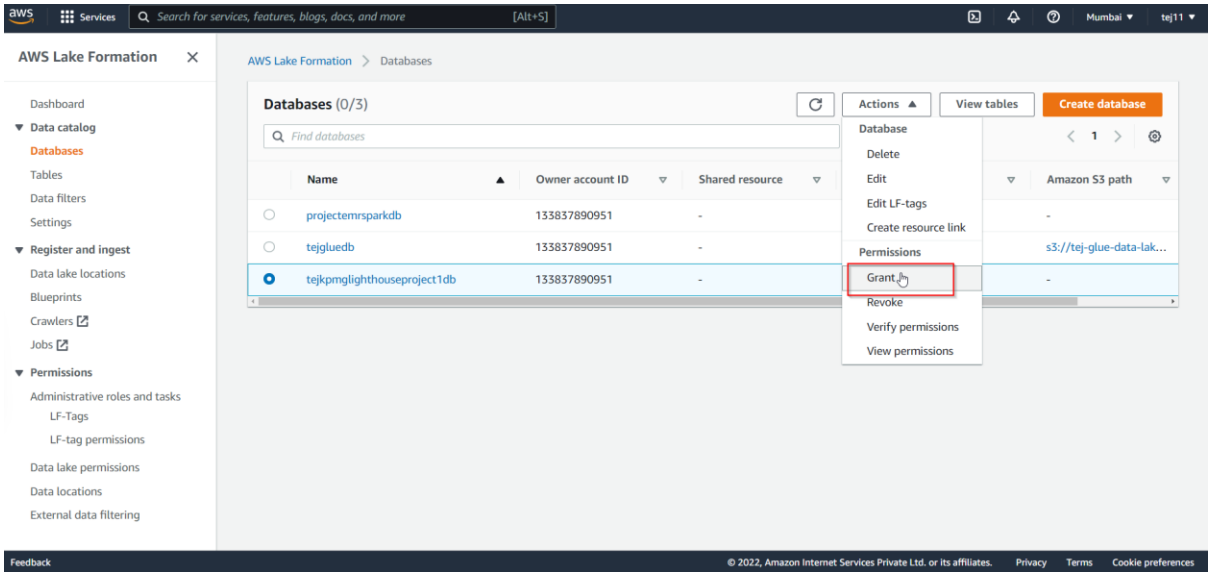


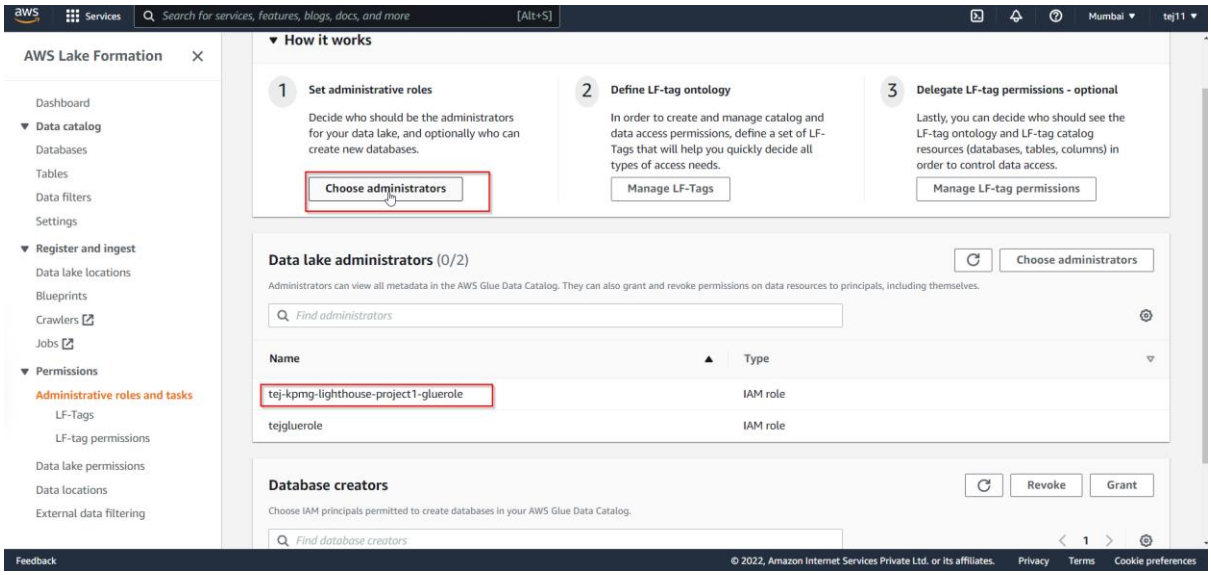


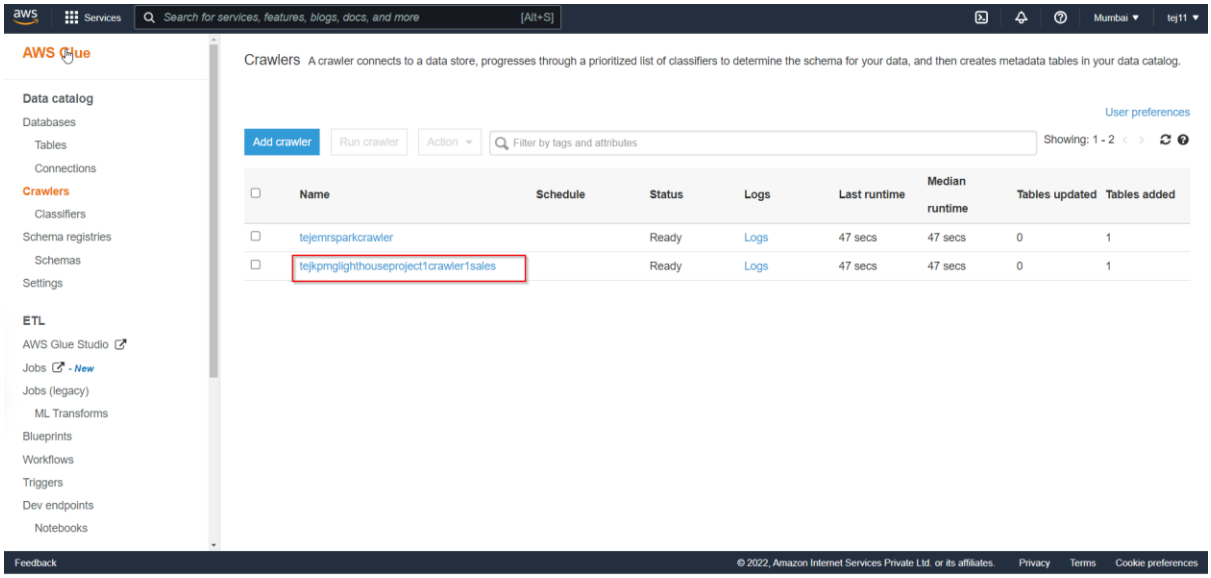


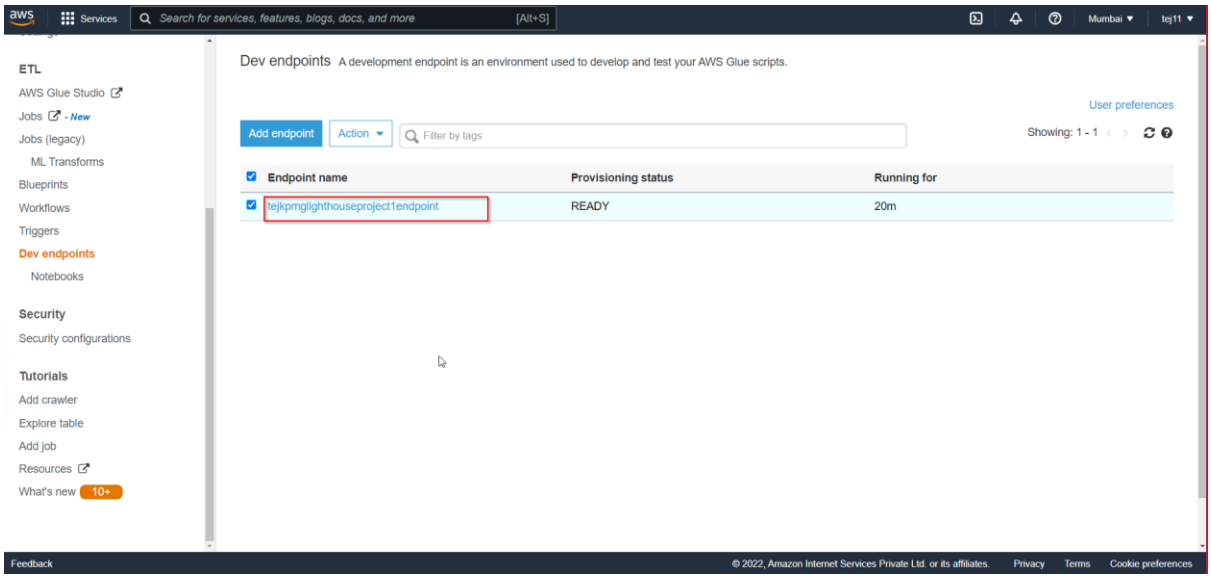
Glue database

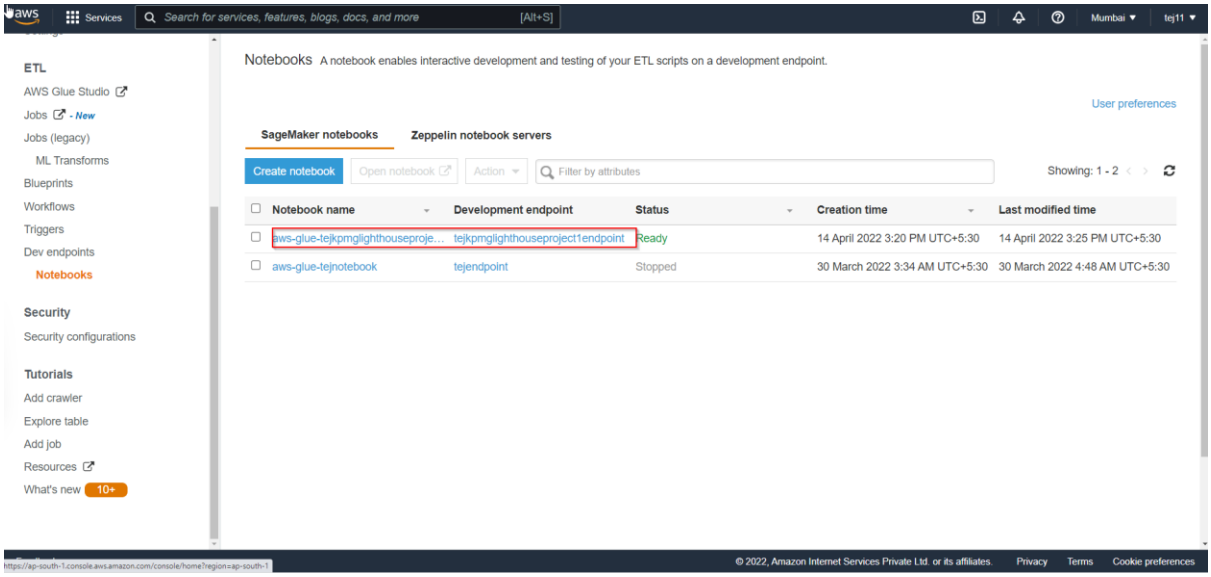



AWS Lake Formation
Dashboard
Data catalog
Databases
Tables
Data filters
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
LF-Tags
LF-tag permissions
Data lake permissions
Data locations
External data filtering
AWS Lake Formation > Databases
Databases (0/3)
Actions
View tables
Create database
Find databases
Name
Owner account ID
Shared resource
Amazon S3 path
projectemrsparkdb 133837890951 -
tejgluedb 133837890951 - s3://tej-glue-data-lak...
tejkpmglighthouseproject1db 133837890951 -
Database
Delete
Edit
Edit LF-tags
Create resource link
Permissions
Grant
Revoke
Verify permissions
View permissions
Feedback
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


AWS Lake Formation
How it works
1 Set administrative roles
Decide who should be the administrators for your data lake, and optionally who can create new databases.
Choose administrators
2 Define LF-tag ontology
In order to create and manage catalog and data access permissions, define a set of LF-Tags that will help you quickly decide all types of access needs.
Manage LF-Tags
3 Delegate LF-tag permissions - optional
Lastly, you can decide who should see the LF-tag ontology and LF-tag catalog resources (databases, tables, columns) in order to control data access.
Manage LF-tag permissions
Data lake administrators (0/2)
Choose administrators
Administrators can view all metadata in the AWS Glue Data Catalog. They can also grant and revoke permissions on data resources to principals, including themselves.
Find administrators
Name
Type
tej-kpmg-lighthouse-project1-gluerole IAM role
tejgluerole IAM role
Database creators
Revoke
Grant
Choose IAM principals permitted to create databases in your AWS Glue Data Catalog.
Find database creators
Feedback
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Mumbai
tej11
AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
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.
User preferences
Add crawler
Run crawler
Action
Filter by tags and attributes
Showing: 1 - 2
Name
Schedule
Status
Logs
Last runtime
Median runtime
Tables updated
Tables added
tejemrsparkcrawler
Ready
Logs
47 secs
47 secs
0
1
tejkpmglighthouseproject1crawler1sales
Ready
Logs
47 secs
47 secs
0
1
Feedback
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Mumbai
tej11
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Resources
What's new 10+
Dev endpoints A development endpoint is an environment used to develop and test your AWS Glue scripts.
User preferences
Add endpoint
Action
Filter by tags
Showing: 1 - 1
Endpoint name
Provisioning status
Running for
tejkpmglighthouseproject1endpoint
READY
20m
Feedback
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences



1)The Data for the Workshop

In the workshop, you learn about using PySpark to create Glue Job. PySpark is used to process and transform the data. Before you start the workshop, let's understand the data to be configured in the data lake for the workshop.

You will configure data lake with two data sets - **sales** and **customers**. The data sets are stored in Amazon S3. The AWS Glue and AWS Lake Formation services are used to create the data lake.

The following are the schema of the data sets:

Uber eats customers data set fields: {CUSTOMERID, CUSTOMERNAME, EMAIL, CITY, COUNTRY, TERRITORY, CONTACTFIRSTNAME, CONTACTLASTNAME}

Uber eats sales data set fields: {ORDERNUMBER, QUANTITYORDERED, PRICEEACH, ORDERLINENUMBER, SALES, ORDERDATE, STATUS, QTR_ID, MONTH_ID, YEAR_ID, PRODUCTLINE, MSRP, PRODUCTCODE, DEALSIZE, CUSTOMERID}

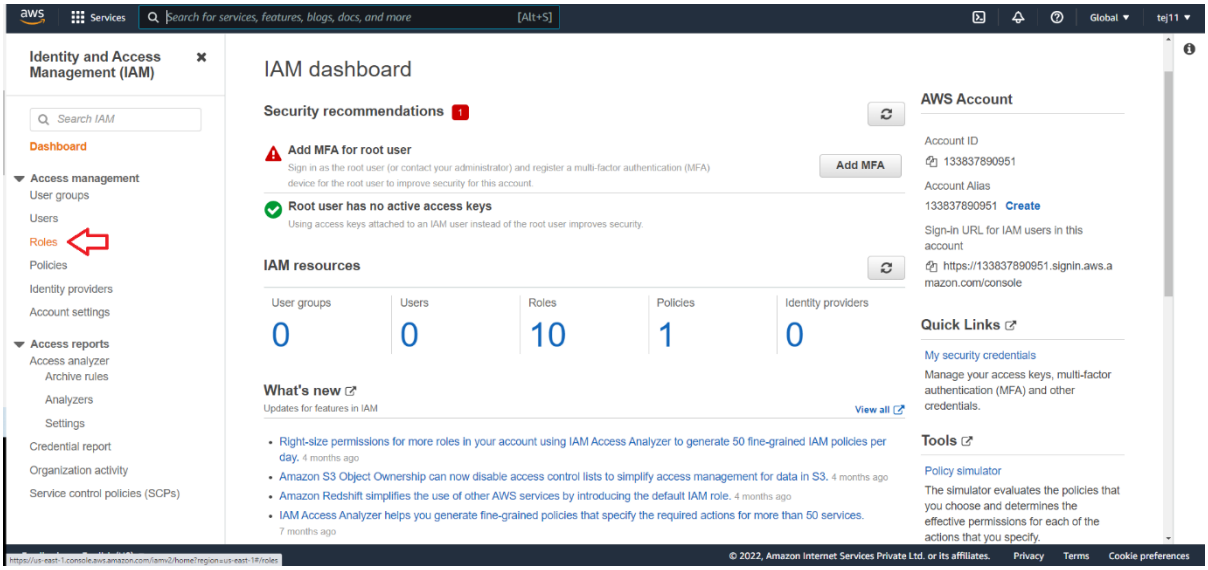
Let's start building now.

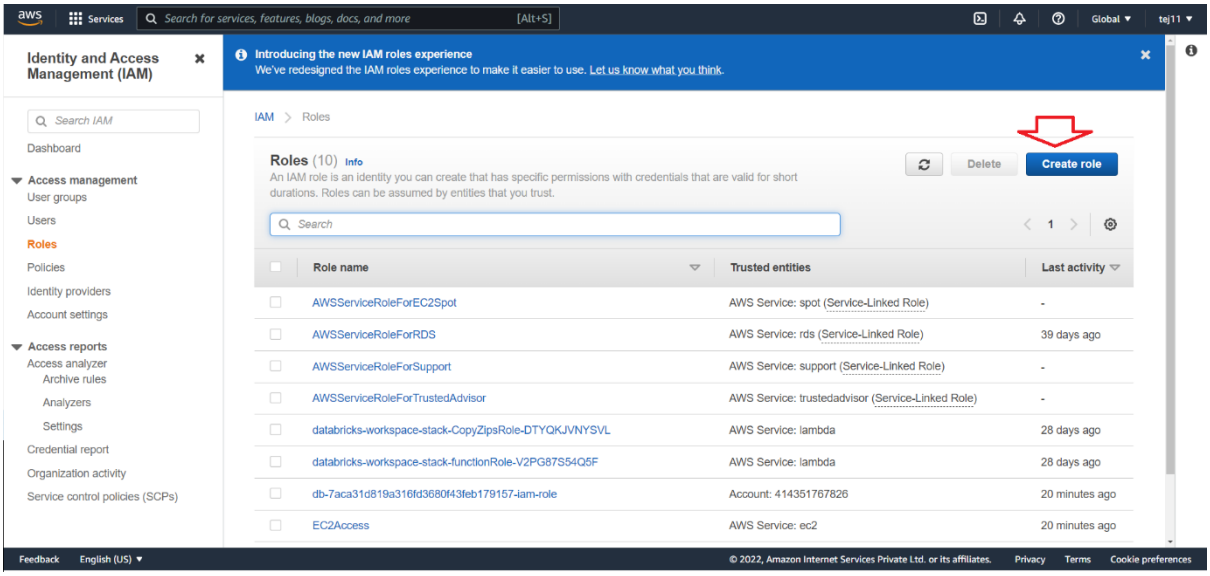
2)

Create IAM Role

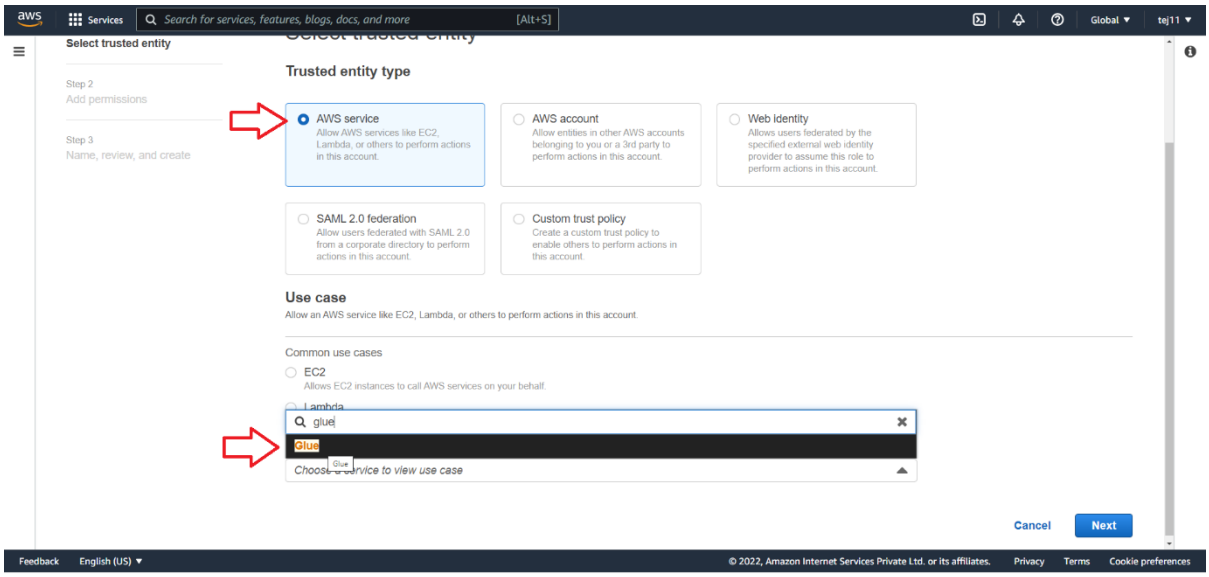
You now create IAM Role which is used by the AWS Glue crawler, job and developer endpoint to perform tasks like data catalog creation, data transformation / processing.

1. Login to your AWS account and go to IAM Management Console. Click on the **Roles** menu in the left side and then click on the **Create role** button.

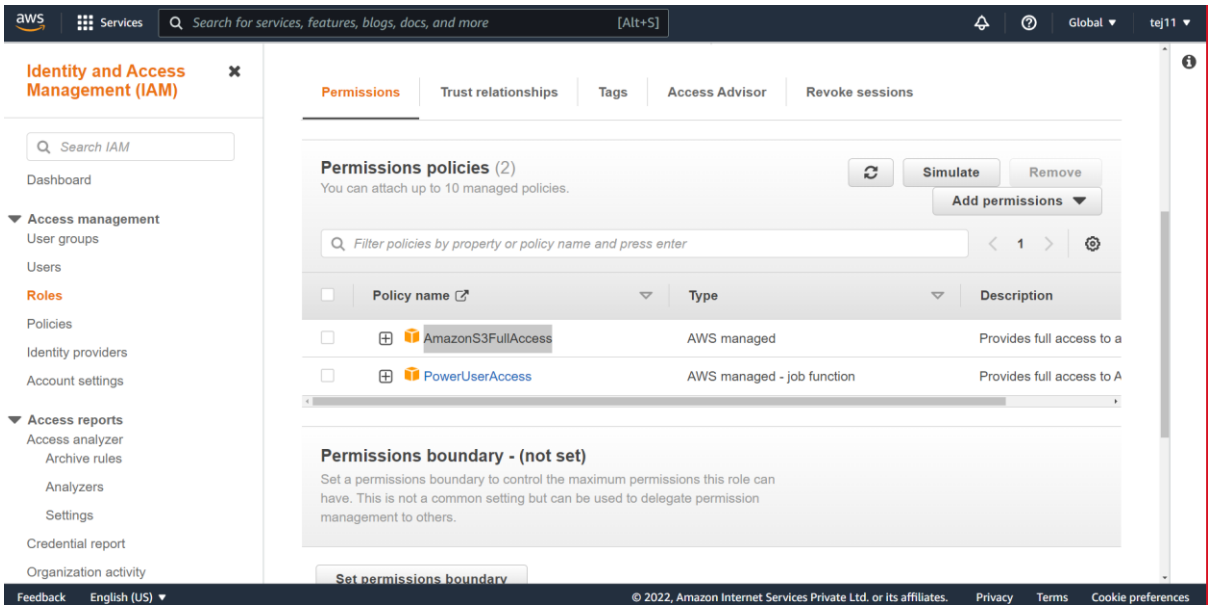


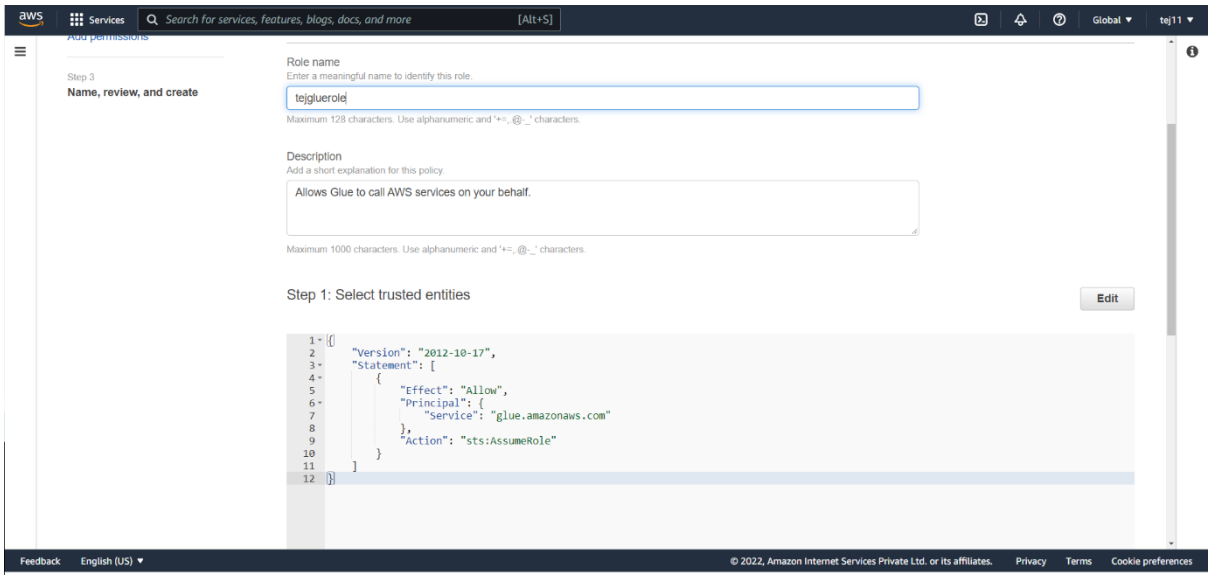


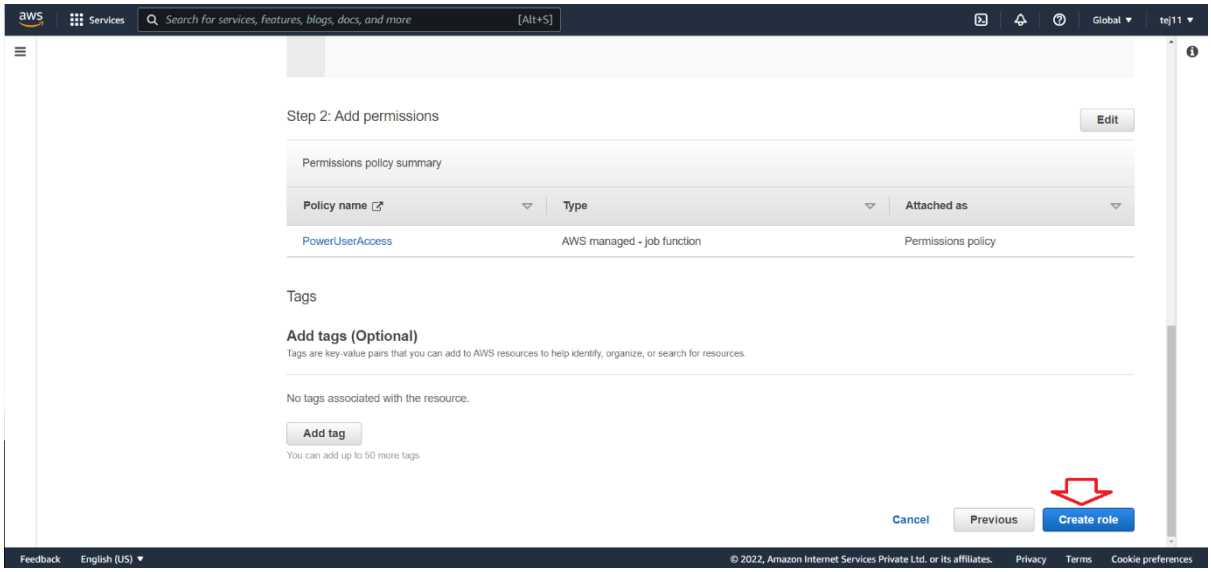
On the next screen, select **Glue** as the AWS Service. Click on the **Next: Permission** button.



