



Data Engineering Lab

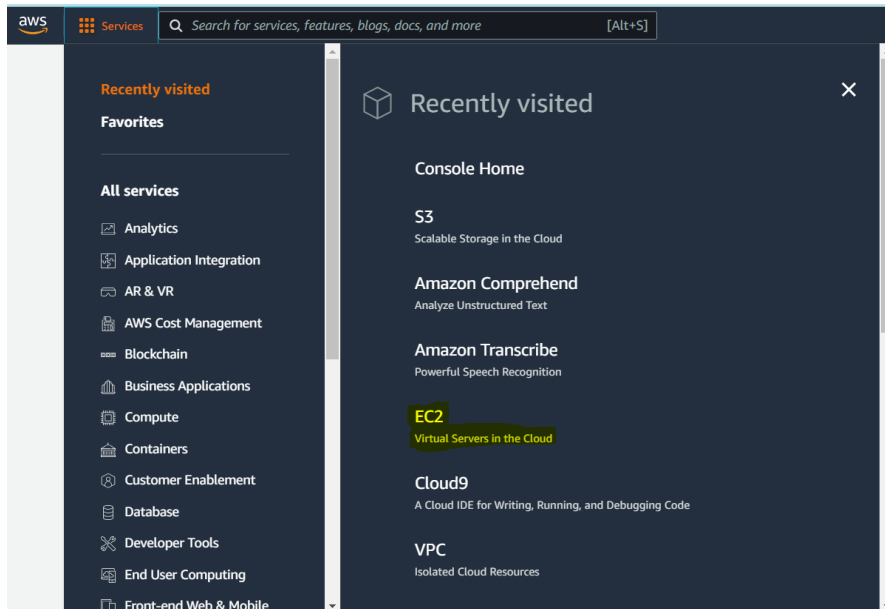
Gain hands-on experience on Scalable Data
Engineering on AWS Cloud!

Kinesis to S3

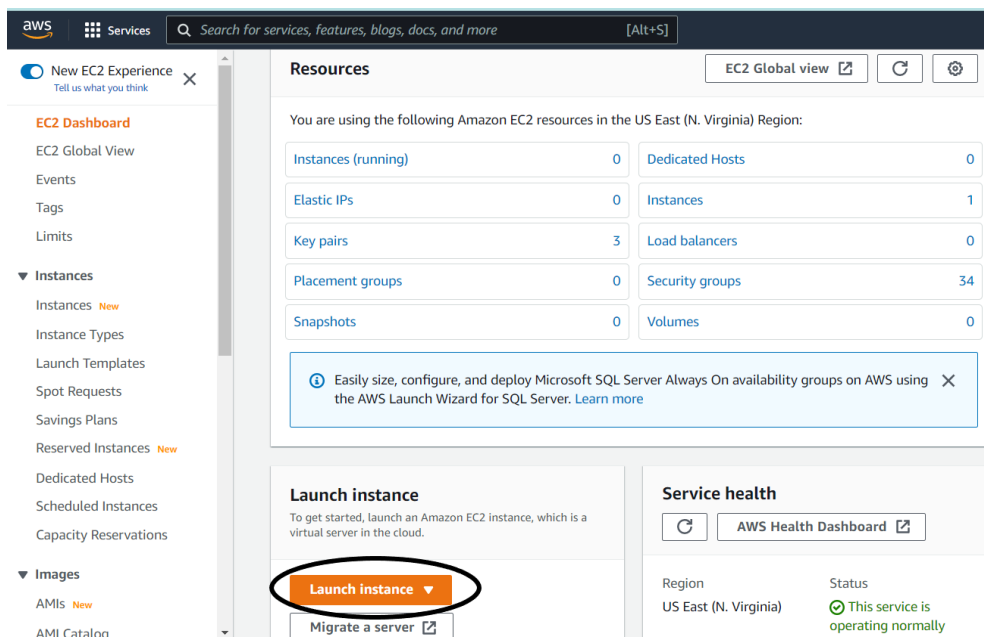
Step 1: Launch EC2 Instance

1. Go to the AWS console and click on “Services”.

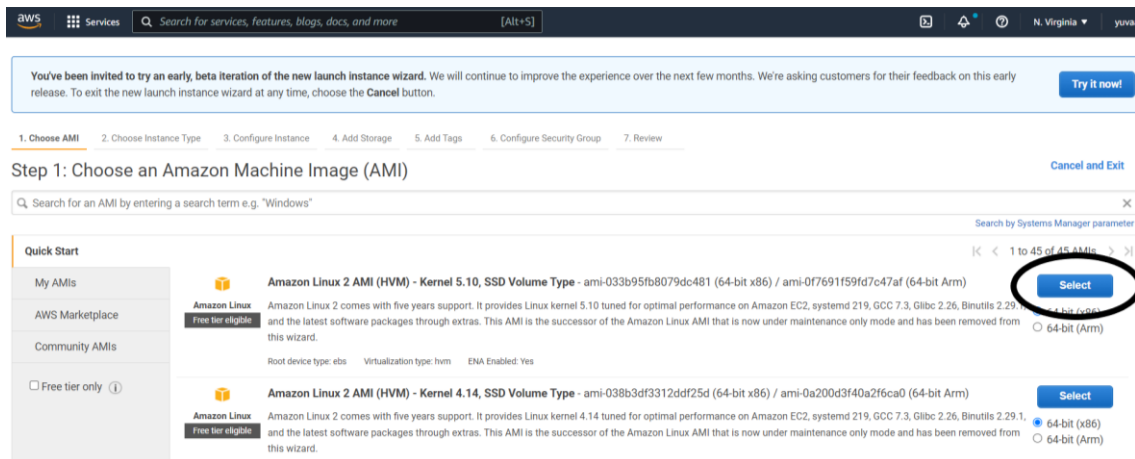
Then choose EC2 from the dropdown.



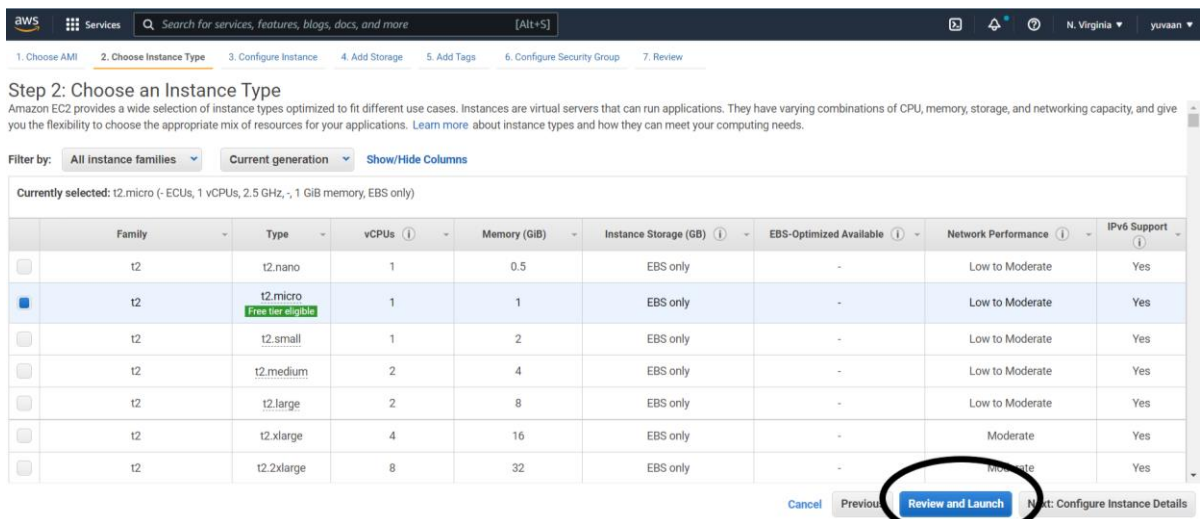
2. On the EC2 Dashboard, Click on the “Launch Instance” button



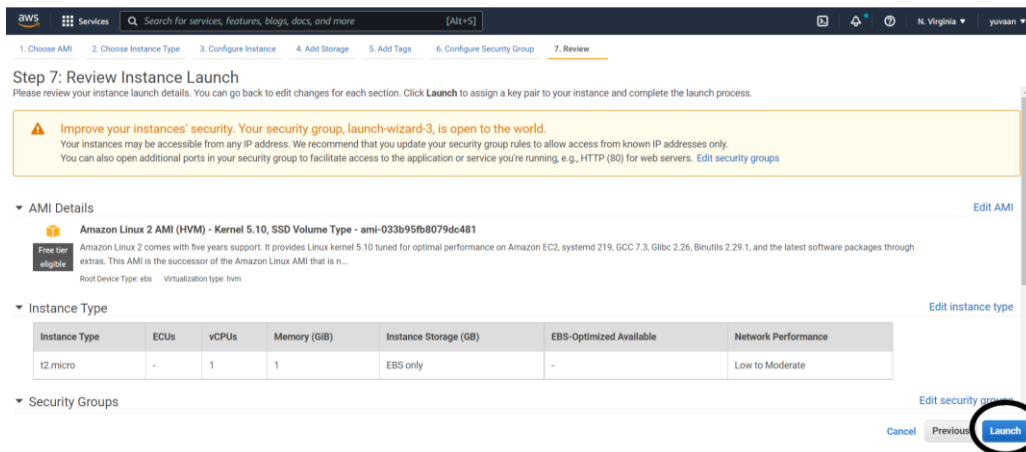
3. On the next screen of the launch wizard, select an AMI (Amazon Linux 2 AMI)



4. On the next screen of the launch wizard, select an instance family. For this demo, the default selected instance family “t2.micro” will work fine. Then click on the ‘Review and launch’ button.