On the next screen, select **PowerUserAccess/**AmazonS3FullAccess as the policies. The workshop is using this policy to simplify the authorization. However, in the production implementation, you select minimum required permission for the role. Click on the **Next: Tags** button.

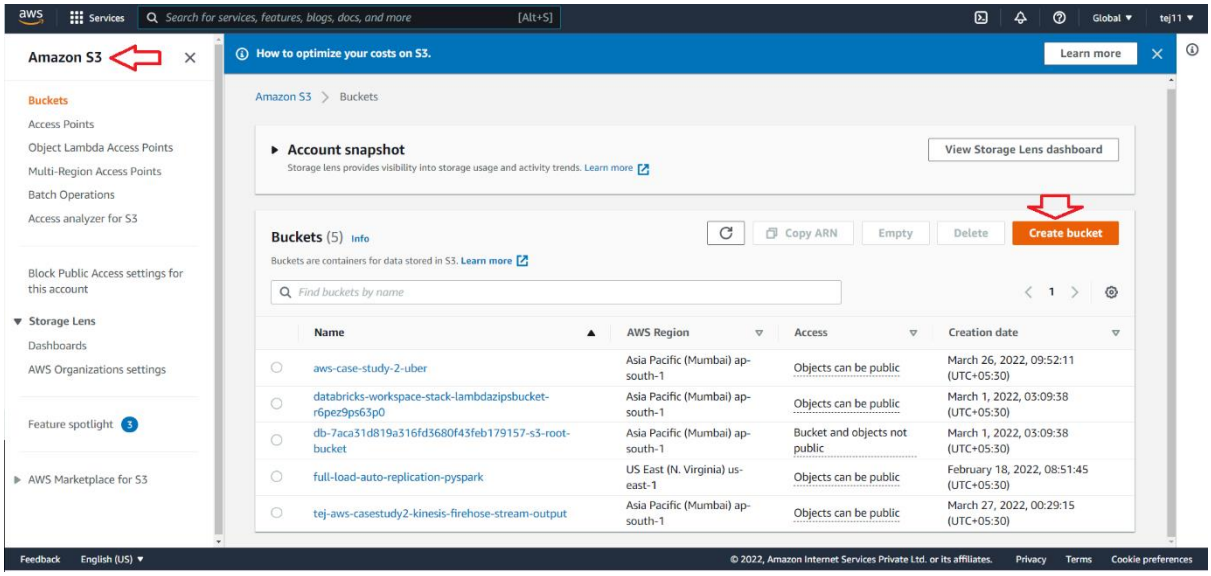




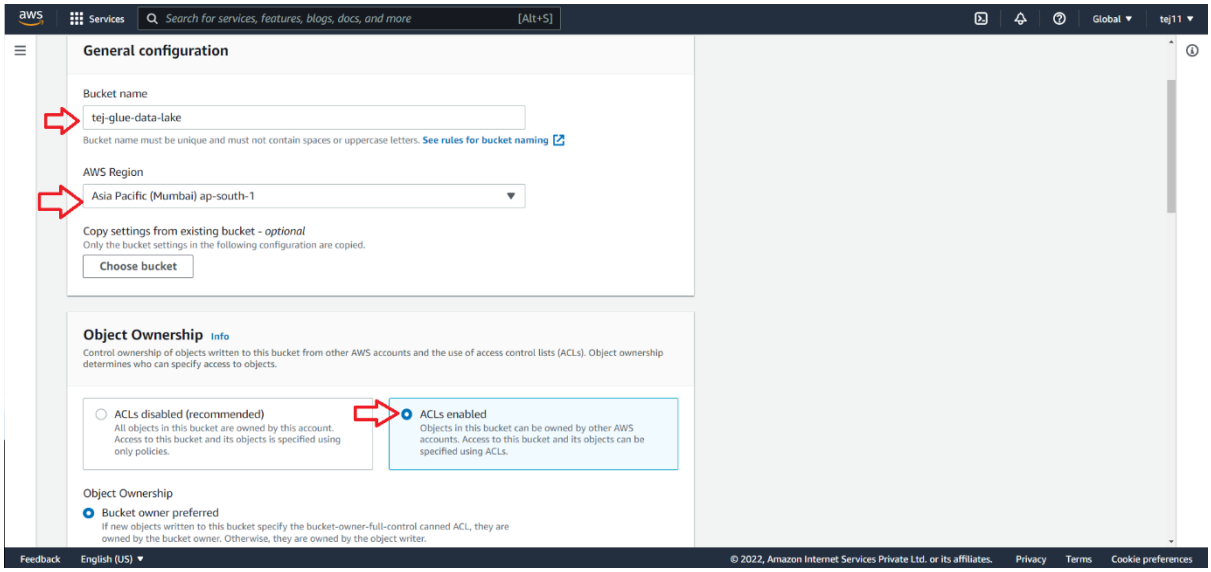


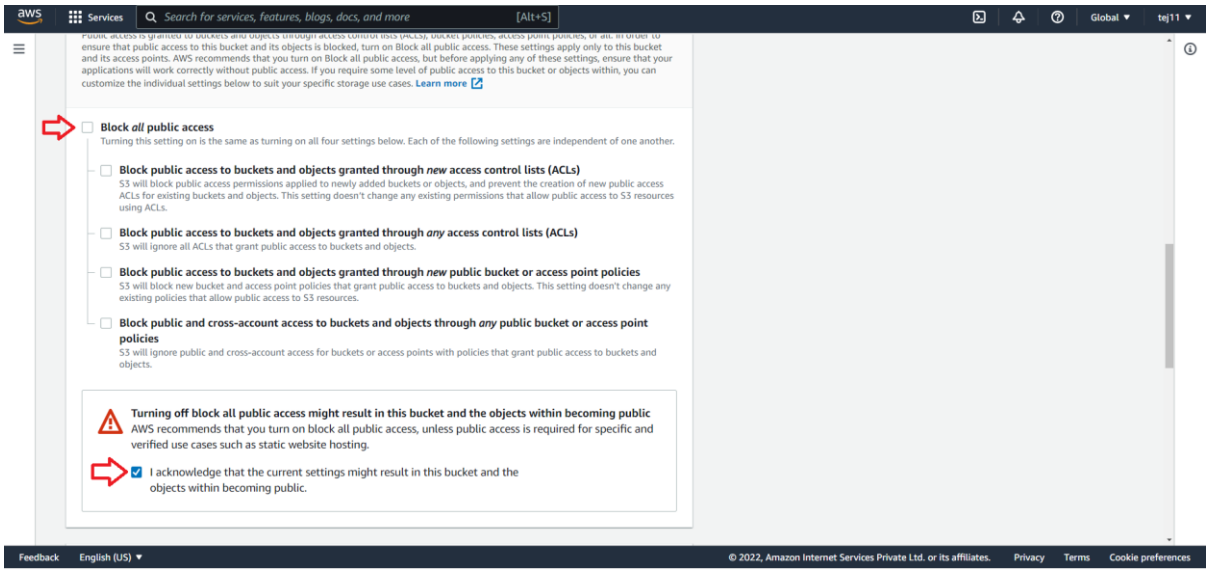
1. The role is created in no time. The next task is to create the S3 bucket for the data lake.

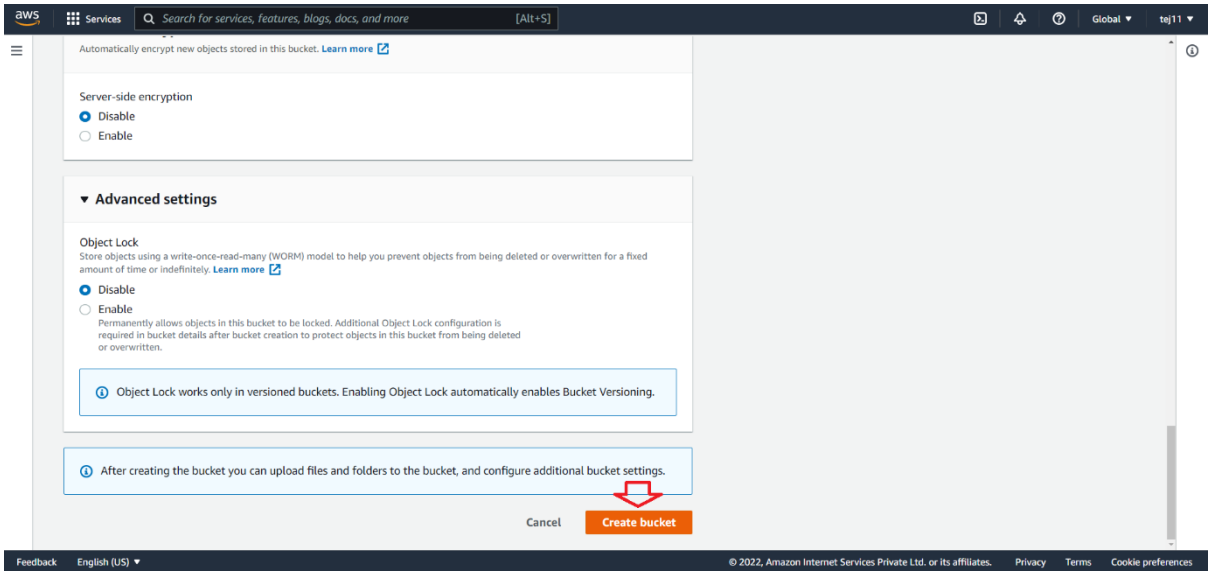
3

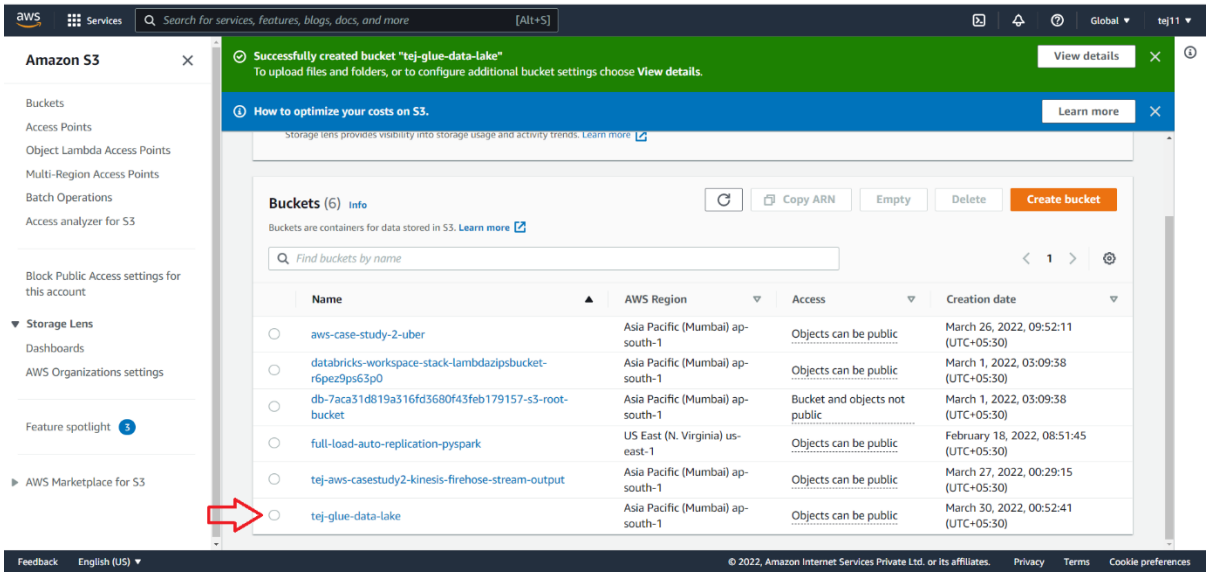


Create bucket button to create a new bucket with name **tej-glue-data-lake**. If the bucket name is not available, then use a different name which is available.

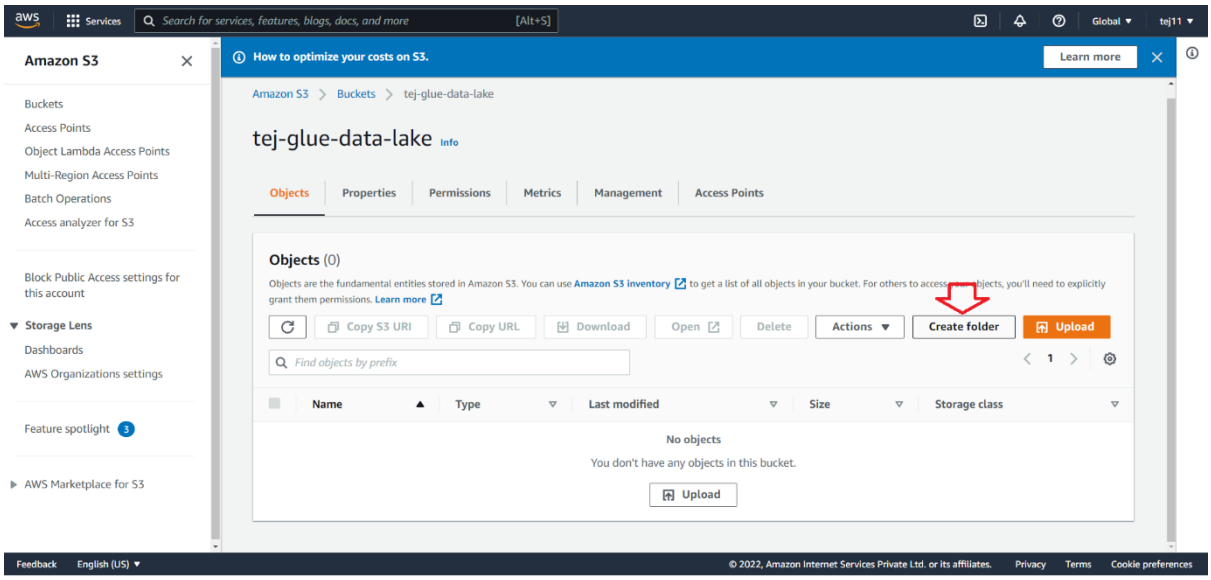

Services
Search for services, features, blogs, docs, and more [Alt+S]
Global
tej11
General configuration
Bucket name
tej-glue-data-lake
Bucket name must be unique and must not contain spaces or uppercase letters. See rules for bucket naming
AWS Region
Asia Pacific (Mumbai) ap-south-1
Copy settings from existing bucket - optional
Only the bucket settings in the following configuration are copied.
Choose bucket
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.
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.
Object Ownership
Bucket owner preferred
If new objects written to this bucket specify the bucket-owner-full-control canned ACL, they are owned by the bucket owner. Otherwise, they are owned by the object writer.
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

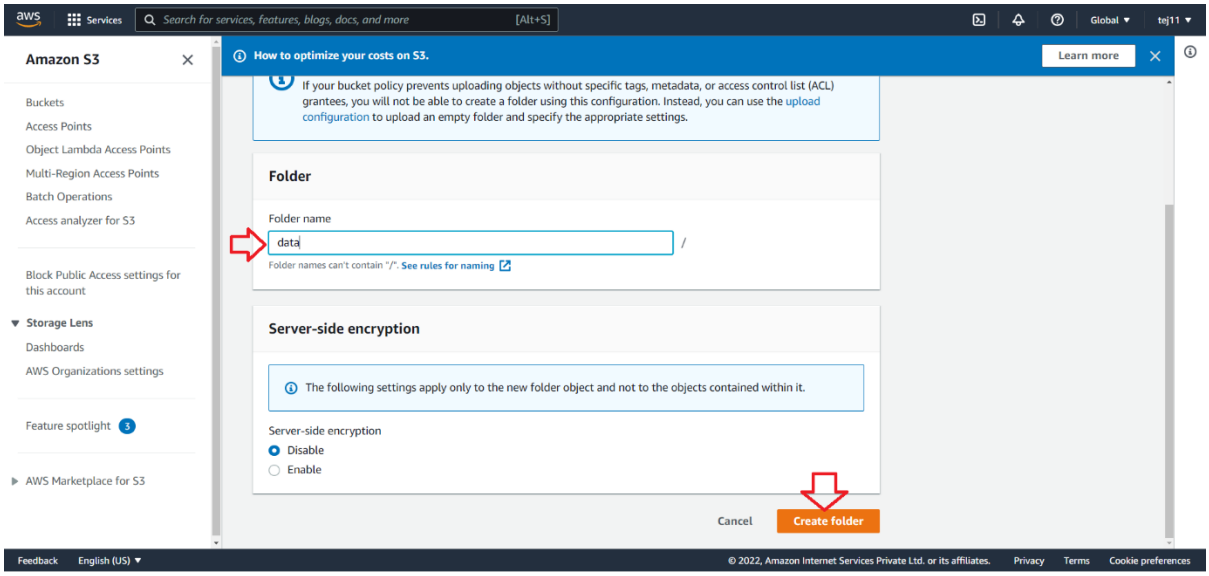

Services
Search for services, features, blogs, docs, and more [Alt+S]
Global
tej11
ensure that public access to this bucket and its objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, 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 this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more
Block all public access
Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.
Block public access to buckets and objects granted through new access control lists (ACLs)
S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.
Block public access to buckets and objects granted through any access control lists (ACLs)
S3 will ignore all ACLs that grant public access to buckets and objects.
Block public access to buckets and objects granted through new public bucket or access point policies
S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.
Block public and cross-account access to buckets and objects through any public bucket or access point policies
S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.
Turning off block all public access might result in this bucket and the objects within becoming public
AWS recommends that you turn on block all public access, unless public access is required for specific and verified use cases such as static website hosting.
I acknowledge that the current settings might result in this bucket and the objects within becoming public.
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

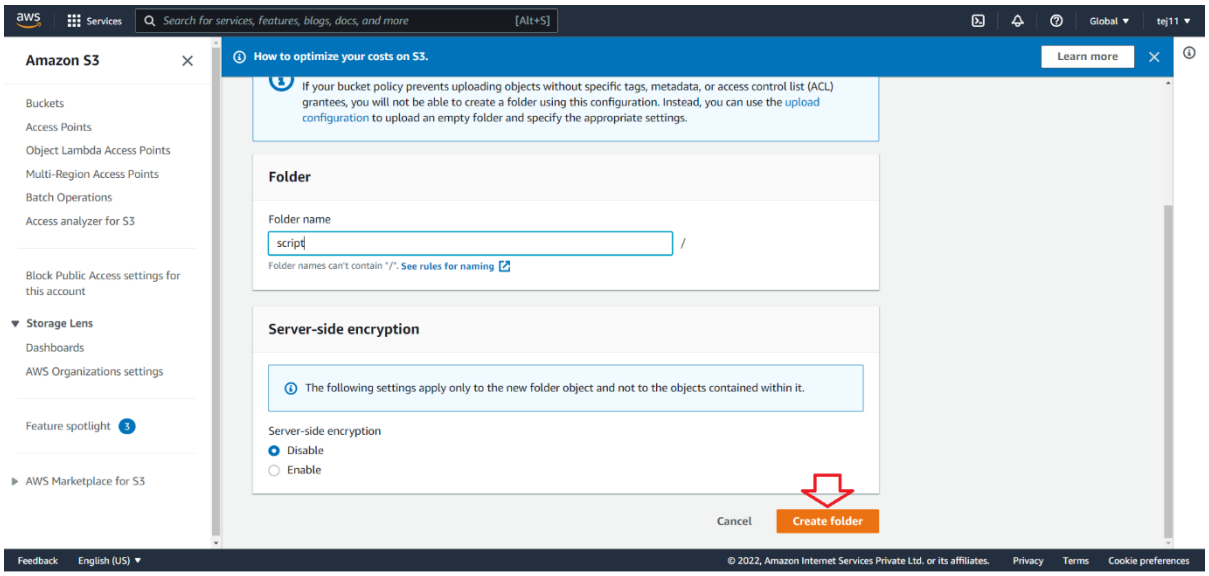

Services
Search for services, features, blogs, docs, and more [Alt+S]
Global
tej11
Automatically encrypt new objects stored in this bucket. Learn more
Server-side encryption
Disable
Enable
Advanced settings
Object Lock
Store objects using a write-once-read-many (WORM) model to help you prevent objects from being deleted or overwritten for a fixed amount of time or indefinitely. Learn more
Disable
Enable
Permanently allows objects in this bucket to be locked. Additional Object Lock configuration is required in bucket details after bucket creation to protect objects in this bucket from being deleted or overwritten.
Object Lock works only in versioned buckets. Enabling Object Lock automatically enables Bucket Versioning.
After creating the bucket you can upload files and folders to the bucket, and configure additional bucket settings.
Cancel
Create bucket
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

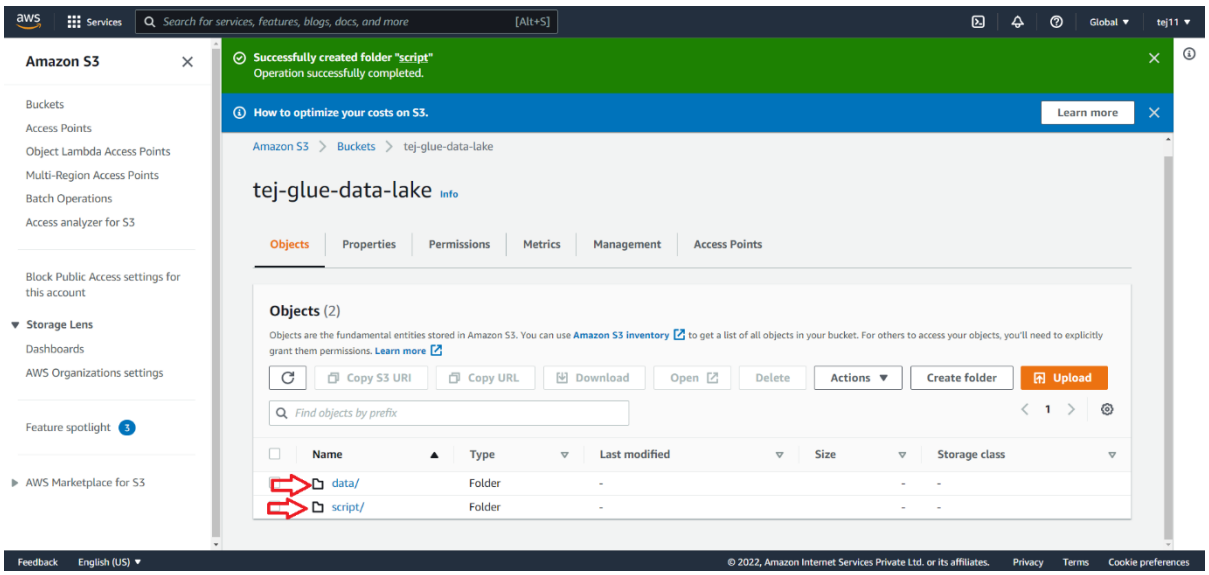


Click on the **tej-glue-data-lake** bucket to open it. Within the **tej-glue-data-lake**, create two folders **data** and **script** using the **+ Create folder** button. The **data** folder is used to keep the data lake data while the **script** folder is used by Glue job to store the script.

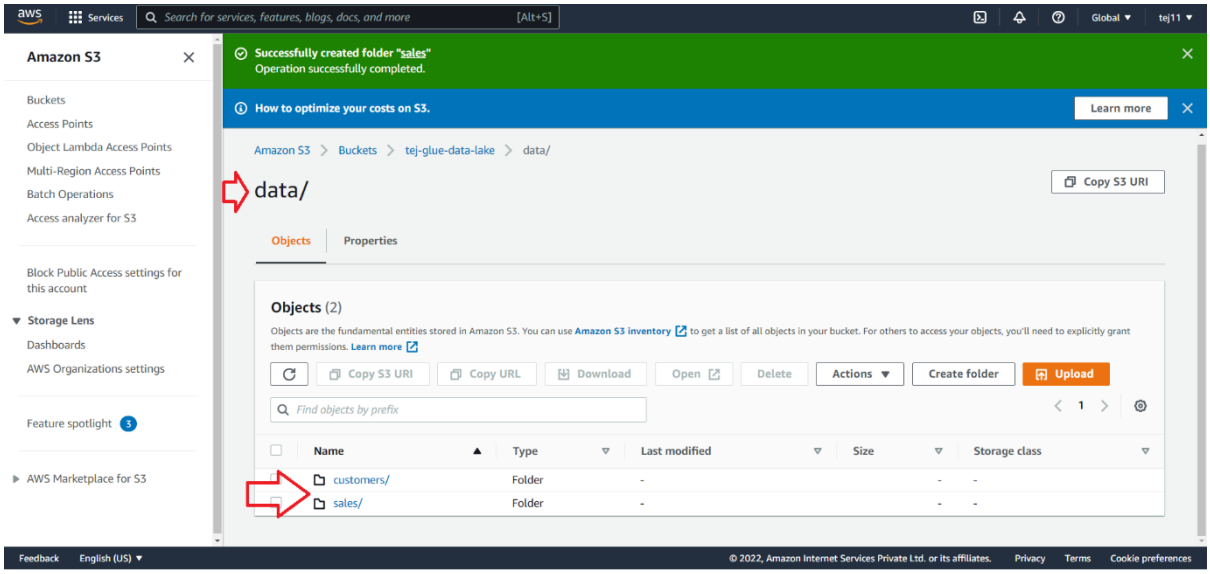



Amazon S3
How to optimize your costs on S3.
Learn more
If your bucket policy prevents uploading objects without specific tags, metadata, or access control list (ACL) grantees, you will not be able to create a folder using this configuration. Instead, you can use the upload configuration to upload an empty folder and specify the appropriate settings.
Folder
Folder name
data
/
Folder names can't contain "/". See rules for naming
Server-side encryption
The following settings apply only to the new folder object and not to the objects contained within it.
Server-side encryption
Disable
Enable
Cancel
Create folder
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon S3
How to optimize your costs on S3.
Learn more
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
script
/
Folder names can't contain "/". See rules for naming
Server-side encryption
The following settings apply only to the new folder object and not to the objects contained within it.
Server-side encryption
Disable
Enable
Cancel
Create folder
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon S3
Successfully created folder "script"
Operation successfully completed.
How to optimize your costs on S3.
Learn more
Amazon S3 > Buckets > tej-glue-data-lake
tej-glue-data-lake info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (2)
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
data/
Folder
script/
Folder
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

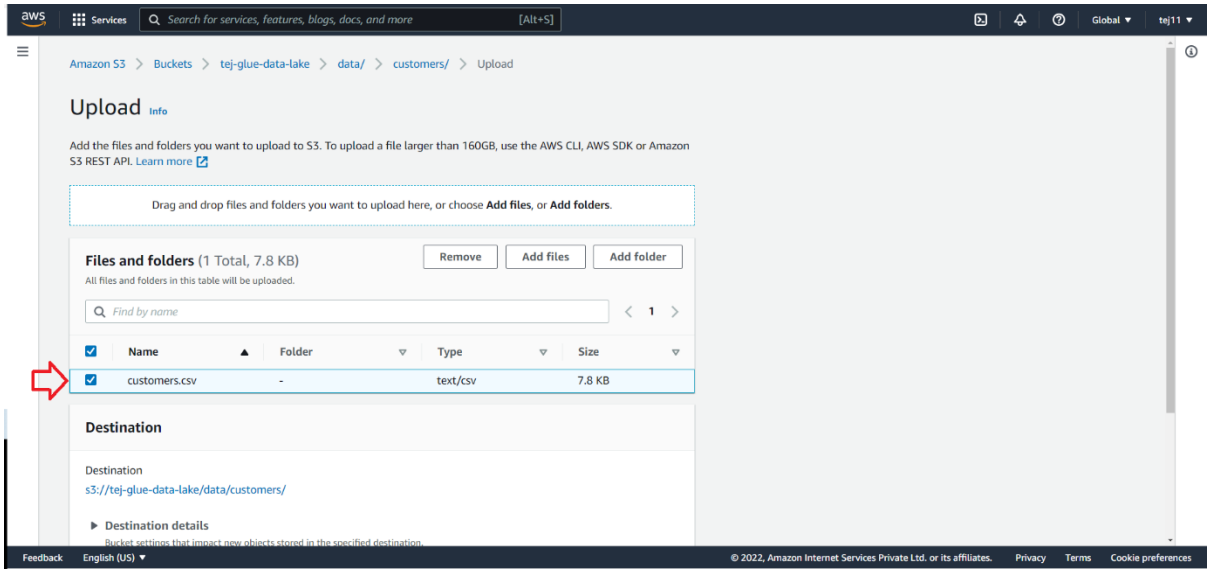
Click on the **data** folder to open it. Within the **data** folder, create two folders **customers** and **sales** using the **+ Create folder** button.

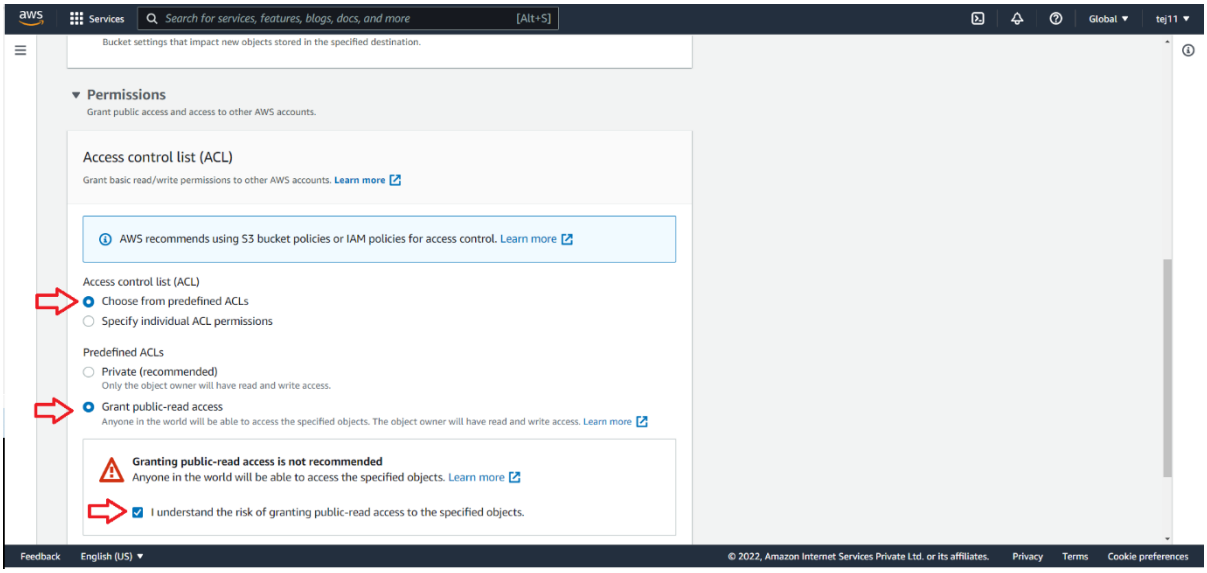


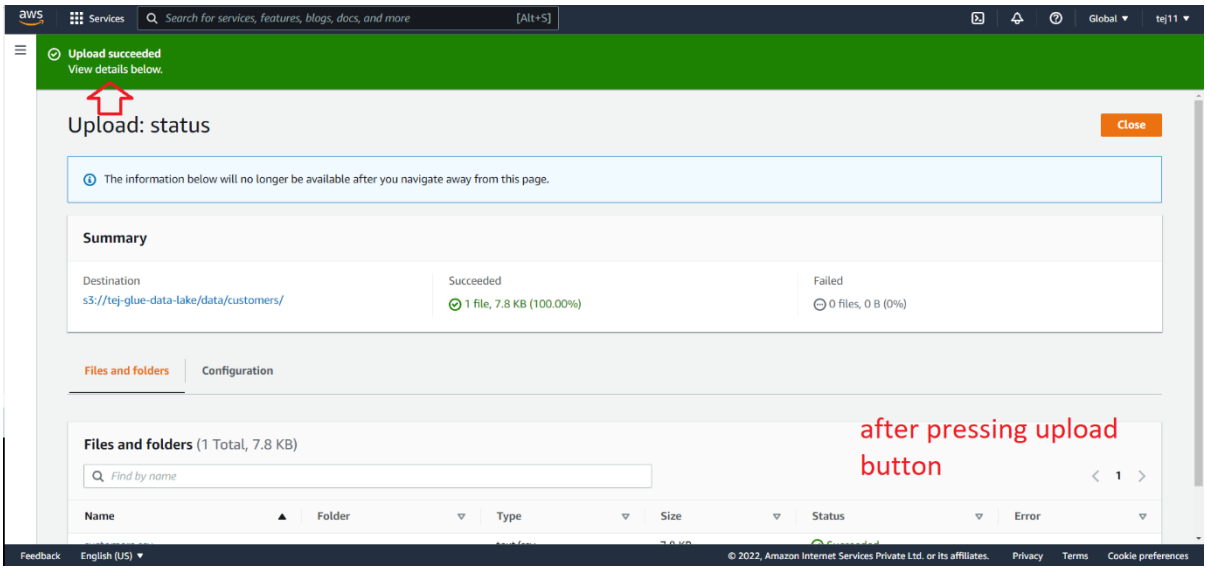
Click to open **customers** and **sales** folders one by one and upload **customers.csv** and **sales.csv** files in the **customers** and **sales** folders respectively. Use **Upload** button to upload the files. The **customers.csv** and **sales.csv** files are available for download using the following links –

https://filepost.io/d/XuX9yGpRfP

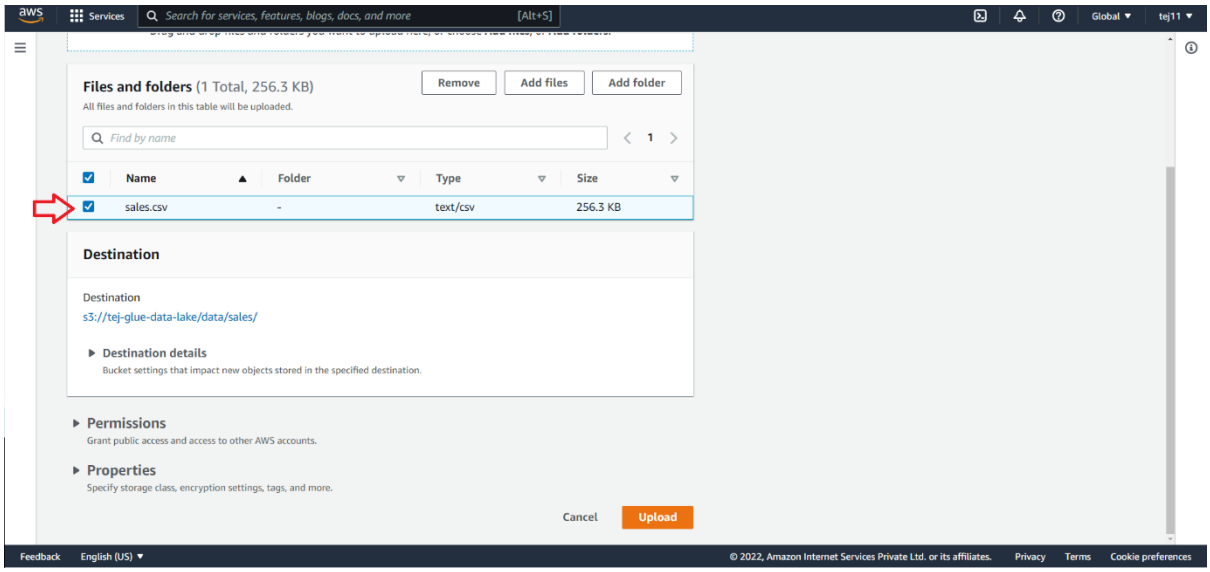
upload steps

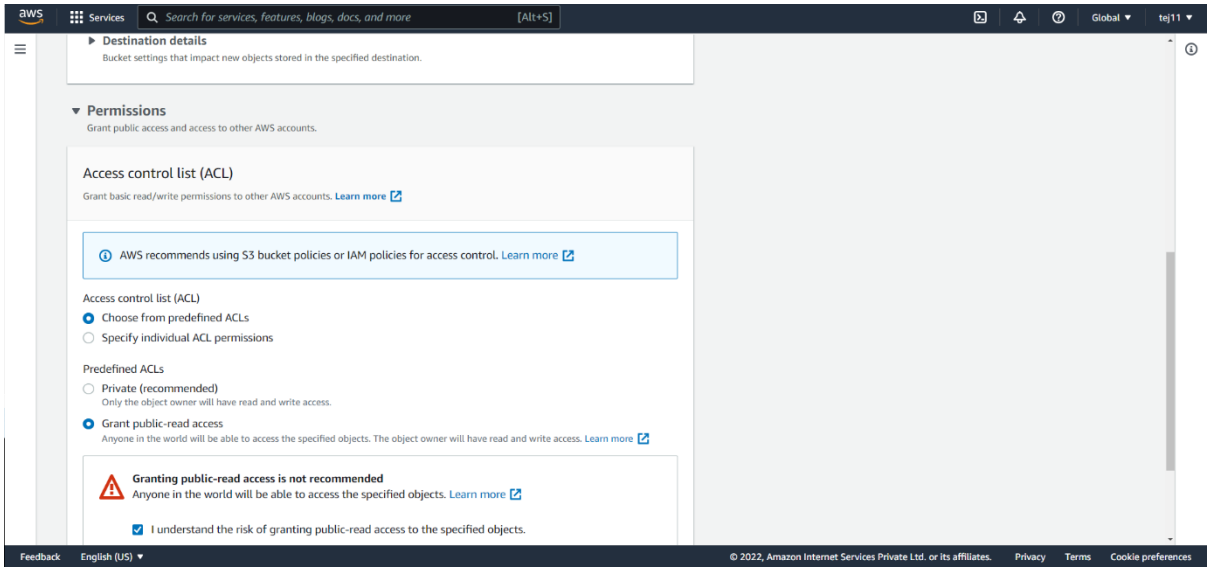


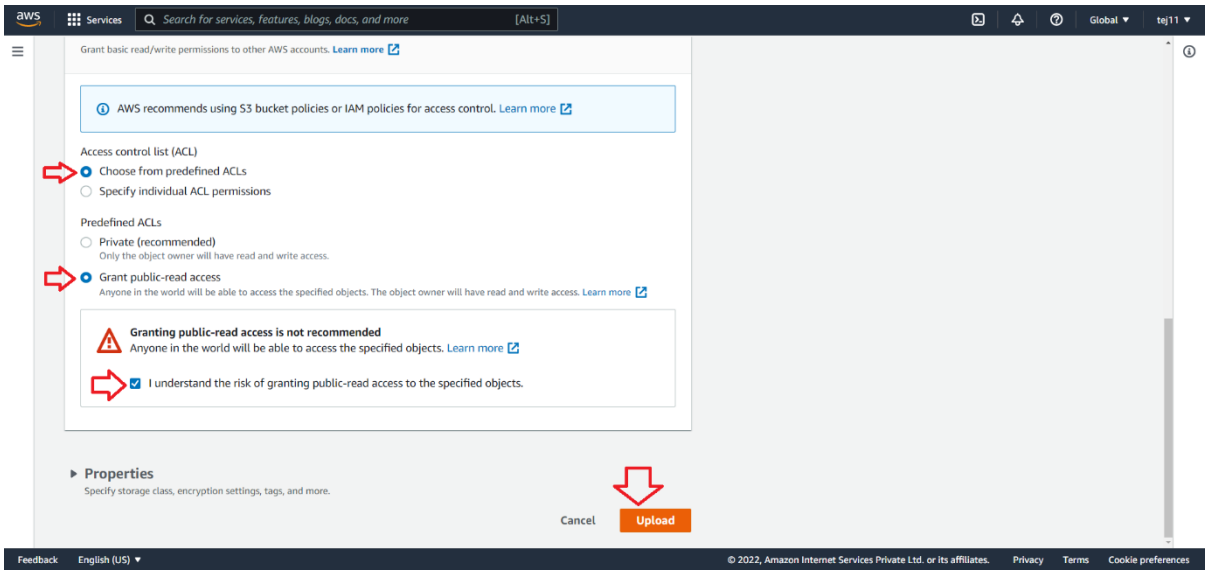




Next uploading sales data in sales folder.





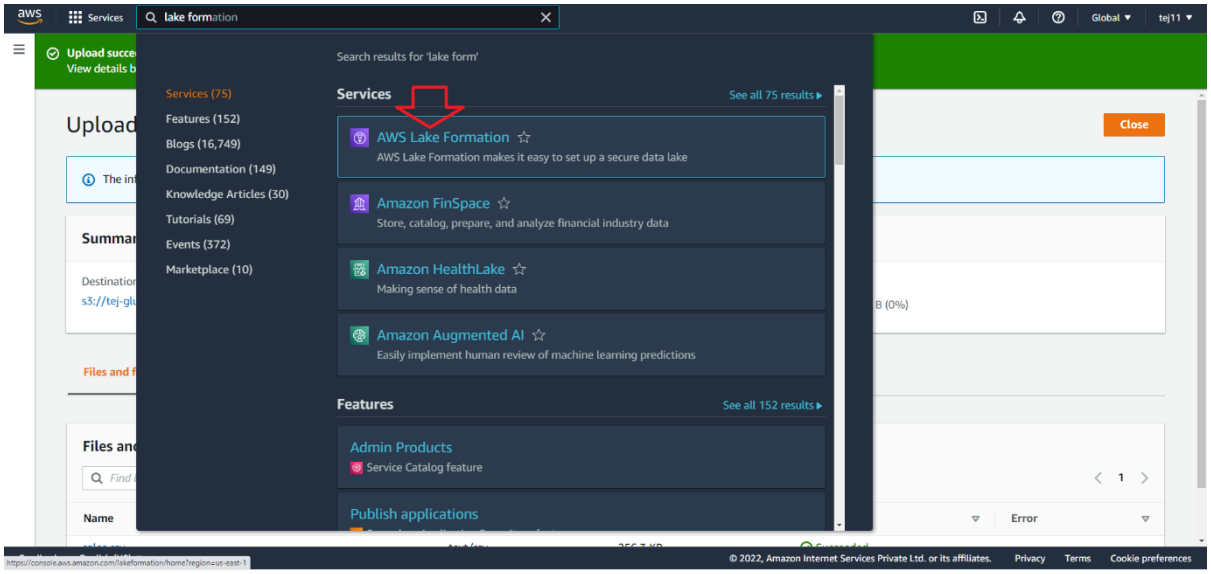


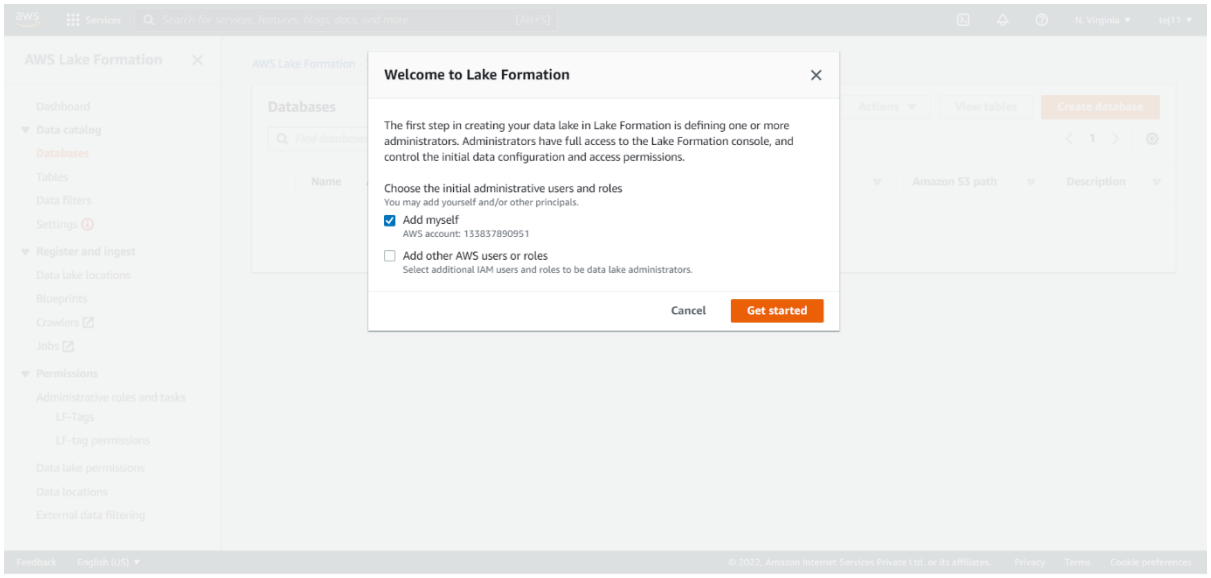
1. *The data is ready. Let’s start with the data lake configuration.

4

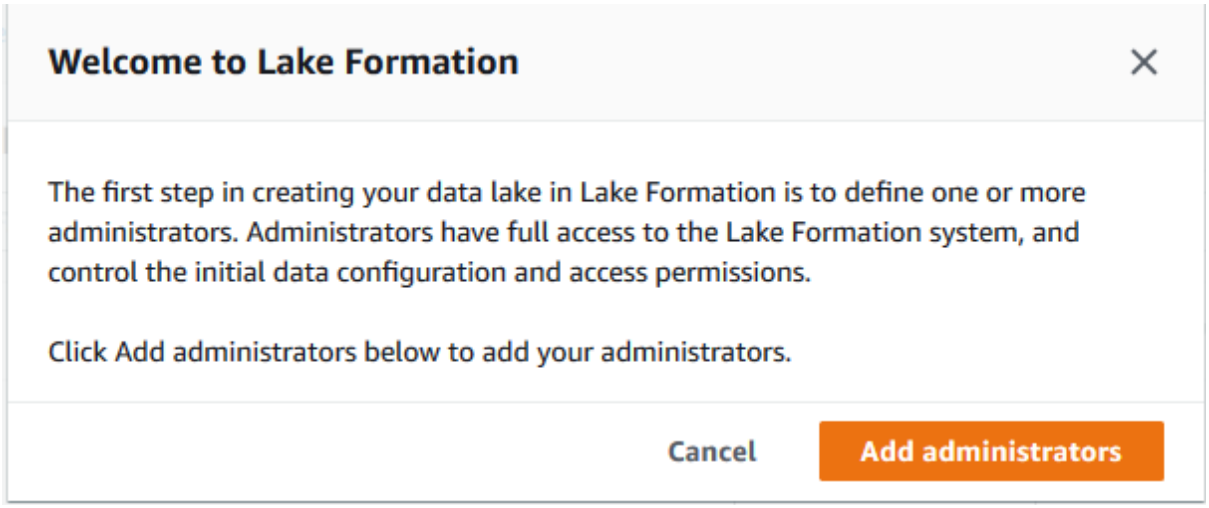
Configure Data Lake

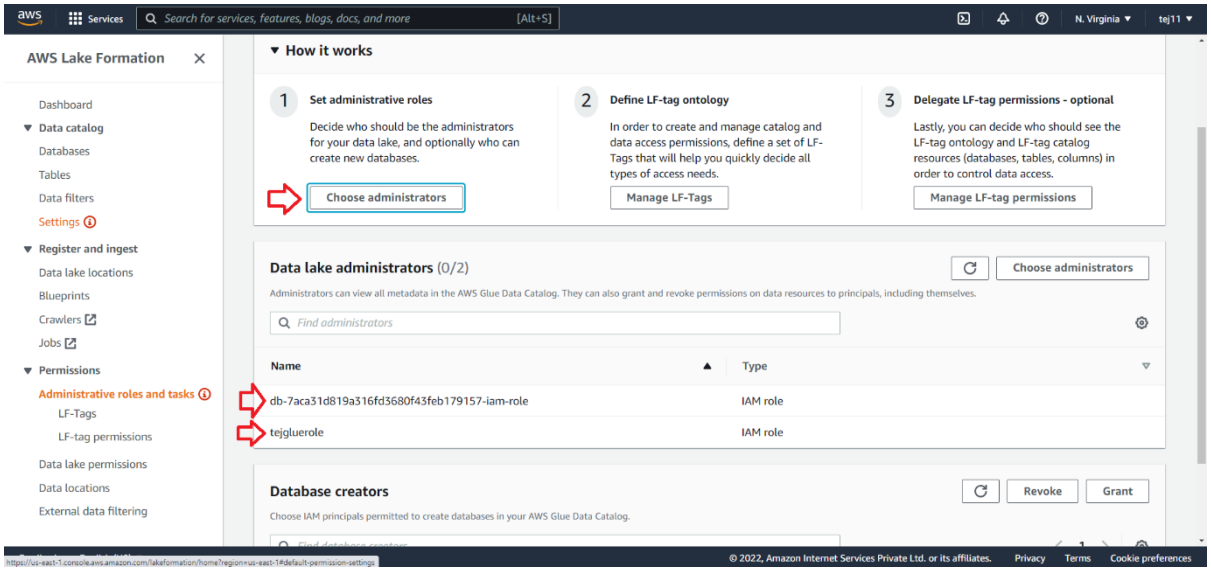
The data is ready in Amazon S3 bucket. Let’s start configuration of the data lake. In data lake, you create a database and configure it with Amazon S3 bucket location. A database is used to organize data catalog tables in the data lake. You will later configure crawler to automatically create the data catalog tables in the data lake.