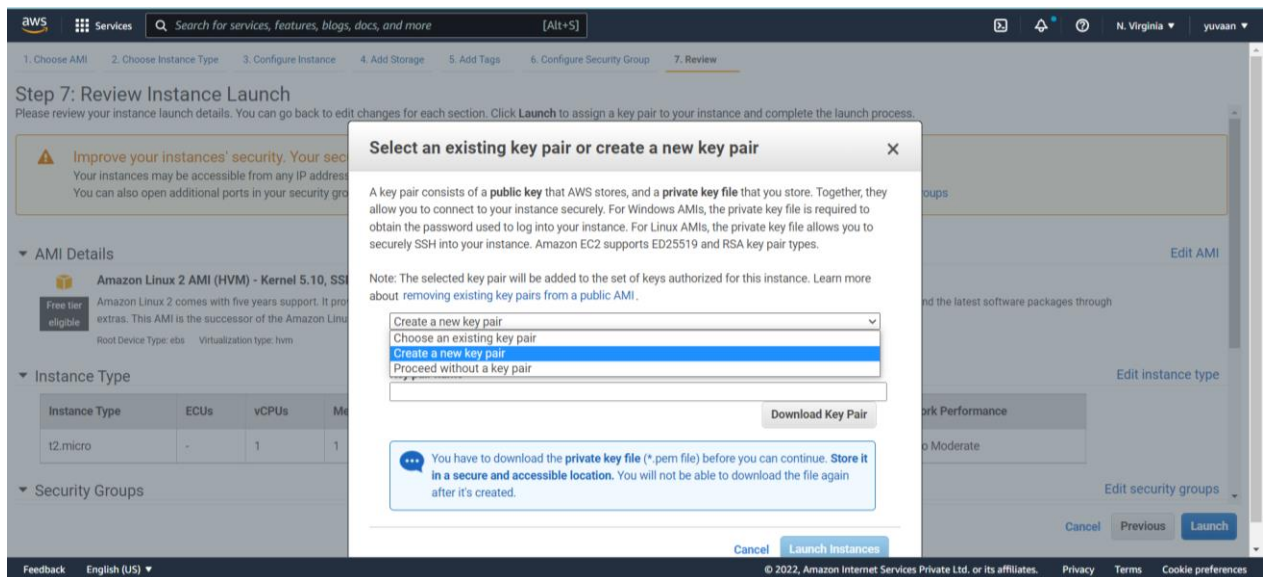


5. On the next screen, click on the ‘Launch’ button to start an EC2 instance.

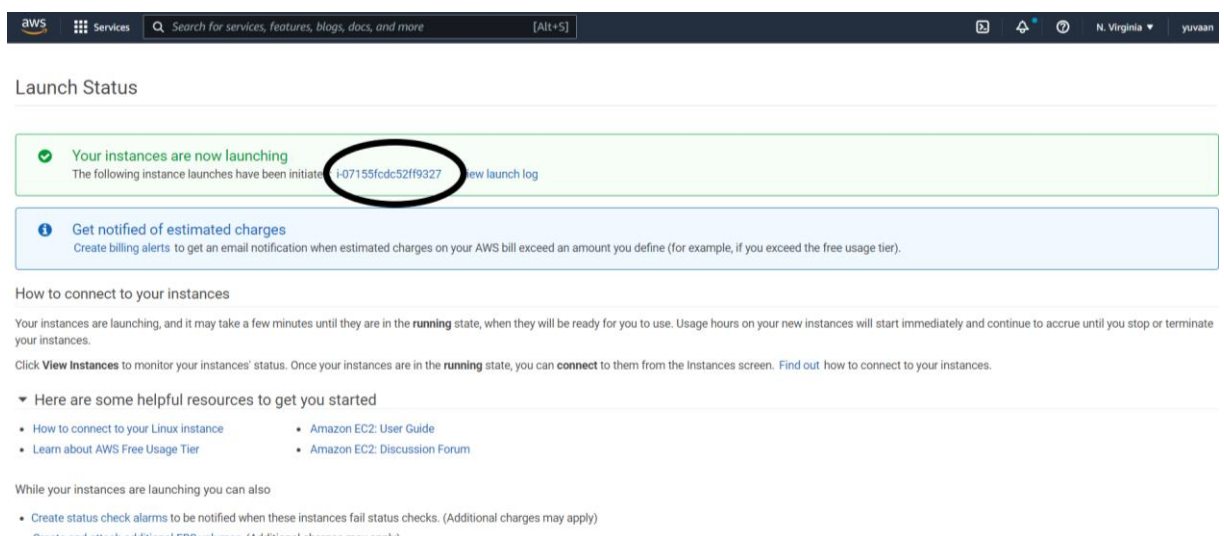


6. On the next Screen, select “Create a new key-pair” from the dropdown and click on

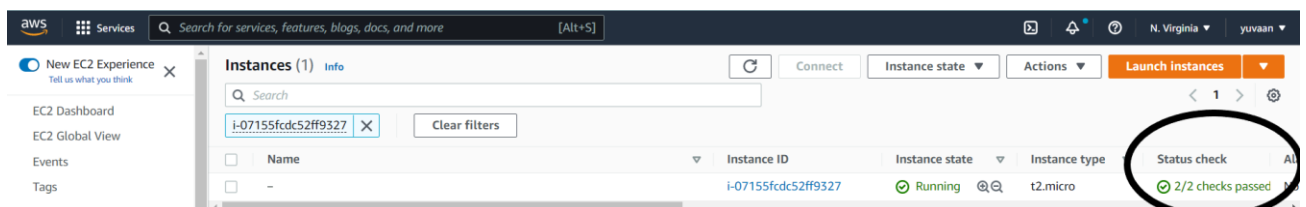
the “Download Key Pair” button. After the download finishes, click on the “Launch Instance” button.



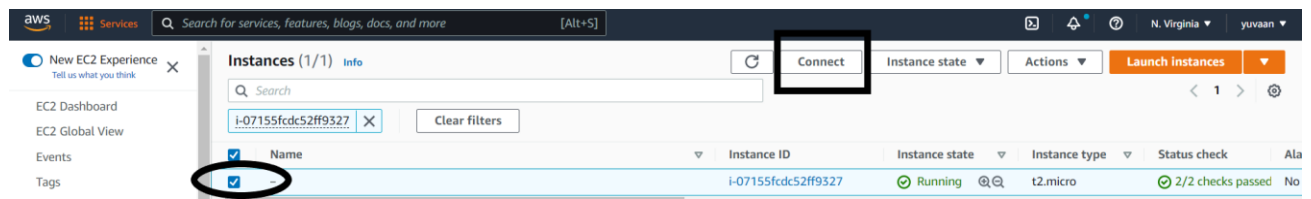
- After you launch the instance, “Launch Status” will appear/ click on the instance id(a hexadecimal string) to view the running instance.



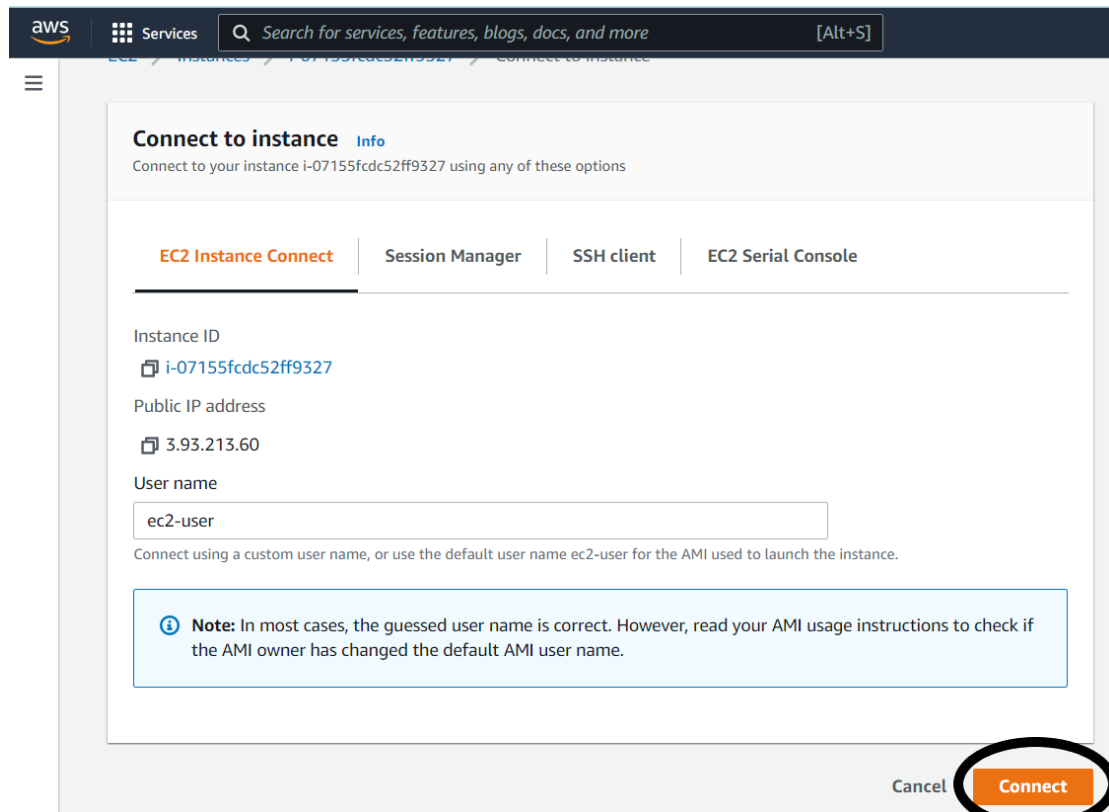
- This screen will show you the instance that you just launched. Before connecting to the instance, wait for instance status to show passed.



- Now select the checkbox next to the instance name and then click on the ‘Connect’ button.

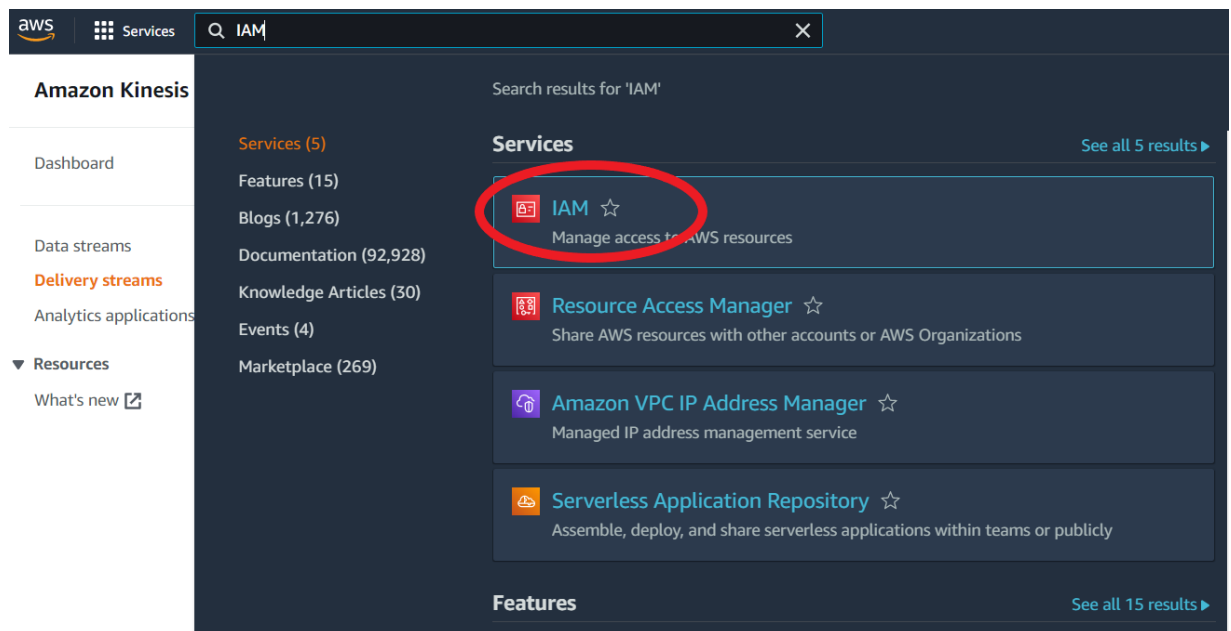


10. On the next screen, click on 'Connect' once again.

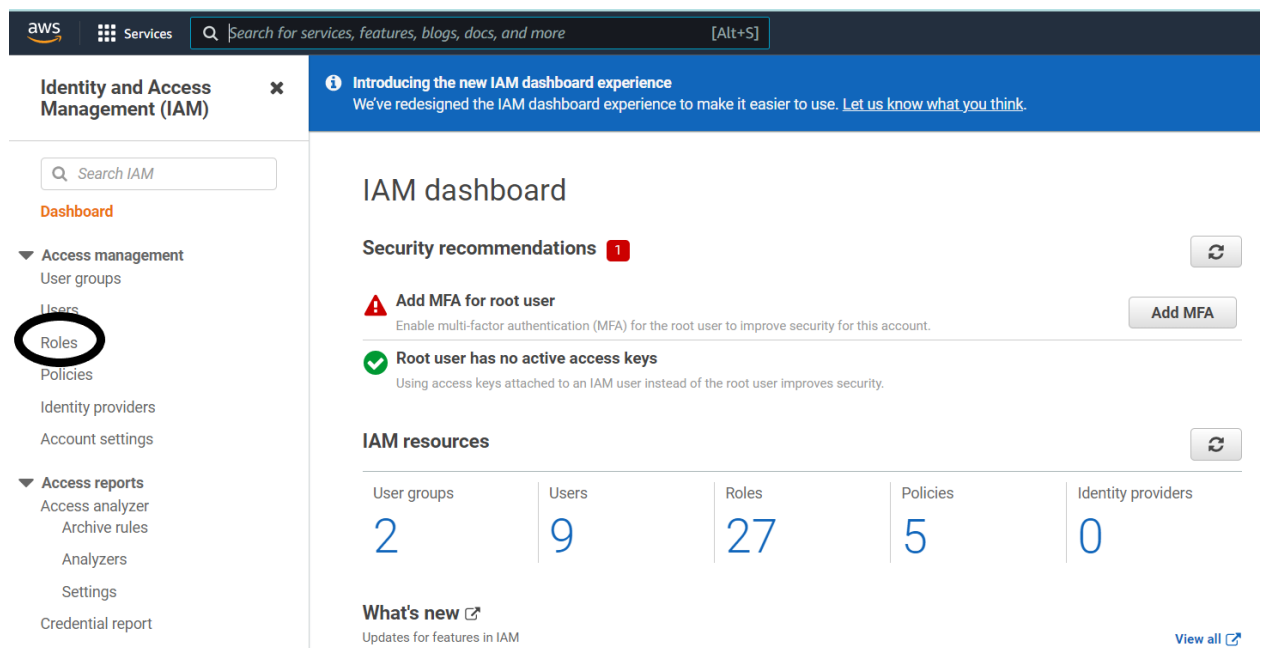


Step 2: Create an IAM role for EC2 instance

1. Go to “Services” and search for IAM.



2. On the IAM Dashboard, select “Roles” from the left hand side navigation menu.



3. Now click on the “Create Role” button.

Introducing the new IAM roles experience

We've redesigned the IAM roles experience to make it easier to use. [Let us know what you think.](#)

IAM > Roles

Roles (27) [Info](#)

↺

Delete

Create role

Q Search

< 1 2 > ⚙

<input type="checkbox"/>	Role name	Trusted entities	Last activity
<input type="checkbox"/>	AWSGlueServiceRole-hshdhdh	AWS Service: glue	125 days ago
<input type="checkbox"/>	AWSGlueServiceRole-s3crawler	AWS Service: glue	125 days ago
<input type="checkbox"/>	AWSGlueServiceRole-s3role	AWS Service: glue	125 days ago
<input type="checkbox"/>	AWSServiceRoleForAmazonEKS	AWS Service: eks (Service-Linked Role)	129 days ago

4. On the next screen, check the “EC2” radio button and click on Next.

Select trusted entity

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.

Common use cases

☒ **EC2**
Allows EC2 instances to call AWS services on your behalf.

☐ **Lambda**
Allows Lambda functions to call AWS services on your behalf.

Use cases for other AWS services:

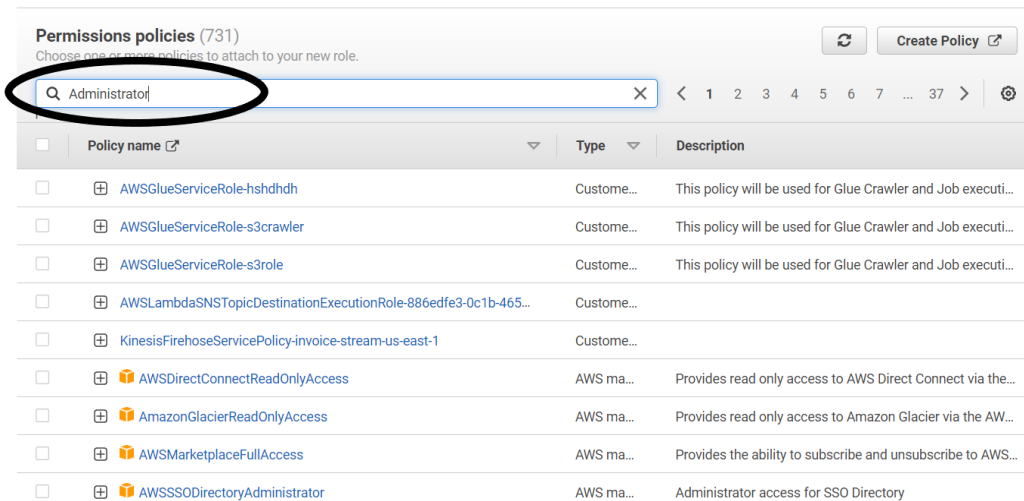
Choose a service to view use case

Cancel

Next

5. In the “Permissions policies” window, search for “Administrator.”

Add permissions

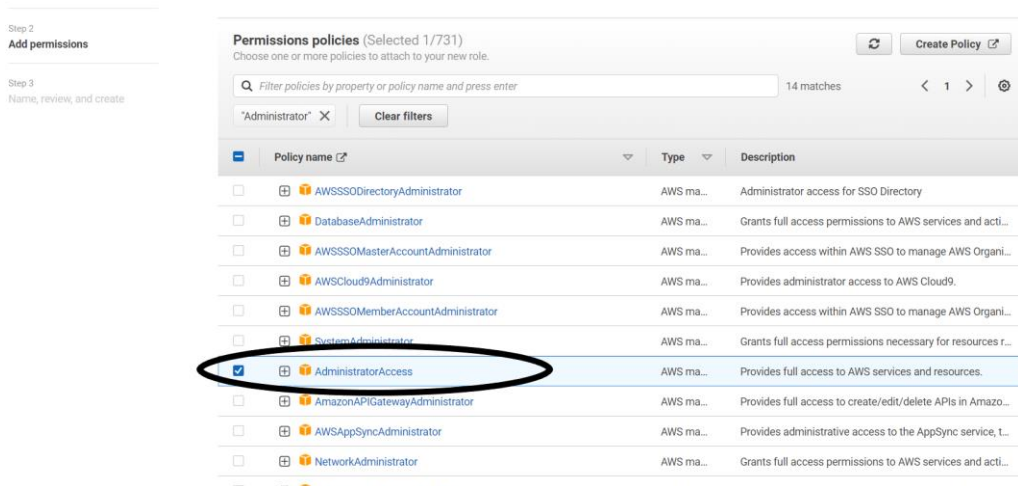


Permissions policies (731)
Choose one or more policies to attach to your new role.

Search: Administrator

<input type="checkbox"/>	Policy name	Type	Description
<input type="checkbox"/>	AWSGlueServiceRole-hshdhdh	Custom...	This policy will be used for Glue Crawler and Job executi...
<input type="checkbox"/>	AWSGlueServiceRole-s3crawler	Custom...	This policy will be used for Glue Crawler and Job executi...
<input type="checkbox"/>	AWSGlueServiceRole-s3role	Custom...	This policy will be used for Glue Crawler and Job executi...
<input type="checkbox"/>	AWSLambdaSNSTopicDestinationExecutionRole-886edfe3-0c1b-465...	Custom...	
<input type="checkbox"/>	KinesisFirehoseServicePolicy-invoice-stream-us-east-1	Custom...	
<input type="checkbox"/>	AWSDirectConnectReadOnlyAccess	AWS ma...	Provides read only access to AWS Direct Connect via the...
<input type="checkbox"/>	AmazonGlacierReadOnlyAccess	AWS ma...	Provides read only access to Amazon Glacier via the AW...
<input type="checkbox"/>	AWSMarketplaceFullAccess	AWS ma...	Provides the ability to subscribe and unsubscribe to AWS...
<input type="checkbox"/>	AWSSSODirectoryAdministrator	AWS ma...	Administrator access for SSO Directory

6. Select the “Administrator Access” from the list and click Next



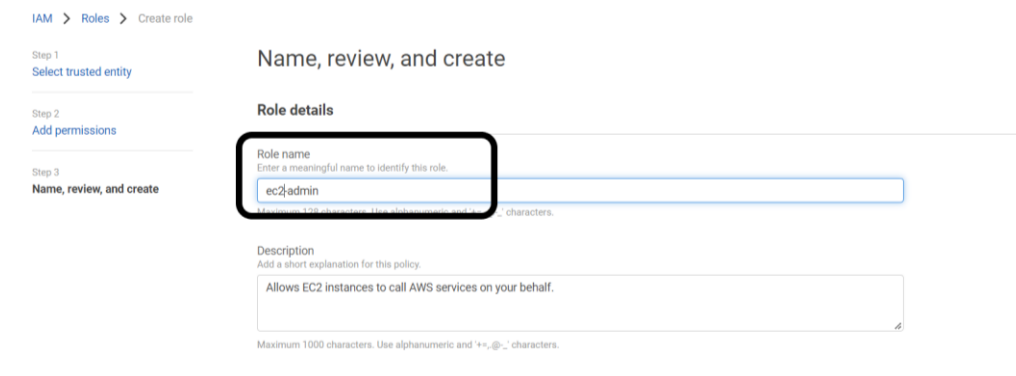
Permissions policies (Selected 1/731)
Choose one or more policies to attach to your new role.

Filter policies by property or policy name and press enter: 14 matches

Administrator X Clear filters

<input type="checkbox"/>	Policy name	Type	Description
<input type="checkbox"/>	AWSSSODirectoryAdministrator	AWS ma...	Administrator access for SSO Directory
<input type="checkbox"/>	DatabaseAdministrator	AWS ma...	Grants full access permissions to AWS services and acti...
<input type="checkbox"/>	AWSSSOMasterAccountAdministrator	AWS ma...	Provides access within AWS SSO to manage AWS Organi...
<input type="checkbox"/>	AWSCloud9Administrator	AWS ma...	Provides administrator access to AWS Cloud9.
<input type="checkbox"/>	AWSMemberAccountAdministrator	AWS ma...	Provides access within AWS SSO to manage AWS Organi...
<input type="checkbox"/>	Custom Administrator	AWS ma...	Grants full access permissions necessary for resources r...
<input checked="" type="checkbox"/>	AdministratorAccess	AWS ma...	Provides full access to AWS services and resources.
<input type="checkbox"/>	AmazonAPIGatewayAdministrator	AWS ma...	Provides full access to create/edit/delete APIs in Amaz...
<input type="checkbox"/>	AWSAppSyncAdministrator	AWS ma...	Provides administrative access to the AppSync service, t...
<input type="checkbox"/>	NetworkAdministrator	AWS ma...	Grants full access permissions to AWS services and acti...

7. On the next screen, you have to provide a role name. You can provide any name.



IAM > Roles > Create role

Step 1: Select trusted entity
Step 2: Add permissions
Step 3: Name, review, and create

Name, review, and create

Role details

Role name
Enter a meaningful name to identify this role.
ec2admin

Description
Add a short explanation for this policy.
Allows EC2 instances to call AWS services on your behalf.

Maximum 1000 characters. Use alphanumeric and '+', '@', '-' characters.

8. Scroll down to the bottom of the screen and click on the “Create role” button.

Tags

Add tags (Optional)
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.

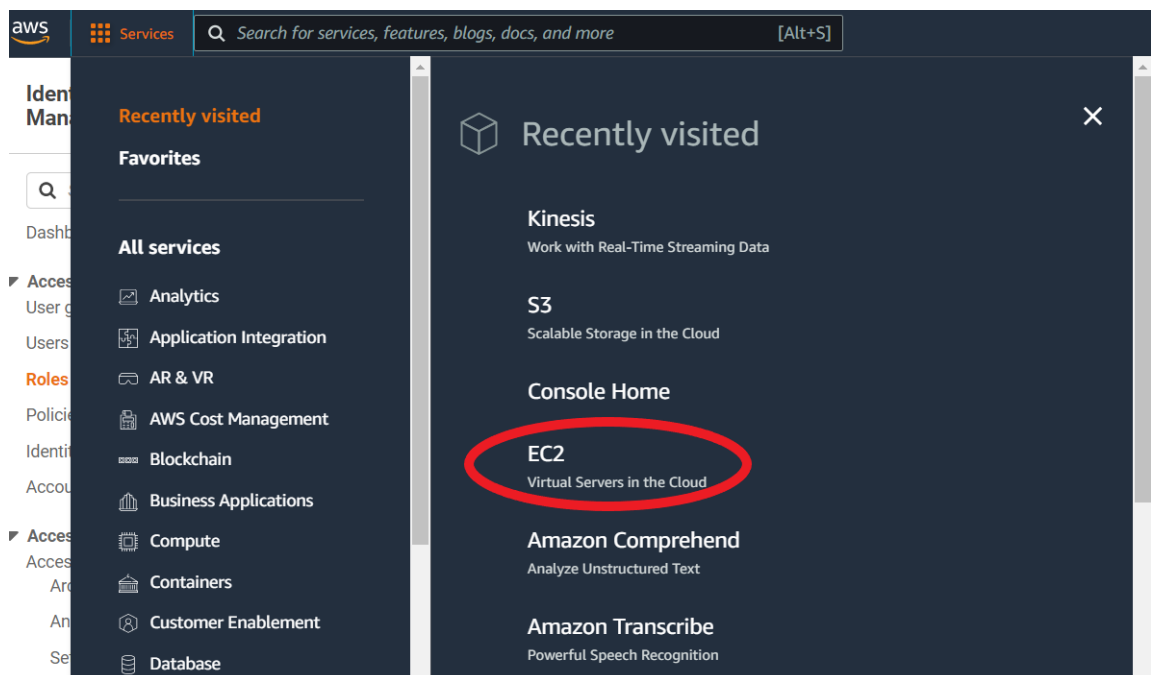
Add tag

You can add up to 50 more tags

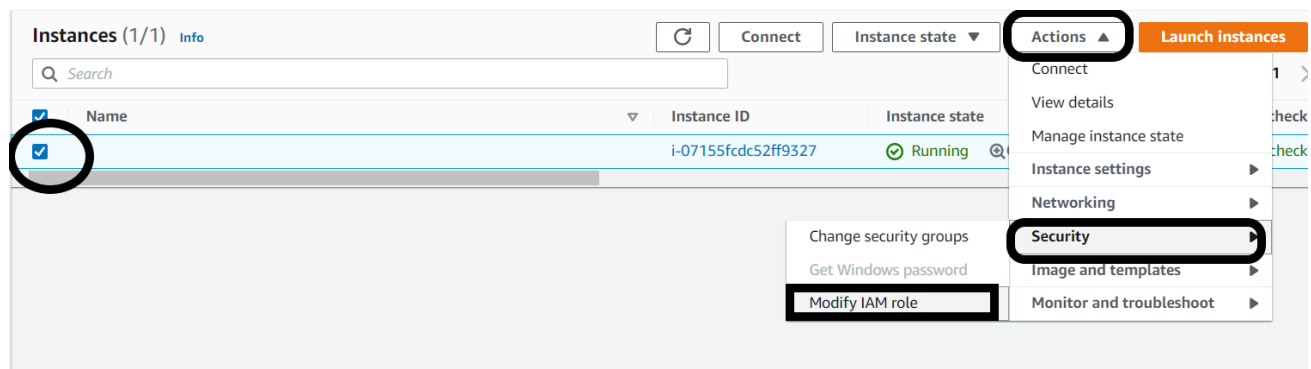
[Cancel](#) [Previous](#) [Create role](#)