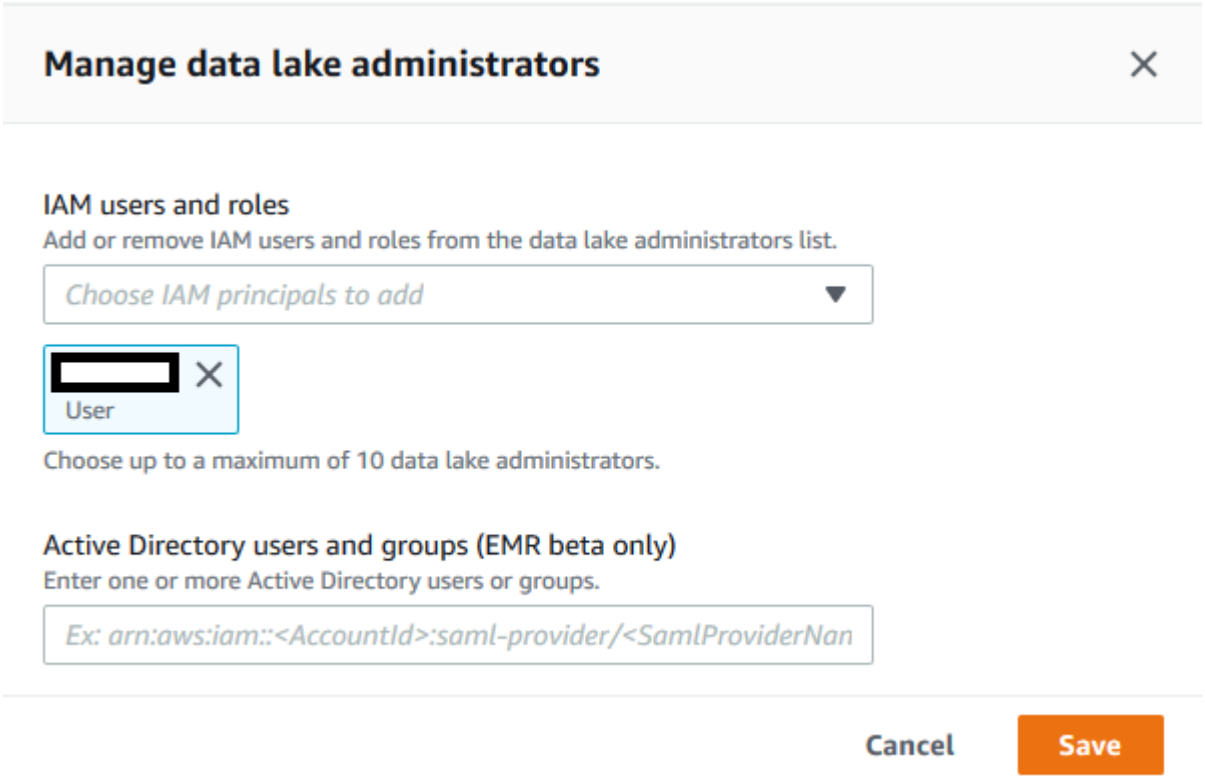


Open the AWS Lake Formation console. If you are using Lake Formation for the first time in the region, it will ask you to create a data lake administrator. A data lake administrator is an IAM user or IAM role that performs administrative tasks on the data lake. For the first time user, it will pop-up a message to add administrators. Click on the **Add administrators** button to create administrators for your Data Lake.



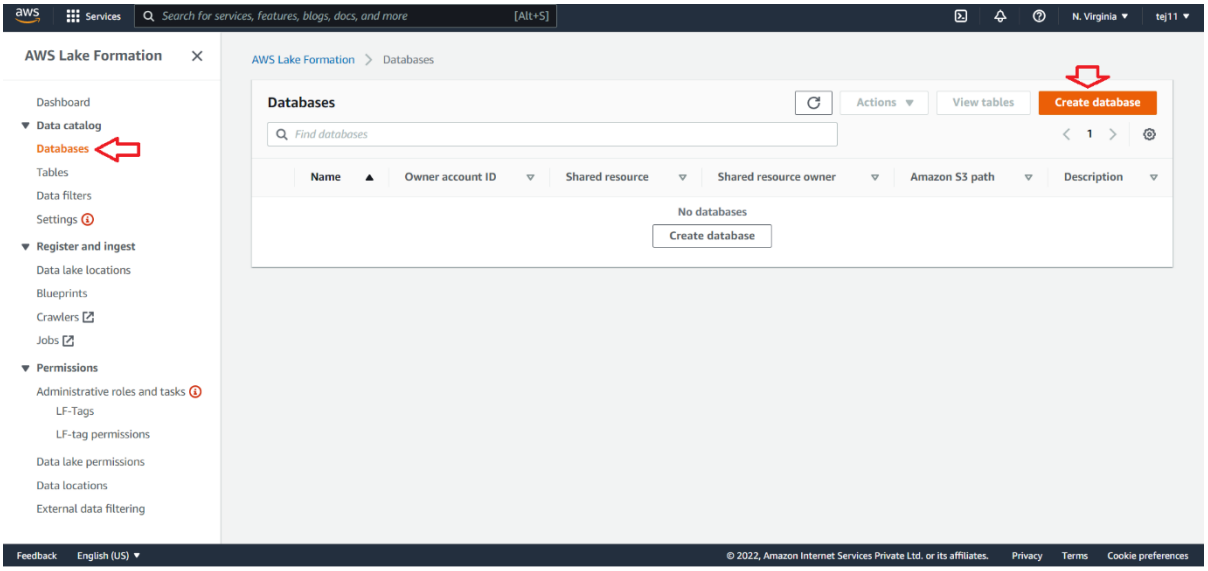


Select **your AWS logged-in IAM user** from the drop down list. For the rest of workshop, the user will be considered as a data lake administrator and will have full access to the data lake. Click on the **Save** button.

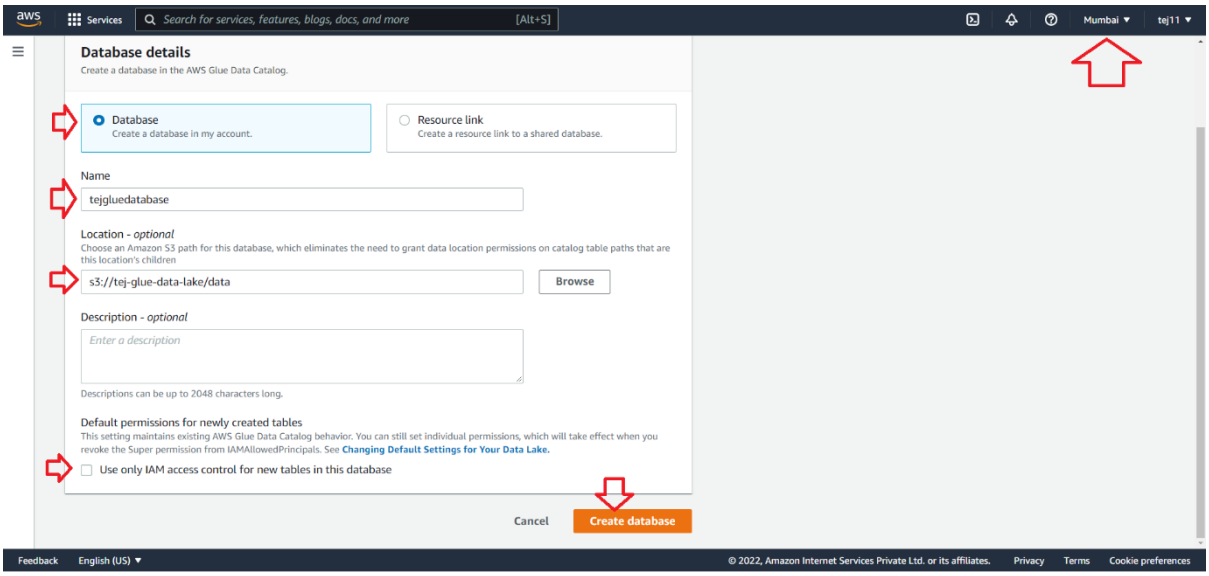


Note: if you did not get the popup then that means the data lake already has an administrator. You can check that by clicking on the "Admins and database creators" menu in the left. If you see that your logged-in IAM username is listed as the "Data Lake Administrator" then you are ok to move to the next step. Otherwise, click on the "Grant" button to add "your AWS logged-in IAM user" as the administrator of the data lake.

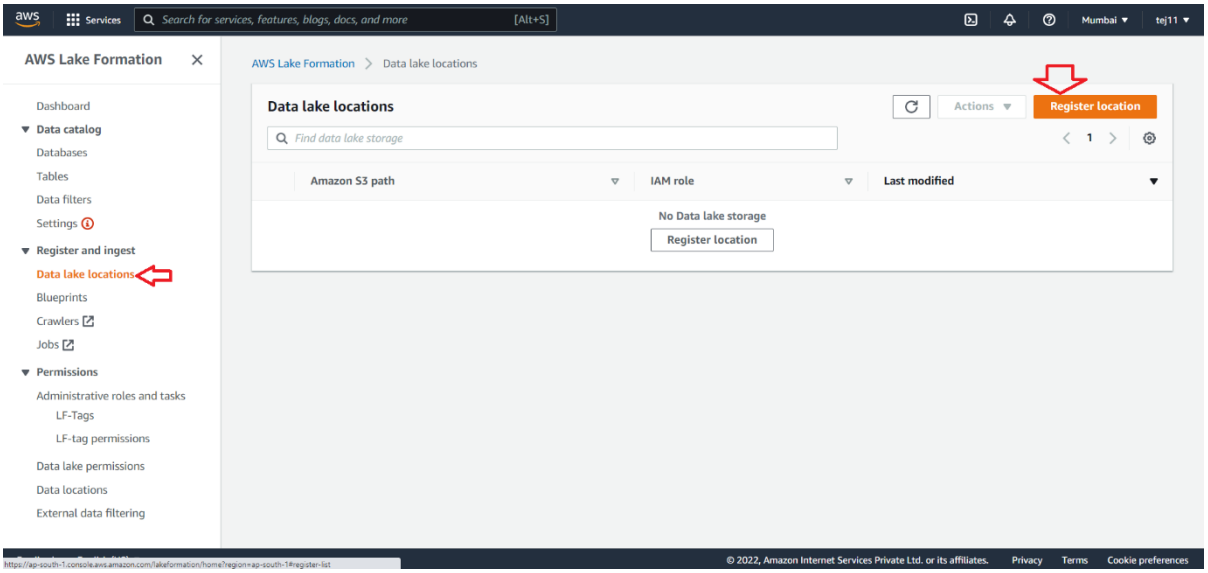
After adding the administrator, you will create the database. On the AWS Lake Formation console, click on the **Databases** option on the left menu and then click on **Create database** button.



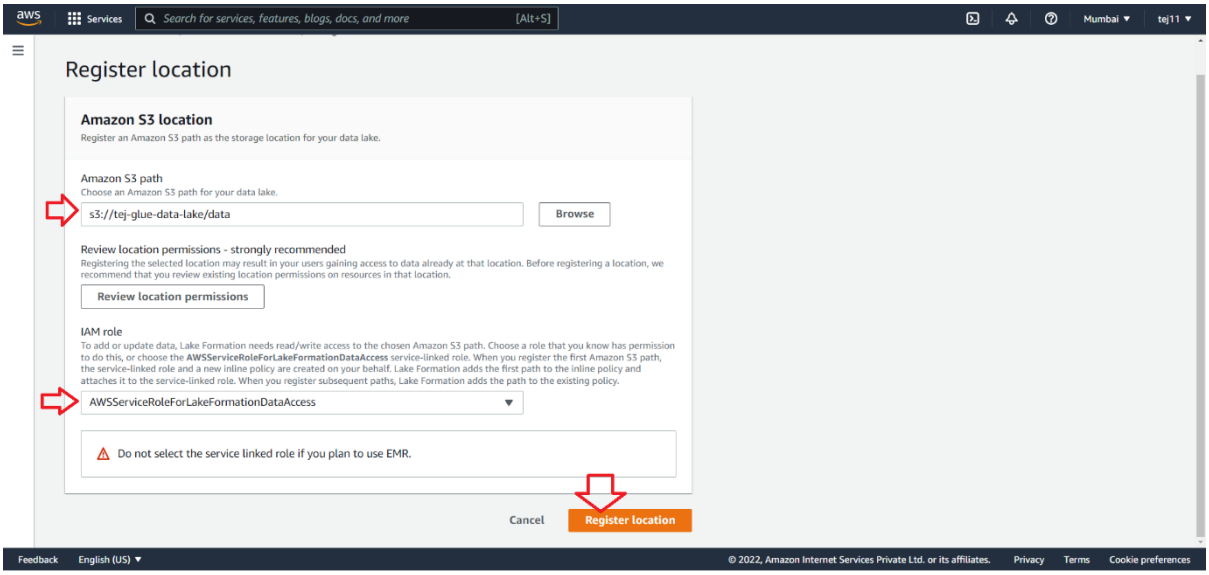
On the next screen, enter **tejgluedatabase** as the **Name**. On the **Location box**, select the S3 data lake path as **s3://tej-glue-data-lake/data**. If you created the bucket with different name, then you replace **tej-glue-data-lake** part with that name. Make sure you **Uncheck** the option - **Use only IAM access control for new tables in this database**. Leave rest of the options as default and click on **Create database** button.



The database is added in no time. Now register the Amazon S3 bucket as your data lake storage. In the navigation pane, choose **Data lake locations**, and then choose **Register location**.



Enter a path to the existing S3 bucket **s3://tej-glue-data-lake/data**. If you created the bucket with a different name, then you replace **tej-glue-data-lake** part with that name. For the IAM role, select the **AWSServiceRoleForLakeFormationDataAccess** role. Click **Register location** to save it.



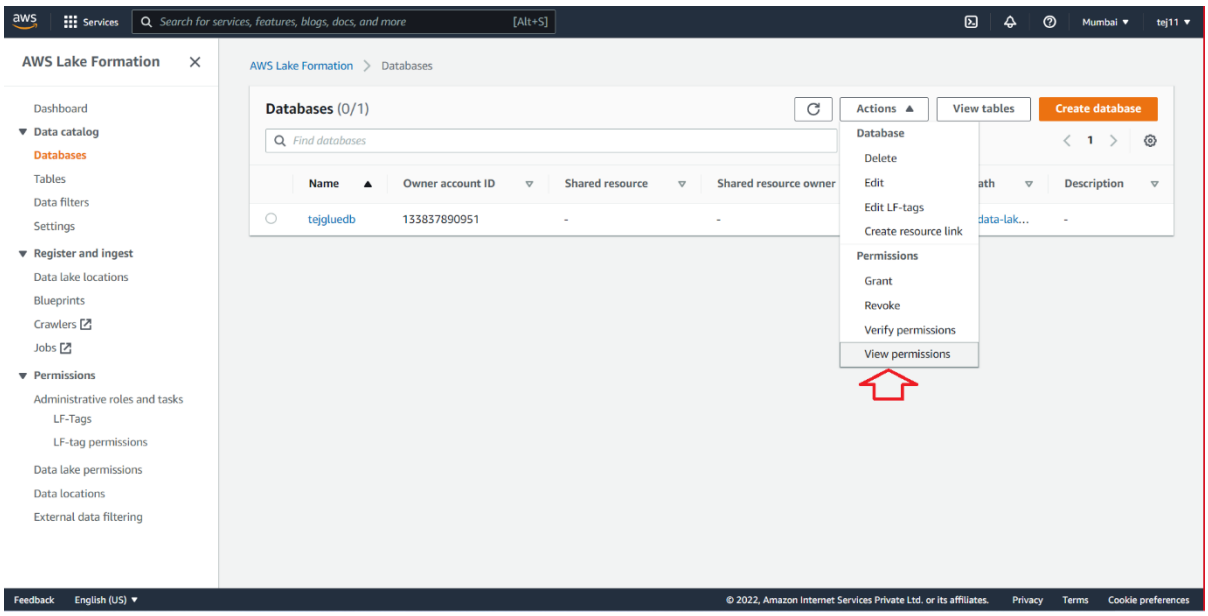
It is time to use crawler to catalog the S3 bucket data in the database.

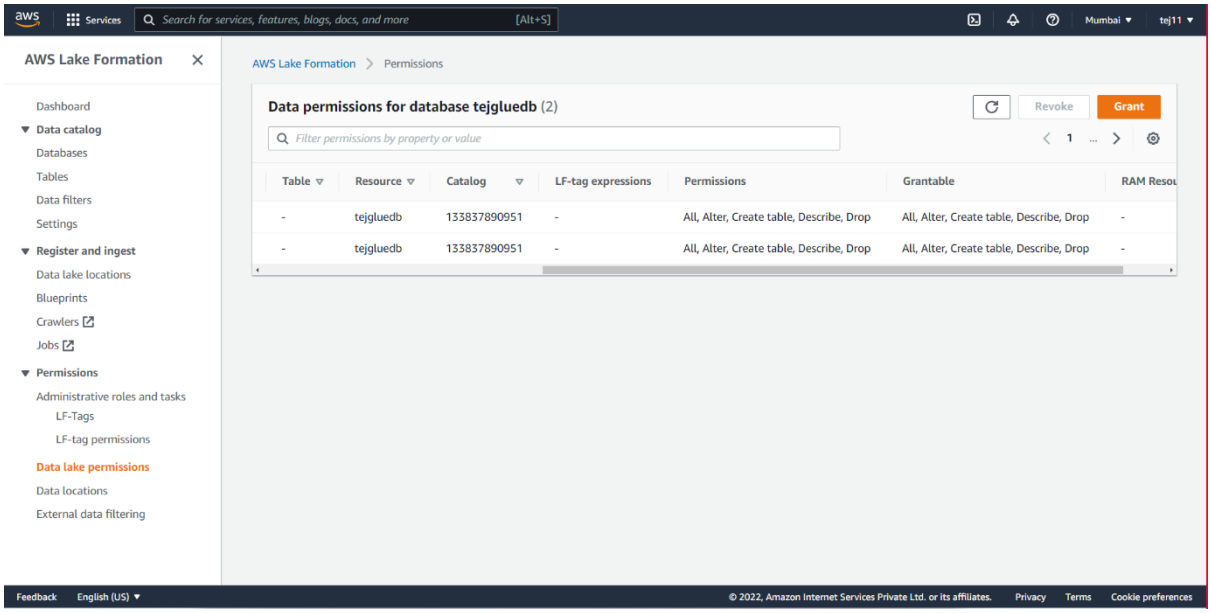
5

Configure and Run Crawler

One of the fundamental principle of building the data lake is that every data in the data lake should be catalogued. The catalog is automated using crawlers in AWS Glue. The crawler uses role based authorization to create catalog in the data lake database. You created an IAM Role **tejgluerole** in the earlier task which the crawler will use to create data catalog in the database. You need to assign database permission for this role. After the permission configuration, you will create and run crawler to catalog the data.

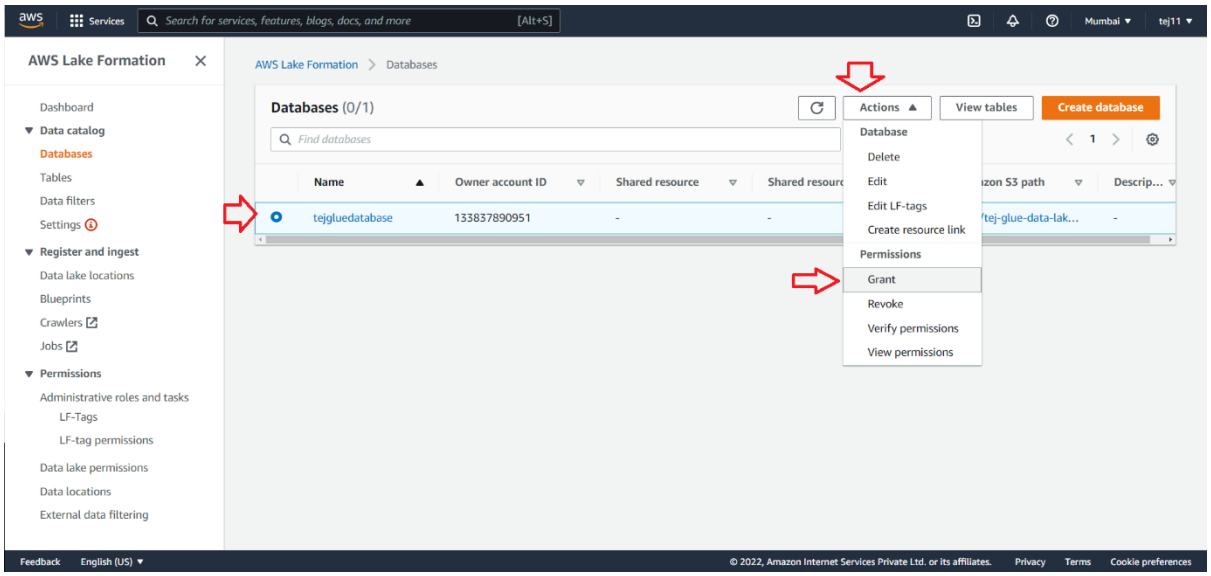
Before granting , the permissions , please vie permission to already see , if permissions are assigned .

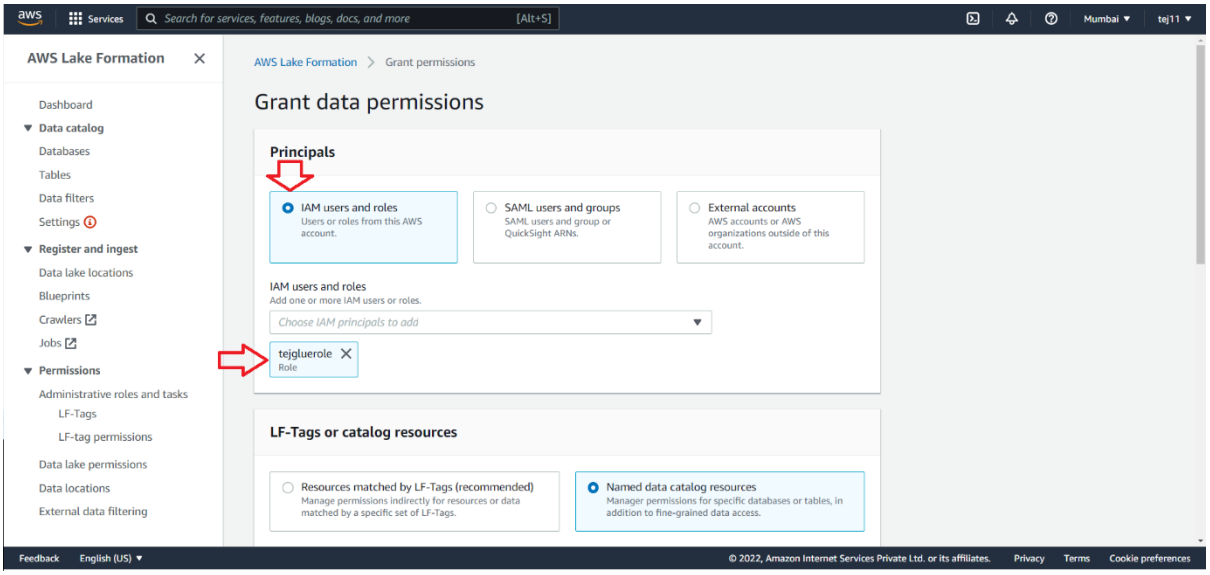


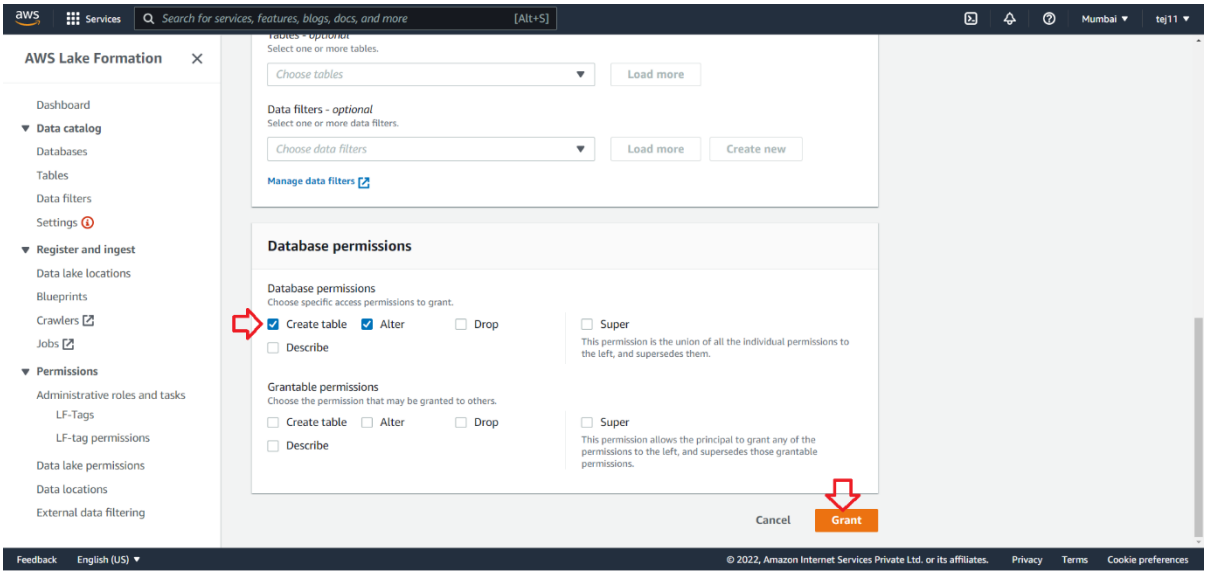


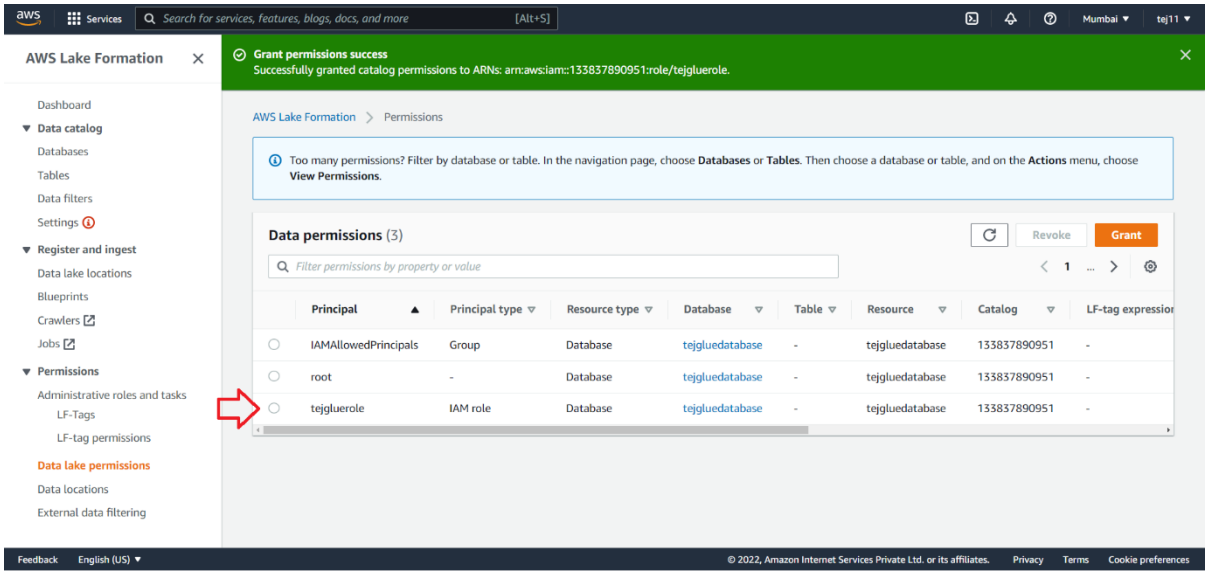
Open the AWS Lake Formation console, click on the **Databases** option on the left. You will see **tejgluedatabase** database listed.