Step 3: Attach IAM role to EC2

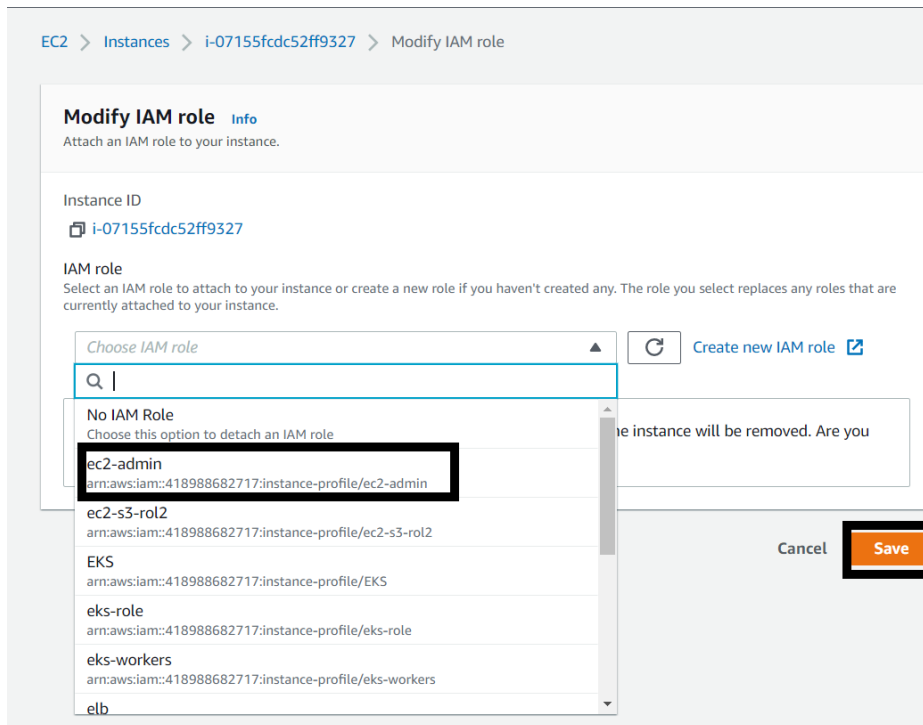
1. Choose "EC2" from the services and go to the EC2 Dashboard.



2. Select your EC2 instance and click on "Actions". Under actions, click on "Security" and choose "Modify IAM Role" from the dropdown.



3. On the next screen, choose the IAM role created in Step 2 from the dropdown and click on the "Save" button.



Step 4: Download Simulation APP on EC2

1. After connecting to EC2, download the simulation application on EC2.

Run the command given below to download the code.

```
wget https://invoice-generator-  
di.s3.amazonaws.com/InvoiceGenerator.zip
```

2. Extract the files using the “unzip” command. The unzip command given below will create a folder named “InvoiceGenerator” containing two files (customer_retails.csv and InvoiceGenerator.py)

```
unzip InvoiceGenerator.zip
```

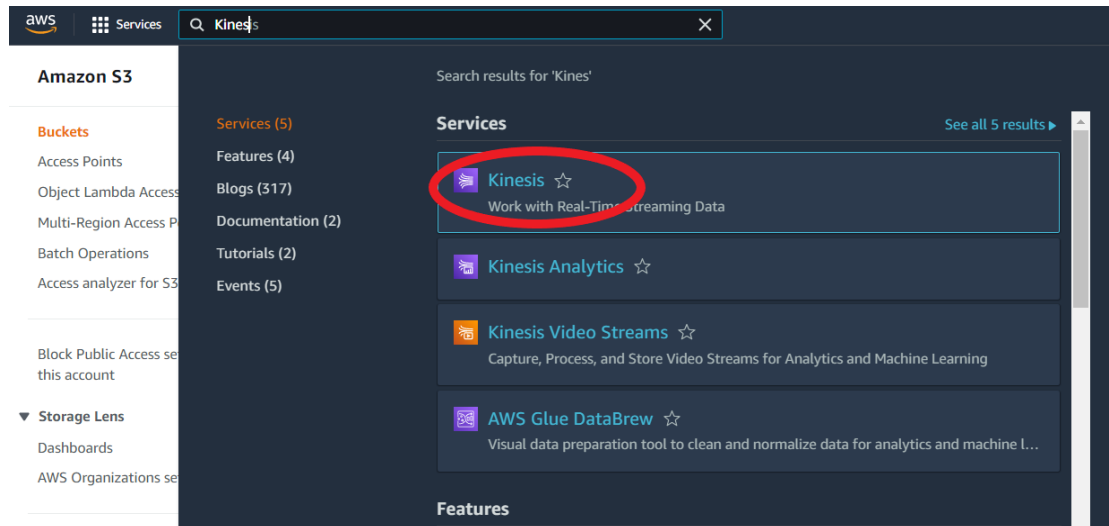
3. On running, the InvoiceGenerator.py will produce invoices logs at the path “/var/logs/invoices”. So, you must give execute permission to the file InvoiceGenerator.py and create the logs directory as below:

```
chmod +x InvoiceGenerator.py  
mkdir /var/logs/invoices
```

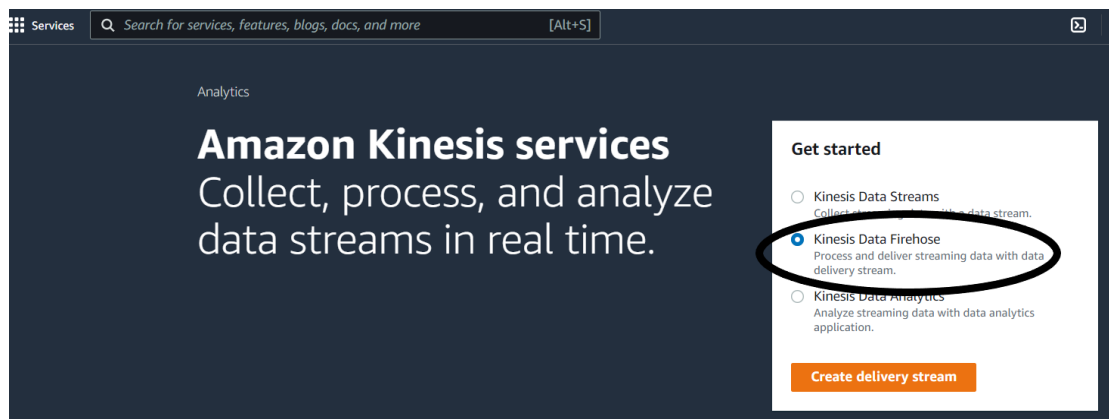
Step 4: Create Kinesis Firehose on AWS

Amazon Kinesis Data Firehose is an extract, transform, and load (ETL) service that reliably captures, transforms, and delivers streaming data to data lakes, data stores, and analytics services.

1. First search for Kinesis in AWS Services and double click on “Kinesis” from the dropdown



2. Select the “Kinesis Firehose” radio button and click on the “Create Delivery Stream” button.



3. On the next screen ,select “Direct PUT” in the source. This is because in this exercise, we will directly put data from the EC2 instance into this Kinesis Firehose using the Kinesis Agent.

aws Services Search for services, features, blogs, docs, and more [Alt+S]

Amazon Kinesis > Delivery streams > Create delivery stream

Create a delivery stream [Info](#)

► **Amazon Kinesis Data Firehose: How it works**

Choose source and destination
Specify the source and the destination for your delivery stream. You cannot change the source and destination of your delivery stream once it has been created.

Source [Info](#)

Choose a source

Q |

Amazon Kinesis Data Streams
Choose this option if you want to use Kinesis Data Streams as the data source for your delivery

Direct PUT
Choose this option to create a Kinesis Data Firehose delivery stream that producer applications write to directly.

Create delivery stream

4. In the “Destination” dropdown, select “AWS S3.” We will be pushing our raw logs there, that are generated by the Invoice Generator Simulation App on EC2.

aws Services Search for services, features, blogs, docs, and more [Alt+S]

Amazon Kinesis > Delivery streams > Create delivery stream

Q |

Amazon OpenSearch Service
An open-source search and analytics engine for use cases such as log analytics, real-time application monitoring, and click stream analytics.

Amazon Redshift
An enterprise-level, petabyte scale, fully managed data warehousing service.

Amazon S3
Object storage built to store and retrieve any amount of data from anywhere. When you choose Amazon S3 as the destination, you can configure dynamic partitioning for your delivery stream.

Datadog
Monitor your servers, containers, databases, and third-party services.

Dynatrace
Stream metrics into Dynatrace for root cause analysis.

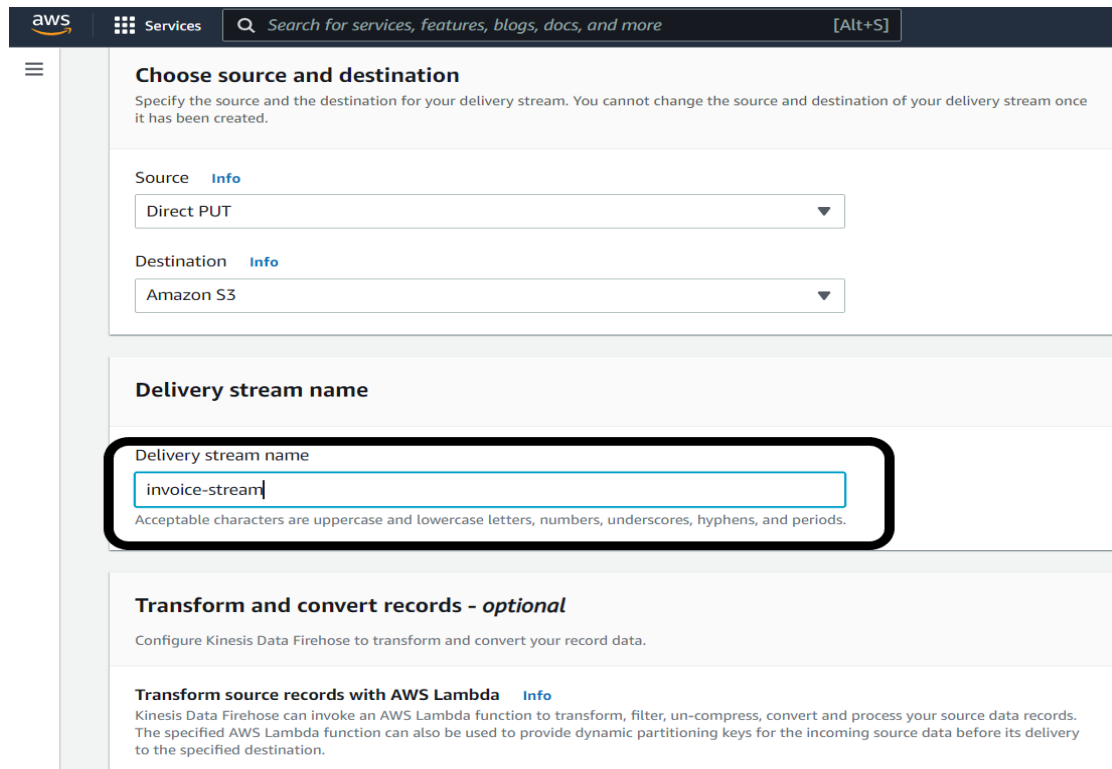
HTTP Endpoint
A way to deliver data to your custom destination.

Honeycomb
Observability for a distributed world-designed for high cardinality data and collaborative

Choose a destination

Cancel **Create delivery stream**

5. Provide a name to your delivery stream, like “invoice-stream”.



Choose source and destination
Specify the source and the destination for your delivery stream. You cannot change the source and destination of your delivery stream once it has been created.

Source [Info](#)
Direct PUT

Destination [Info](#)
Amazon S3

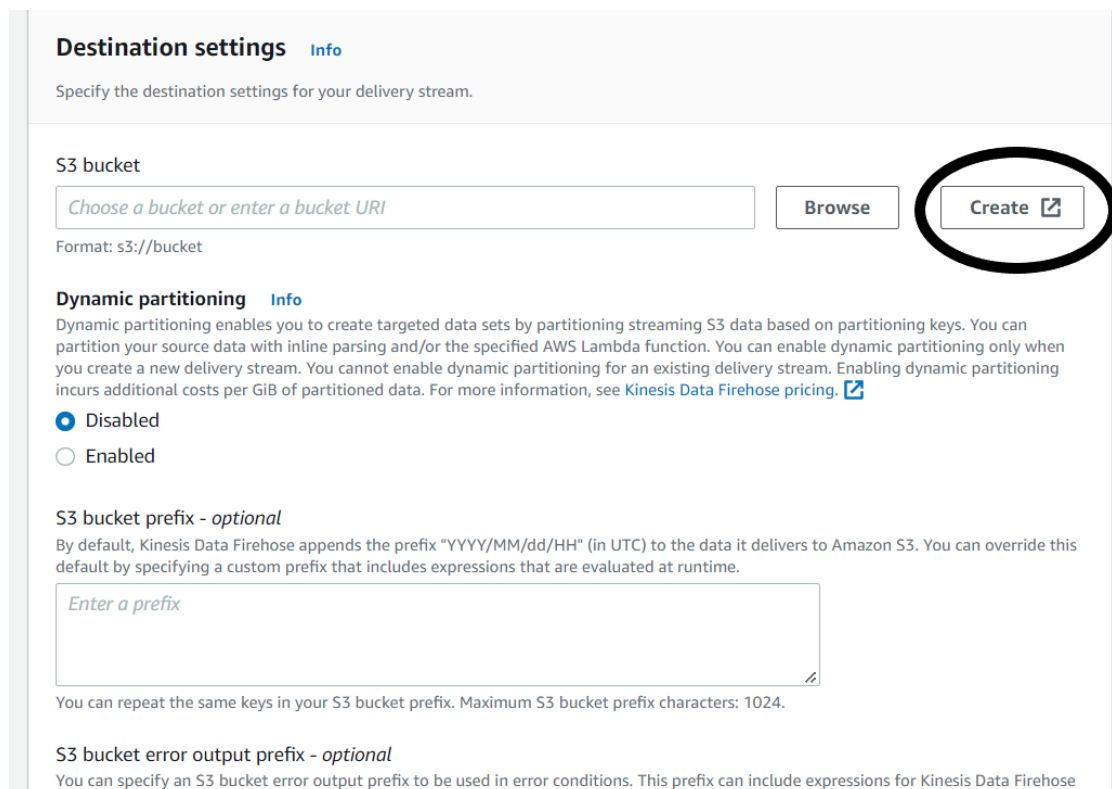
Delivery stream name

Delivery stream name
invoice-stream
Acceptable characters are uppercase and lowercase letters, numbers, underscores, hyphens, and periods.

Transform and convert records - optional
Configure Kinesis Data Firehose to transform and convert your record data.

Transform source records with AWS Lambda [Info](#)
Kinesis Data Firehose can invoke an AWS Lambda function to transform, filter, un-compress, convert and process your source data records. The specified AWS Lambda function can also be used to provide dynamic partitioning keys for the incoming source data before its delivery to the specified destination.

6. Scroll down and under “Destination settings”, click on the “Create” button. It will open an S3 dashboard to choose where data will be loaded on AWS.



Destination settings [Info](#)
Specify the destination settings for your delivery stream.

S3 bucket

Choose a bucket or enter a bucket URI

Format: s3://bucket

Browse Create [↗](#)

Dynamic partitioning [Info](#)
Dynamic partitioning enables you to create targeted data sets by partitioning streaming S3 data based on partitioning keys. You can partition your source data with inline parsing and/or the specified AWS Lambda function. You can enable dynamic partitioning only when you create a new delivery stream. You cannot enable dynamic partitioning for an existing delivery stream. Enabling dynamic partitioning incurs additional costs per GiB of partitioned data. For more information, see [Kinesis Data Firehose pricing](#).

☒ Disabled
☐ Enabled

S3 bucket prefix - optional
By default, Kinesis Data Firehose appends the prefix "YYYY/MM/dd/HH" (in UTC) to the data it delivers to Amazon S3. You can override this default by specifying a custom prefix that includes expressions that are evaluated at runtime.

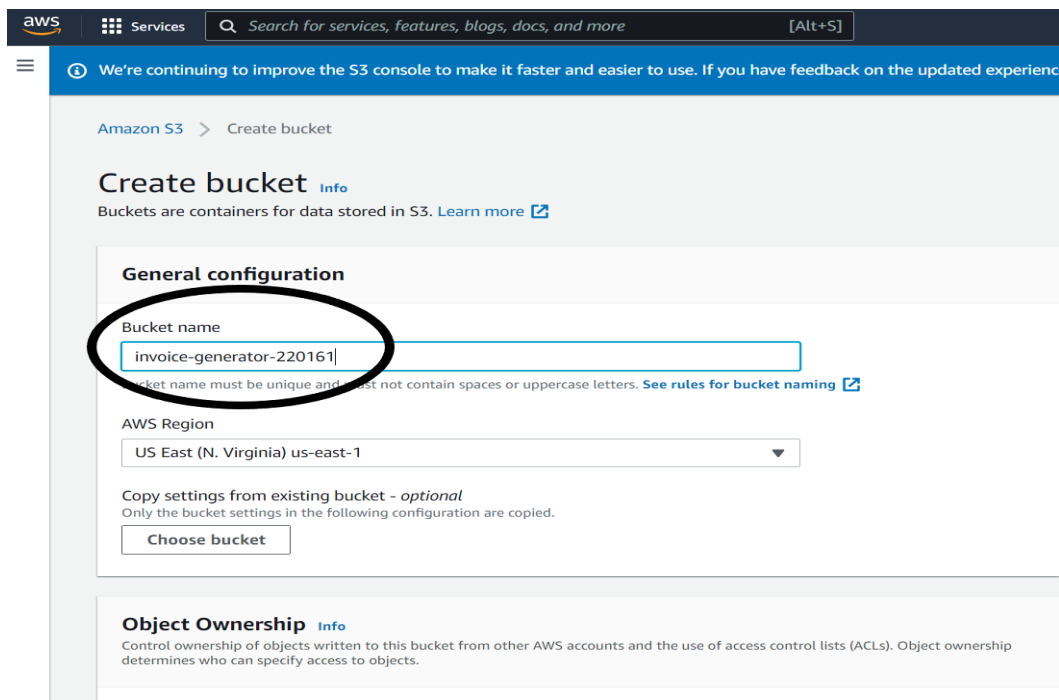
Enter a prefix

You can repeat the same keys in your S3 bucket prefix. Maximum S3 bucket prefix characters: 1024.

S3 bucket error output prefix - optional
You can specify an S3 bucket error output prefix to be used in error conditions. This prefix can include expressions for Kinesis Data Firehose

7. In the S3 Dashboard, provide a unique name to your S3 bucket. Keep the default

configuration, scroll down, and click on the “Create Bucket” button.



aws Services Search for services, features, blogs, docs, and more [Alt+S]

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience

Amazon S3 > Create bucket

Create bucket [Info](#)

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

General configuration

Bucket name

invoice-generator-220161

Bucket name must be unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)

AWS Region

US East (N. Virginia) us-east-1

Copy settings from existing bucket - optional

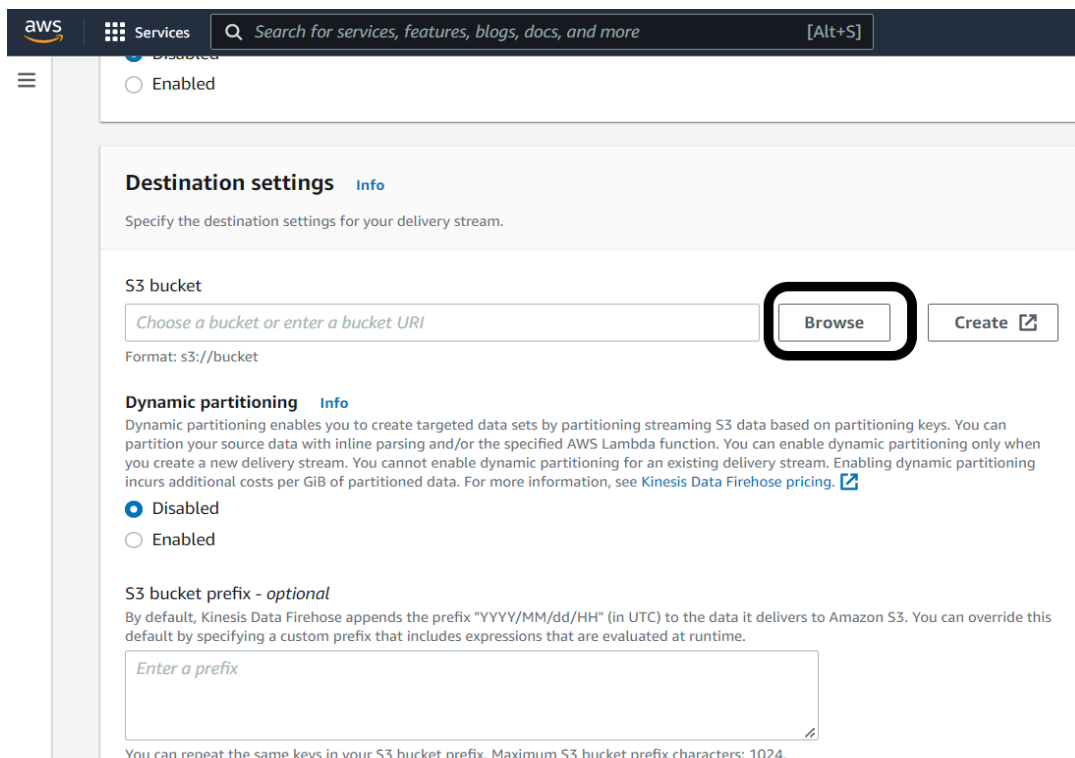
Only the bucket settings in the following configuration are copied.

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.

- Now go back to the Kinesis Dashboard. Under the “Destination settings”, locate the “S3 bucket” path, click on “Browse”, and choose the S3 bucket you just created.. Finally, scroll down, and click on the “Create Delivery Stream” button.



aws Services Search for services, features, blogs, docs, and more [Alt+S]

Disabled

Enabled

Destination settings [Info](#)

Specify the destination settings for your delivery stream.

S3 bucket

Choose a bucket or enter a bucket URI

Browse Create

Format: s3://bucket

Dynamic partitioning [Info](#)

Dynamic partitioning enables you to create targeted data sets by partitioning streaming S3 data based on partitioning keys. You can partition your source data with inline parsing and/or the specified AWS Lambda function. You can enable dynamic partitioning only when you create a new delivery stream. You cannot enable dynamic partitioning for an existing delivery stream. Enabling dynamic partitioning incurs additional costs per GiB of partitioned data. For more information, see [Kinesis Data Firehose pricing](#).