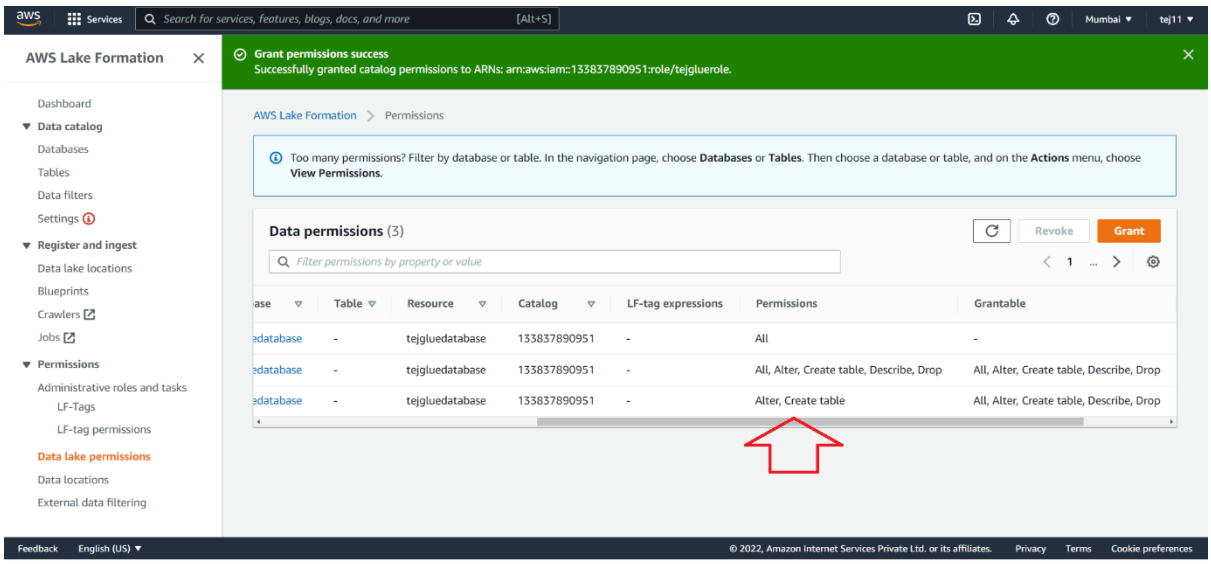
Select the **tejgluedatabase** database and click on the **Grant** menu option under the **Action** dropdown menu.



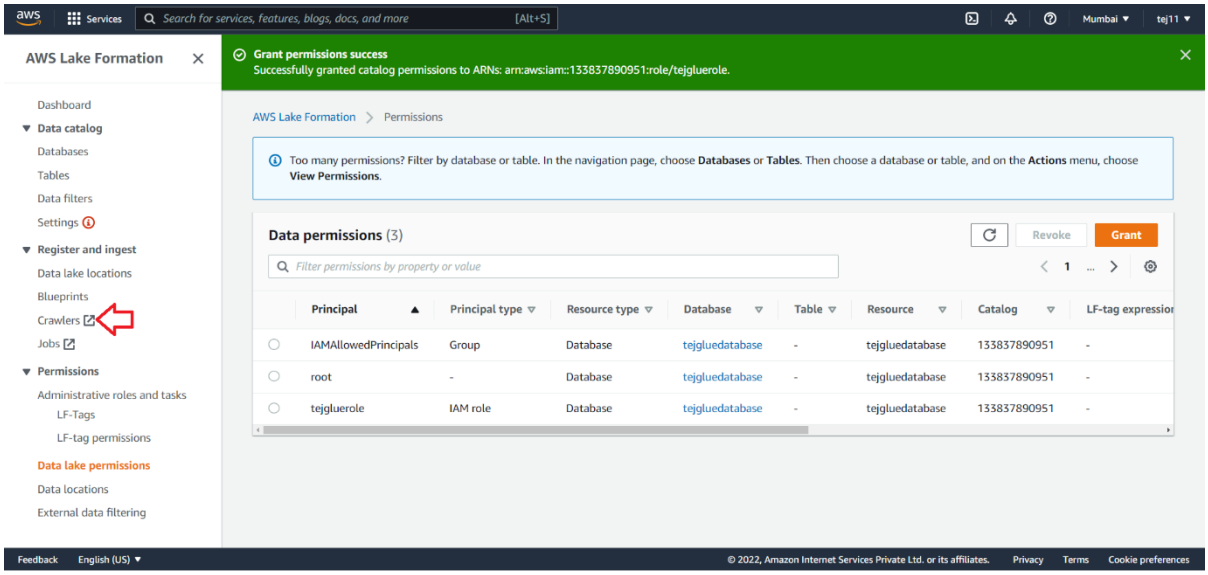

aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Mumbai
tej11
AWS Lake Formation
Dashboard
Data catalog
Databases
Tables
Data filters
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
LF-Tags
LF-tag permissions
Data lake permissions
Data locations
External data filtering
AWS Lake Formation > Grant permissions
Grant data permissions
Principals
IAM users and roles
Users or roles from this AWS account.
SAML users and groups
SAML users and group or QuickSight ARNs.
External accounts
AWS accounts or AWS organizations outside of this account.
IAM users and roles
Add one or more IAM users or roles.
Choose IAM principals to add
tejgluerole
Role
LF-Tags or catalog resources
Resources matched by LF-Tags (recommended)
Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.
Named data catalog resources
Manager permissions for specific databases or tables, in addition to fine-grained data access.
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

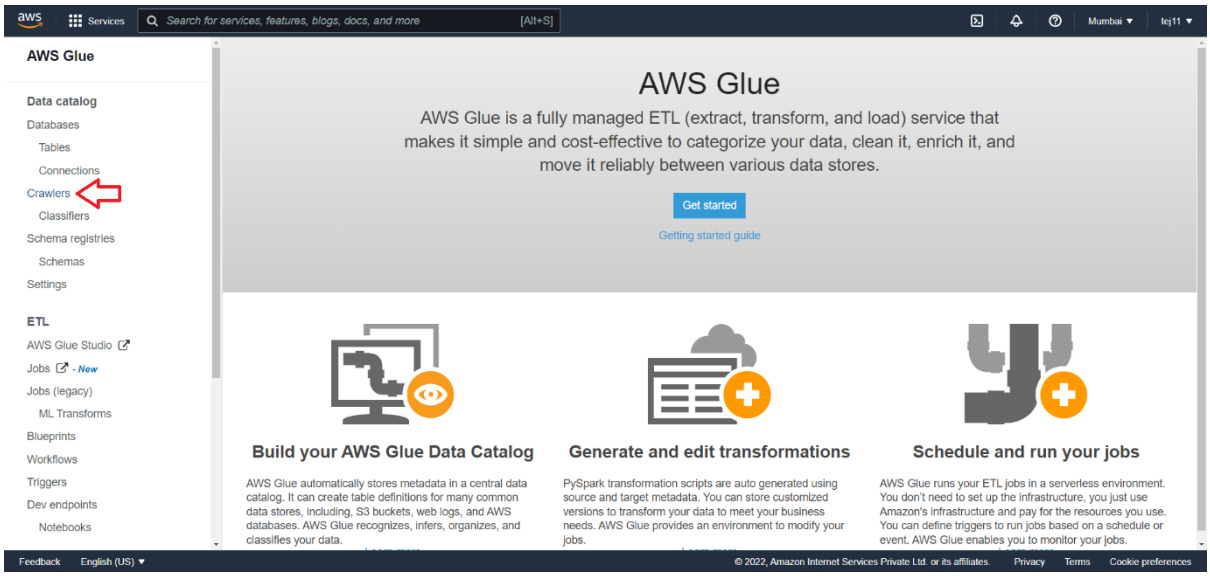

aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Mumbai
tej11
AWS Lake Formation
Dashboard
Data catalog
Databases
Tables
Data filters
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
LF-Tags
LF-tag permissions
Data lake permissions
Data locations
External data filtering
Select one or more tables.
Choose tables
Load more
Data filters - optional
Select one or more data filters.
Choose data filters
Load more
Create new
Manage data filters
Database permissions
Database permissions
Choose specific access permissions to grant.
Create table
Alter
Drop
Describe
Super
This permission is the union of all the individual permissions to the left, and supersedes them.
Grantable permissions
Choose the permission that may be granted to others.
Create table
Alter
Drop
Describe
Super
This permission allows the principal to grant any of the permissions to the left, and supersedes those grantable permissions.
Cancel
Grant
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

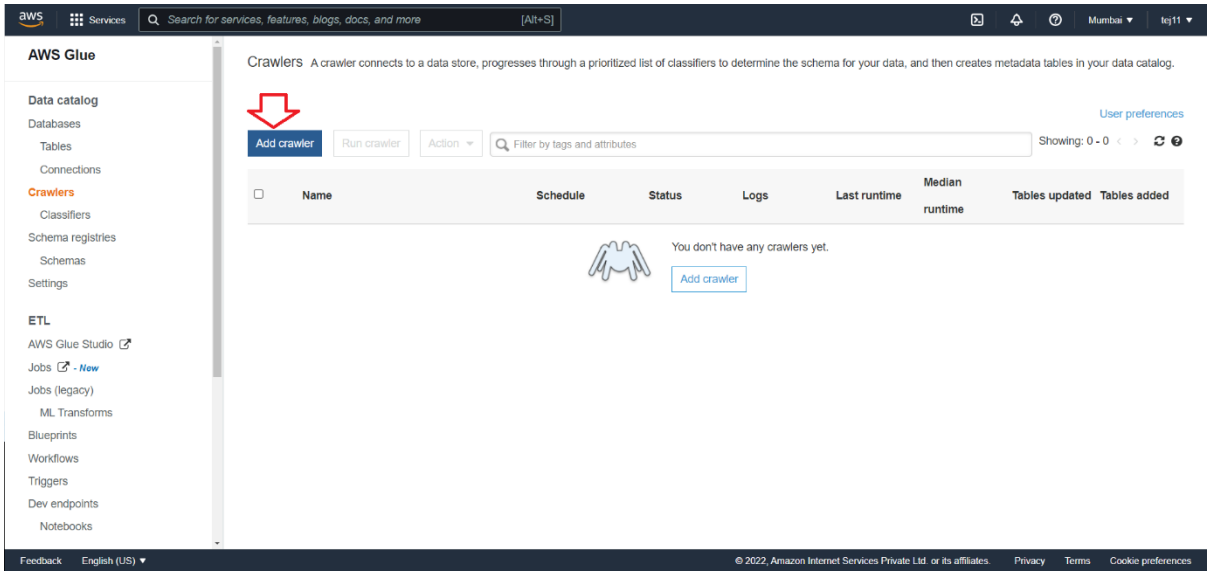


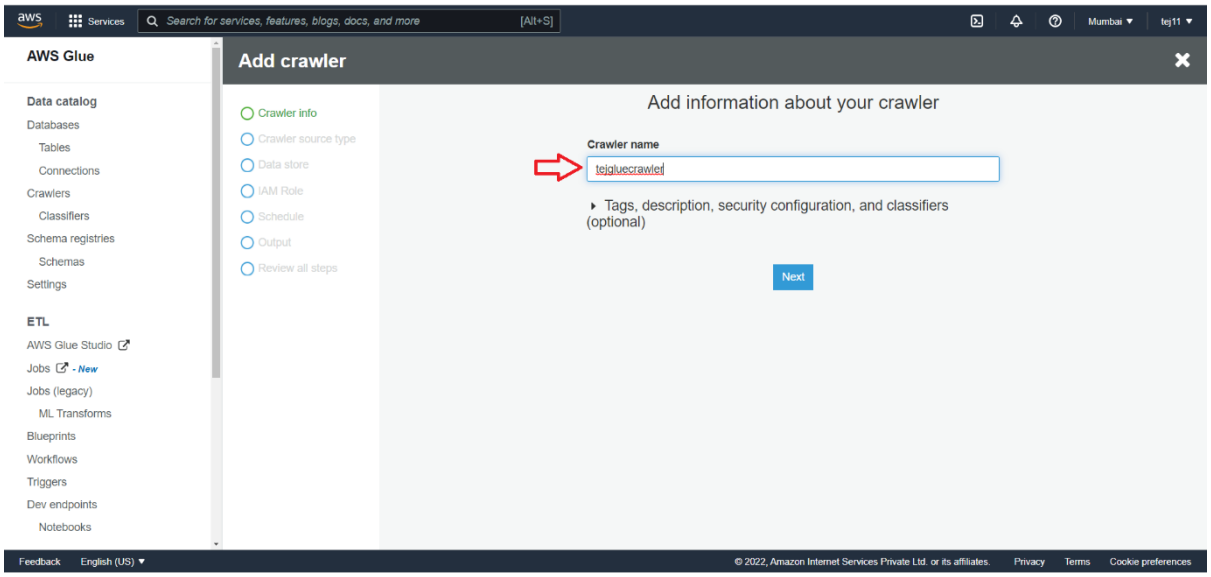


After assigning permission, time to configure and run crawler. Open the AWS Glue console. Click on the **Crawlers** menu in the left and then click on the **Add crawler** button.

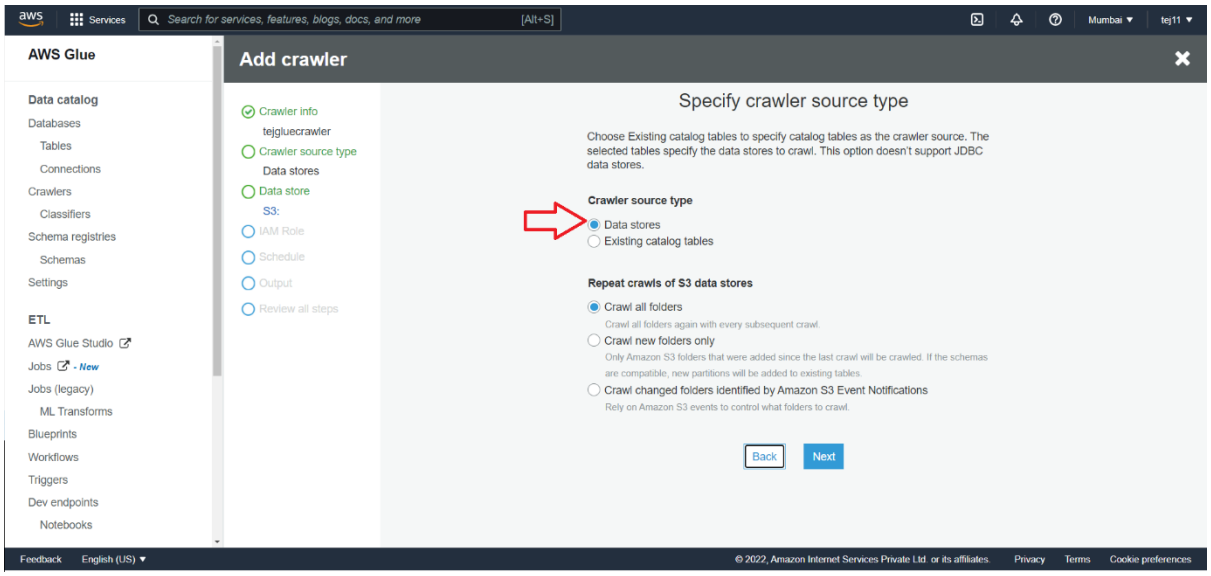

AWS Lake Formation
Grant permissions success
Successfully granted catalog permissions to ARNs: arn:aws:iam::133837890951:role/tejgluerole.
AWS Lake Formation > Permissions
Too many permissions? Filter by database or table. In the navigation page, choose Databases or Tables. Then choose a database or table, and on the Actions menu, choose View Permissions.
Data permissions (3)
Revoke
Grant
Dashboard
Data catalog
Databases
Tables
Data filters
Settings
Register and ingest
Data lake locations
Blueprints
Crawlers
Jobs
Permissions
Administrative roles and tasks
LF-Tags
LF-tag permissions
Data lake permissions
Data locations
External data filtering
Principal
Principal type
Resource type
Database
Table
Resource
Catalog
IAMAllowedPrincipals Group Database tejgluedatabase - tejgluedatabase 133837890951 -
root - Database tejgluedatabase - tejgluedatabase 133837890951 -
tejgluerole IAM role Database tejgluedatabase - tejgluedatabase 133837890951 -

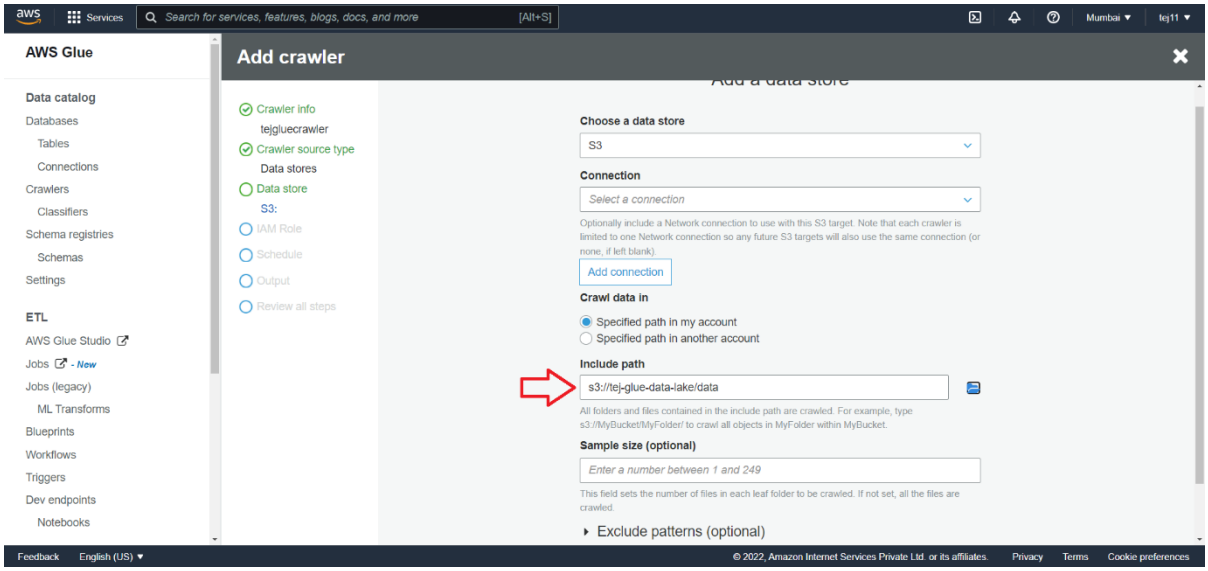

AWS Glue
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
AWS Glue
AWS Glue is a fully managed ETL (extract, transform, and load) service that makes it simple and cost-effective to categorize your data, clean it, enrich it, and move it reliably between various data stores.
Get started
Getting started guide
Build your AWS Glue Data Catalog
AWS Glue automatically stores metadata in a central data catalog. It can create table definitions for many common data stores, including, S3 buckets, web logs, and AWS databases. AWS Glue recognizes, infers, organizes, and classifies your data.
Generate and edit transformations
PySpark transformation scripts are auto generated using source and target metadata. You can store customized versions to transform your data to meet your business needs. AWS Glue provides an environment to modify your jobs.
Schedule and run your jobs
AWS Glue runs your ETL jobs in a serverless environment. You don't need to set up the infrastructure, you just use Amazon's infrastructure and pay for the resources you use. You can define triggers to run jobs based on a schedule or event. AWS Glue enables you to monitor your jobs.

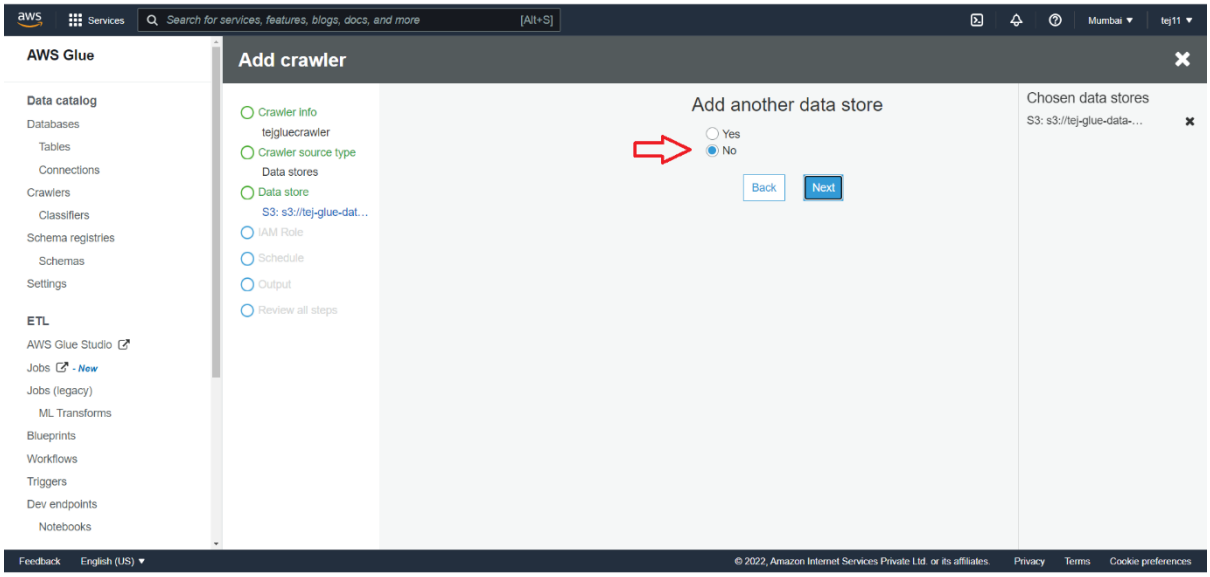

AWS Glue
Crawlers A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.
Add crawler
Run crawler
Action
Filter by tags and attributes
Showing: 0 - 0
User preferences
Name
Schedule
Status
Logs
Last runtime
Median runtime
Tables updated
Tables added
You don't have any crawlers yet.
Add crawler
Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks

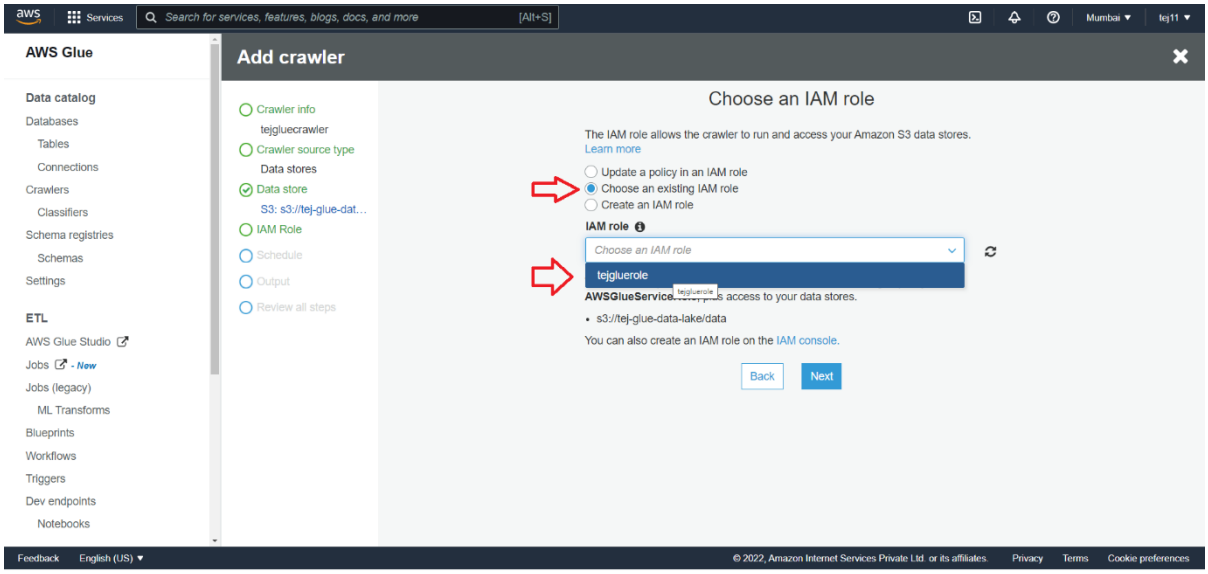


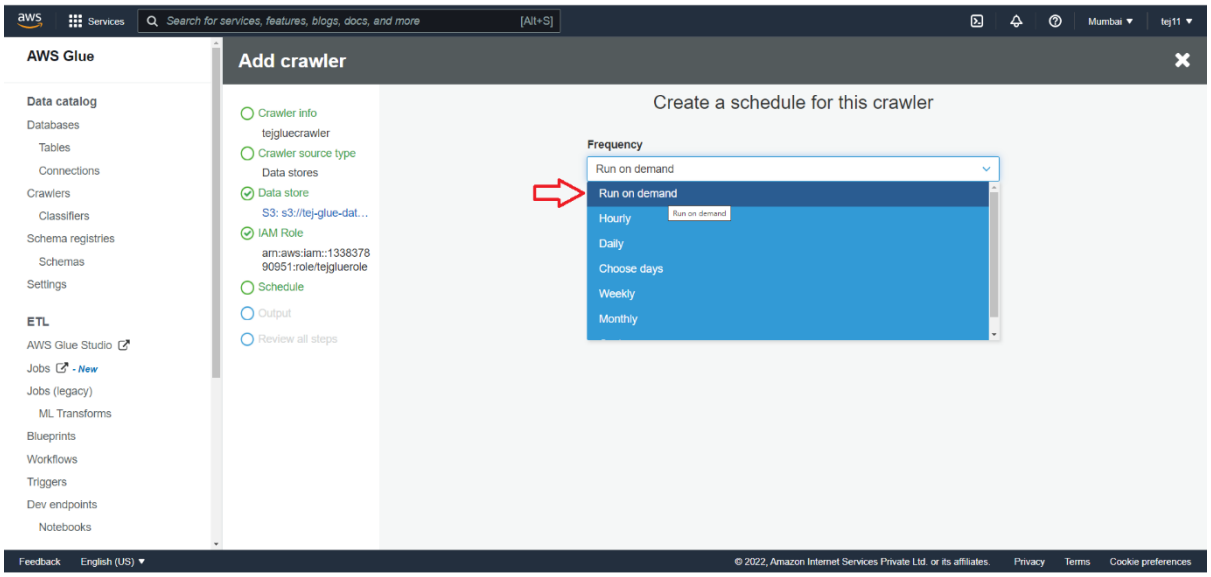
On the next screen, select **Data stores** as the Crawler source type and click **Next**.