Disabled

Enabled

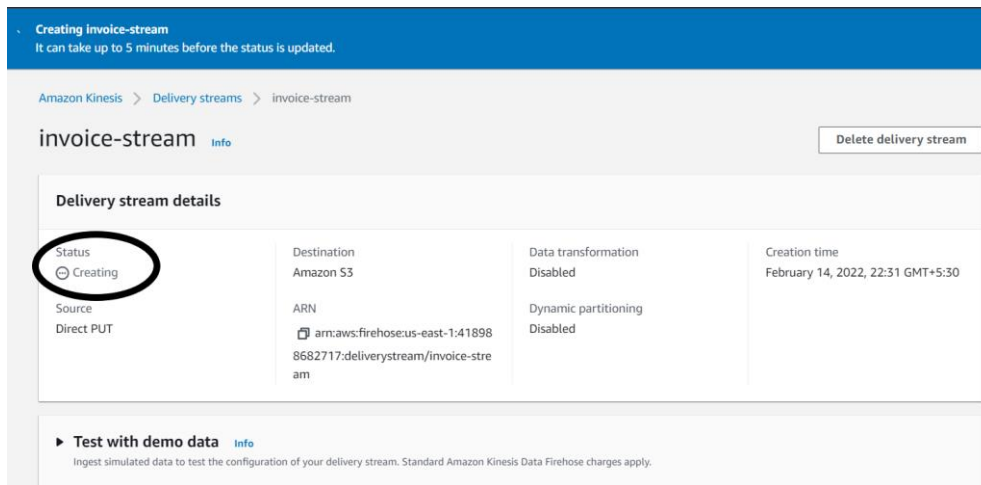
S3 bucket prefix - optional

By default, Kinesis Data Firehose appends the prefix "YYYY/MM/dd/HH" (in UTC) to the data it delivers to Amazon S3. You can override this default by specifying a custom prefix that includes expressions that are evaluated at runtime.

Enter a prefix

You can repeat the same keys in your S3 bucket prefix. Maximum S3 bucket prefix characters: 1024.

9. You have successfully created a Kinesis Firehose stream. You may have to wait for stream status to change to “Active”.



Step 5: Run Kinesis Agent on EC2

Kinesis Agent is a **stand-alone Java software application** that offers an easy way to collect and send data to Kinesis data streams and Kinesis Firehose.

1. Before running the Simulation app, first download the AWS Kinesis Agent on EC2.

```
sudo yum install -y aws-kinesis-agent
```

2. Let us configure the Kinesis-agent to listen to the logs generated by our InvoiceGenerator simulation app at the location “/var/logs/invoices”.

```
cd /etc/aws-kinesis
sudo nano agent.json
```

```
{
  "cloudwatch.emitMetrics": true,
  "kinesis.endpoint": "",
  "firehose.endpoint": "firehose.us-east-1.amazonaws.com",
  "flows": [
    {
      "filePattern": "/var/logs/invoices/*.log",
      "deliveryStream": "invoice-stream"
    }
  ]
}
```

3. Start the Kinesis Agent and go back to the home directory to start the simulation application to generate logs.

```
sudo service aws-kinesis-agent start
sudo chkconfig aws-kinesis-agent on
cd ~
```

Step 6: Run InvoiceGenerator Simulation Application

1. Start the application and generate 10000 lines in the log.

```
cd ~/InvoiceGenerator/
sudo ./InvoiceGenerator.py 10000
```

2. Check the logs directory

```
cd /var/logs/invoices
```

3. Check Kinesis logs

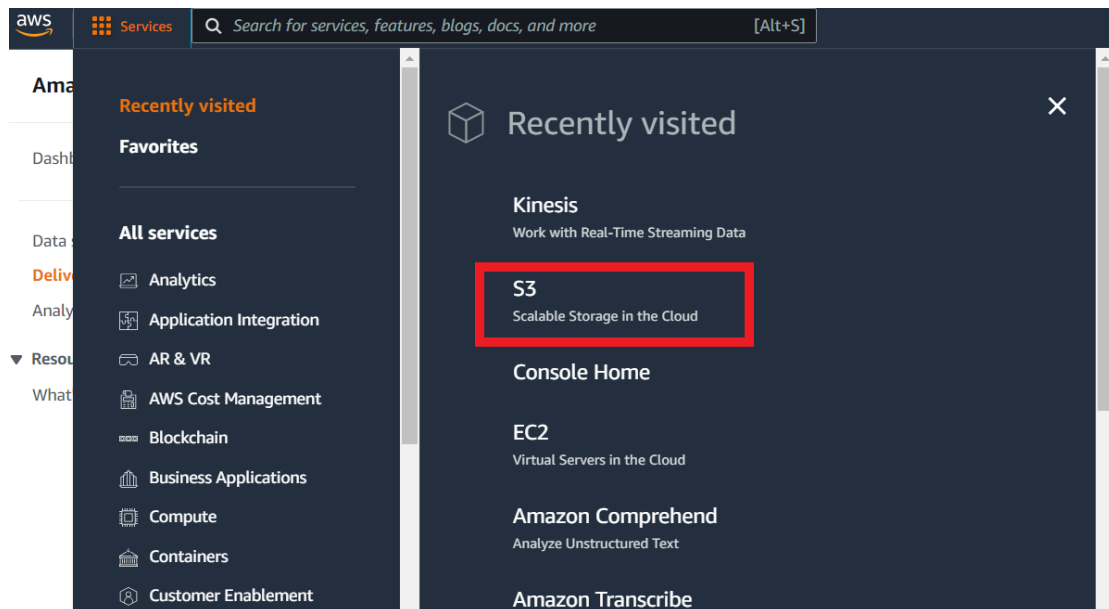
```
tail -f /var/log/aws-kinesis-agent/aws-kinesis-agent.log
```

```
and 0 records sent successfully to destinations. Uptime: 30000ms
2022-02-14 20:19:54.919+0000 (FileTailor[fh:invoice-stream:/var/log/invoices/*.log].MetricsEmitter RUNNING) com.amazon.kinesis.streaming.agent.tailing.FileTailor [INFO] FileTailor[fh:invoice-stream:/var/log/invoices/*.log]: Tailor Progress: Tailor has parsed 0 records (834295 bytes), transformed 0 records, skipped 0 records, and has successfully sent 0 records to destination.
2022-02-14 20:19:54.922+0000 (Agent.MetricsEmitter RUNNING) com.amazon.kinesis.streaming.agent.Agent [INFO] Agent: Progress: 0 records parsed (834295 bytes), and 0 records sent successfully to destinations. Uptime: 120086ms
2022-02-14 20:20:18.712+0000 (FileTailor[fh:invoice-stream:/var/log/invoices/*.log]) com.amazon.kinesis.streaming.agent.tailing.FirehoseParser [INFO] FirehoseParser[fh:invoice-stream:/var/log/invoices/*.log]: Continuing to parse /var/log/invoices/20220214-202018.log.
2022-02-14 20:20:18.932+0000 (sender-0) com.amazon.kinesis.streaming.agent.UserDefinedCredentialsProvider [INFO] No custom implementation of credentials provider present in the config file
2022-02-14 20:20:19.792+0000 (sender-9) com.amazon.kinesis.streaming.agent.UserDefinedCredentialsProvider [INFO] No custom implementation of credentials provider present in the config file
2022-02-14 20:20:24.919+0000 (FileTailor[fh:invoice-stream:/var/log/invoices/*.log].MetricsEmitter RUNNING) com.amazon.kinesis.streaming.agent.tailing.FileTailor [INFO] FileTailor[fh:invoice-stream:/var/log/invoices/*.log]: Tailor Progress: Tailor has parsed 10000 records (1681457 bytes), transformed 0 records, skipped 0 records, and has successfully sent 10000 records to destinations.
2022-02-14 20:20:24.922+0000 (Agent.MetricsEmitter RUNNING) com.amazon.kinesis.streaming.agent.Agent [INFO] Agent: Progress: 10000 records parsed (1681457 bytes), and 10000 records sent successfully to destinations. Uptime: 150086ms
```

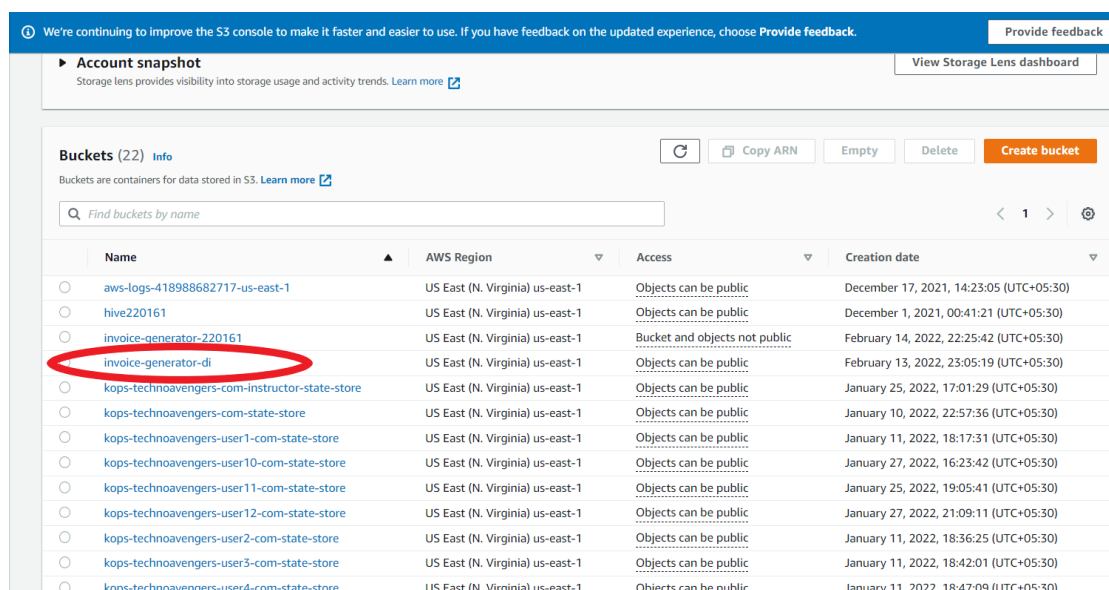

Step 7: Check S3

Please note that Firehose has a buffer time of minimum 1 minute to push data to the destination.

1. Search “S3” on the services dropdown.



2. Choose the bucket that we have configured as “Destination” in our Firehose.



3. Finally check whether data is published in S3.

The screenshot shows the Amazon S3 console interface. At the top, the breadcrumb navigation indicates the path: Amazon S3 > invoice-generator-220161 > 2022/ > 02/ > 14/ > 20/. The main heading is '20/'. On the right, there is a 'Copy S3 URI' button. Below the heading, there are two tabs: 'Objects' (selected) and 'Properties'. The 'Objects' section shows 'Objects (1)' and a description: '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](#)'. Below this, there is a toolbar with buttons: Refresh, Copy S3 URI, Copy URL, Download, Open, Delete, Actions, Create folder, and Upload. A search bar with the placeholder 'Find objects by prefix' is also present. Below the toolbar, there is a table with one object listed:

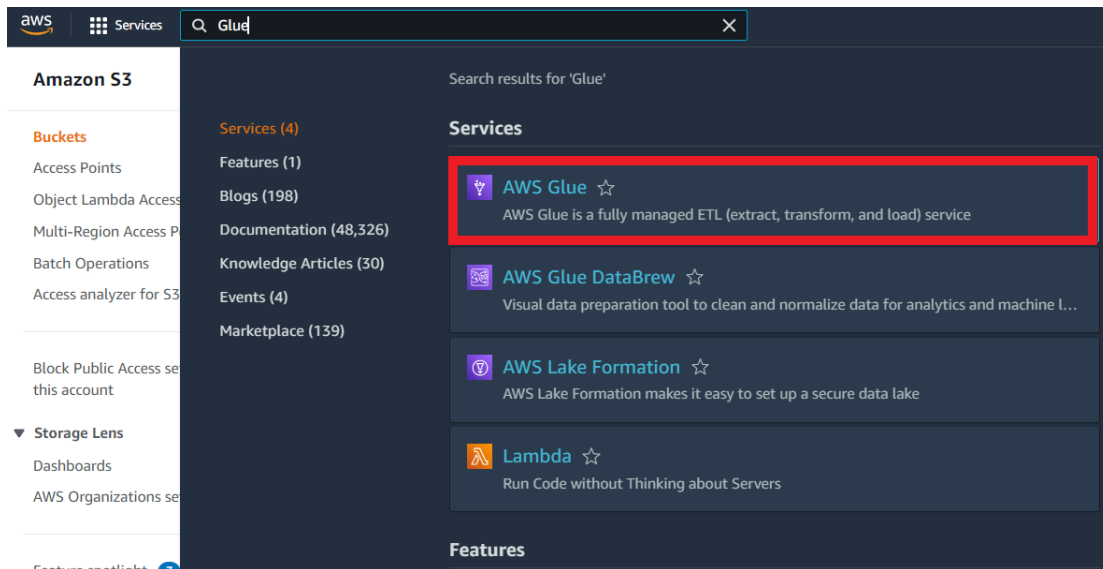
<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	invoice-stream-1-2022-02-14-20-20-21-d6ecbb71-b2e2-4a4c-a99b-e63270869500	-	February 15, 2022, 01:55:24 (UTC+05:30)	827.3 KB	Standard

Data Catalog Using Glue

Once the data is in your data lake such as S3, you can create a Data Catalog from it using Glue.

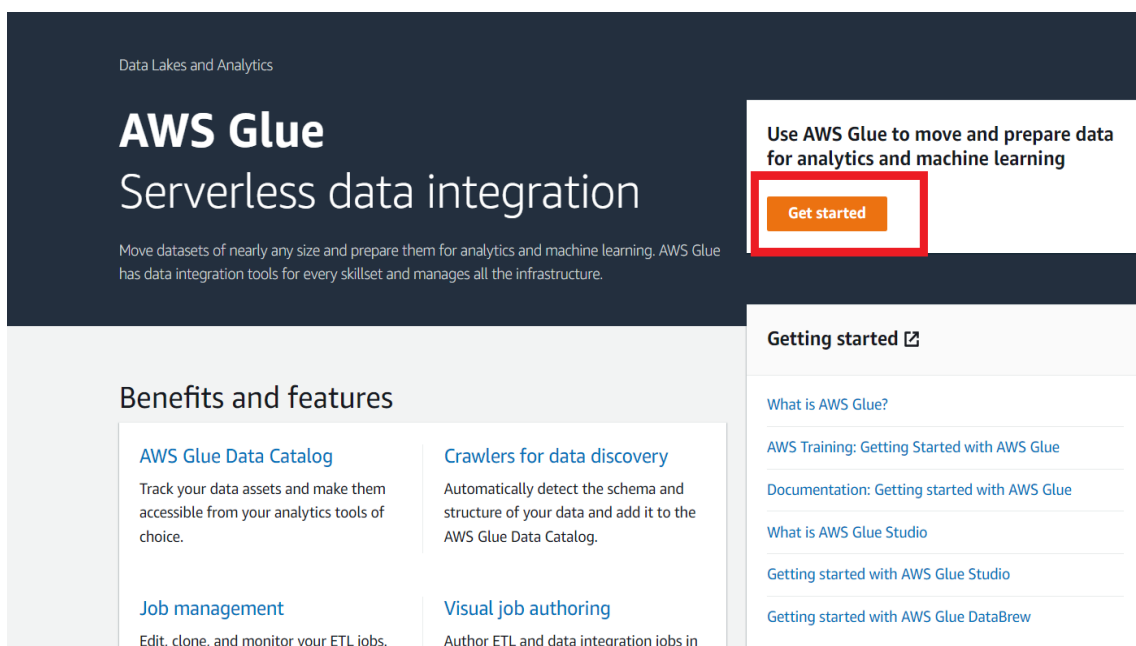
Step 1: Start Glue

Type “Glue” in Services and select “AWS Glue” from the list.



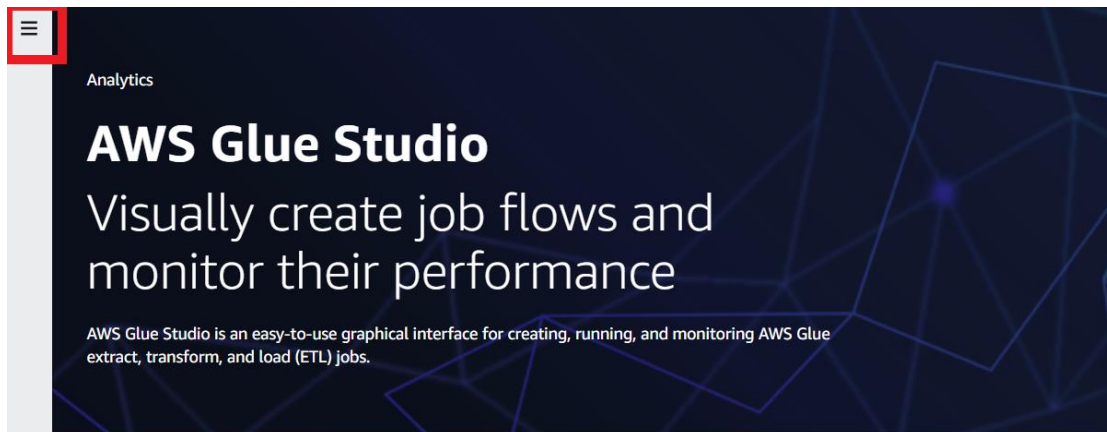
Step 2: Get

Click on “Get Started” to move forward.

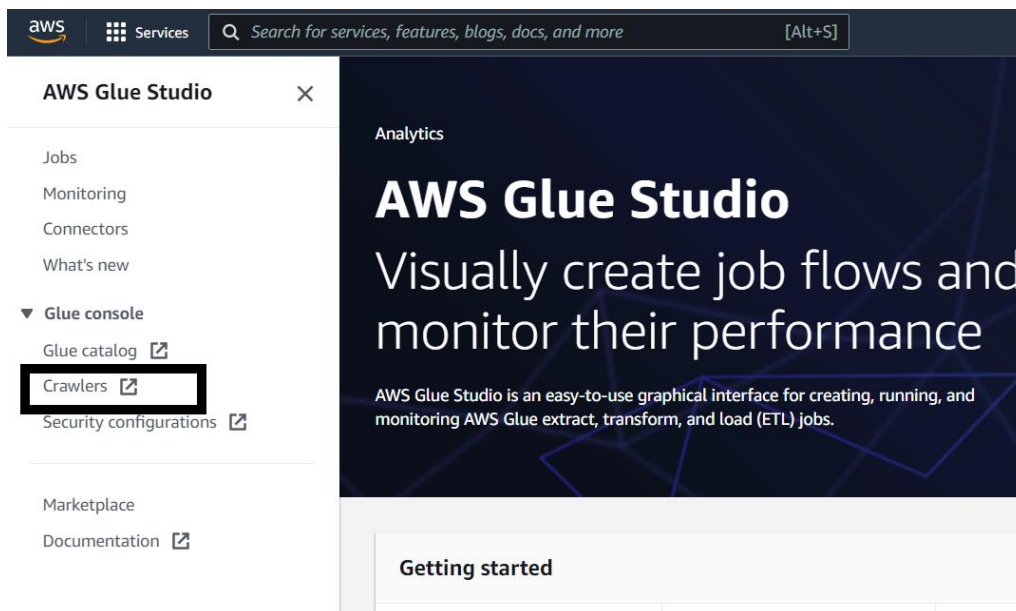


Step 3: Create Crawler

1. Click on the left hand side menu symbol to open the Glue Dashboard.

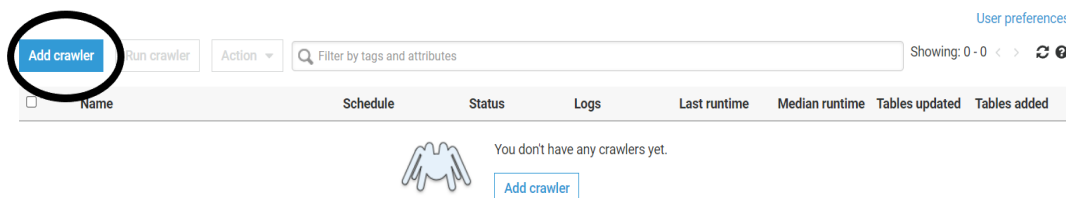


2. Click on "Crawlers" in the menu.



3. In the next screen, click on the "Add Crawler" button.

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.



4. Give a name to your crawler (in this example *s3-crawler*) and click on Next.

The screenshot shows the 'Add crawler' wizard. On the left, a sidebar lists the steps: 'Crawler info' (selected with a green circle), 'Crawler source type', 'Data store', 'IAM Role', 'Schedule', 'Output', and 'Review all steps'. The main area is titled 'Add information about your crawler'. It features a text input field for 'Crawler name' containing 's3-crawler', which is highlighted with a black rectangular box. Below the input field, there is a note: 'Tags, description, security configuration, and classifiers (optional)'. At the bottom right of the main area is a blue 'Next' button.

5. On the next screen, you have to define the crawler source type. Choose “Data Stores” under “Crawler source type”. Keep the default option “Crawl all folders” under “Repeat crawls of S3 data stores”. Click next.

The screenshot shows the 'Specify crawler source type' wizard. The left sidebar is updated: 'Crawler info' now has 's3-crawler' below it, and 'Crawler source type' is selected with a green circle. The main area is titled 'Specify crawler source type' and includes a descriptive paragraph: 'Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores.' Below this, there are two sections. The first, 'Crawler source type', has two radio buttons: 'Data stores' (selected with a blue dot) and 'Existing catalog tables'. The second section, 'Repeat crawls of S3 data stores', has three radio buttons: 'Crawl all folders' (selected with a blue dot), 'Crawl new folders only', and 'Crawl changed folders identified by Amazon S3 Event Notifications'. Each option has a brief description. At the bottom right, there are two buttons: a light blue 'Back' button and a blue 'Next' button, which is highlighted with a black rectangular box.

6. On the next screen, choose the S3 bucket from which you would like to create a data catalog in “Include path” and click next.

- On next screen asking to add another data store, keep the default “No” option and click Next

- On the next screen, select “Create an IAM role” from the options available and provide a name to this role. Then click Next.

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 ⓘ

AWSGlueServiceRole- s3

To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named "AWSGlueServiceRole-rolename" and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://invoice-generator-220161

You can also create an IAM role on the [IAM console](#).

[Back](#) [Next](#)

9. On the next screen, we need to define the frequency at which the crawler will crawl S3. For now, let's keep it "Run on demand" and click Next.

Create a schedule for this crawler

Frequency

Run on demand

[Back](#) [Next](#)

10. On the next screen, we need to define a database. Click on "Add database", to open a new prompt. Provide any name for your database and click on Next.

Configure the crawler's output

Database ⓘ

Choose a database to contain tables
 ▼

Select a database

Add database

Prefix added to tables (optional) ⓘ

Type a prefix added to table names

▶ Grouping behavior for S3 data (optional)
 ▶ Configuration options (optional)

Back
Next

11. On the final screen, click on the “Finish” button to add this Crawler.

Add crawler

- ✓ **Crawler info**
s3-crawler
- ✓ **Crawler source type**
Data stores
- ✓ **Data store**
S3: s3://invoice-gene...
- ✓ **IAM Role**
arn:aws:iam::418988682717:role/service-role/AWSGlueServiceRole-s33
- ✓ **Schedule**
Run on demand
- ✓ **Output**
invoicedb
- **Review all steps**

IAM role

IAM role	arn:aws:iam::418988682717:role/service-role/AWSGlueServiceRole-s33
-----------------	--

Schedule

Schedule	Run on demand
-----------------	---------------

Output

Database	invoicedb
Prefix added to tables (optional)	
Create a single schema for each S3 path	false
Table level (optional)	
▶ Configuration options	

Back
Finish

Step 4: Run Crawler

1. Now we have to run our crawler to crawl all folders defined under S3 bucket. This will create a table from the crawler.