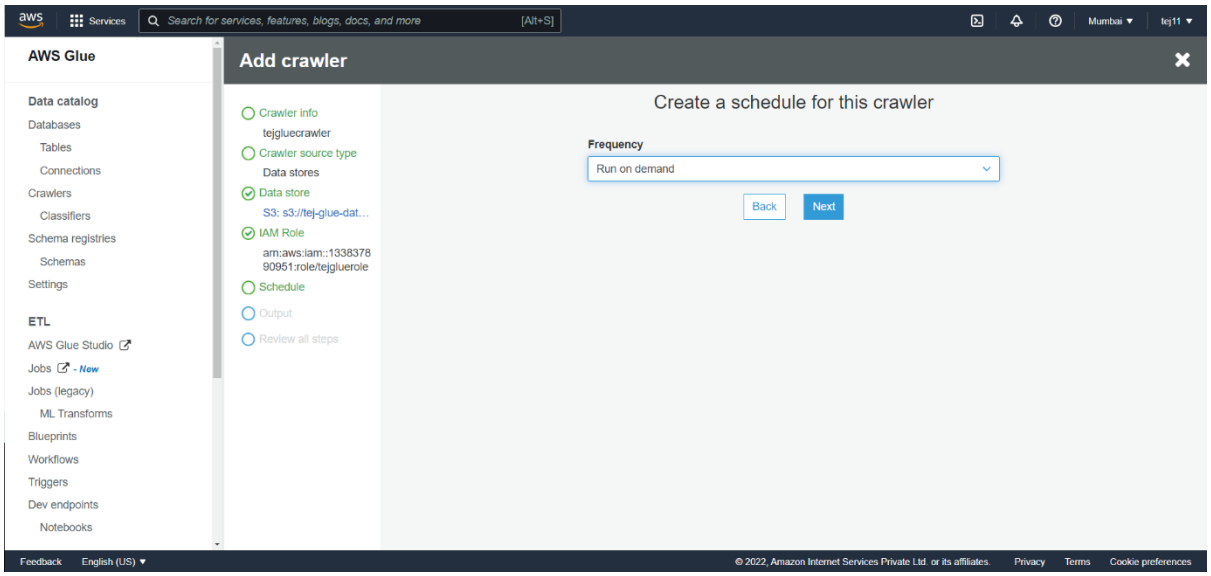


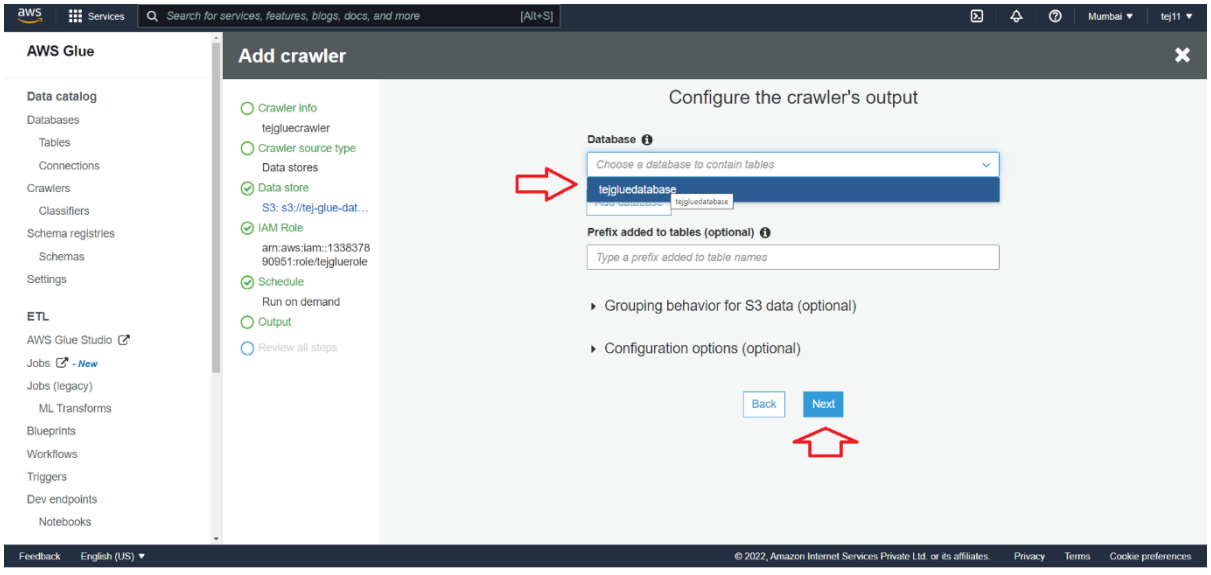

Add crawler
Crawler info
tejgluecrawler
Crawler source type
Data stores
Data store
S3:
IAM Role
Schedule
Output
Review all steps
Choose a data store
S3
Connection
Select a connection
Add connection
Crawl data in
Specified path in my account
Specified path in another account
Include path
s3://tej-glue-data-lake/data
Sample size (optional)
Enter a number between 1 and 249
Exclude patterns (optional)

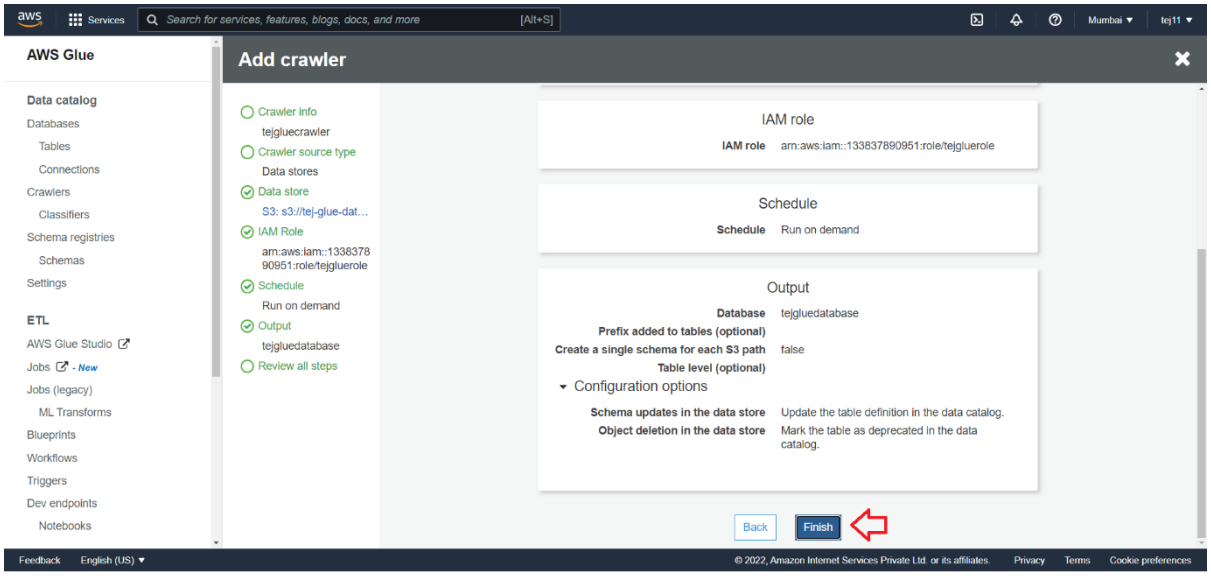

Add crawler
Add another data store
Yes
No
Back
Next
Chosen data stores
S3: s3://tej-glue-data-...


Add crawler
Choose an IAM role
The IAM role allows the crawler to run and access your Amazon S3 data stores.
Learn more
Update a policy in an IAM role
Choose an existing IAM role
Create an IAM role
IAM role
Choose an IAM role
tejgluerole
s3://tej-glue-data-lake/data
You can also create an IAM role on the IAM console.
Back
Next

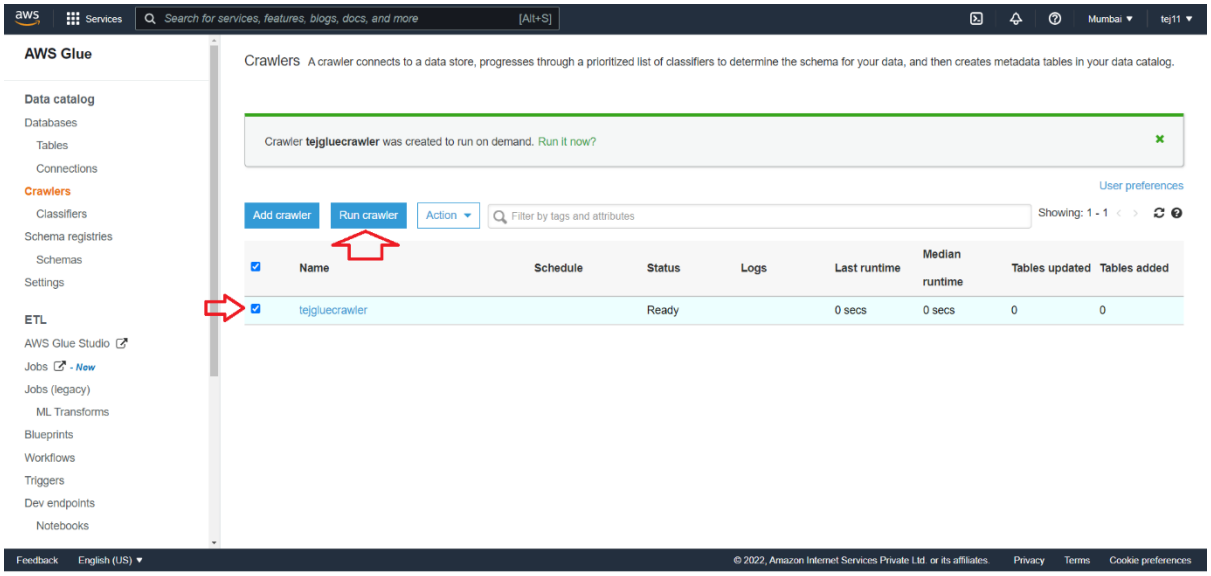

Add crawler
Create a schedule for this crawler
Frequency
Run on demand
Hourly
Daily
Choose days
Weekly
Monthly
Crawler info
tejgluecrawler
Crawler source type
Data stores
Data store
IAM Role
Schedule
Output
Review all steps

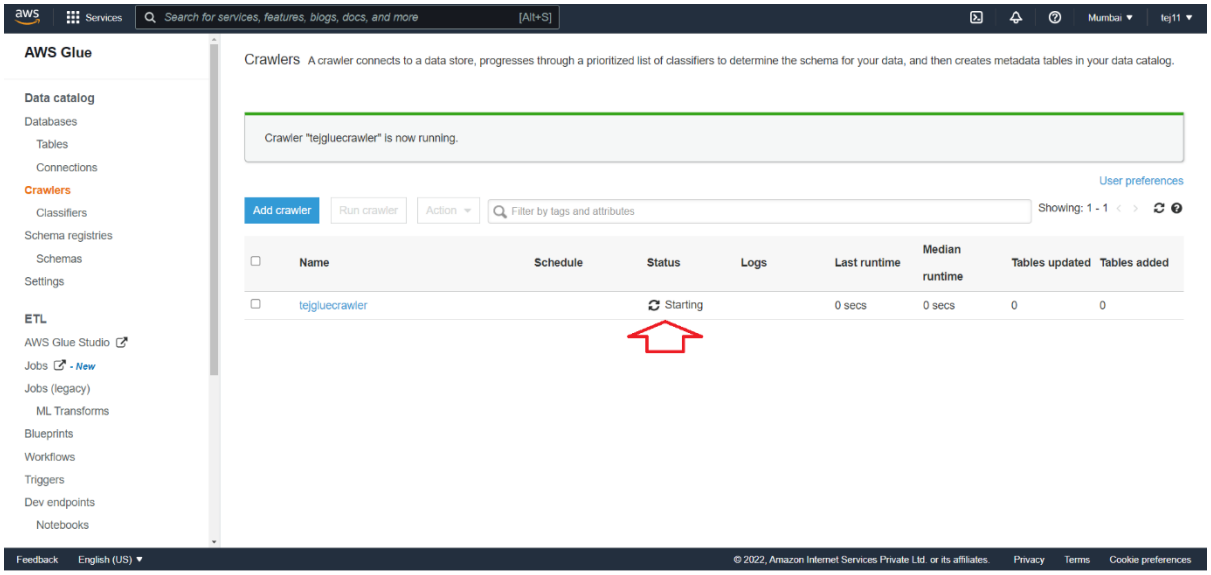

Add crawler
Create a schedule for this crawler
Frequency
Run on demand
Back
Next


Add crawler
Configure the crawler's output
Database
Choose a database to contain tables
tejgluedatabase
Prefix added to tables (optional)
Type a prefix added to table names
Grouping behavior for S3 data (optional)
Configuration options (optional)
Back
Next

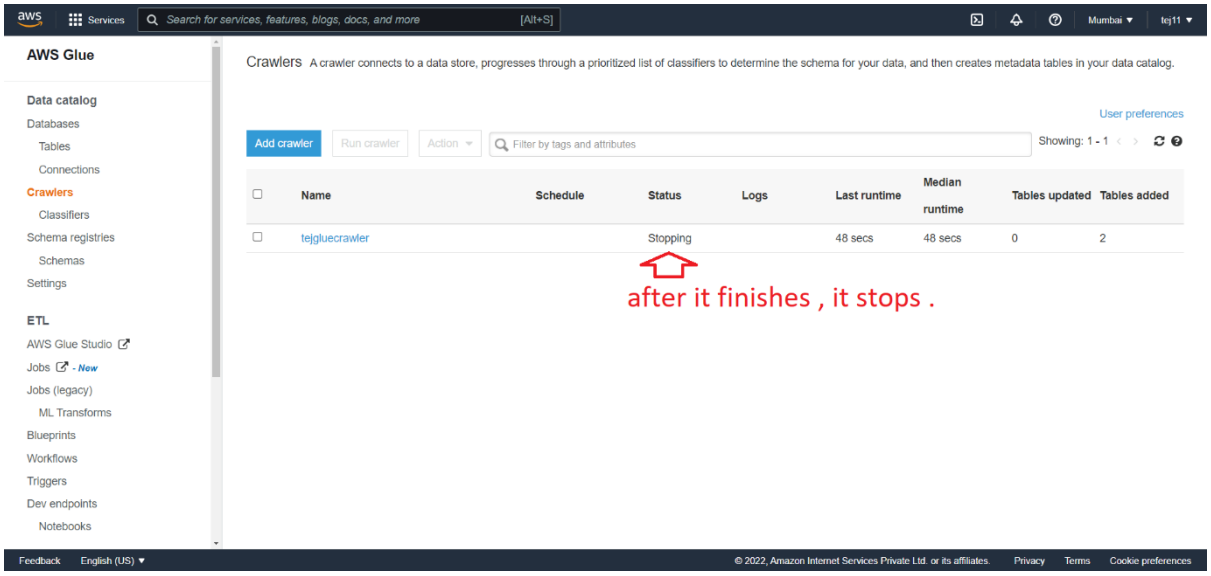


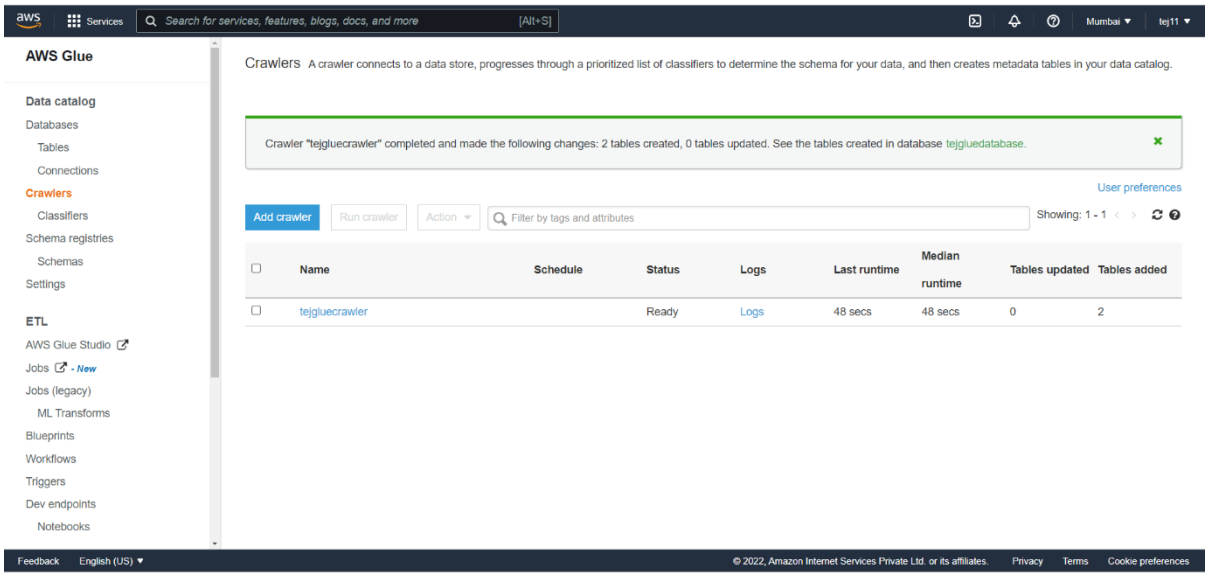
The crawler is created in no time. Click on **Run it now?** link. Alternatively, you can select the crawler and run the crawler from the **Action** menu.





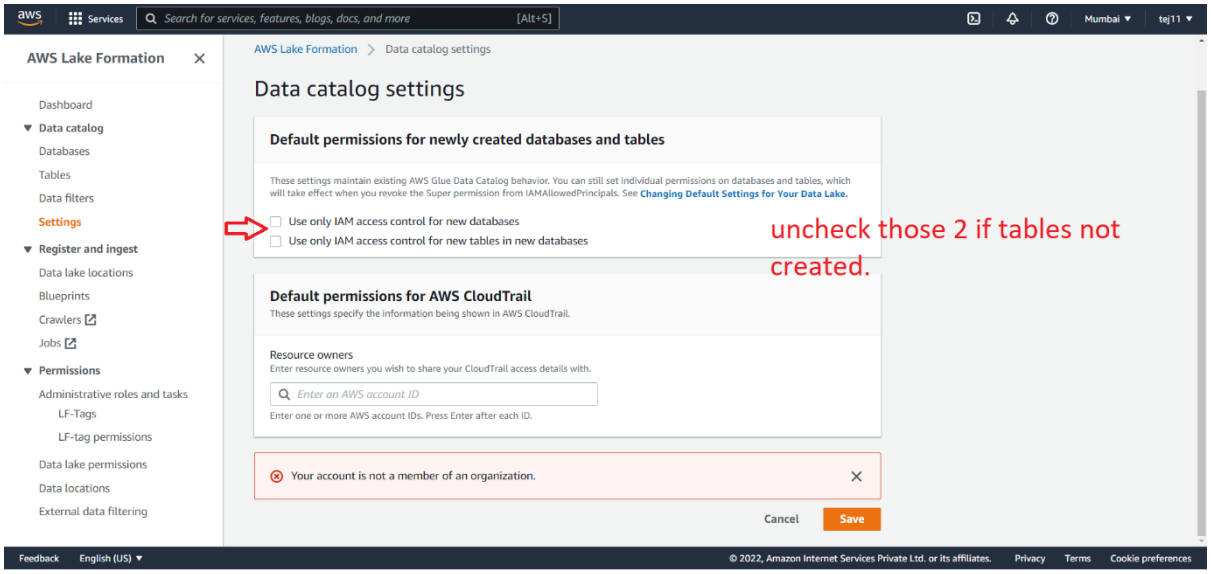
It may take up to a couple of minutes for the crawler to finish crawling the bucket. You should be able to see a success message that there are two tables created by the crawler in **tejgluedatabase** database. These are two tables for the **sales** and **customers** data in Amazon S3.

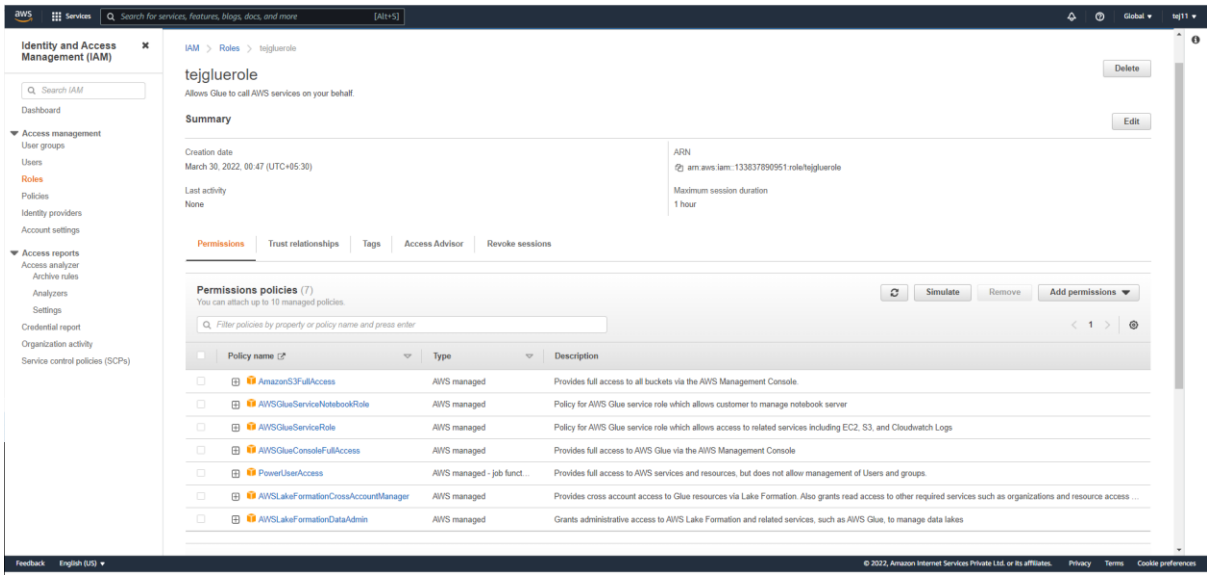


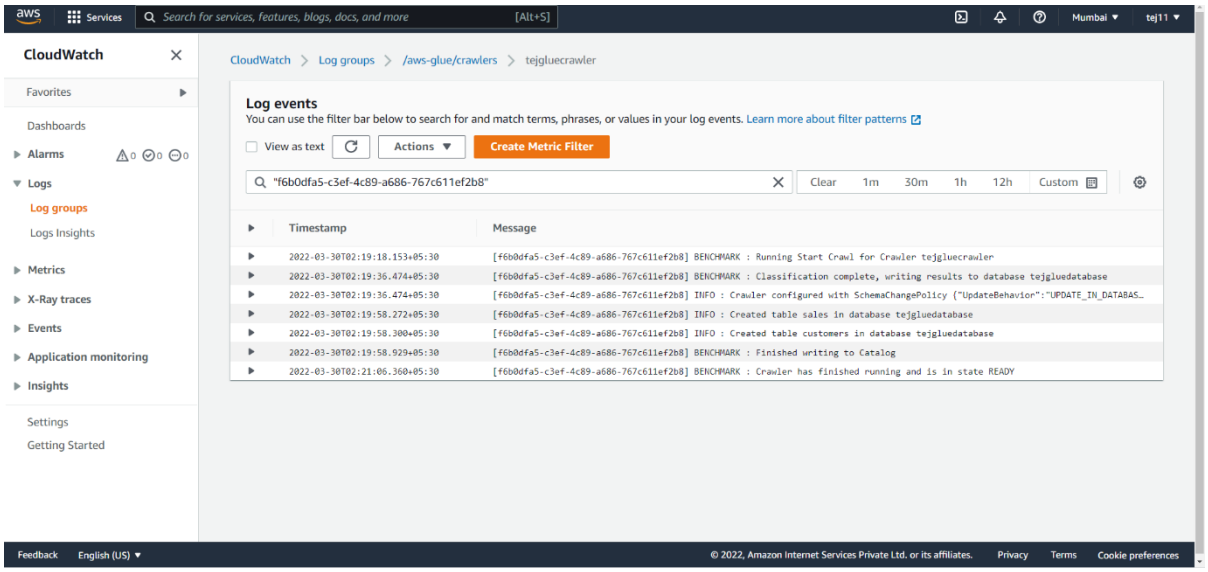


It may take up to a couple of minutes for the crawler to finish crawling the bucket. You should be able to see a success message that there are two tables created by the crawler in **tejgluedatabase** database. These are two tables for the **sales** and **customers** data in Amazon S3.

Go back to the AWS Lake Formation console, click on the **Tables** menu in the left. You can see two tables **sales** and **customers** created.







Role Permission to the Catalog

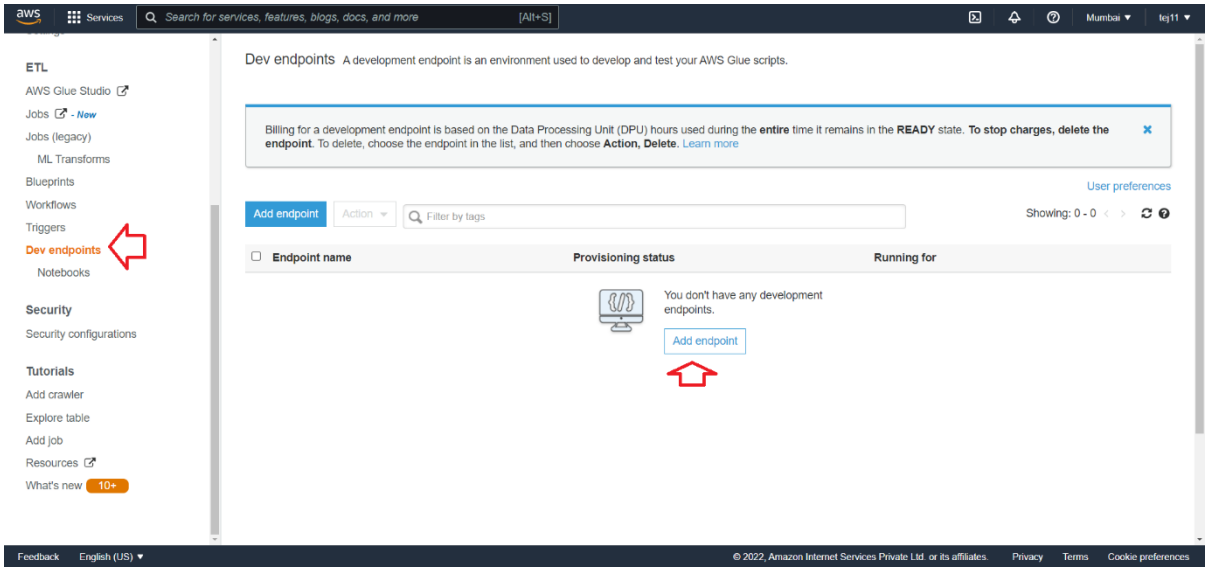
, but datalake is expericening a bug , which will be rectified soon . if you are not able to see the tables .

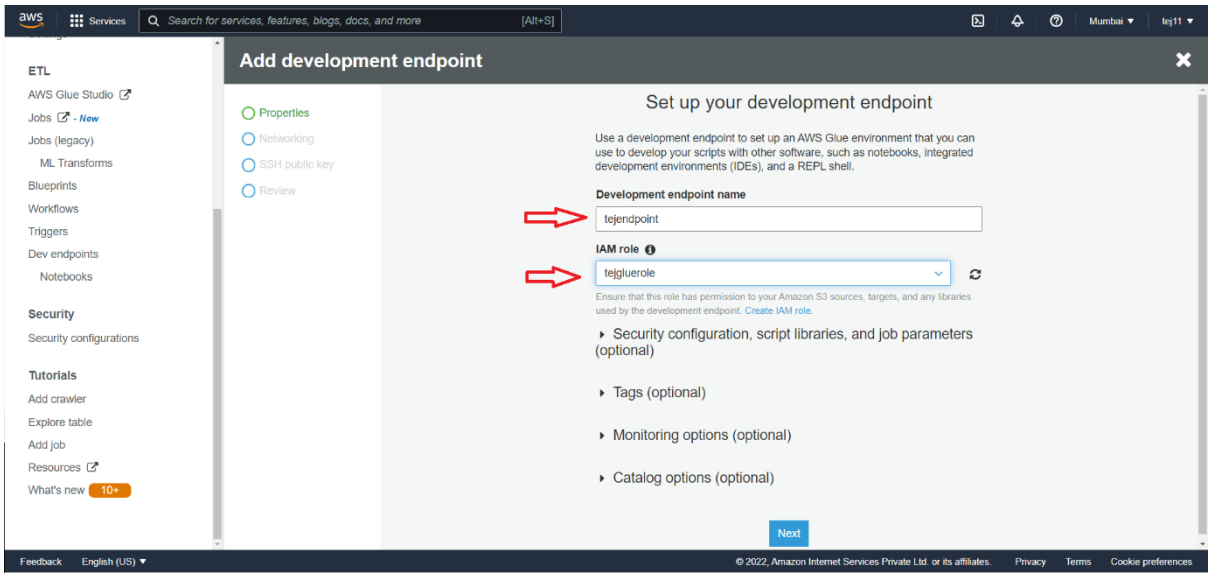
6

Create Developer Endpoint

Developer endpoint provides development environment to create Glue Job using languages and frameworks like PySpark. In this task, you create a developer endpoint which you will use to code with PySpark.

1. Goto the AWS Glue console, click on the **Dev endpoints** option in the left menu and then click on the **Add endpoint** button.

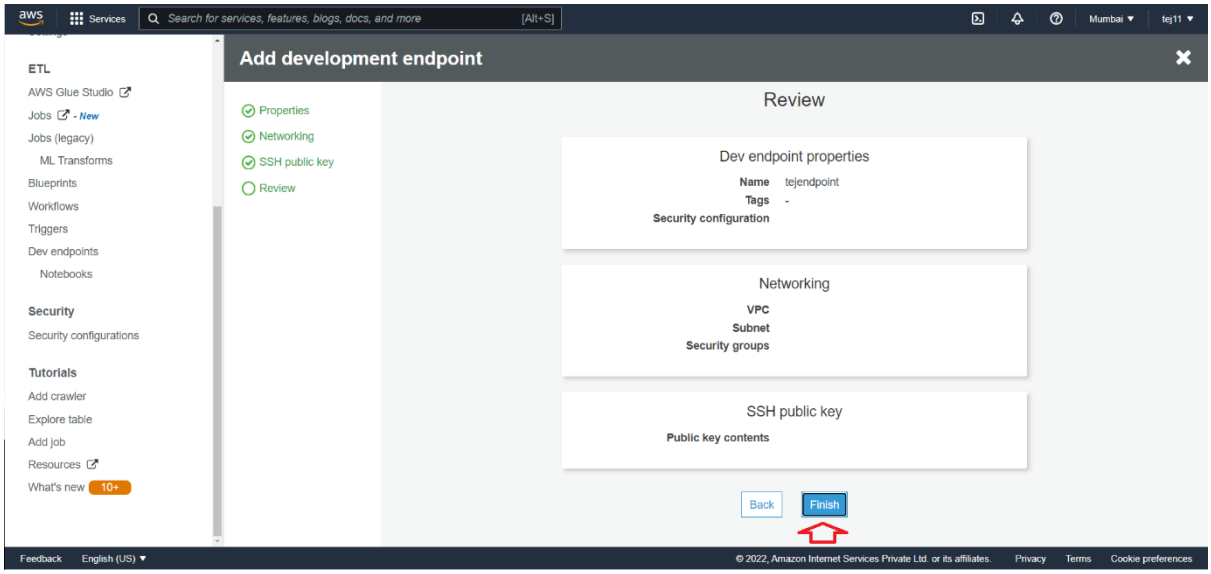




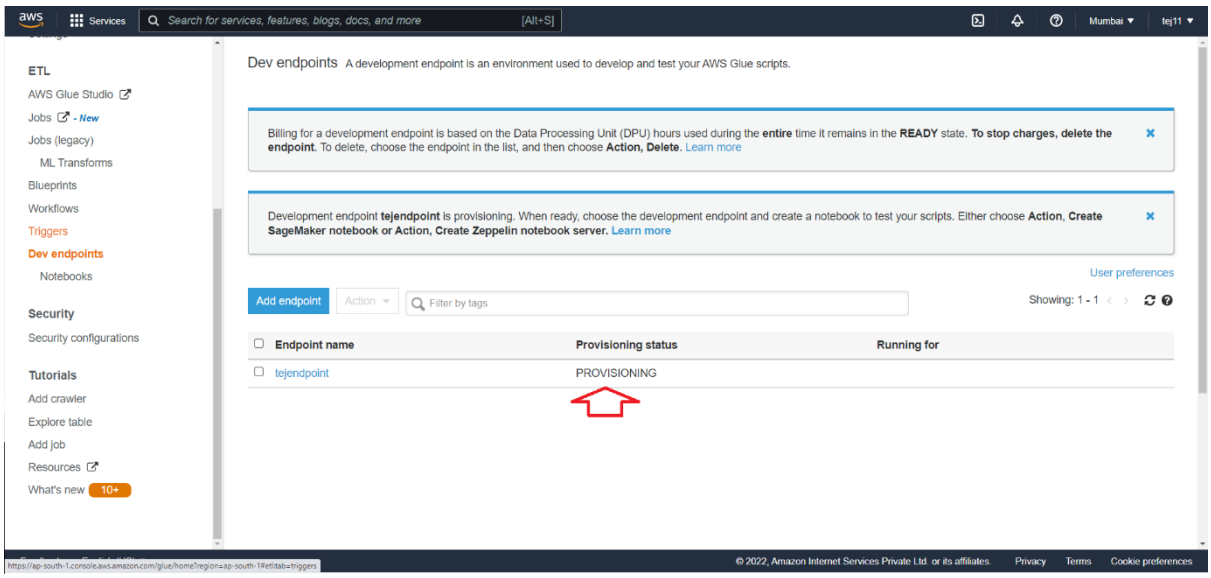
On the next screen, select **Skip networking information** as the option and click on the **Next** button.

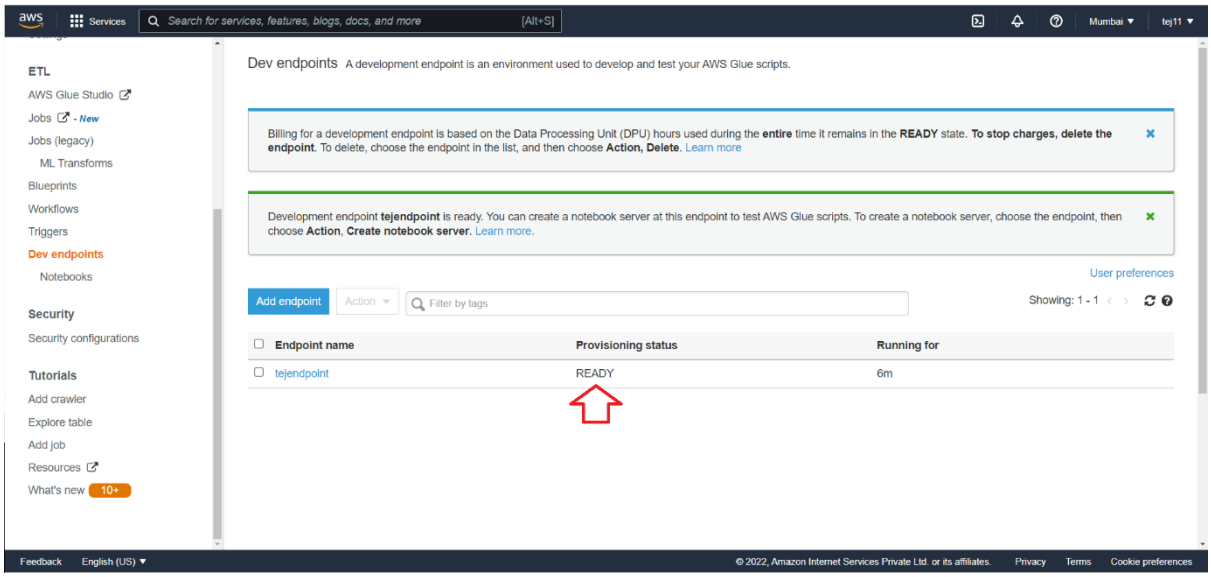
On the next **Add an SSH public key (Optional)** screen, click on the **Next** button.

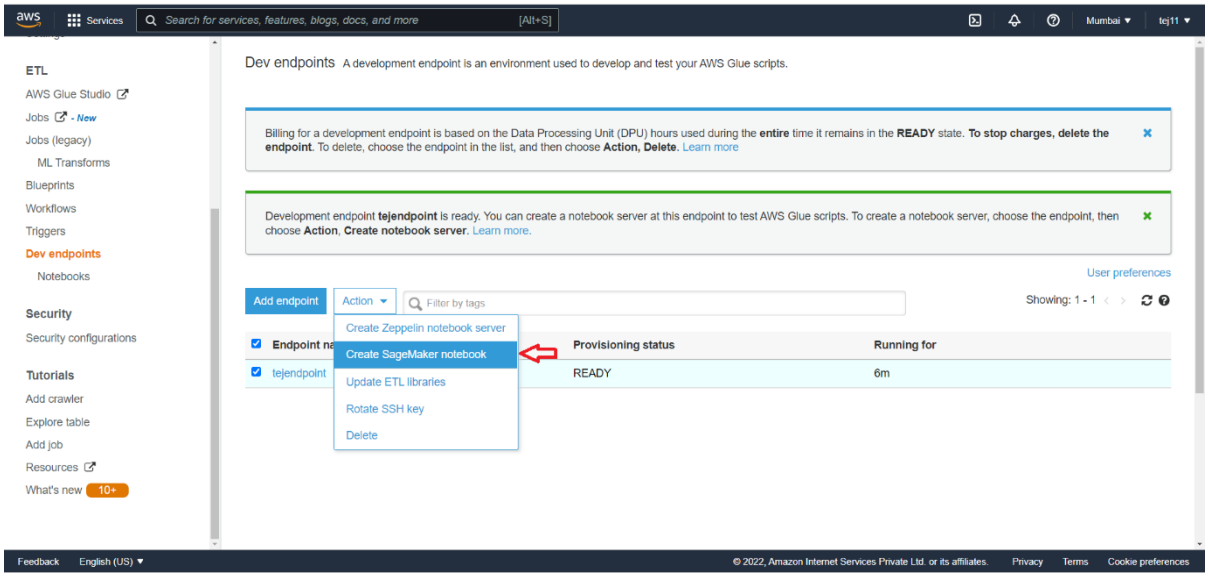
On the next **Review** screen, click on the **Finish** button. The endpoint creation will start.

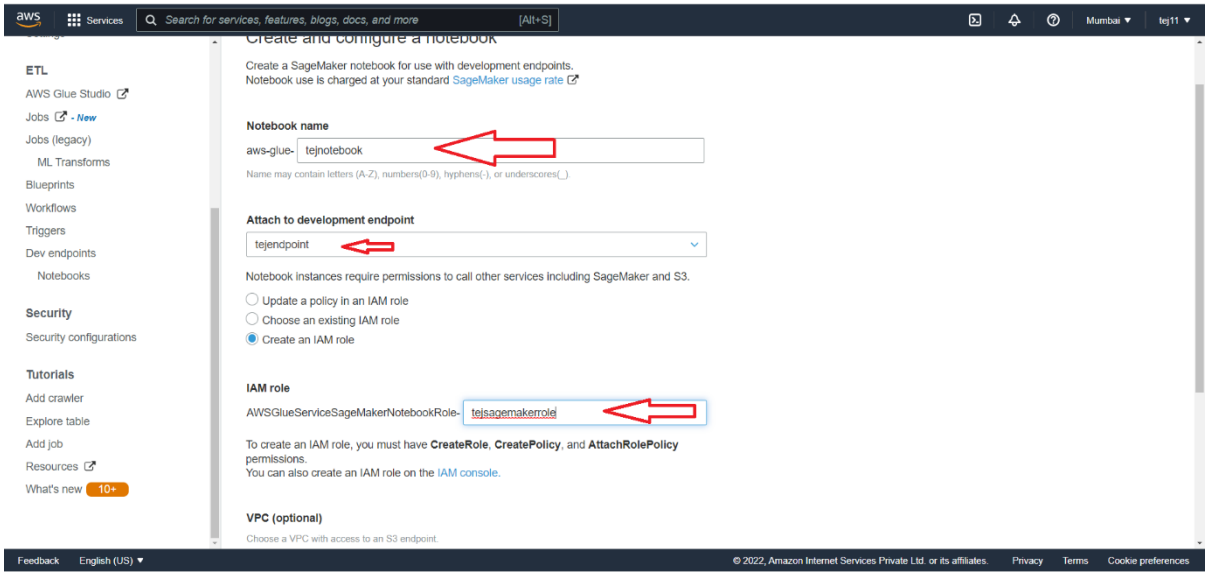


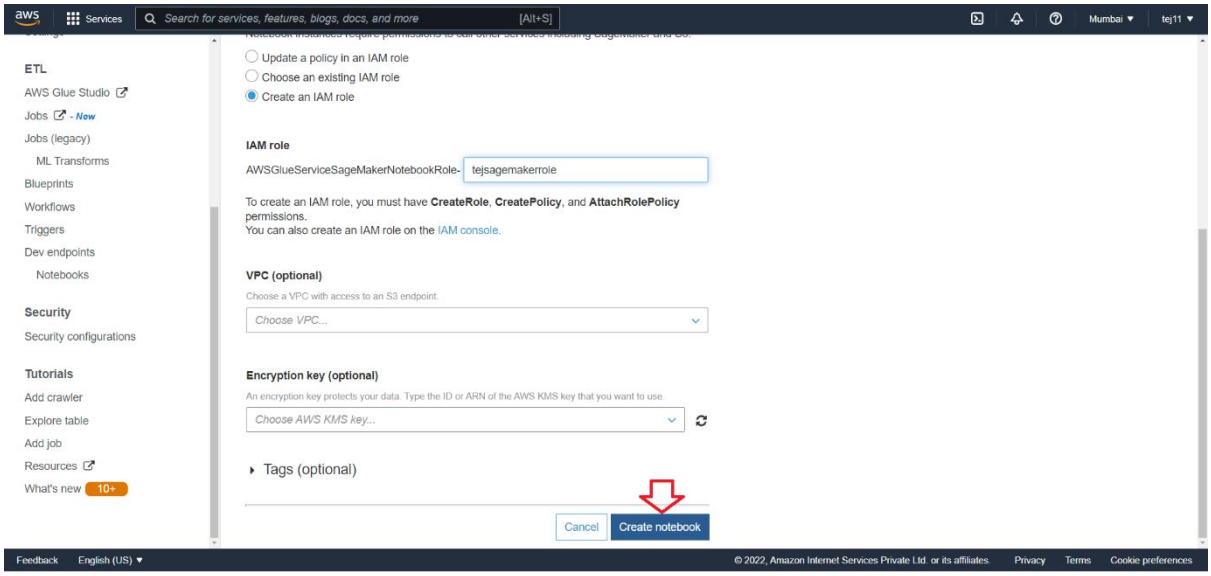
It will take some 8-10 mins for the developer endpoint to be ready. Wait till the status changes to **READY**.

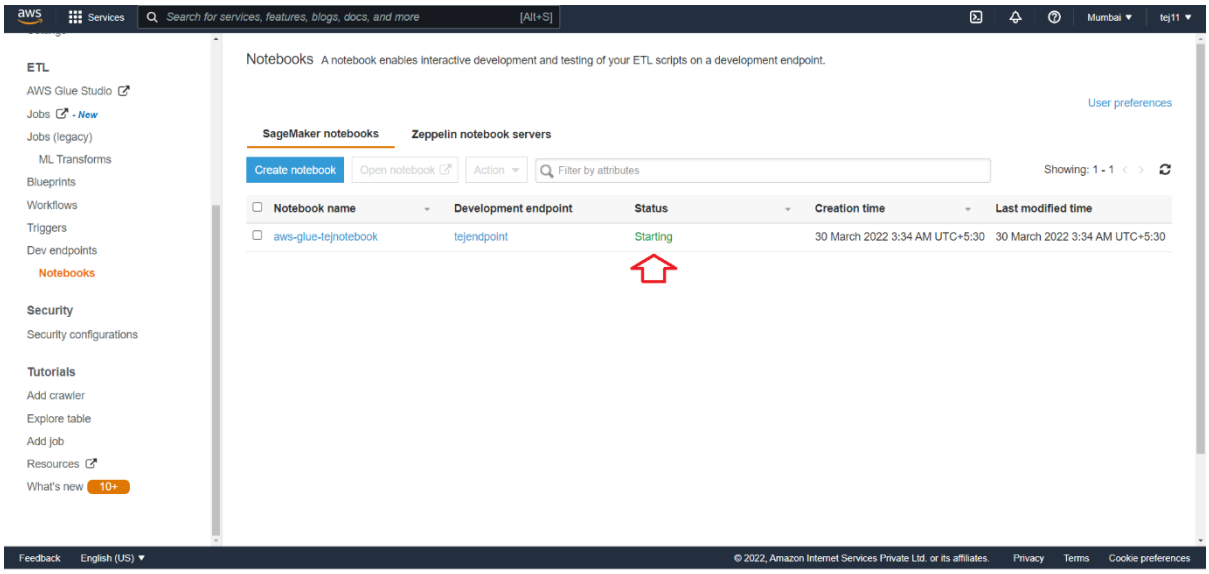



Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai
tej11
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Resources
What's new 10+
Dev endpoints A development endpoint is an environment used to develop and test your AWS Glue scripts.
Billing for a development endpoint is based on the Data Processing Unit (DPU) hours used during the entire time it remains in the READY state. To stop charges, delete the endpoint. To delete, choose the endpoint in the list, and then choose Action, Delete. Learn more
Development endpoint tejendpoint is ready. You can create a notebook server at this endpoint to test AWS Glue scripts. To create a notebook server, choose the endpoint, then choose Action, Create notebook server. Learn more.
User preferences
Add endpoint
Action
Filter by tags
Showing: 1 - 1
Endpoint name
Provisioning status
Running for
tejendpoint
READY
6m
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai
tej11
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Resources
What's new 10+
Dev endpoints A development endpoint is an environment used to develop and test your AWS Glue scripts.
Billing for a development endpoint is based on the Data Processing Unit (DPU) hours used during the entire time it remains in the READY state. To stop charges, delete the endpoint. To delete, choose the endpoint in the list, and then choose Action, Delete. Learn more
Development endpoint tejendpoint is ready. You can create a notebook server at this endpoint to test AWS Glue scripts. To create a notebook server, choose the endpoint, then choose Action, Create notebook server. Learn more.
User preferences
Add endpoint
Action
Filter by tags
Showing: 1 - 1
Create Zeppelin notebook server
Create SageMaker notebook
Update ETL libraries
Rotate SSH key
Delete
Endpoint name
Provisioning status
Running for
tejendpoint
READY
6m
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai
tej11
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Add job
Resources
What's new 10+
Create a SageMaker notebook for use with development endpoints.
Notebook use is charged at your standard SageMaker usage rate
Notebook name
aws-glue-
tejnotebook
Name may contain letters (A-Z), numbers(0-9), hyphens(-), or underscores(_).
Attach to development endpoint
tejendpoint
Notebook instances require permissions to call other services including SageMaker and S3.
Update a policy in an IAM role
Choose an existing IAM role
Create an IAM role
IAM role
AWSGlueServiceSageMakerNotebookRole-
tejsagemakerrole
To create an IAM role, you must have CreateRole, CreatePolicy, and AttachRolePolicy permissions.
You can also create an IAM role on the IAM console.
VPC (optional)
Choose a VPC with access to an S3 endpoint.
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences





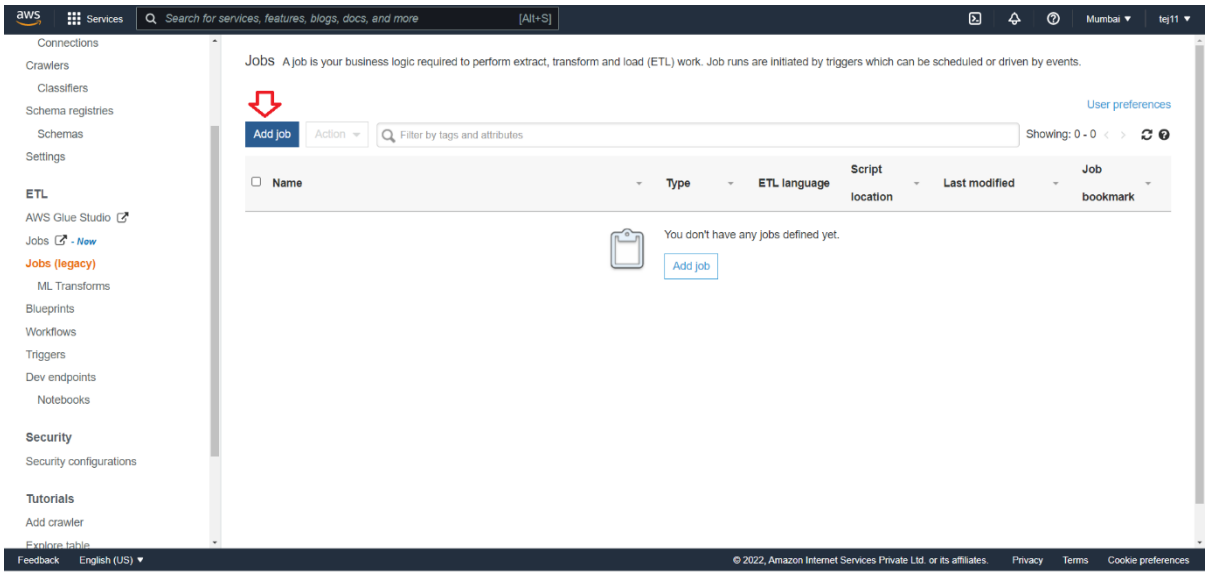
The development environment is ready. Let's do PySpark programming in notebook which then you use later to create a Glue job.