User preferences

Add crawler

Run crawler

Action

Filter by tags and attributes

Showing: 1 - 1

<input checked="" type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input checked="" type="checkbox"/>	s3-crawler		Ready		0 secs	0 secs	0	0

2. After the crawler is successfully executed, you may see 1 under the “Tables added” column.

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	s3-crawler		Ready	Logs	1 min	1 min	0	1

3. To see what is added in the table, click on the “Tables” option in the left side navigation menu.

AWS Glue

Data catalog
Databases
Tables
Connections
Crawlers
Classifiers
Schema registries
Schemas
Settings

ETL
AWS Glue Studio

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the :

[Add crawler](#) [Run crawler](#) [Action](#)

<input type="checkbox"/>	Name	Schedule	Status	Logs
<input type="checkbox"/>	s3-crawler		Ready	Logs

4. In the tables screen, you would see a table added. Click on the table to check more details.

Add tables

Action

Filter by attributes or search by keyword

Save view

Showing: 1 - 1

<div><input type="checkbox"/></div> Name	Database	Location	Classification	Last updated	Deprecated
<div><input type="checkbox"/></div> invoice_generator_220161	invoicedb	s3://invoice-generator-220161/	csv	15 February 2022 2:21 AM UT...	

5. Upon clicking the table name , you will see metadata fetched by the crawler from the underlying S3 bucket.

Name

invoice_generator_220161

Description

Database

invoicedb

Classification

csv

Location

s3://invoice-generator-220161/

Connection

Deprecated

No

Last updated

Tue Feb 15 02:21:34 GMT+530 2022

Input format

org.apache.hadoop.mapred.TextInputFormat

Output format

org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat

Serde serialization lib

org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Serde parameters

field.delim

,

sizeKey

847162

objectCount

1

UPDATED_BY_CRAWLER

s3-crawler

CrawlerSchemaSerializerVersion

1.0

Table properties

recordCount

9012

averageRecordSize

94

CrawlerSchemaDeserializerVersion

1.0

compressionType

none

columnsOrdered

true

areColumnsQuoted

false

delimiter

,

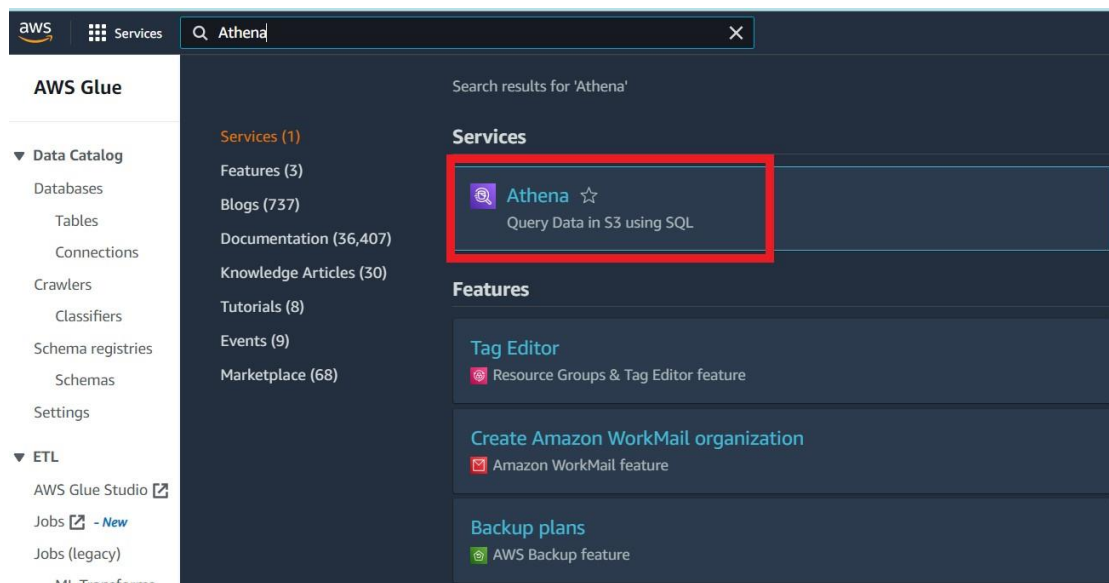
typeOfData

file

Analytics Using Athena

Amazon Athena is an **interactive query service** that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run. This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.

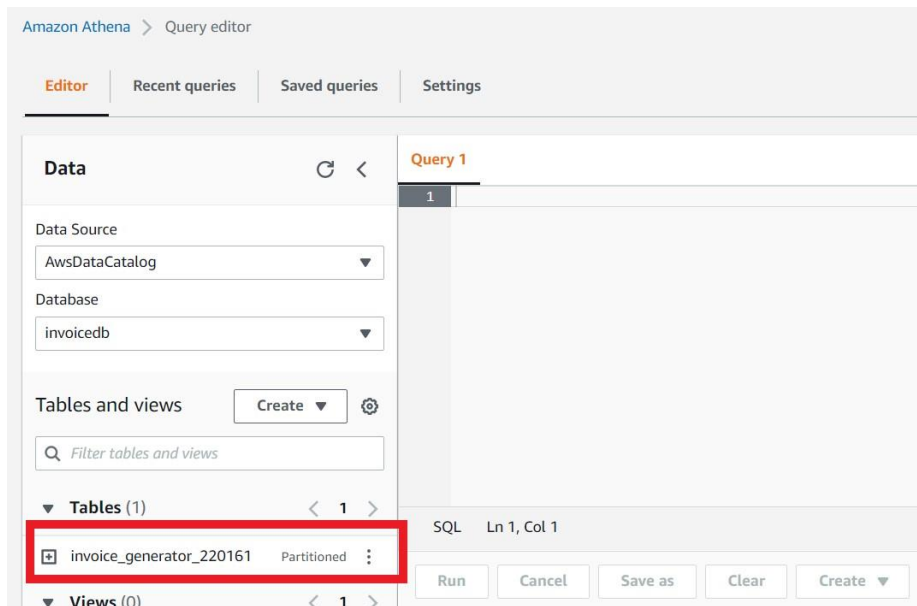
1. Search for “Athena” in Services.



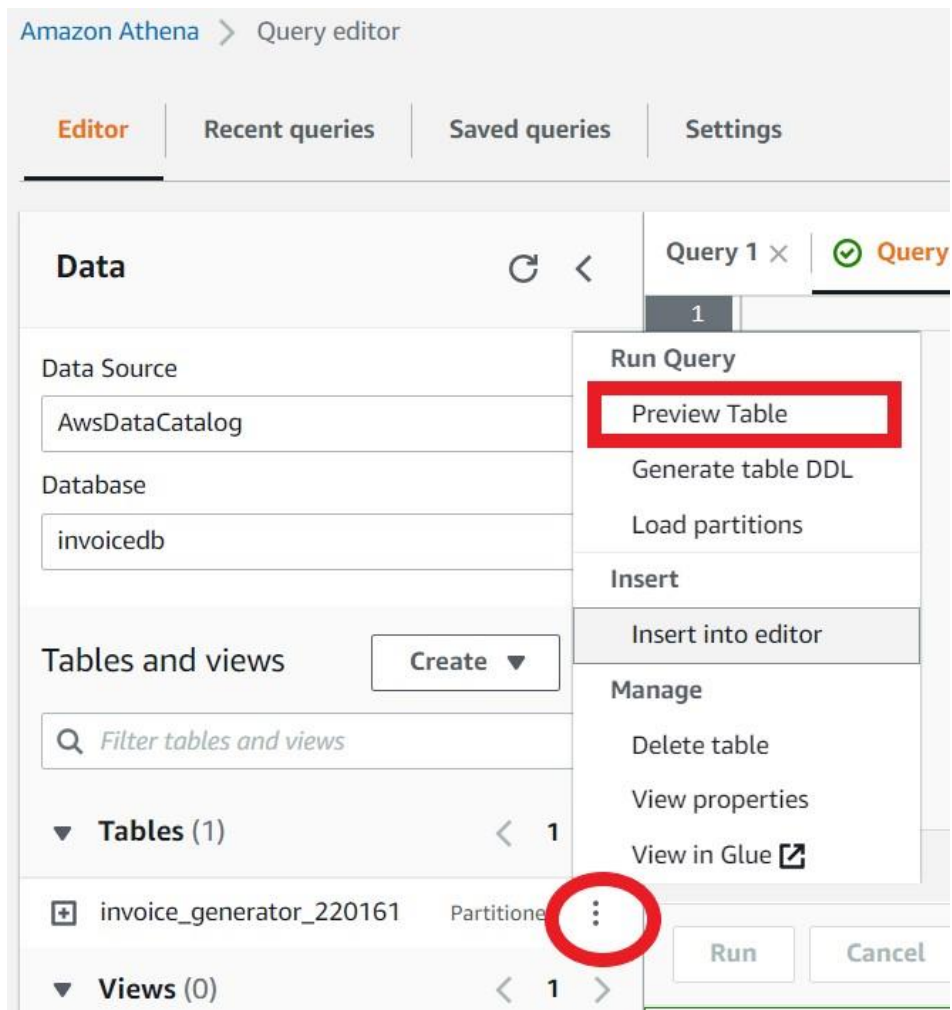
2. Click on the “Explore the query editor” button.



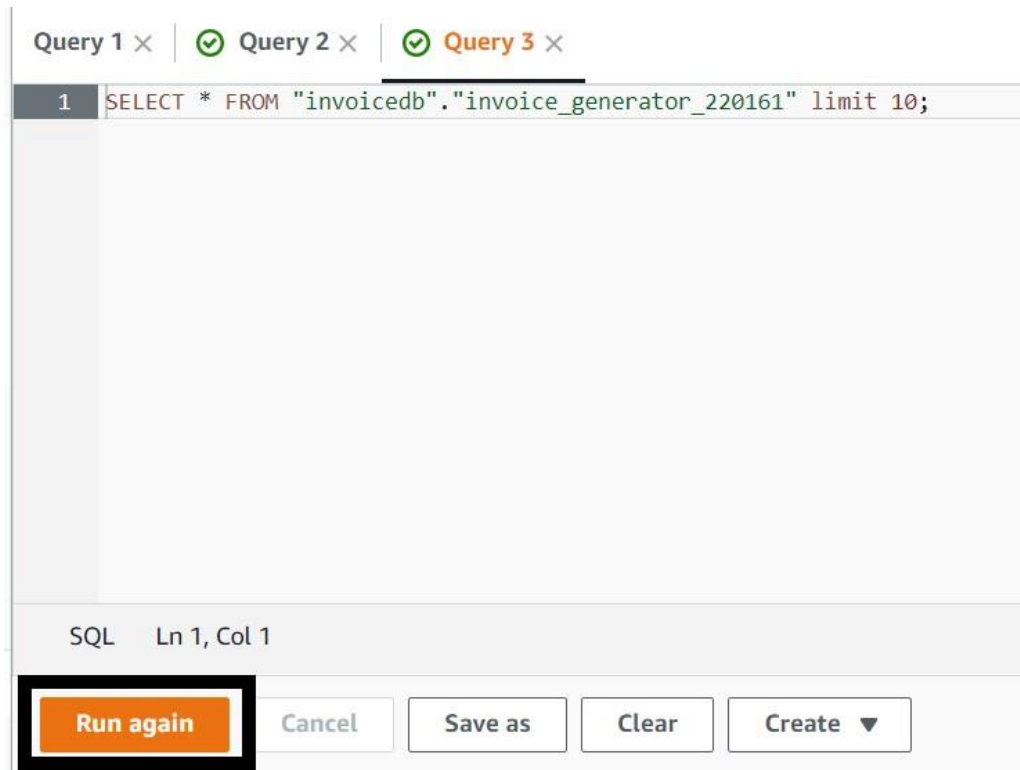
3. You will see the same table listed here that we created using Glue from S3 bucket.



4. Click on the “Preview Table” option to see the contents of the table.



5. It will create a SELECT SQL query to run against the data saved on S3. Click on the “Run again” button to run this query.



6. Wonderful, here are the results.

Results (10)

Copy

Download results

Search rows

<

1

>

#	col0	col1	col2	col3	col4
1	C538341	22943	CHRISTMAS LIGHTS 10 VINTAGE BAUBLES	-6	12/10/201
2	538346	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	24	12/10/201
3	538346	22923	FRIDGE MAGNETS LES ENFANTS ASSORTED	12	12/10/201
4	538307	22909	SET OF 20 VINTAGE CHRISTMAS NAPKINS	2	12/10/201
5	538309	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	1	12/10/201
6	538312	22576	SWALLOW WOODEN CHRISTMAS DECORATION	3	12/10/201