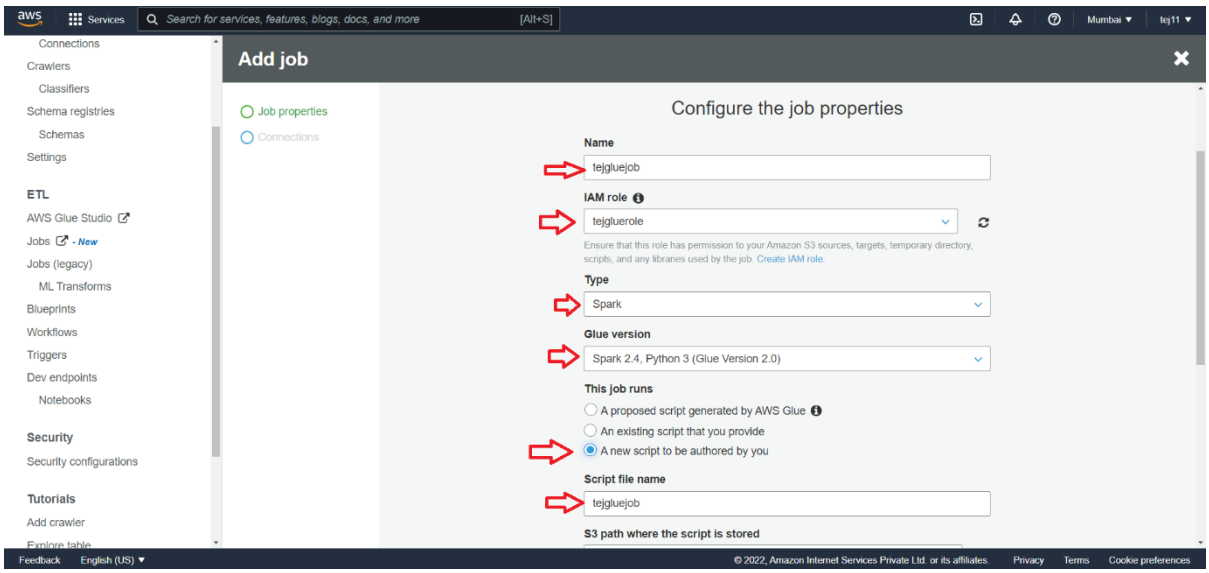
Glue notebook job

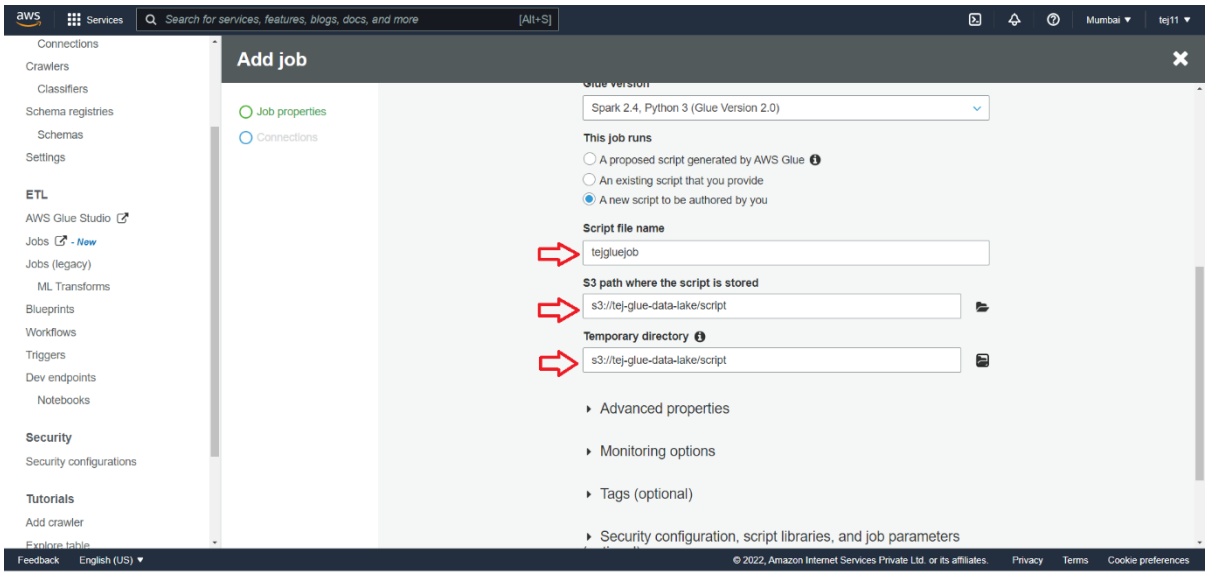
```
import sys

from awsglue.transforms import *

from awsglue.utils import getResolvedOptions

from pyspark.context import SparkContext

from awsglue.context import GlueContext
```

```
from awsglue.job import Job


glueContext = GlueContext(SparkContext.getOrCreate())

salesDF = glueContext.create_dynamic_frame.from_catalog(
        database="tejgluedb",
        table_name="sales")
customerDF = glueContext.create_dynamic_frame.from_catalog(
        database="tejgluedb",
        table_name="customers")

salesDF.printSchema()
customerDF.printSchema()

customersalesDF=Join.apply(salesDF, customerDF, 'customerid', 'customerid')
customersalesDF.printSchema()

customersalesDF = customersalesDF.drop_fields(['`.customerid`'])
customersalesDF.printSchema()

glueContext.write_dynamic_frame.from_options(customersalesDF, connection_type =
"s3", connection_options = {"path": "s3://tej-glue-data-lake/data/customersales"},
format = "json")
```

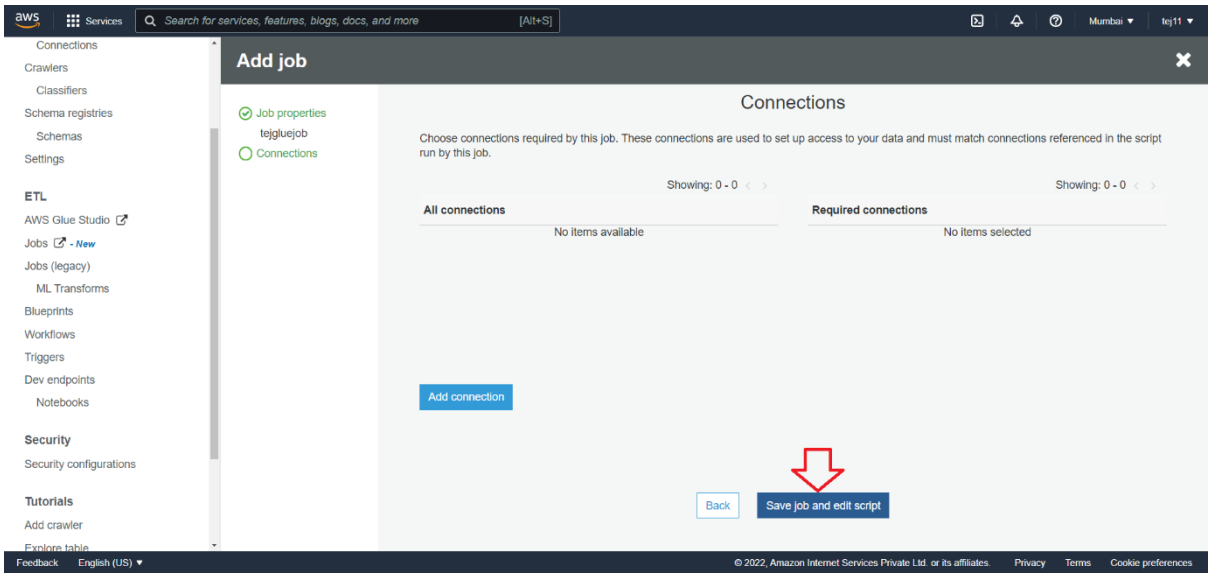
Glue job


Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai
tej11
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings
ETL
AWS Glue Studio
Jobs - New
Jobs (legacy)
ML Transforms
Blueprints
Workflows
Triggers
Dev endpoints
Notebooks
Security
Security configurations
Tutorials
Add crawler
Explore table
Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.
User preferences
Add job
Action
Filter by tags and attributes
Showing: 0 - 0
Name
Type
ETL language
Script location
Last modified
Job bookmark
You don't have any jobs defined yet.
Add job
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


Add job
Job properties
Connections
Configure the job properties
Name
tejgluejob
IAM role
tejgluerole
Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. Create IAM role.
Type
Spark
Glue version
Spark 2.4, Python 3 (Glue Version 2.0)
This job runs
A proposed script generated by AWS Glue
An existing script that you provide
A new script to be authored by you
Script file name
tejgluejob
S3 path where the script is stored


Add job
Job properties
Connections
Spark 2.4, Python 3 (Glue Version 2.0)
This job runs
A proposed script generated by AWS Glue
An existing script that you provide
A new script to be authored by you
Script file name
tejgluejob
S3 path where the script is stored
s3://tej-glue-data-lake/script
Temporary directory
s3://tej-glue-data-lake/script
Advanced properties
Monitoring options
Tags (optional)
Security configuration, script libraries, and job parameters

On the next **Connections** page, click on the **Save job and edit script** button. You come to the page where you can write your code. It is the same code you wrote in the previous task in the notebook.



In the editor

```
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


glueContext = GlueContext(SparkContext.getOrCreate())


salesDF = glueContext.create_dynamic_frame.from_catalog(

        database="tejgluedb",

        table_name="sales")

customerDF = glueContext.create_dynamic_frame.from_catalog(
```

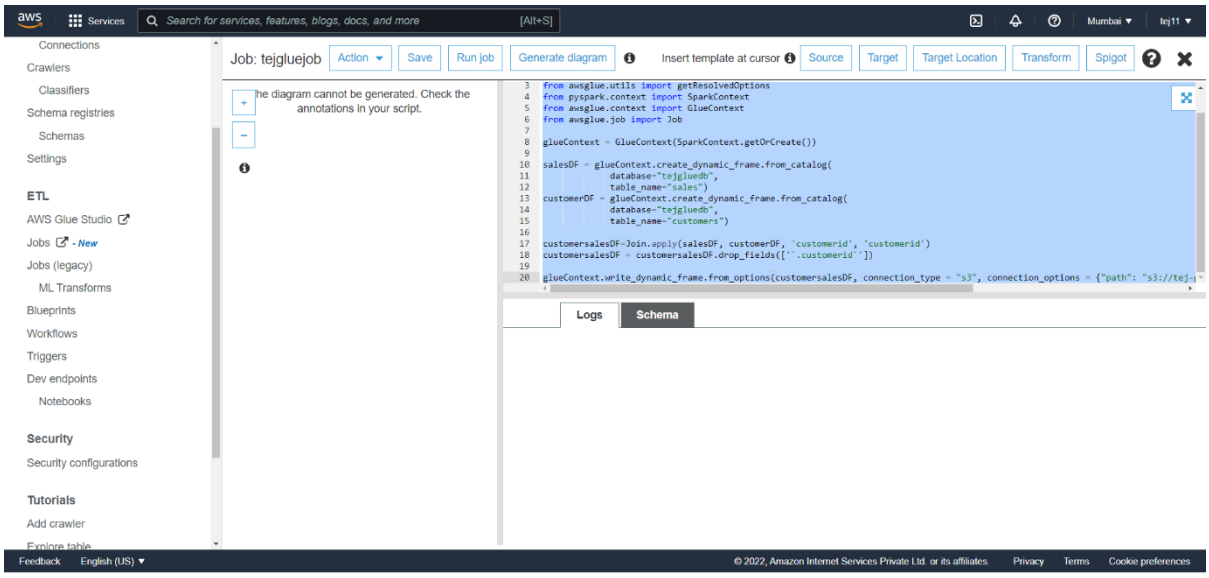
database="tejgluedb",

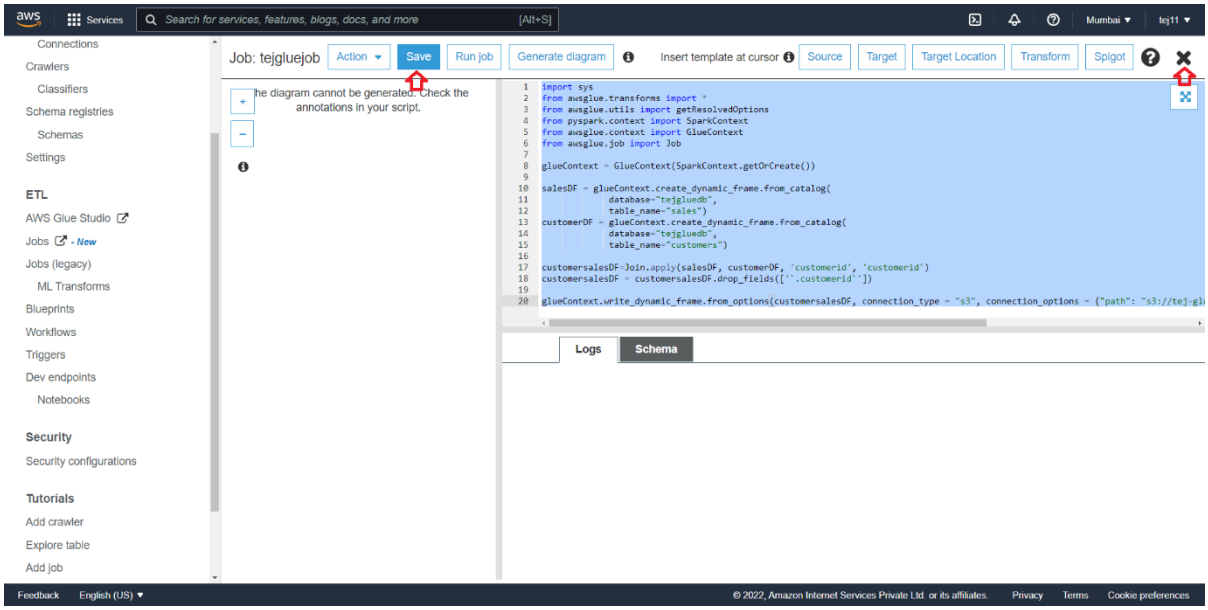
table_name="customers")

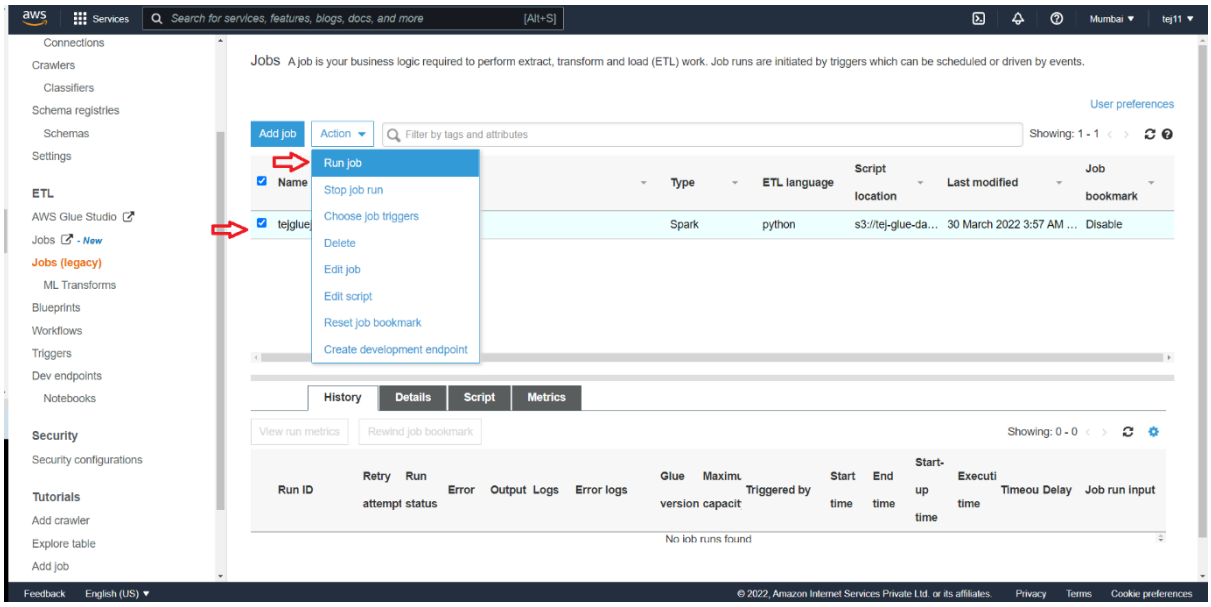
customersalesDF=Join.apply(salesDF, customerDF, 'customerid', 'customerid')

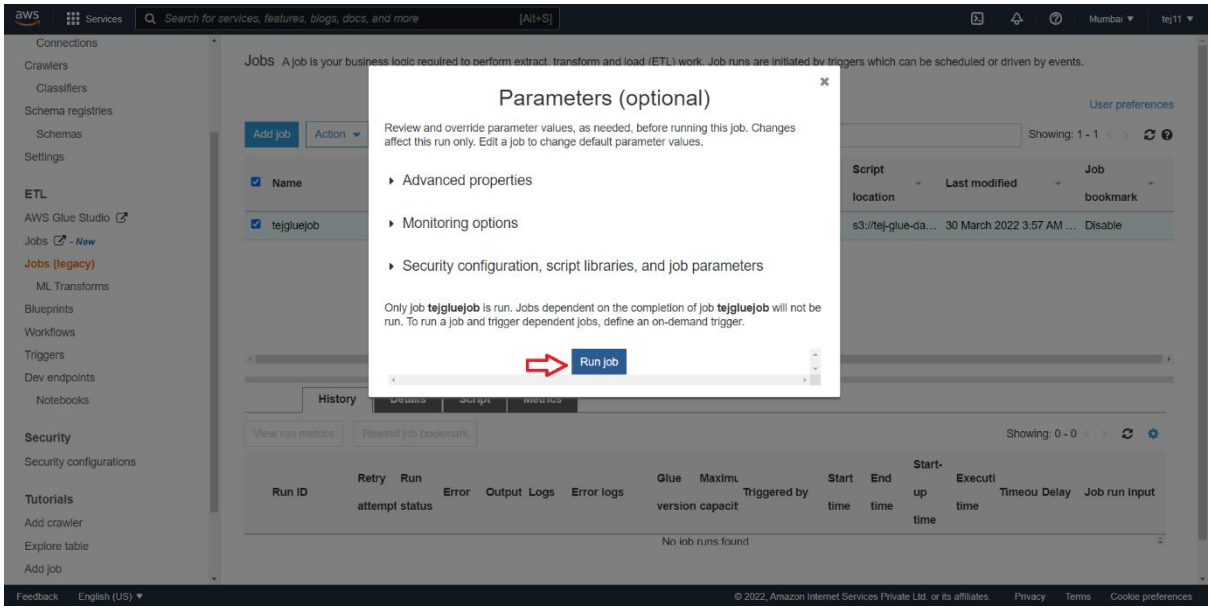
customersalesDF = customersalesDF.drop_fields(['`.customerid`'])

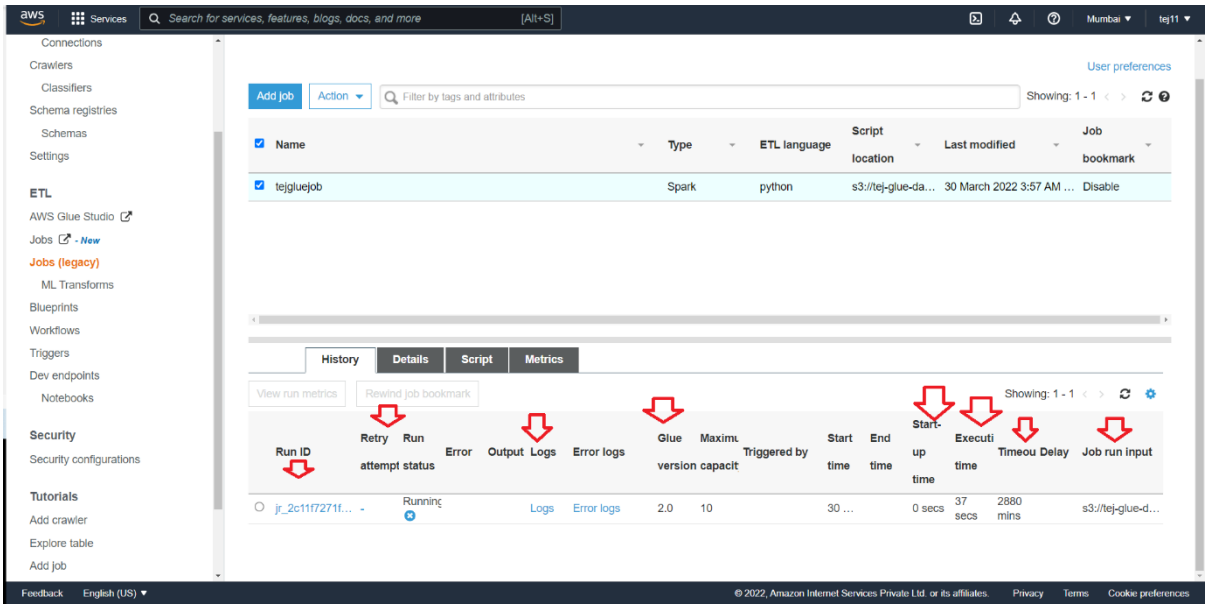
glueContext.write_dynamic_frame.from_options(customersalesDF, connection_type = "s3", connection_options = {"path": "s3://tej-glue-data-lake/data/customersalesjobsscript"}, format = "json")

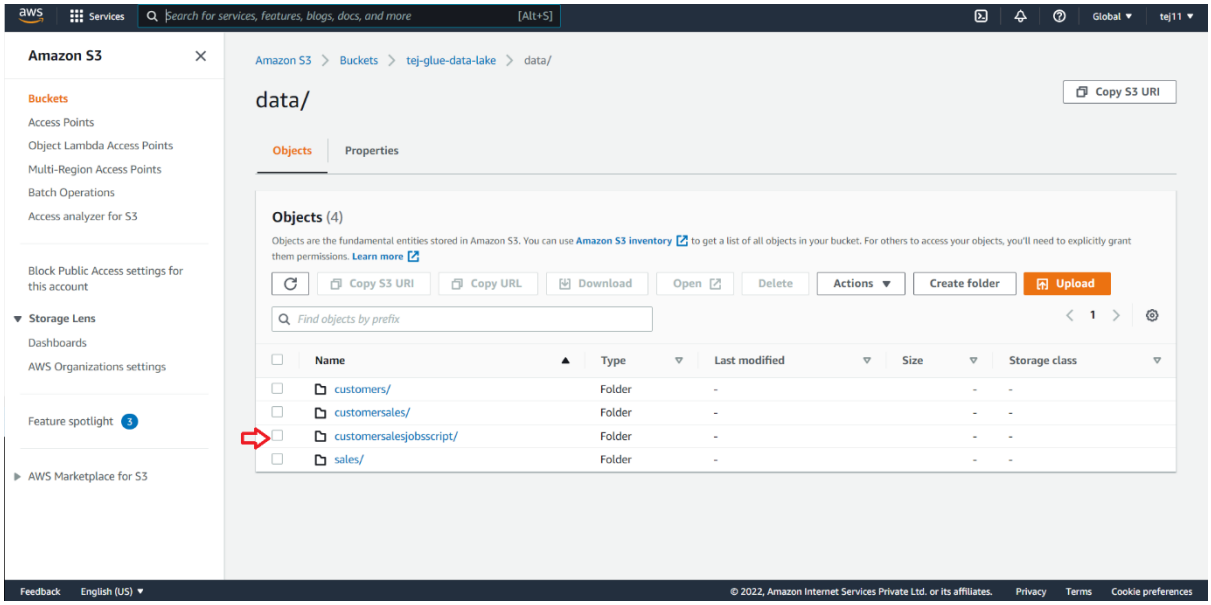



Job: tejgluejob
Action
Save
Run job
Generate diagram
Insert template at cursor
Source
Target
Target Location
Transform
Spigot
The diagram cannot be generated. Check the annotations in your script.
Logs
Schema


Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.
User preferences
Add job
Action
Run job
Stop job run
Choose job triggers
Delete
Edit job
Edit script
Reset job bookmark
Create development endpoint
History
Details
Script
Metrics
No job runs found


Parameters (optional)
Review and override parameter values, as needed, before running this job. Changes affect this run only. Edit a job to change default parameter values.
Advanced properties
Monitoring options
Security configuration, script libraries, and job parameters
Only job tejgluejob is run. Jobs dependent on the completion of job tejgluejob will not be run. To run a job and trigger dependent jobs, define an on-demand trigger.
Run job


History
Details
Script
Metrics
Run ID
Retry attempt
Run status
Error
Output
Logs
Error logs
Glue version
Maximum capacity
Triggered by
Start time
End time
Start-up time
Execution time
Timeout
Delay
Job run input
No job runs found


Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
tej11
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
3
AWS Marketplace for S3
Amazon S3 > Buckets > tej-glue-data-lake > data/
data/
Copy S3 URI
Objects
Properties
Objects (4)
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
1
Name
Type
Last modified
Size
Storage class
customers/
Folder
customersales/
Folder
customersalesjobsscript/
Folder
sales/
Folder
Feedback
English (US)
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences