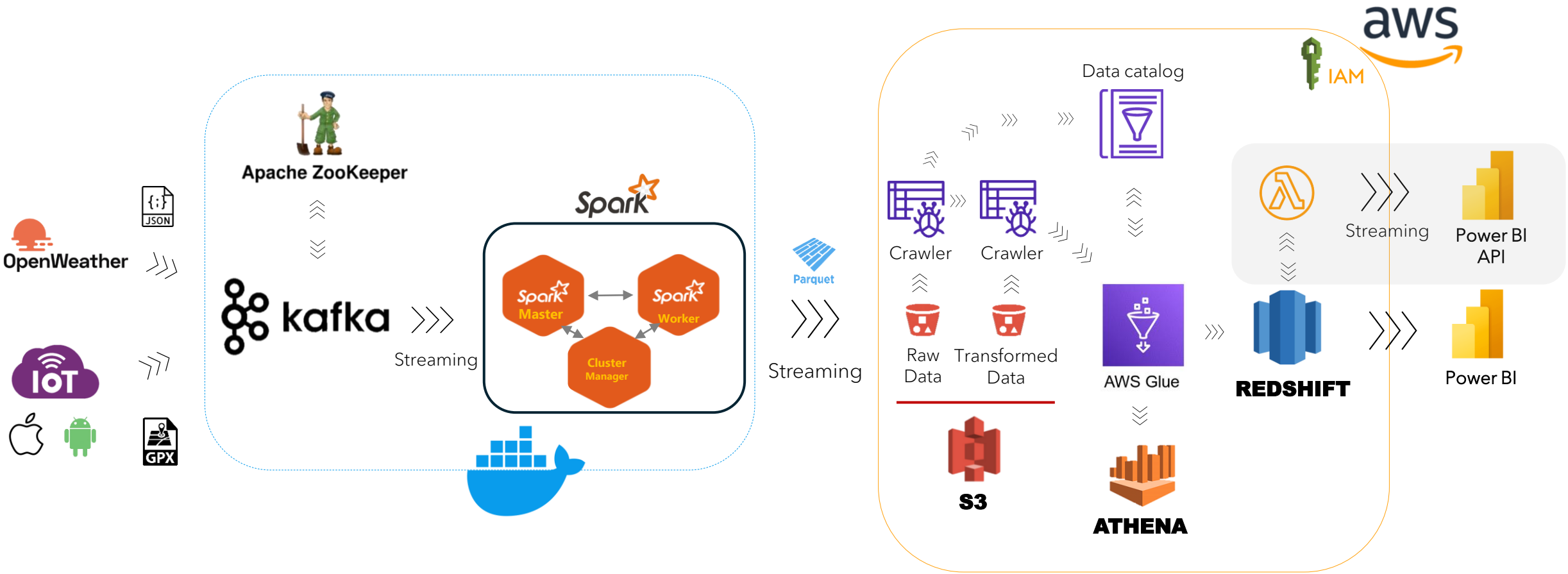
AWS Data Processing End-to-end pipeline | Street Drive lessons realtime monitoring

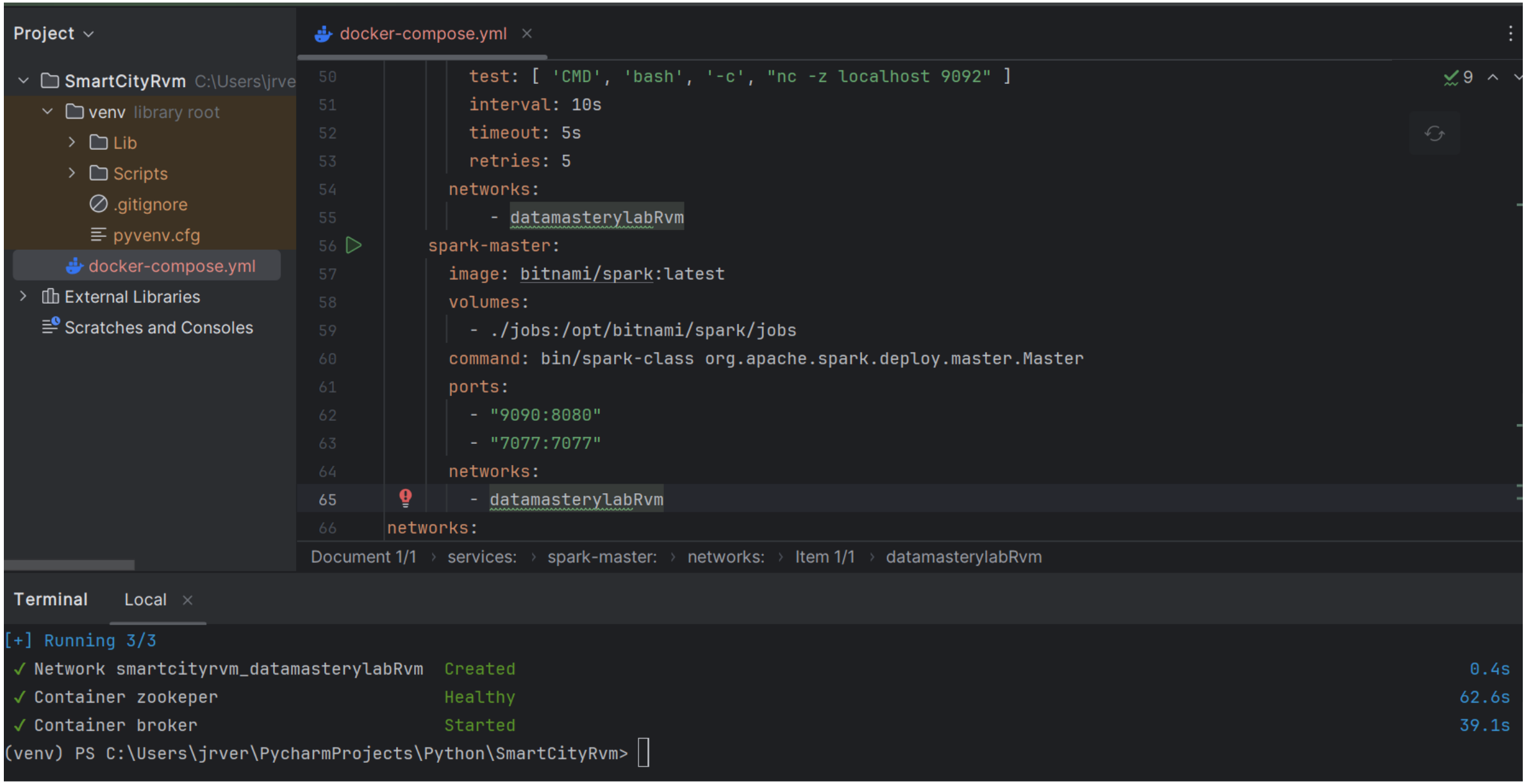


Project Setup

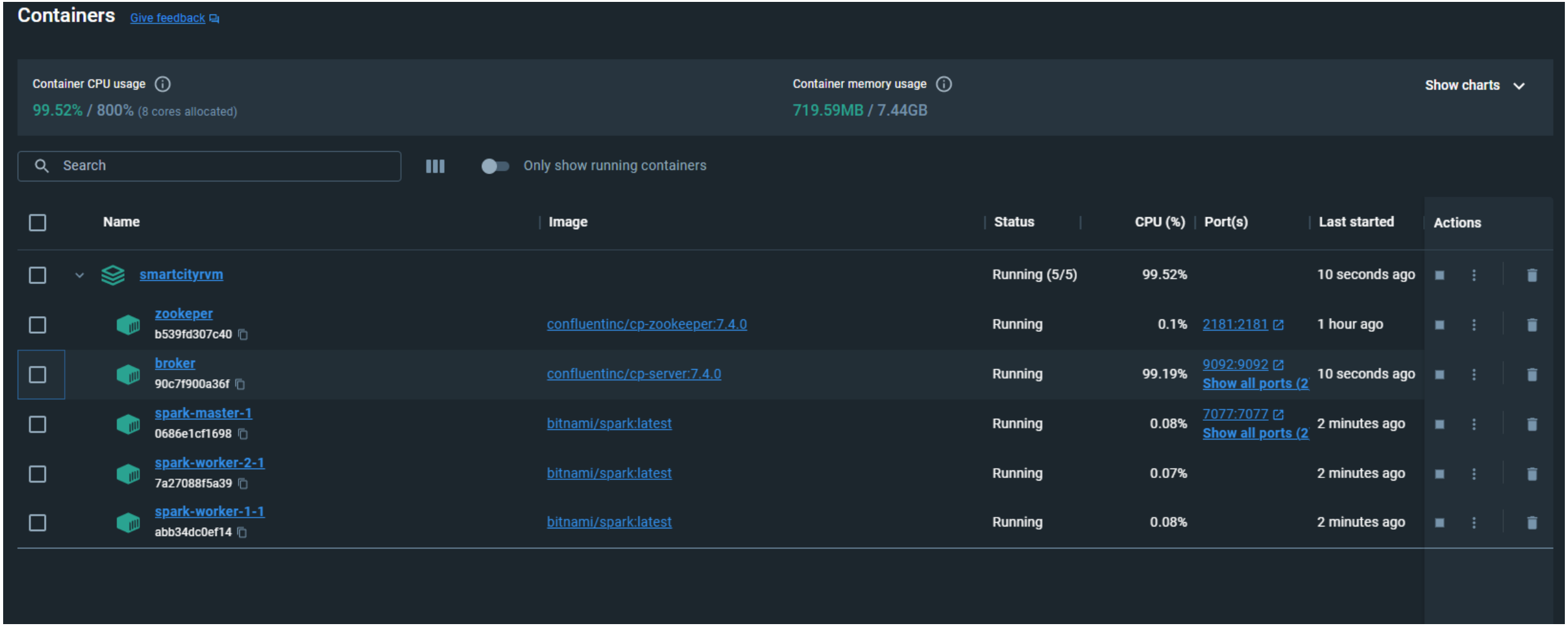
Docker

Kafka and Spark code and configuration

Creating zookeeper and Broker docker containers



Zookeeper, Broker, Spark master and workers docker containers creation



Checking Spark cluster creation

Spark Master at spark://172.18.0.3:7077

URL: spark://172.18.0.3:7077
Alive Workers: 2
Cores in use: 4 Total, 0 Used
Memory in use: 2.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240223184144-172.18.0.6-32827	172.18.0.6:32827	ALIVE	2 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20240223184145-172.18.0.5-46291	172.18.0.5:46291	ALIVE	2 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Openweather

Registration and endpoint Discovery

(after this go to the code and implement data extraction)

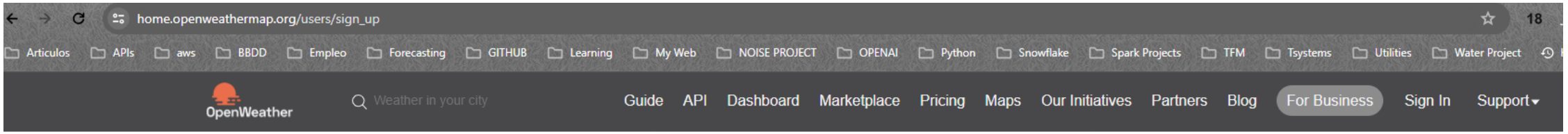
https://openweathermap.org/api

Free tier weather data:

https://openweathermap.org/current

Free tier air polution data:

https://openweathermap.org/api/air-pollution



Create New Account

Username

Enter email

Password

Repeat Password

We will use information you provided for management and administration purposes, and for keeping you informed by mail, telephone, email and SMS of other products and services from us and our partners. You can proactively manage your preferences or opt-out of communications with us at any time using Privacy Centre. You have the right to access your data held by us or to request your data to be deleted. For full details please see the OpenWeather Privacy Policy.

☐ I am 16 years old and over

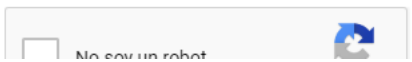
☐ I agree with Privacy Policy, Terms and conditions of sale and Websites terms and conditions of use

I consent to receive communications from OpenWeather Group of Companies and their partners:

☐ System news (API usage alert, system update, temporary system shutdown, etc)

☐ Product news (change to price, new product features, etc)

☐ Corporate news (our life, the launch of a new service, etc)



OpenWeather

Weather in your city

Guide API Dashboard Marketplace Pricing Maps Our Initiatives Partners Blog For Business Rvm Support

My services
My API keys
My payments
My profile
Logout

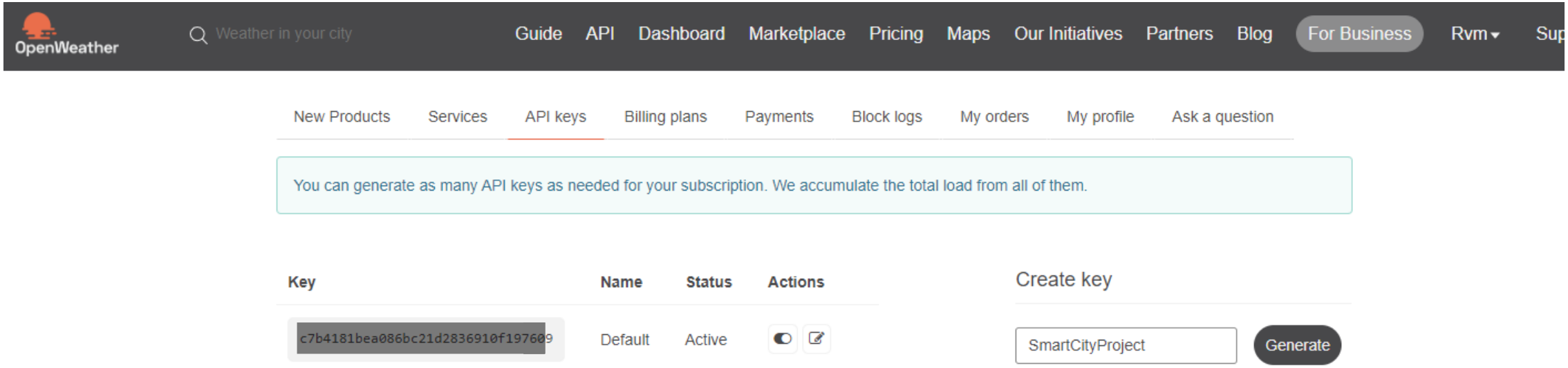
New Products | Services | API keys | Billing plans | Payments | Block logs | My orders | My profile | Ask a question

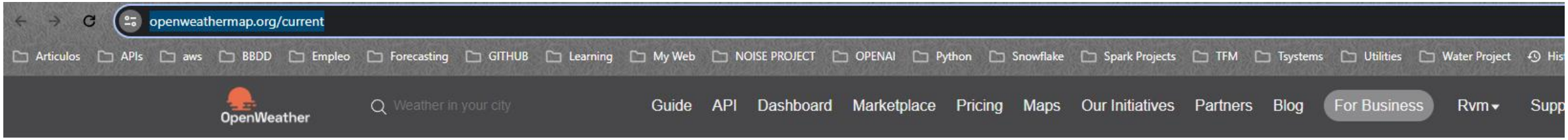
Historical weather for any location

Our new technology, Time Machine, has allowed us to enhance the data in the Historical Weather Collection.

- Historical weather data available for **ANY** coordinate

Api Key to call the services from your code





Current weather data

Product concept

Access current weather data for any location on Earth! We collect and process weather data from different sources such as global and local weather models, satellites, radars and a vast network of weather stations. Data is available in JSON, XML, or HTML format.

Call current weather data

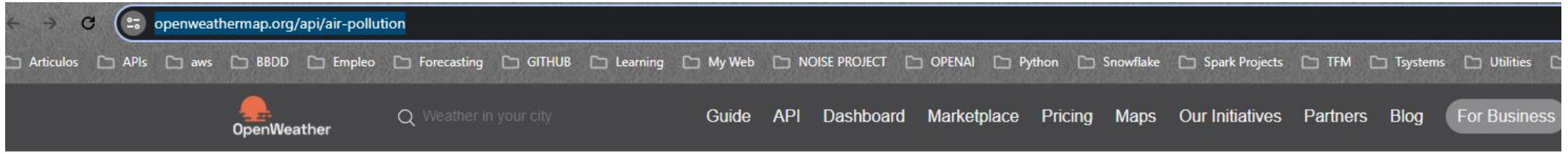
How to make an API call

API call

```
https://api.openweathermap.org/data/2.5/weather?lat=
{lat}&lon={lon}&appid={API key}
```

Parameters

`lat`	required	Latitude. If you need the geocoder to automatic convert city names and zip-codes to geo coordinates and the other way



Air Pollution API

Air Pollution API concept

Air Pollution API provides current, forecast and historical air pollution data for any coordinates on the globe.

Besides basic Air Quality Index, the API returns data about polluting gases, such as Carbon monoxide (CO), Nitrogen monoxide (NO), Nitrogen dioxide (NO_2), Ozone (O_3), Sulphur dioxide (SO_2), Ammonia (NH_3), and particulates ($PM_{2.5}$ and PM_{10}).

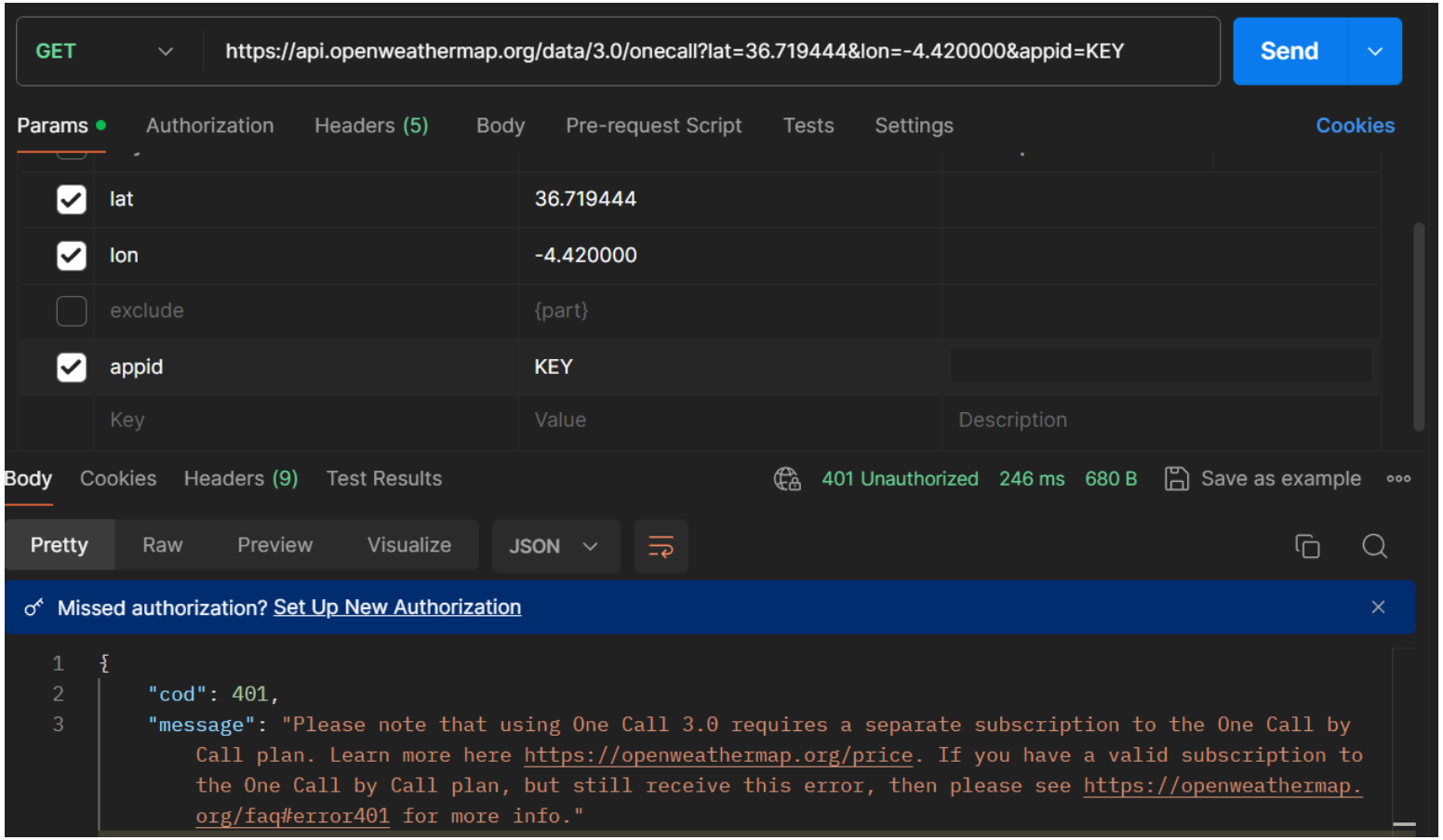
Air pollution forecast is available for 4 days with hourly granularity. Historical data is accessible from 27th November 2020.

Here is a description of OpenWeather scale for Air Quality Index levels:

Qualitative name	**Index**	**Pollutant concentration in μg/m^3**					
		SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO
Good	1	[0; 20)	[0; 40)	[0; 20)	[0; 10)	[0; 60)	[0; 4400)
Fair	2	[20; 80)	[40; 70)	[20; 50)	[10; 25)	[60; 100)	[4400; 9400)

OpenWeather API Endpoints Discovery
It takes up to 2 hours been able to get a 200.Ok response.

POSTMAN COLLECTION within my Repository



Frequently Asked Questions

General overview

About the company: what is OpenWeather? +

Get started

How to get an API key +

What products and types of data can I request? +

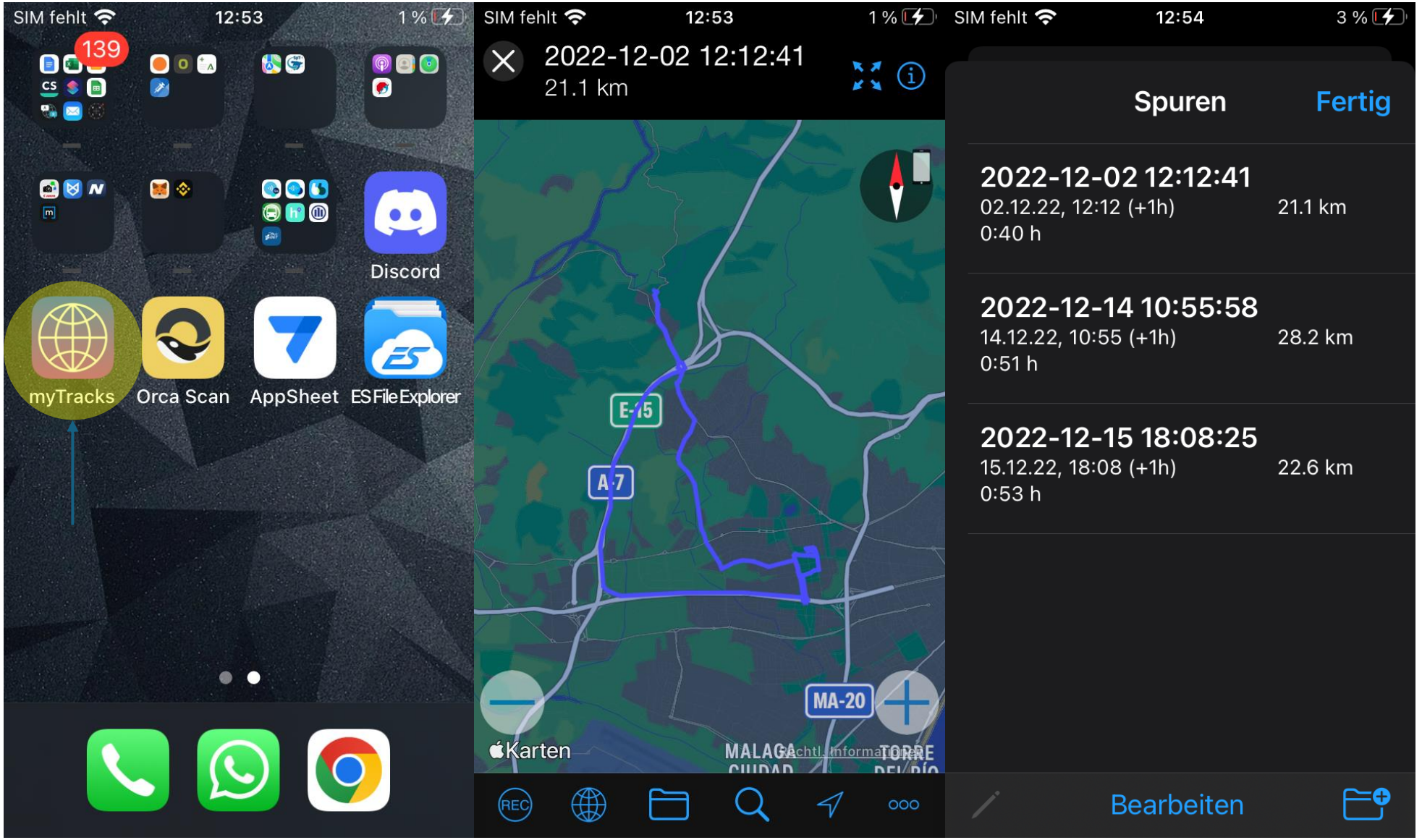
Do I need to activate my API key? +

> Your API key will be activated automatically, up to 2 hours after your successful registration. We invite you to read the API documentation that explains how to use our APIs.

Getting location information from smatphone
GPS Tracker: My Tracker

Note: this was my old Iphone SE with the language set in German, because I was there for five years. I still keep it and I used it to track my Motorbike driving lessons.

The records are from 2 years ago since then I was learning how to create an End-to-End project with the data.


SIM fehlt
12:53
1 %
139
Discord
myTracks
Orca Scan
AppSheet
ESFileExplorer
SIM fehlt
12:53
1 %
2022-12-02 12:12:41
21.1 km
E-15
A-7
MA-20
Karten
MALAGA
SIM fehlt
12:54
3 %
Spuren
Fertig
2022-12-02 12:12:41
02.12.22, 12:12 (+1h)
21.1 km
0:40 h
2022-12-14 10:55:58
14.12.22, 10:55 (+1h)
28.2 km
0:51 h
2022-12-15 18:08:25
15.12.22, 18:08 (+1h)
22.6 km
0:53 h
Bearbeiten

AWS

- **Create buckets and its Public policies**
  Empty buckets, the folders will be created as you run the code

- **Create access key**

Amazon S3 > Buckets > Create bucket

Create bucket Info

Buckets are containers for data stored in S3.

General configuration

AWS Region

Europe (Paris) eu-west-3

Bucket name Info

spark-streaming-data-rvmBKT

Bucket name must be unique within the global namespace and follow the bucket naming rules. See rules for bucket naming

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Choose bucket

Format: s3://bucket/prefix

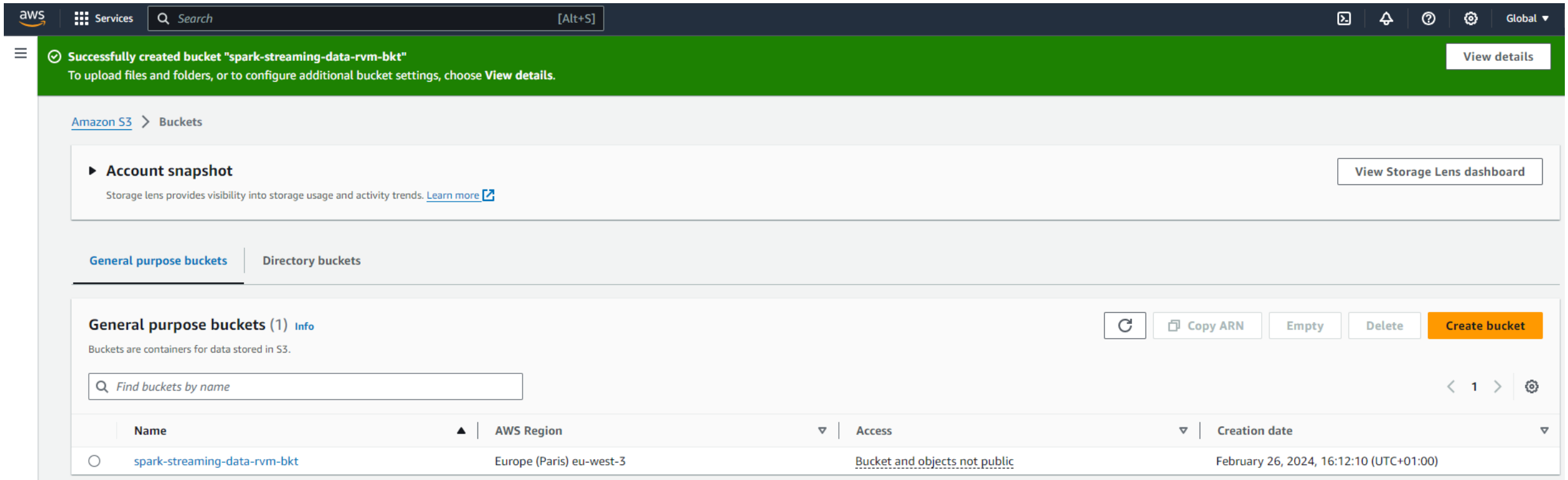
Object Ownership Info

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

- (●) 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 enforced


aws
Services
Search
[Alt+S]
Global
Successfully created bucket "spark-streaming-data-rvm-bkt"
To upload files and folders, or to configure additional bucket settings, choose View details.
View details
Amazon S3
Buckets
Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more
View Storage Lens dashboard
General purpose buckets
Directory buckets
General purpose buckets (1) Info
Buckets are containers for data stored in S3.
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
1
Name
AWS Region
Access
Creation date
spark-streaming-data-rvm-bkt
Europe (Paris) eu-west-3
Bucket and objects not public
February 26, 2024, 16:12:10 (UTC+01:00)

Successfully edited Block Public Access settings for this bucket.

Amazon S3 > Buckets > spark-streaming-data-rvm-bkt

spark-streaming-data-rvm-bkt Info

Objects | Properties | Permissions | Metrics | Management | Access Points

Permissions overview

Access

Bucket and objects not public

Block public access (bucket settings)

Edit

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more

Block *all* public access

Off

▶ **Individual Block Public Access settings for this bucket**

Edit Block public access (bucket settings) Info

Block public access (bucket settings)

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more

- [ ] **Block *all* public access**

  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.

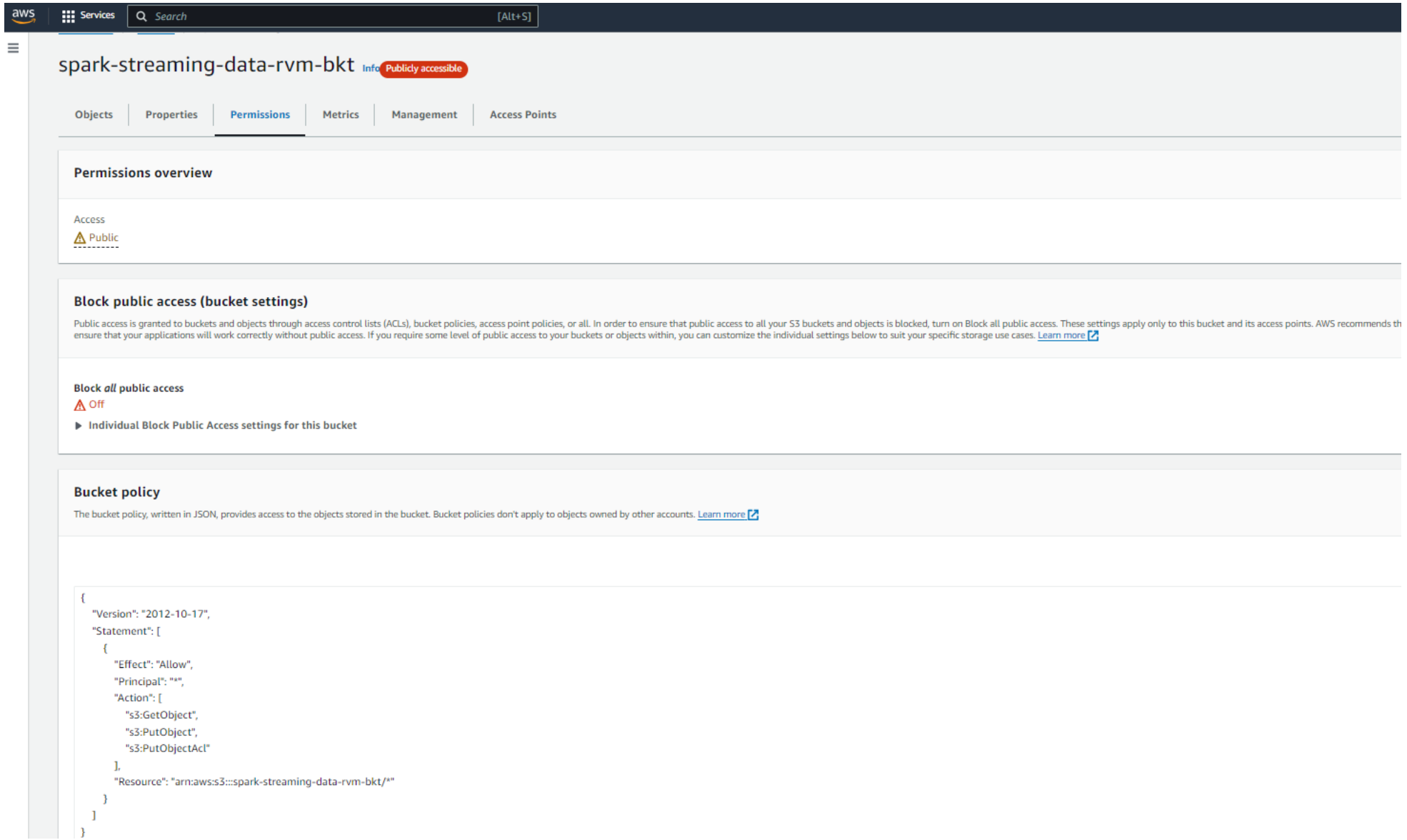
  - [ ] **Block public access to buckets and objects granted through *new* public bucket or access point policies**

    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.

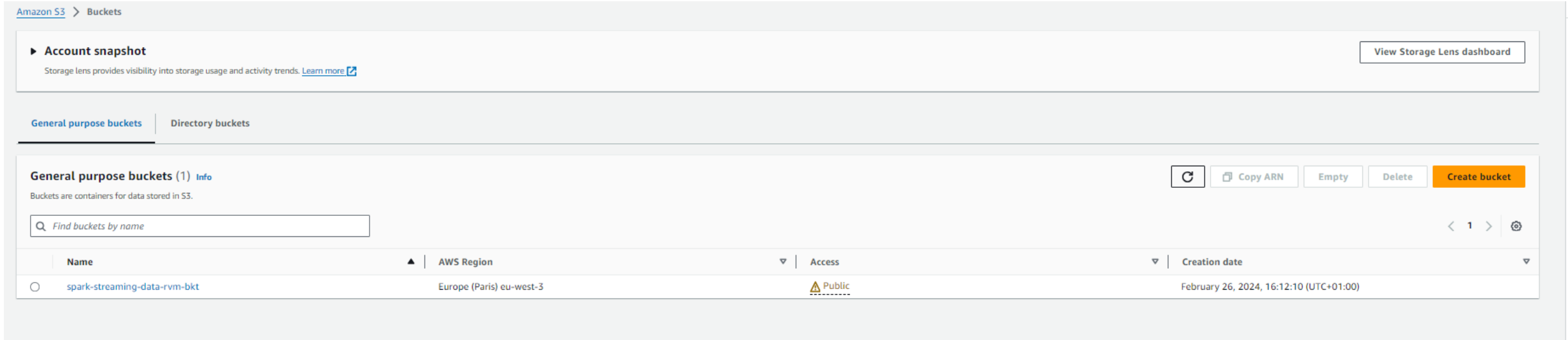
  - [ ] **Block public and cross-account access to buckets and objects through *any* public bucket or access point policies**

    S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.

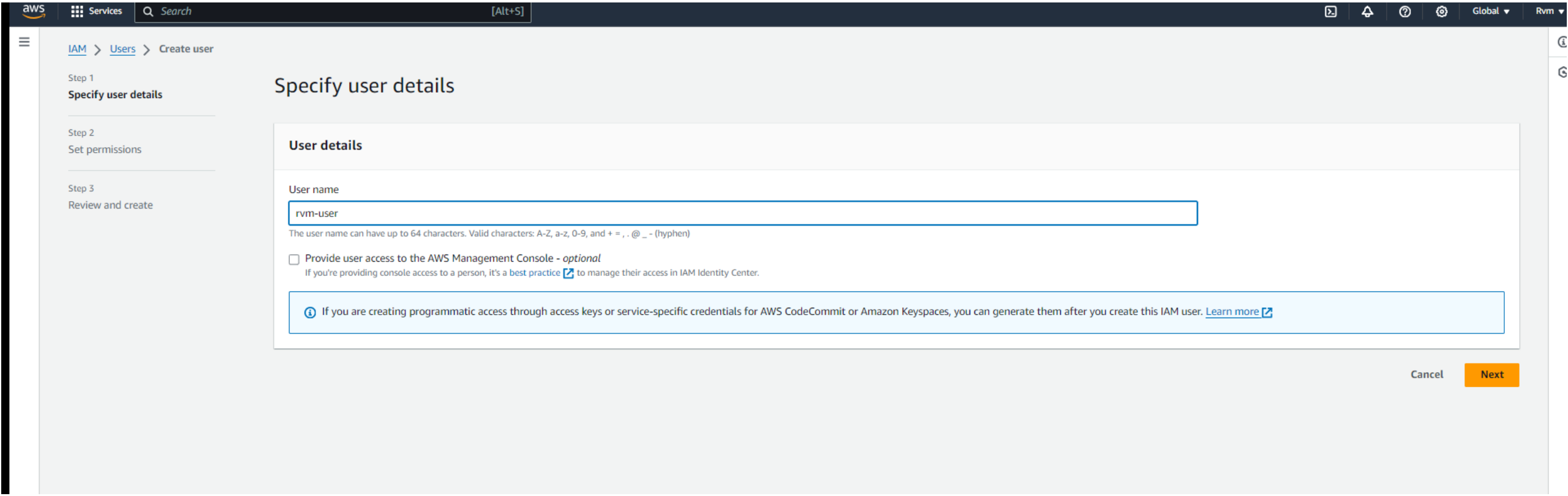
Cancel Save changes


aws
Services
Search
[Alt+S]
spark-streaming-data-rvm-bkt
Info
Publicly accessible
Objects
Properties
Permissions
Metrics
Management
Access Points
Permissions overview
Access
Public
Block public access (bucket settings)
Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends th
ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more
Block all public access
Off
Individual Block Public Access settings for this bucket
Bucket policy
The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts. Learn more
{
"Version": "2012-10-17",
"Statement": [
{
"Effect": "Allow",
"Principal": "*",
"Action": [
"s3:GetObject",
"s3:PutObject",
"s3:PutObjectAcl"
],
"Resource": "arn:aws:s3:::spark-streaming-data-rvm-bkt/*"
}
]
}

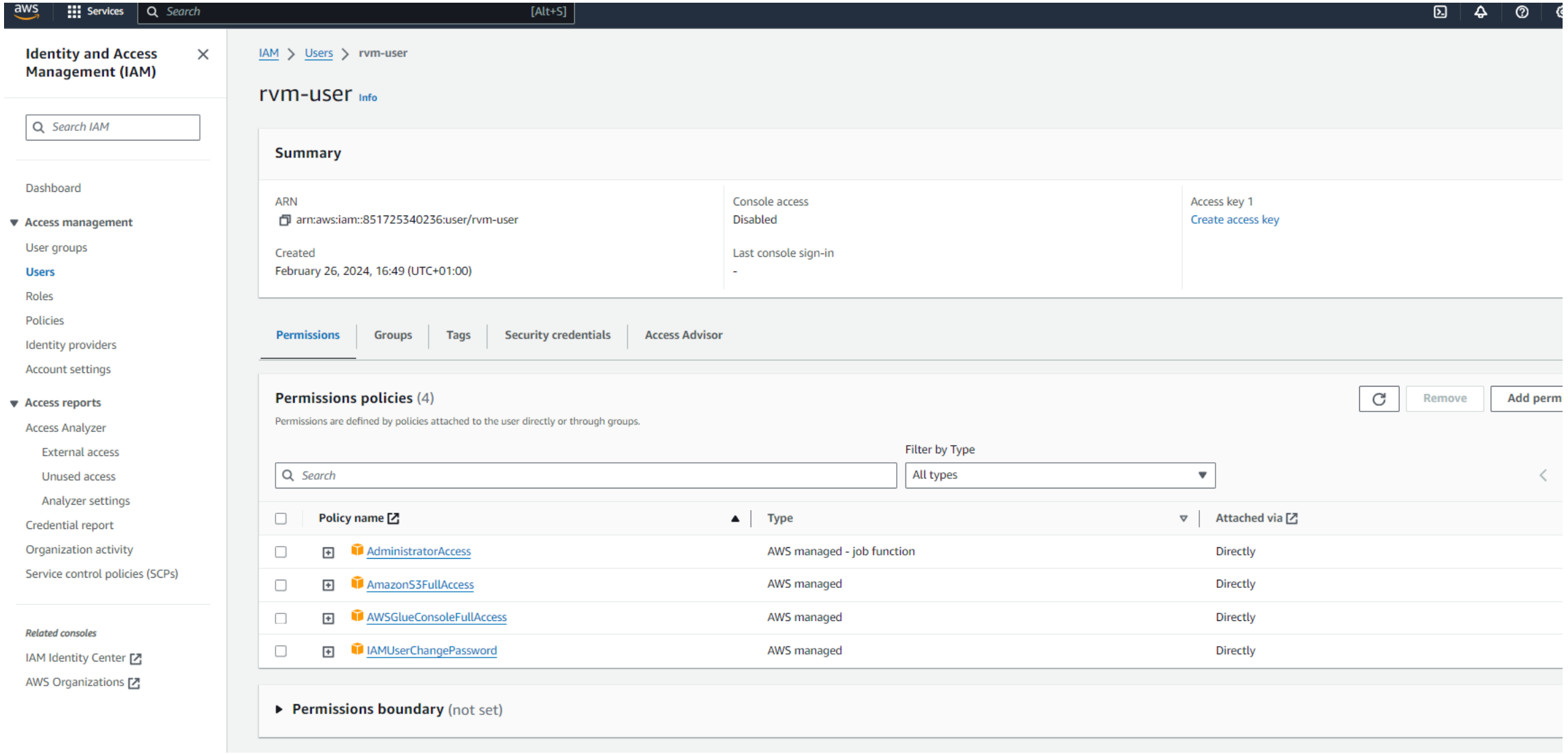
Refresh buckets to update de policy



Create an User

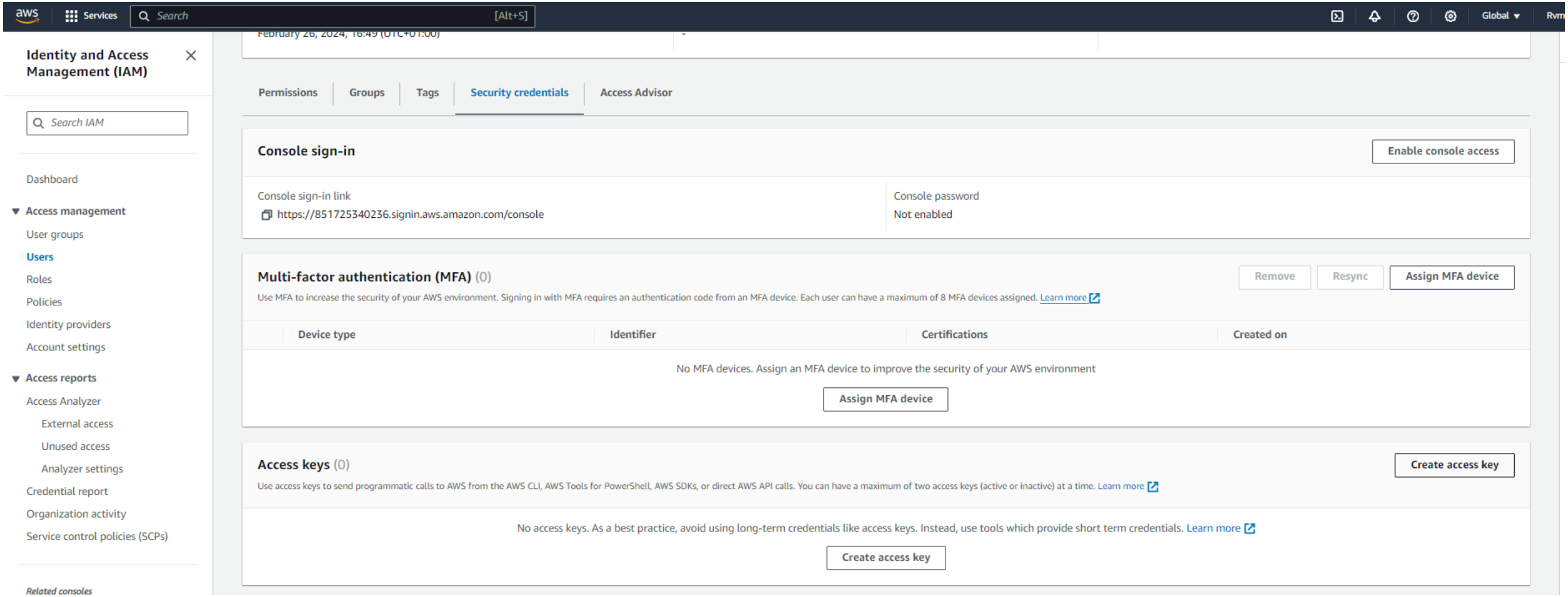


Add permissions to the user



Create Access key to the User.

That enables to load the data from spark (Docker) to the S3 buckets.
The keys will be passed in arguments withtin the code



Step 1
Access key best practices & alternatives

Step 2 - *optional*
Set description tag

Step 3
Retrieve access keys

Access key best practices & alternatives Info

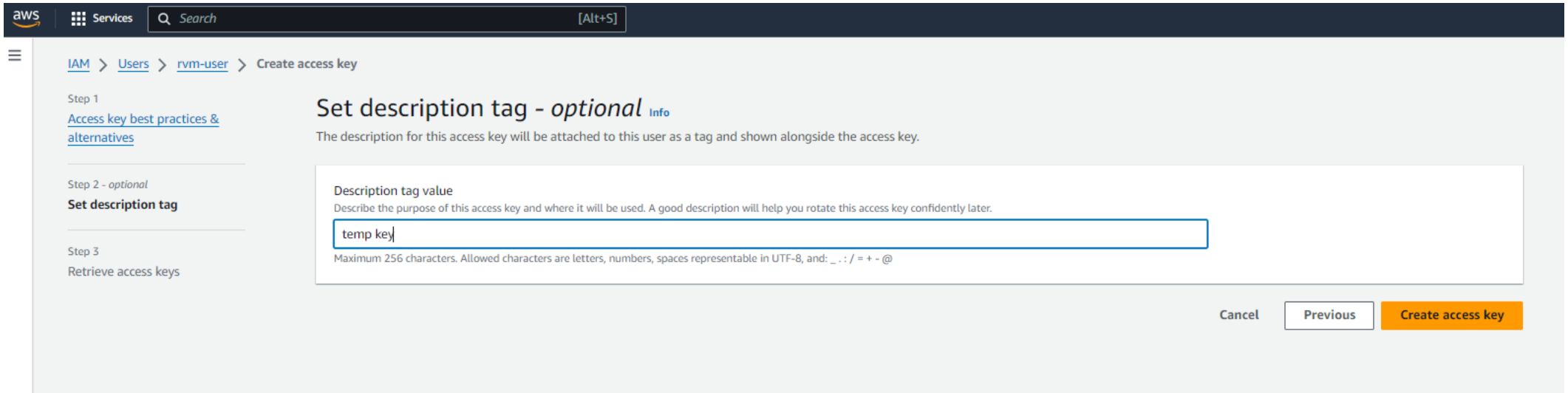
Avoid using long-term credentials like access keys to improve your security. Consider the following use cases and alternatives.

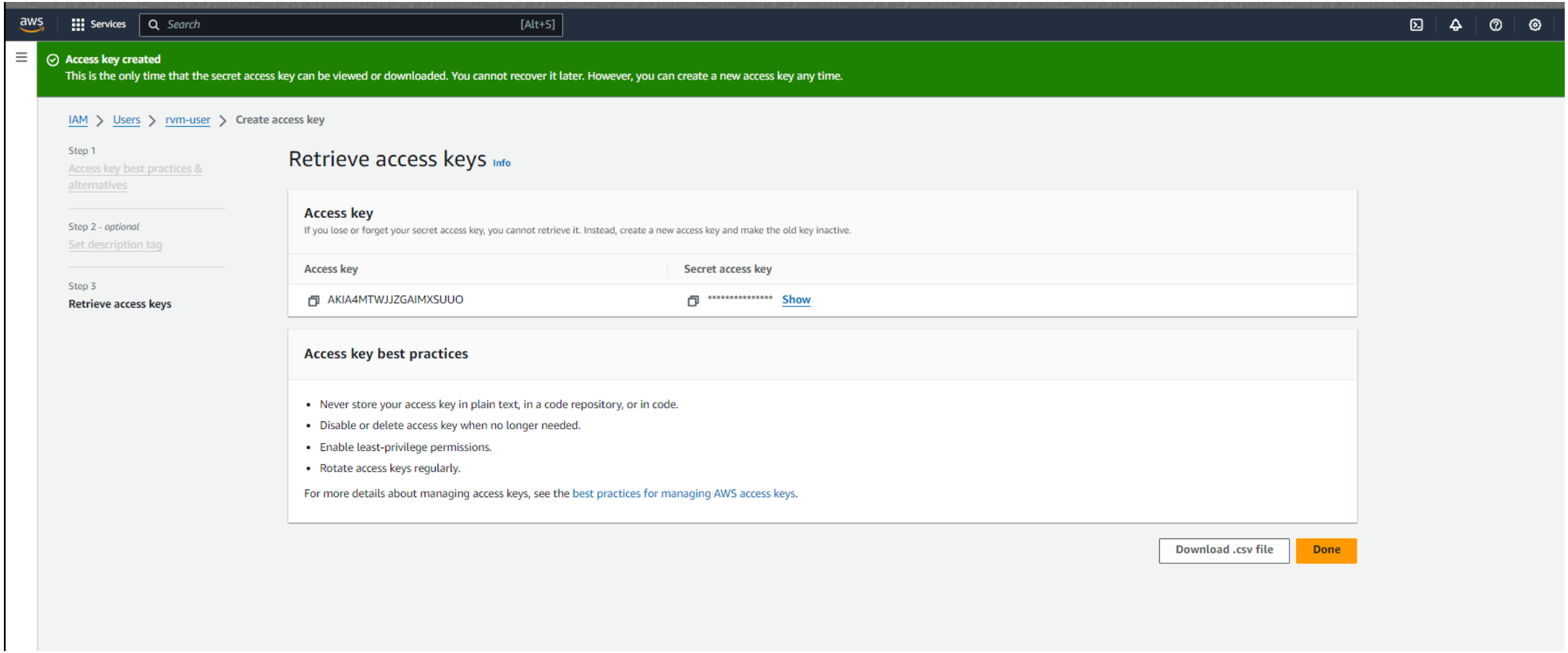
Use case

- ○ Command Line Interface (CLI)
  You plan to use this access key to enable the AWS CLI to access your AWS account.
- ○ Local code
  You plan to use this access key to enable application code in a local development environment to access your AWS account.
- ○ Application running on an AWS compute service
  You plan to use this access key to enable application code running on an AWS compute service like Amazon EC2, Amazon ECS, or AWS Lambda to access your AWS account.
- ○ Third-party service
  You plan to use this access key to enable access for a third-party application or service that monitors or manages your AWS resources.
- ◉ Application running outside AWS
  You plan to use this access key to authenticate workloads running in your data center or other infrastructure outside of AWS that needs to access your AWS resources.
- ○ Other
  Your use case is not listed here.

⚠ **Alternative recommended**
Use IAM Roles Anywhere to generate temporary security credentials for non AWS workloads accessing AWS services. Learn more about providing access for non AWS workloads.

Cancel Next


aws
Services
Search
[Alt+S]
IAM > Users > rvm-user > Create access key
Step 1
Access key best practices & alternatives
Step 2 - optional
Set description tag
Step 3
Retrieve access keys
Set description tag - optional Info
The description for this access key will be attached to this user as a tag and shown alongside the access key.
Description tag value
Describe the purpose of this access key and where it will be used. A good description will help you rotate this access key confidently later.
temp key
Maximum 256 characters. Allowed characters are letters, numbers, spaces representable in UTF-8, and: _ . : / = + - @
Cancel
Previous
Create access key


aws
Services
Search
[Alt+S]
Access key created
This is the only time that the secret access key can be viewed or downloaded. You cannot recover it later. However, you can create a new access key any time.
IAM > Users > rvm-user > Create access key
Step 1
Access key best practices & alternatives
Step 2 - optional
Set description tag
Step 3
Retrieve access keys
Retrieve access keys Info
Access key
If you lose or forget your secret access key, you cannot retrieve it. Instead, create a new access key and make the old key inactive.
Access key
Secret access key
AKIA4MTWJJZGAIMXSUUO
********** Show
Access key best practices
Never store your access key in plain text, in a code repository, or in code.
Disable or delete access key when no longer needed.
Enable least-privilege permissions.
Rotate access keys regularly.
For more details about managing access keys, see the best practices for managing AWS access keys.
Download .csv file
Done

Trigger the Streaming process

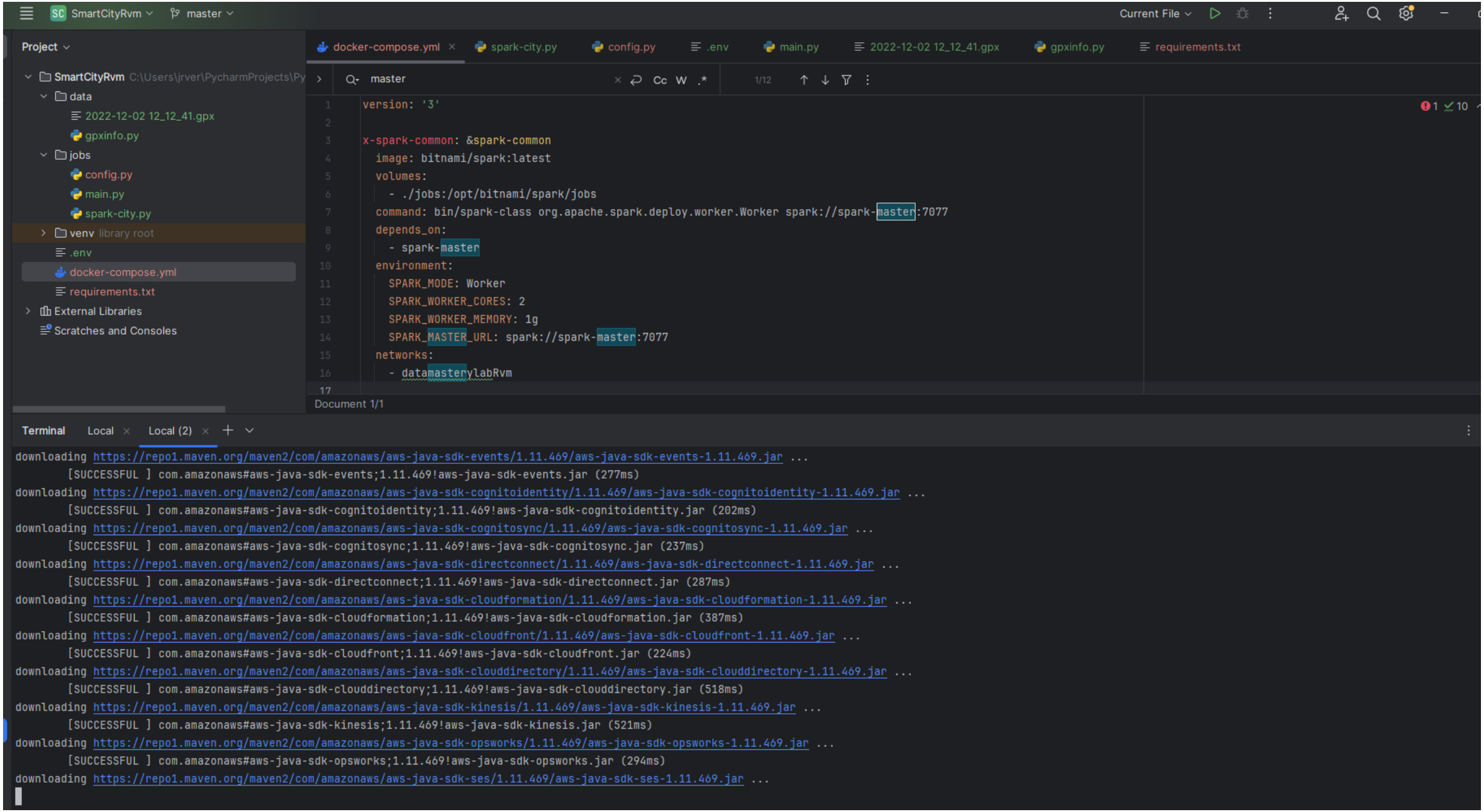
- Docker compose up –d
- Trigger kafka and send data to the topics.

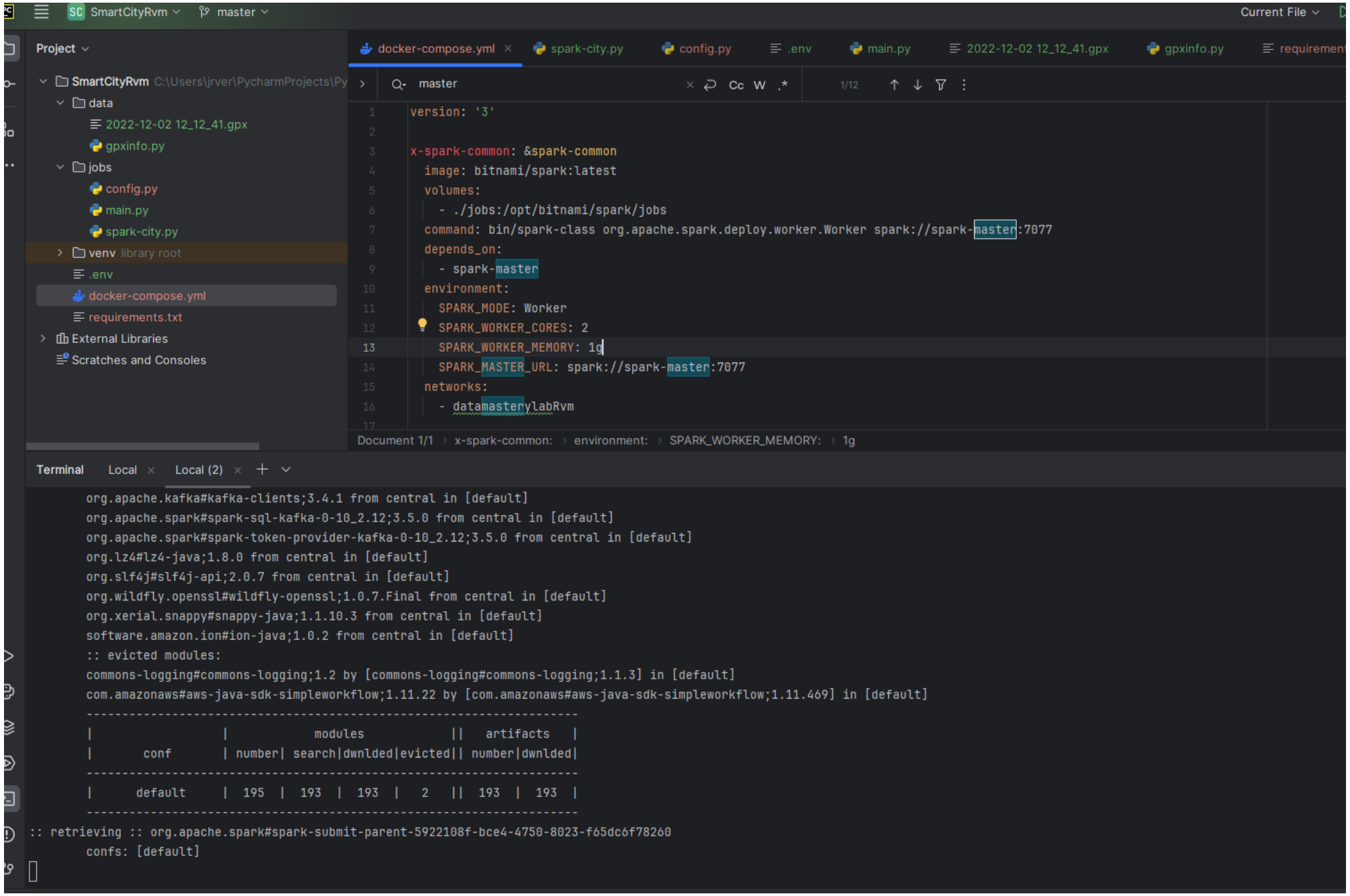
  Run: jobs/main.py

- Trigger Spark to consume topics and send information to S3.

  Run: docker exec -it smartcityrvm-spark-master-1 spark-submit `
--master spark://spark-master:7077 `
--packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.0,org.apache.hadoop:hadoop-aws:3.3.1,com.amazonaws:aws-java-sdk:1.11.469 `
jobs/spark-city.py

All the relevant commands are available into a .txt file in my Repository


SmartCityRvm
master
Current File
Project
SmartCityRvm C:\Users\jrver\PycharmProjects\Py
data
2022-12-02 12_12_41.gpx
gpxinfo.py
jobs
config.py
main.py
spark-city.py
venv library root
.env
docker-compose.yml
requirements.txt
External Libraries
Scratches and Consoles
docker-compose.yml
spark-city.py
config.py
.env
main.py
2022-12-02 12_12_41.gpx
gpxinfo.py
requirements.txt
master
1/12
version: '3'
x-spark-common: &spark-common
image: bitnami/spark:latest
volumes:
- ./jobs:/opt/bitnami/spark/jobs
command: bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077
depends_on:
- spark-master
environment:
SPARK_MODE: Worker
SPARK_WORKER_CORES: 2
SPARK_WORKER_MEMORY: 1g
SPARK_MASTER_URL: spark://spark-master:7077
networks:
- datamasterylabRvm
Document 1/1
Terminal
Local
Local (2)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-events/1.11.469/aws-java-sdk-events-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-events;1.11.469!aws-java-sdk-events.jar (277ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-cognitoidentity/1.11.469/aws-java-sdk-cognitoidentity-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-cognitoidentity;1.11.469!aws-java-sdk-cognitoidentity.jar (202ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-cognitosync/1.11.469/aws-java-sdk-cognitosync-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-cognitosync;1.11.469!aws-java-sdk-cognitosync.jar (237ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-directconnect/1.11.469/aws-java-sdk-directconnect-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-directconnect;1.11.469!aws-java-sdk-directconnect.jar (287ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-cloudformation/1.11.469/aws-java-sdk-cloudformation-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-cloudformation;1.11.469!aws-java-sdk-cloudformation.jar (387ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-cloudfront/1.11.469/aws-java-sdk-cloudfront-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-cloudfront;1.11.469!aws-java-sdk-cloudfront.jar (224ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-clouddirectory/1.11.469/aws-java-sdk-clouddirectory-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-clouddirectory;1.11.469!aws-java-sdk-clouddirectory.jar (518ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-kinesis/1.11.469/aws-java-sdk-kinesis-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-kinesis;1.11.469!aws-java-sdk-kinesis.jar (521ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-opsworks/1.11.469/aws-java-sdk-opsworks-1.11.469.jar ...
[SUCCESSFUL] com.amazonaws#aws-java-sdk-opsworks;1.11.469!aws-java-sdk-opsworks.jar (294ms)
downloading https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-ses/1.11.469/aws-java-sdk-ses-1.11.469.jar ...


SmartCityRvm
master
Current File
Project
SmartCityRvm C:\Users\jrver\PycharmProjects\Py
data
2022-12-02 12_12_41.gpx
gpxinfo.py
jobs
config.py
main.py
spark-city.py
venv library root
.env
docker-compose.yml
requirements.txt
External Libraries
Scratches and Consoles
docker-compose.yml
spark-city.py
config.py
.env
main.py
2022-12-02 12_12_41.gpx
gpxinfo.py
master
1/12
version: '3'
x-spark-common: &spark-common
image: bitnami/spark:latest
volumes:
- ./jobs:/opt/bitnami/spark/jobs
command: bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077
depends_on:
- spark-master
environment:
SPARK_MODE: Worker
SPARK_WORKER_CORES: 2
SPARK_WORKER_MEMORY: 1g
SPARK_MASTER_URL: spark://spark-master:7077
networks:
- datamasterylabRvm
Document 1/1 › x-spark-common: › environment: › SPARK_WORKER_MEMORY: › 1g
Terminal
Local
Local (2)
org.apache.kafka#kafka-clients;3.4.1 from central in [default]
org.apache.spark#spark-sql-kafka-0-10_2.12;3.5.0 from central in [default]
org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.5.0 from central in [default]
org.lz4#lz4-java;1.8.0 from central in [default]
org.slf4j#slf4j-api;2.0.7 from central in [default]
org.wildfly.openssl#wildfly-openssl;1.0.7.Final from central in [default]
org.xerial.snappy#snappy-java;1.1.10.3 from central in [default]
software.amazon.ion#ion-java;1.0.2 from central in [default]
:: evicted modules:
commons-logging#commons-logging;1.2 by [commons-logging#commons-logging;1.1.3] in [default]
com.amazonaws#aws-java-sdk-simpleworkflow;1.11.22 by [com.amazonaws#aws-java-sdk-simpleworkflow;1.11.469] in [default]
conf	modules	artifacts
	number	search	dwnlded	evicted	number	dwnlded
default	195	193	193	2	193	193
:: retrieving :: org.apache.spark#spark-submit-parent-5922108f-bce4-4750-8023-f65dc6f78260
confs: [default]

Spark Master at spark://172.19.0.2:7077

URL: spark://172.19.0.2:7077
Alive Workers: 2
Cores in use: 4 Total, 4 Used
Memory in use: 2.0 GiB Total, 2.0 GiB Used
Resources in use:
Applications: 1 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240226131746-172.19.0.4-35661	172.19.0.4:35661	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240226131746-172.19.0.5-36565	172.19.0.5:36565	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240226165545-0002 (kill)	SmartCityStreaming	4	1024.0 MiB		2024/02/26 16:55:45	spark	RUNNING	4 s

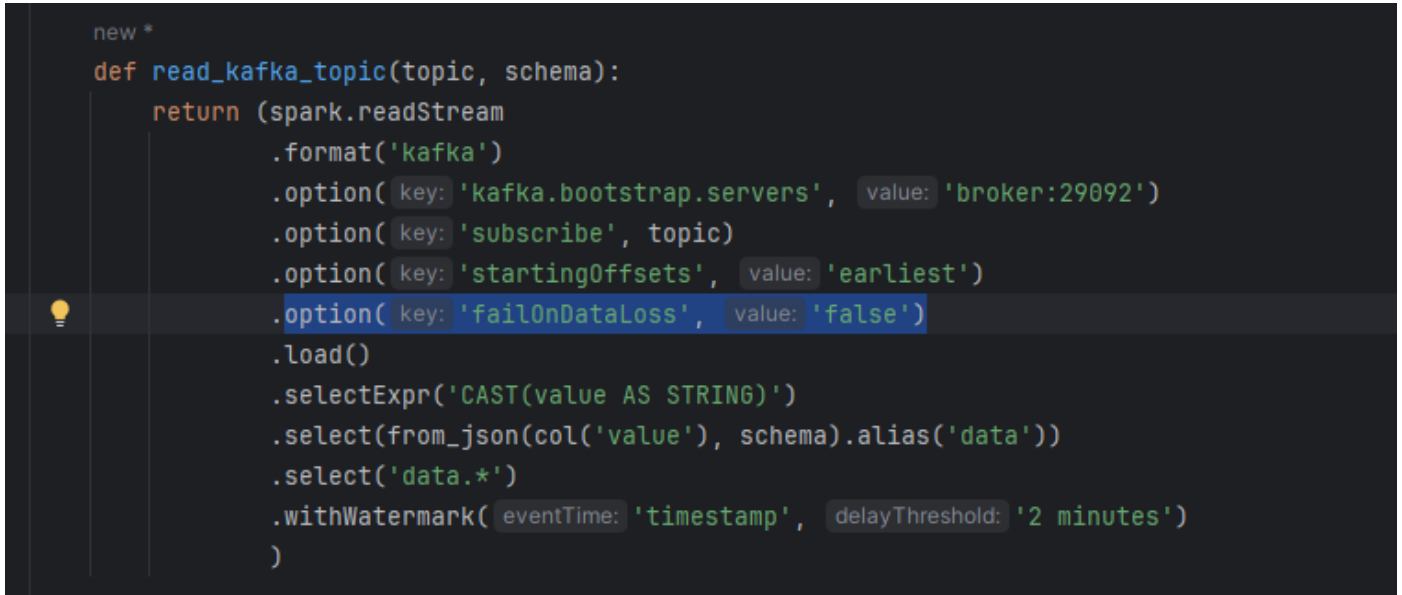
Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240226164242-0001	SmartCityStreaming	4	1024.0 MiB		2024/02/26 16:42:42	spark	FINISHED	6 s
app-20240226163625-0000	SmartCityStreaming	4	1024.0 MiB		2024/02/26 16:36:25	spark	FINISHED	2 s

SOME FAILED TESTS

To avoid the process to fail, **I needed to set this option “failOnDataloss” to “False”.**
Due to the Broker consistency and connection with Spark, it may occurs that few registers are lost, and this option doesnot allow the process to continue if there are some data missing by default.

“The ERROR:
ERROR MicroBatchExecution: Query [id = fac716d9-cdde-491c-bab8-9decd5dfa957, runId = 27662566-449f-41b6-965f-f88387f5d424] terminated with error
java.lang.IllegalStateException: Partition weather_data-0's offset was changed from 30 to 14, some data may have been missed.
Some data may have been lost because they are not available in Kafka any more; either the
data was aged out by Kafka or the topic may have been deleted before all the data in the
topic was processed. If you don't want your streaming query to fail on such cases, set the
source option "failOnDataLoss" to "false". “



Once corrected the issue before, the streaming process from spark to S3 runs properly

```
24/02/26 19:04:37 INFO StandaloneAppClient$ClientEndpoint: Executor updated: app-20240226190437-0000/0 is now RUNNING
24/02/26 19:04:37 INFO StandaloneAppClient$ClientEndpoint: Executor updated: app-20240226190437-0000/1 is now RUNNING
24/02/26 19:04:37 INFO StandaloneSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
24/02/26 19:04:42 WARN MetricsConfig: Cannot locate configuration: tried hadoop-metrics2-s3a-file-system.properties,hadoop-metrics2.properties
24/02/26 19:04:45 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
24/02/26 19:04:50 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
24/02/26 19:04:52 WARN AdminClientConfig: These configurations '[key.deserializer, value.deserializer, enable.auto.commit, max.poll.records, auto.offset.reset]' were supplied but are not used yet.
24/02/26 19:04:54 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
24/02/26 19:04:54 WARN AdminClientConfig: These configurations '[key.deserializer, value.deserializer, enable.auto.commit, max.poll.records, auto.offset.reset]' were supplied but are not used yet.
24/02/26 19:04:58 WARN AdminClientConfig: These configurations '[key.deserializer, value.deserializer, enable.auto.commit, max.poll.records, auto.offset.reset]' were supplied but are not used yet.
24/02/26 19:06:06 WARN FileStreamSinkLog: Compacting took 2779 ms for compact batch 9
24/02/26 19:06:06 WARN FileStreamSinkLog: Compacting took 2859 ms for compact batch 9
24/02/26 19:06:06 WARN FileStreamSinkLog: Compacting took 2873 ms for compact batch 9
24/02/26 19:07:48 WARN FileStreamSinkLog: Compacting took 6337 ms for compact batch 19
24/02/26 19:07:49 WARN FileStreamSinkLog: Compacting took 6139 ms for compact batch 19
24/02/26 19:07:50 WARN FileStreamSinkLog: Compacting took 6243 ms for compact batch 19
```

Spark Master at spark://172.18.0.3:7077

3.5.1

URL: spark://172.18.0.3:7077
Alive Workers: 2
Cores in use: 4 Total, 4 Used
Memory in use: 2.0 GiB Total, 2.0 GiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240227075319-172.18.0.4-45837	172.18.0.4:45837	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240227075319-172.18.0.5-33429	172.18.0.5:33429	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240227111454-0000 (kill)	SmartCityStreaming	4	1024.0 MiB		2024/02/27 11:14:54	spark	RUNNING	6.0 min

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Spark Master at spark://172.18.0.3:7077

3.5.1

URL: spark://172.18.0.3:7077
Alive Workers: 2
Cores in use: 4 Total, 4 Used
Memory in use: 2.0 GiB Total, 2.0 GiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240227075319-172.18.0.4-45837	172.18.0.4:45837	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240227075319-172.18.0.5-33429	172.18.0.5:33429	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240227111454-0000 (kill)	SmartCityStreaming	4	1024.0 MiB		2024/02/27 11:14:54	spark	RUNNING	16 min

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Spark Master at spark://172.18.0.3:7077

3.5.1

URL: spark://172.18.0.3:7077
Alive Workers: 2
Cores in use: 4 Total, 4 Used
Memory in use: 2.0 GiB Total, 2.0 GiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240227075319-172.18.0.4-45837	172.18.0.4:45837	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240227075319-172.18.0.5-33429	172.18.0.5:33429	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	

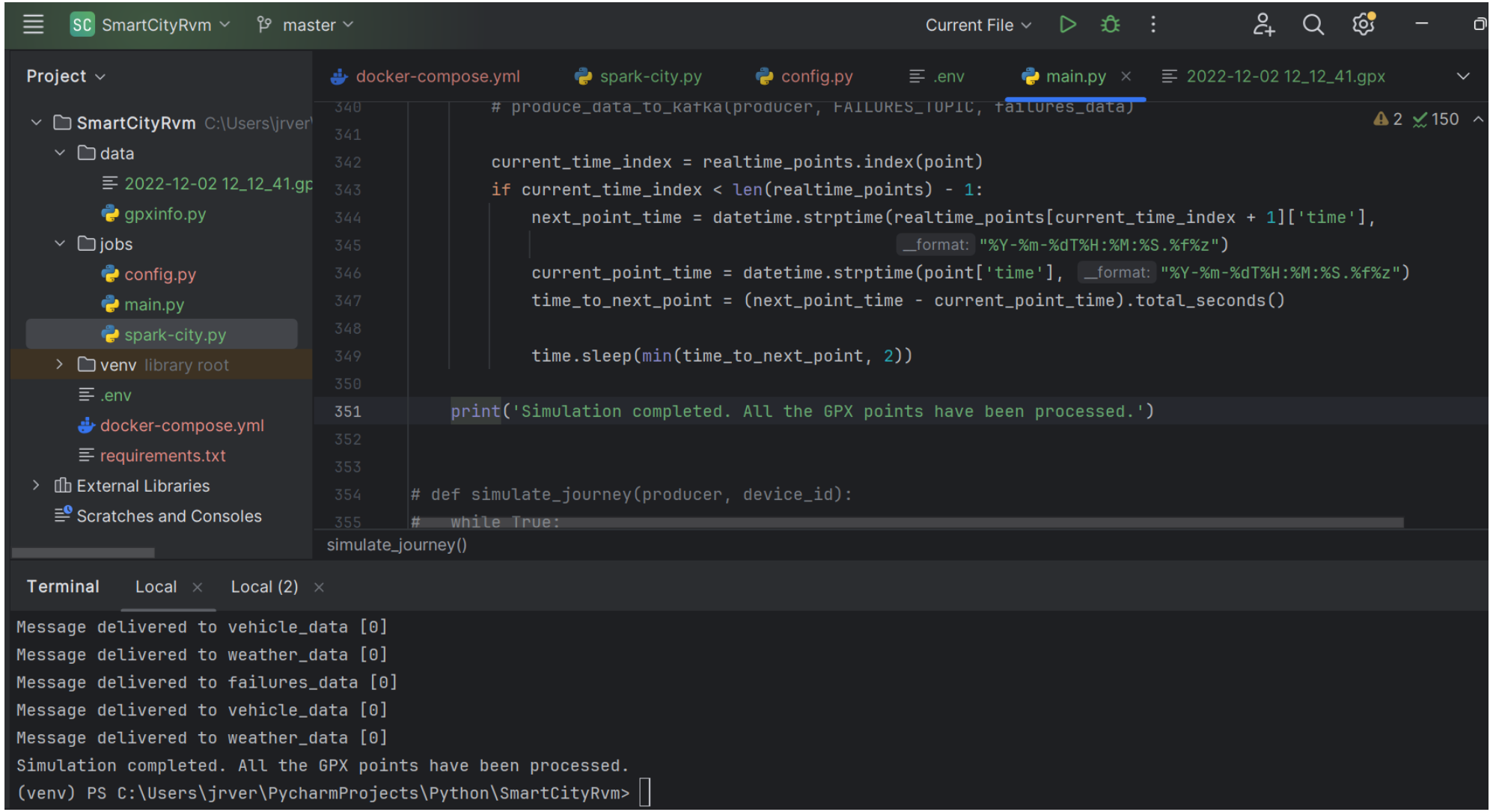
Running Applications (1)

Application ID		Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240227111454-0000	(kill)	SmartCityStreaming	4	1024.0 MiB		2024/02/27 11:14:54	spark	RUNNING	35 min

Completed Applications (0)

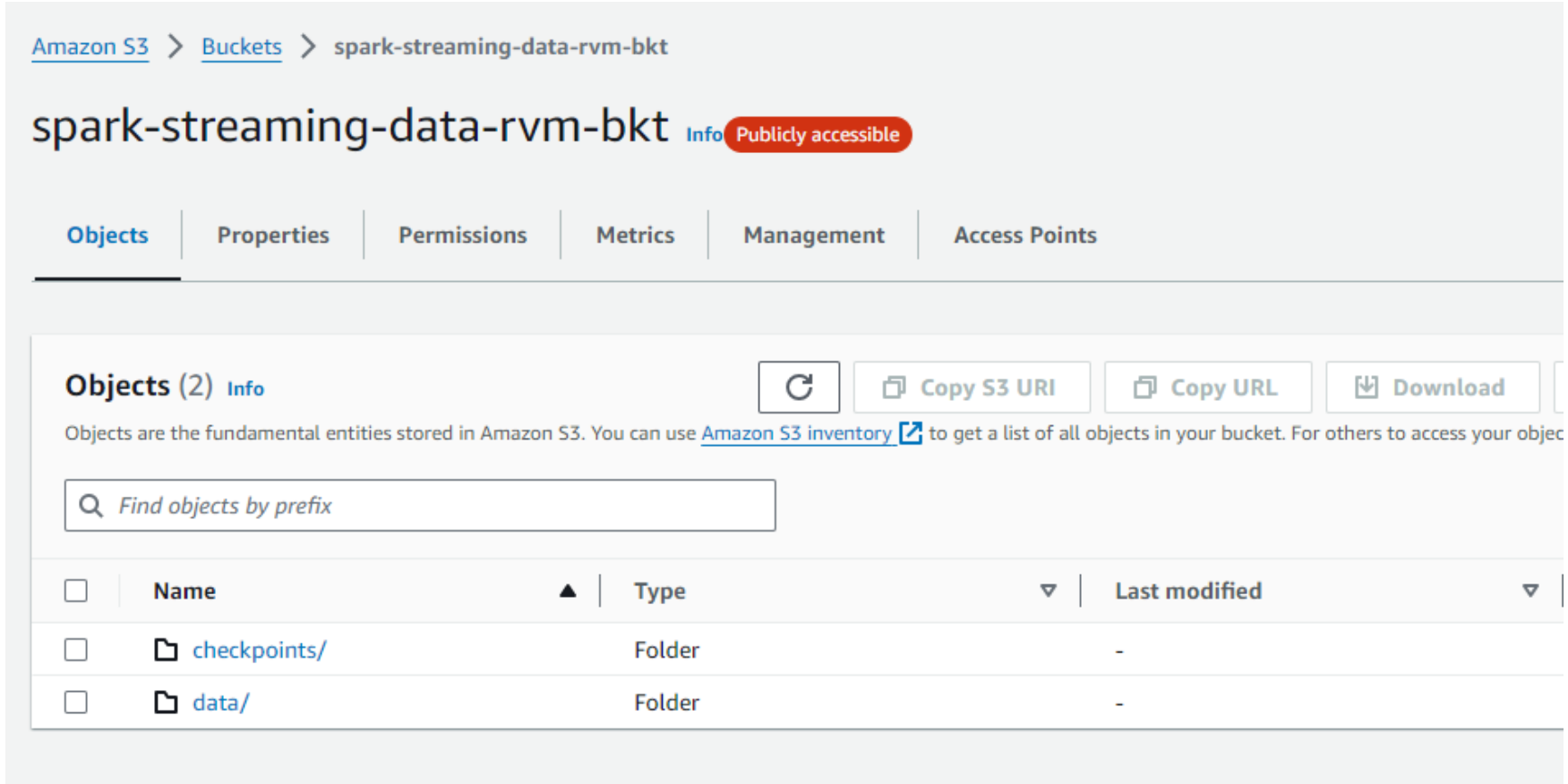
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

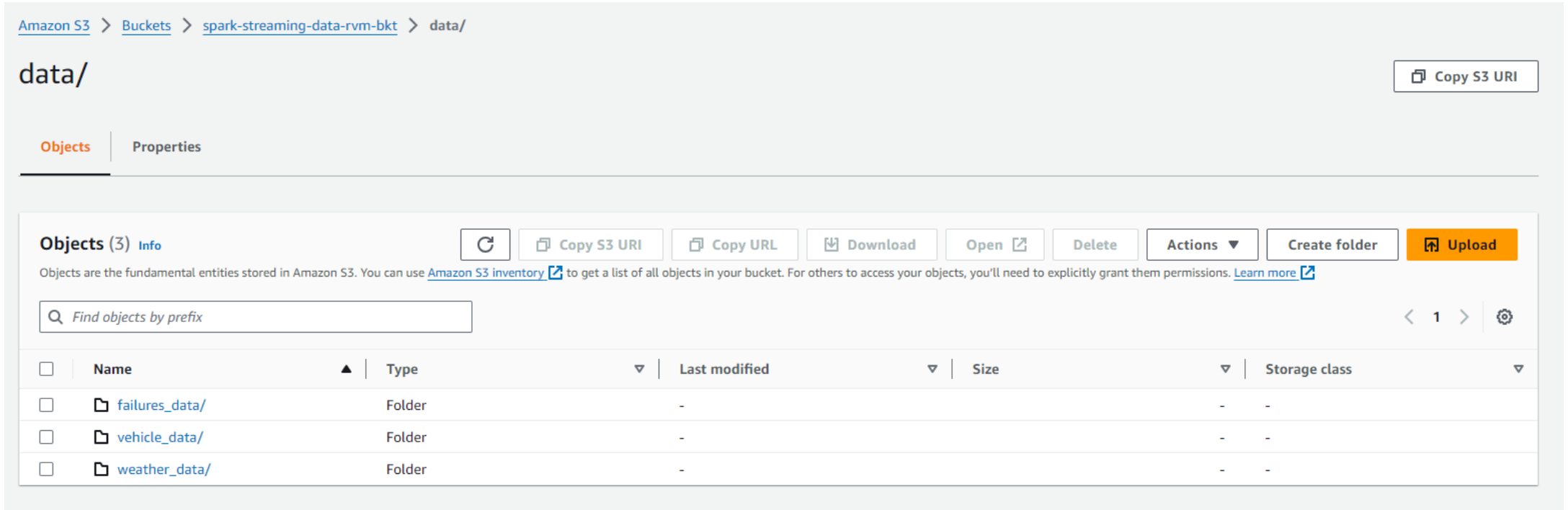
AFTER AROUND 35 MINUTES, ALL THE LINES FROM THE .GPX FILES WERE PROCESSED, FIRST TO KAFKA INTO TOPICS AND AFTER CONSUMED BY SPARK TO S3

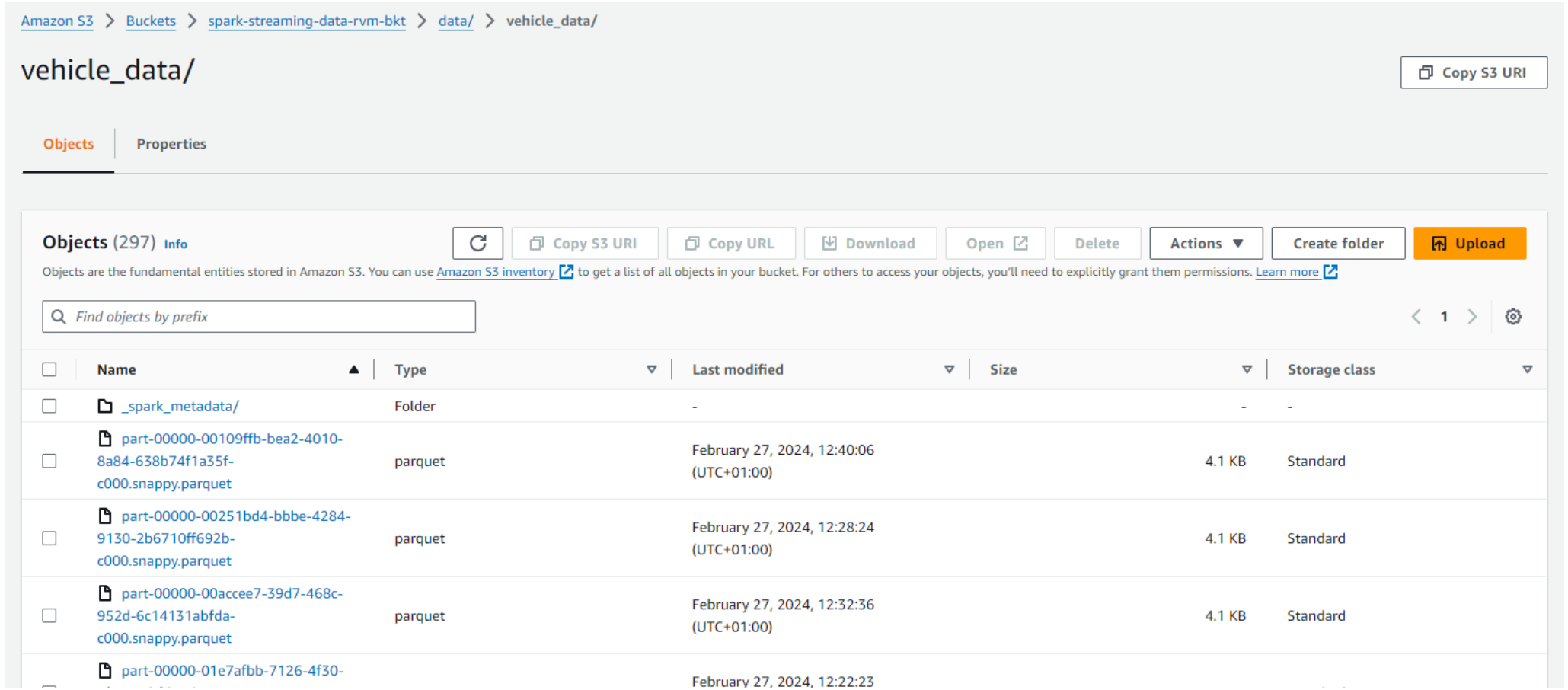


AWS

- S3 Buckets contain all the data (.parquet) and folders defined in the code


Amazon S3 > Buckets > spark-streaming-data-rvm-bkt
spark-streaming-data-rvm-bkt Info Publicly accessible
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (2) Info
Copy S3 URI
Copy URL
Download
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 objec
Find objects by prefix
Name
Type
Last modified
checkpoints/
Folder
-
data/
Folder
-


Amazon S3 > Buckets > spark-streaming-data-rvm-bkt > data/
data/
Copy S3 URI
Objects
Properties
Objects (3) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
failures_data/ Folder - - -
vehicle_data/ Folder - - -
weather_data/ Folder - - -


Amazon S3 > Buckets > spark-streaming-data-rvm-bkt > data/ > vehicle_data/
vehicle_data/
Copy S3 URI
Objects
Properties
Objects (297) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
_spark_metadata/ Folder - - -
part-00000-00109ffb-bea2-4010-8a84-638b74f1a35f-c000.snappy.parquet parquet February 27, 2024, 12:40:06 (UTC+01:00) 4.1 KB Standard
part-00000-00251bd4-bbbe-4284-9130-2b6710ff692b-c000.snappy.parquet parquet February 27, 2024, 12:28:24 (UTC+01:00) 4.1 KB Standard
part-00000-00accee7-39d7-468c-952d-6c14131abfda-c000.snappy.parquet parquet February 27, 2024, 12:32:36 (UTC+01:00) 4.1 KB Standard
part-00000-01e7afbb-7126-4f30-
February 27, 2024, 12:22:23

AWS

- Create crawlers to transform all the .parquet files into an explorable structure as tables in a database.

AWS Glue

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages

Crawlers

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

Crawlers (0) Info

View and manage all available crawlers.

Last updated (UTC)
February 27, 2024 at 12:00:15

Action | Run | Create crawler

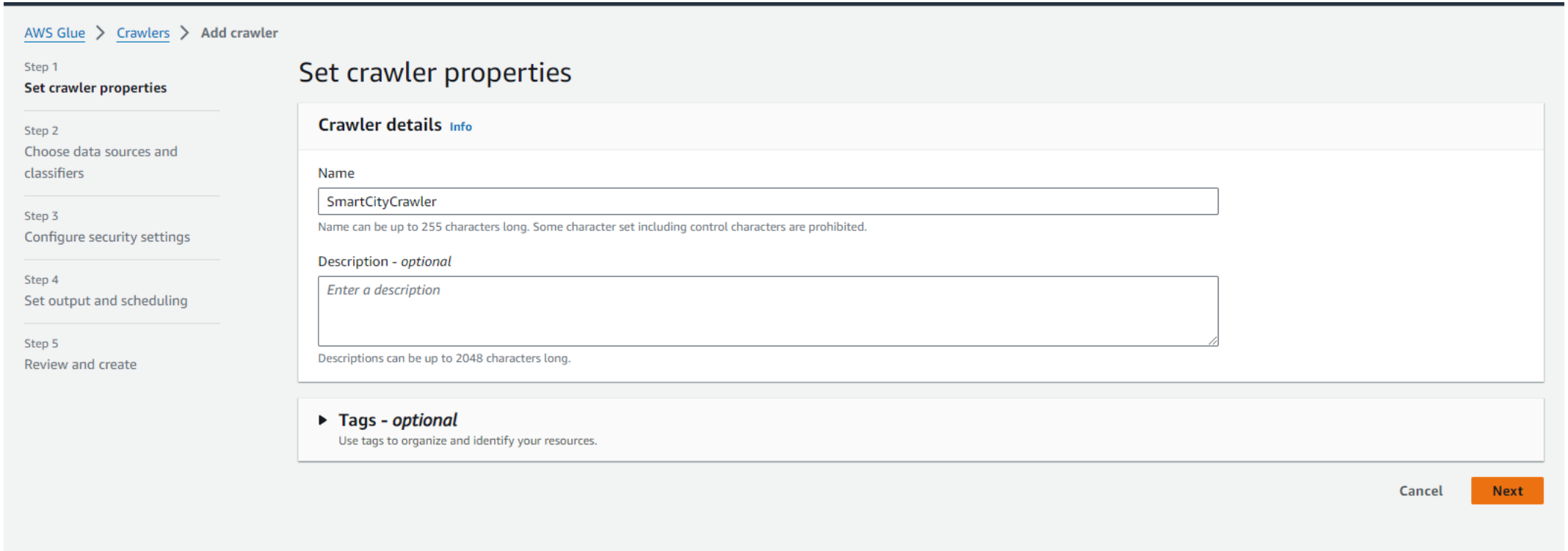
Filter crawlers

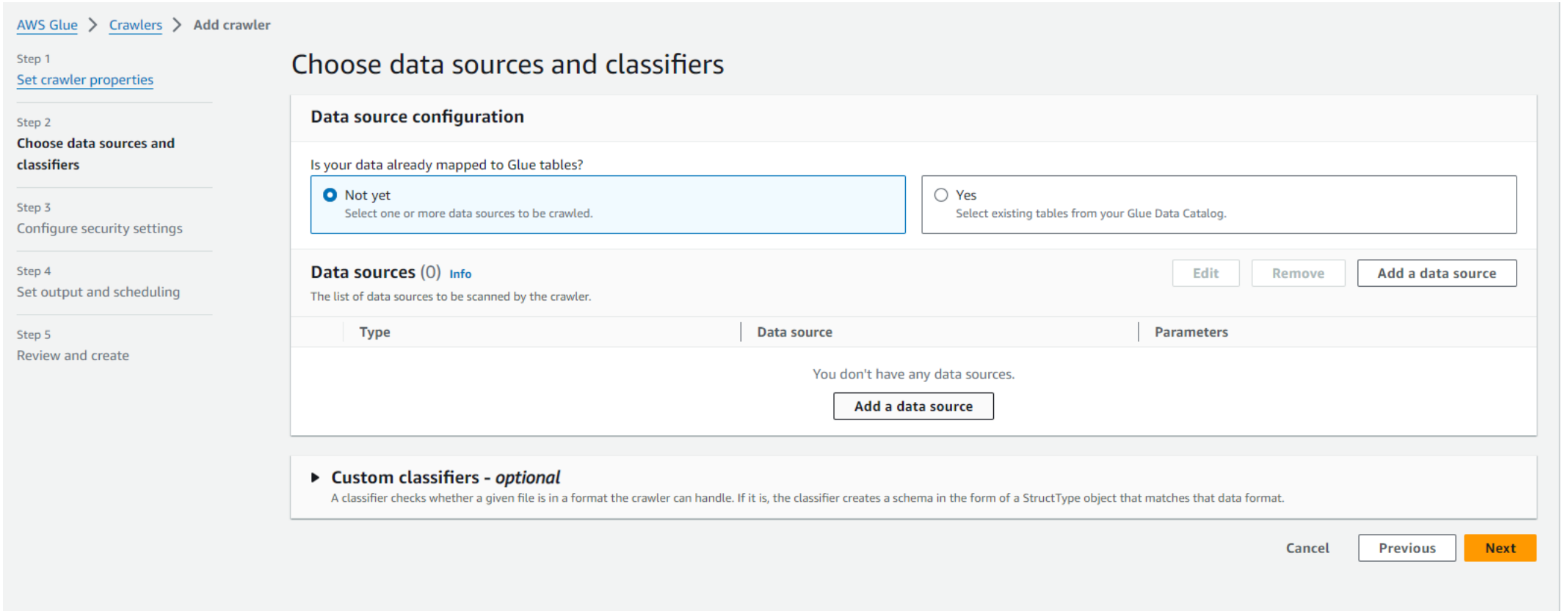
1

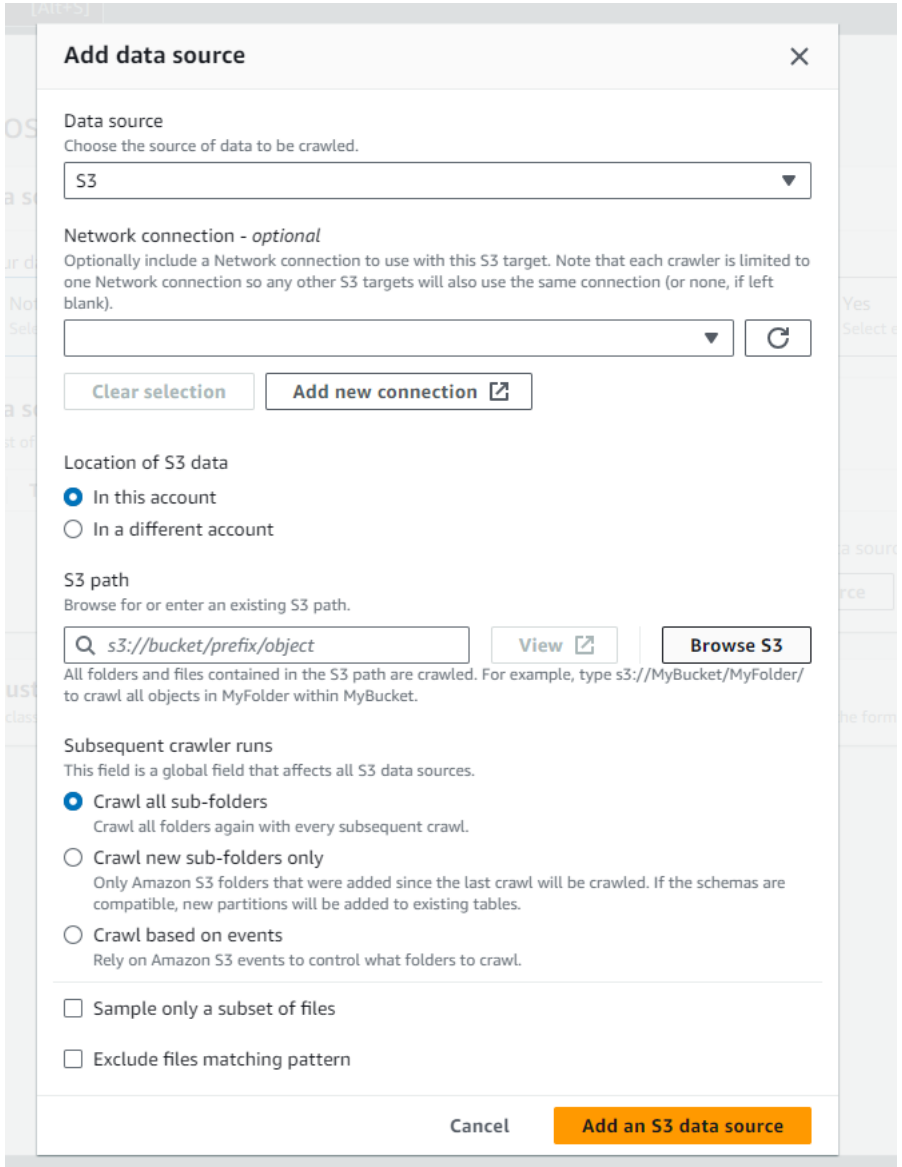
Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last ...

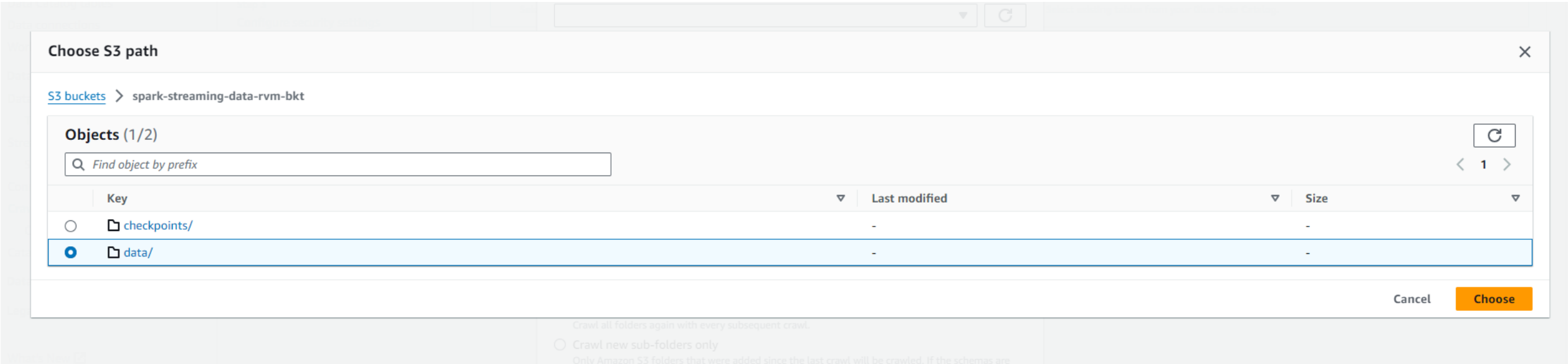
No resources

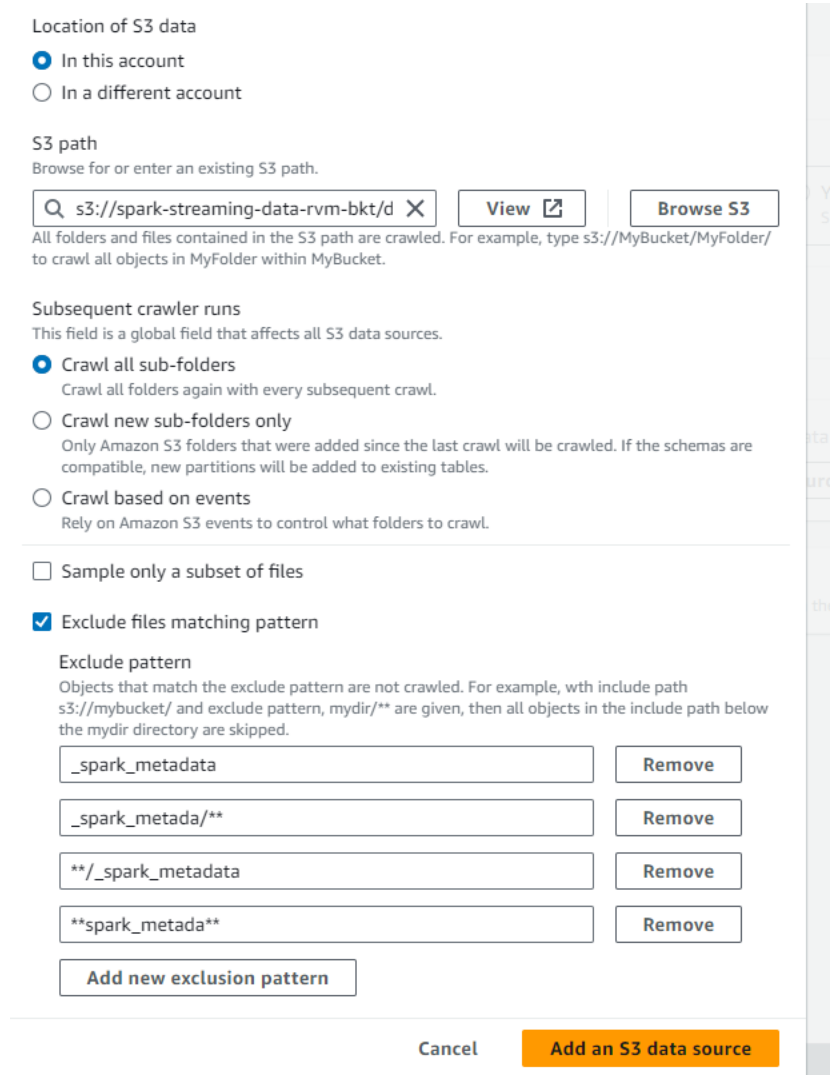
No resources to display.

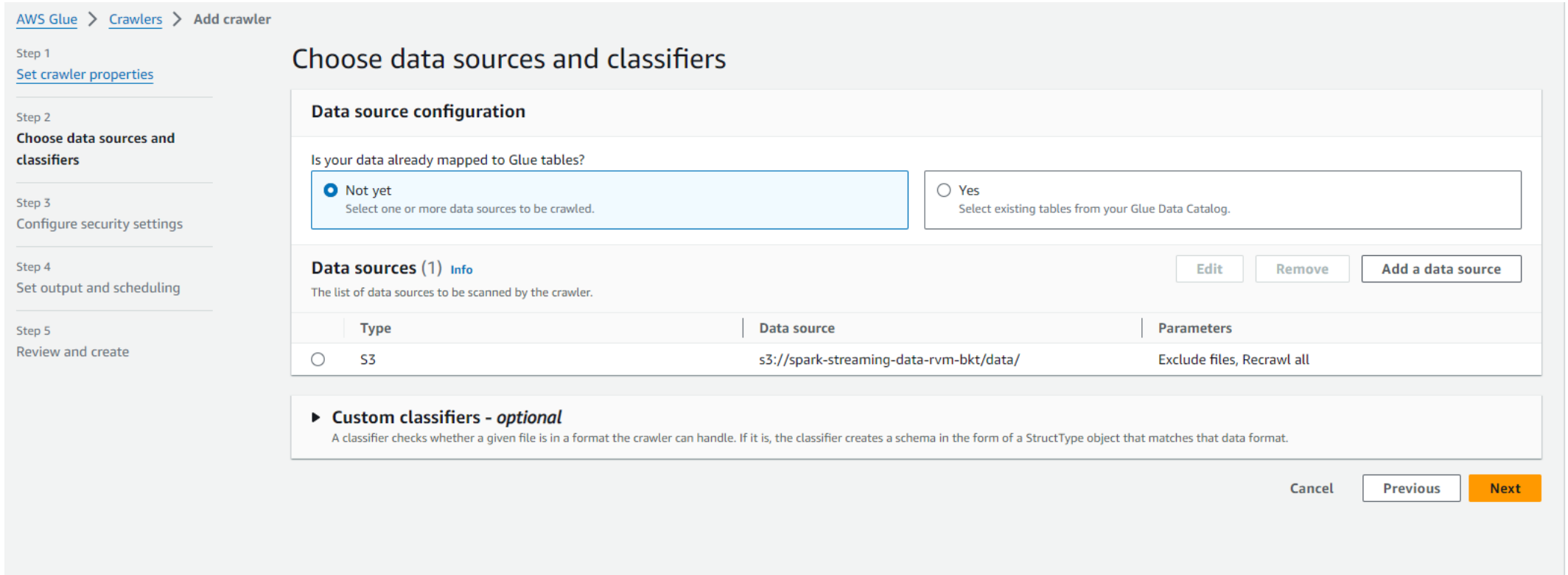

AWS Glue > Crawlers > Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Set crawler properties
Crawler details Info
Name
SmartCityCrawler
Name can be up to 255 characters long. Some character set including control characters are prohibited.
Description - optional
Enter a description
Descriptions can be up to 2048 characters long.
Tags - optional
Use tags to organize and identify your resources.
Cancel
Next


AWS Glue
Crawlers
Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (0) Info
The list of data sources to be scanned by the crawler.
Edit
Remove
Add a data source
Type
Data source
Parameters
You don't have any data sources.
Add a data source
Custom classifiers - optional
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.
Cancel
Previous
Next

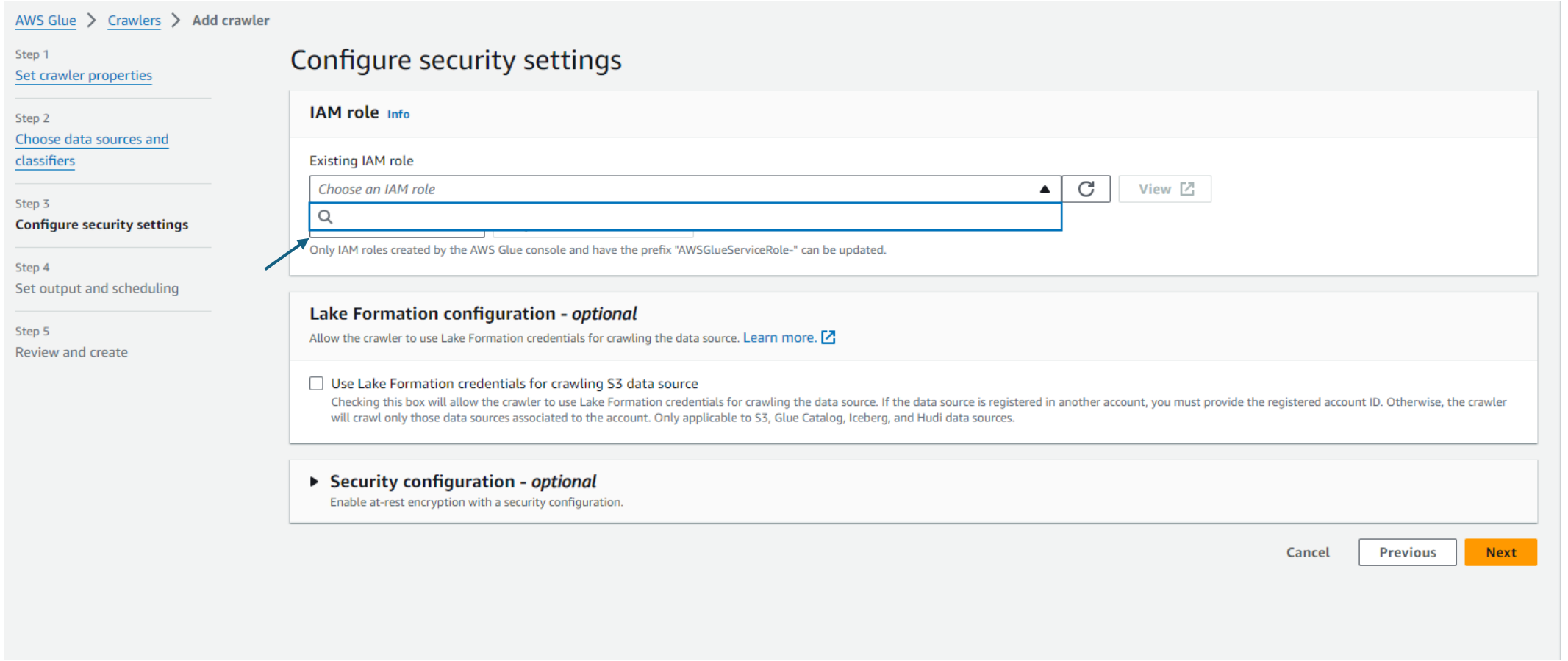

Add data source
Data source
Choose the source of data to be crawled.
S3
Network connection - optional
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).
Clear selection
Add new connection
Location of S3 data
In this account
In a different account
S3 path
Browse for or enter an existing S3 path.
s3://bucket/prefix/object
View
Browse S3
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.
Subsequent crawler runs
This field is a global field that affects all S3 data sources.
Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.
Sample only a subset of files
Exclude files matching pattern
Cancel
Add an S3 data source

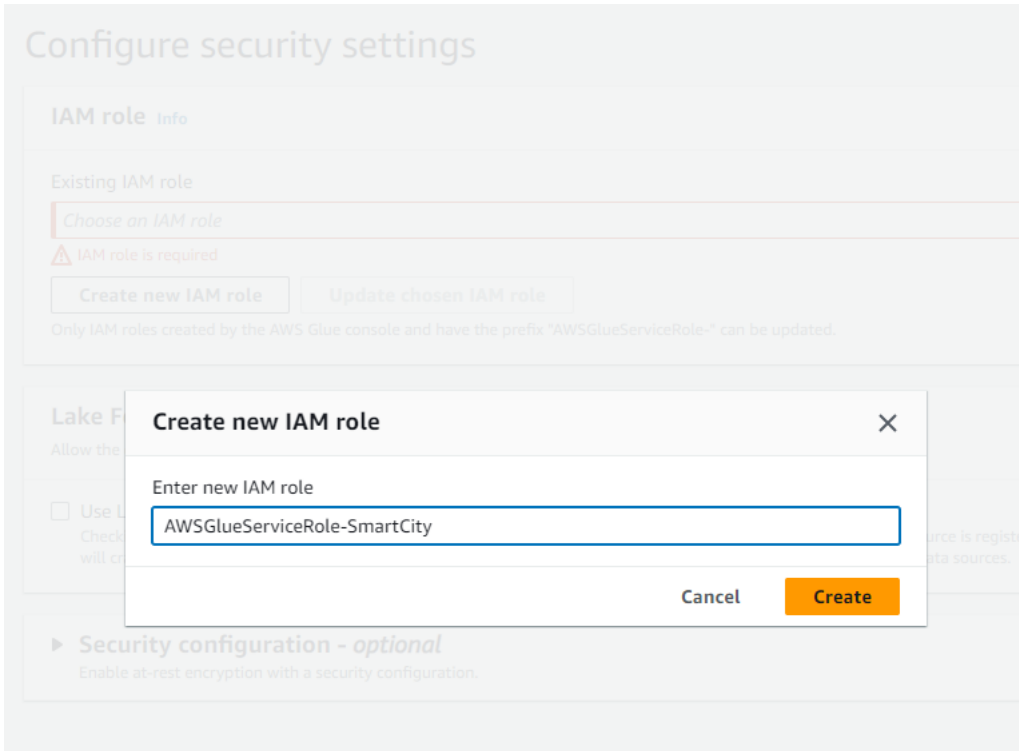

Choose S3 path
S3 buckets
spark-streaming-data-rvm-bkt
Objects (1/2)
Find object by prefix
1
Key
Last modified
Size
checkpoints/
-
-
data/
-
-
Cancel
Choose
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are

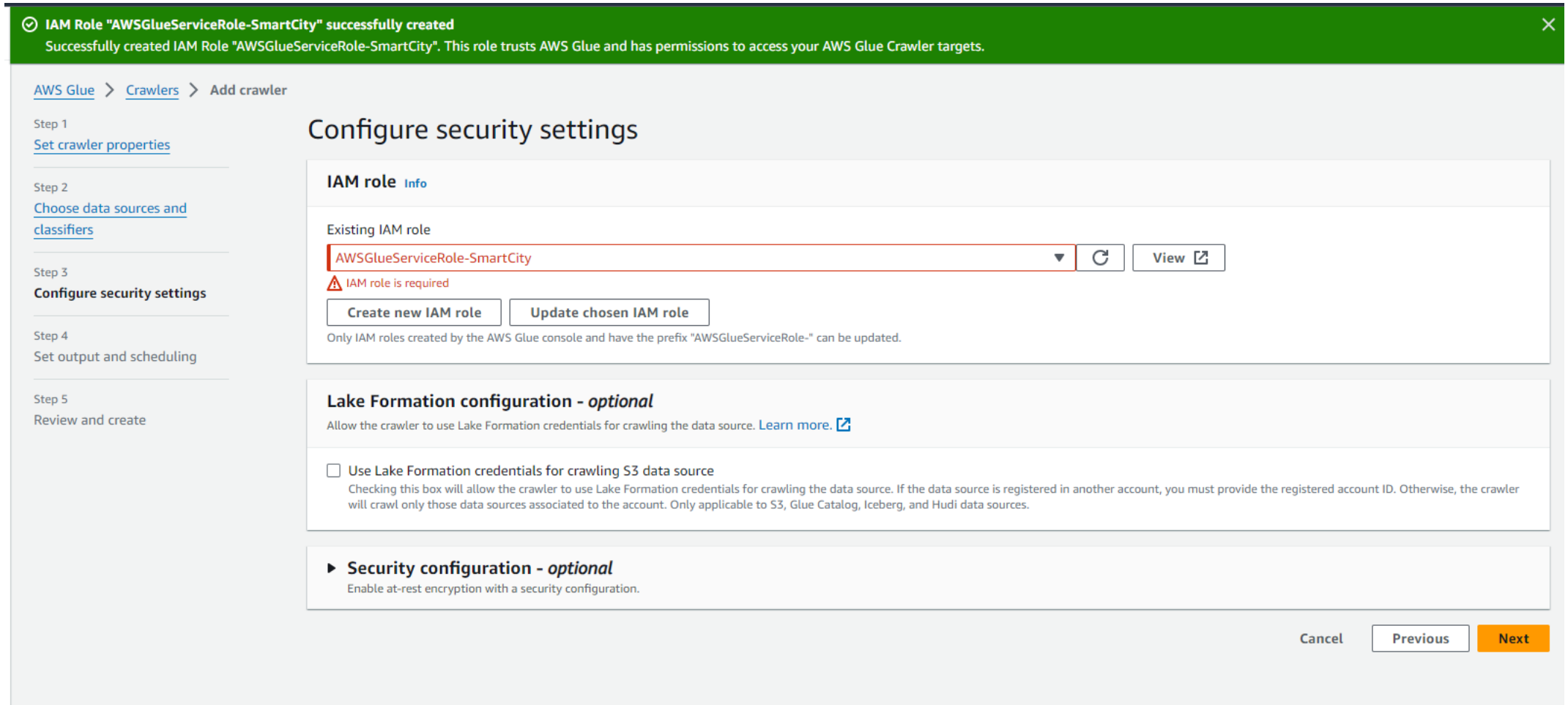

Location of S3 data
In this account
In a different account
S3 path
Browse for or enter an existing S3 path.
s3://spark-streaming-data-rvm-bkt/d
View
Browse S3
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.
Subsequent crawler runs
This field is a global field that affects all S3 data sources.
Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.
Sample only a subset of files
Exclude files matching pattern
Exclude pattern
Objects that match the exclude pattern are not crawled. For example, wth include path s3://mybucket/ and exclude pattern, mydir/** are given, then all objects in the include path below the mydir directory are skipped.
_spark_metadata
Remove
_spark_metada/**
Remove
**/_spark_metadata
Remove
spark_metada
Remove
Add new exclusion pattern
Cancel
Add an S3 data source


AWS Glue > Crawlers > Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (1) Info
The list of data sources to be scanned by the crawler.
Edit
Remove
Add a data source
Type
Data source
Parameters
S3
s3://spark-streaming-data-rvm-bkt/data/
Exclude files, Recrawl all
Custom classifiers - optional
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.
Cancel
Previous
Next

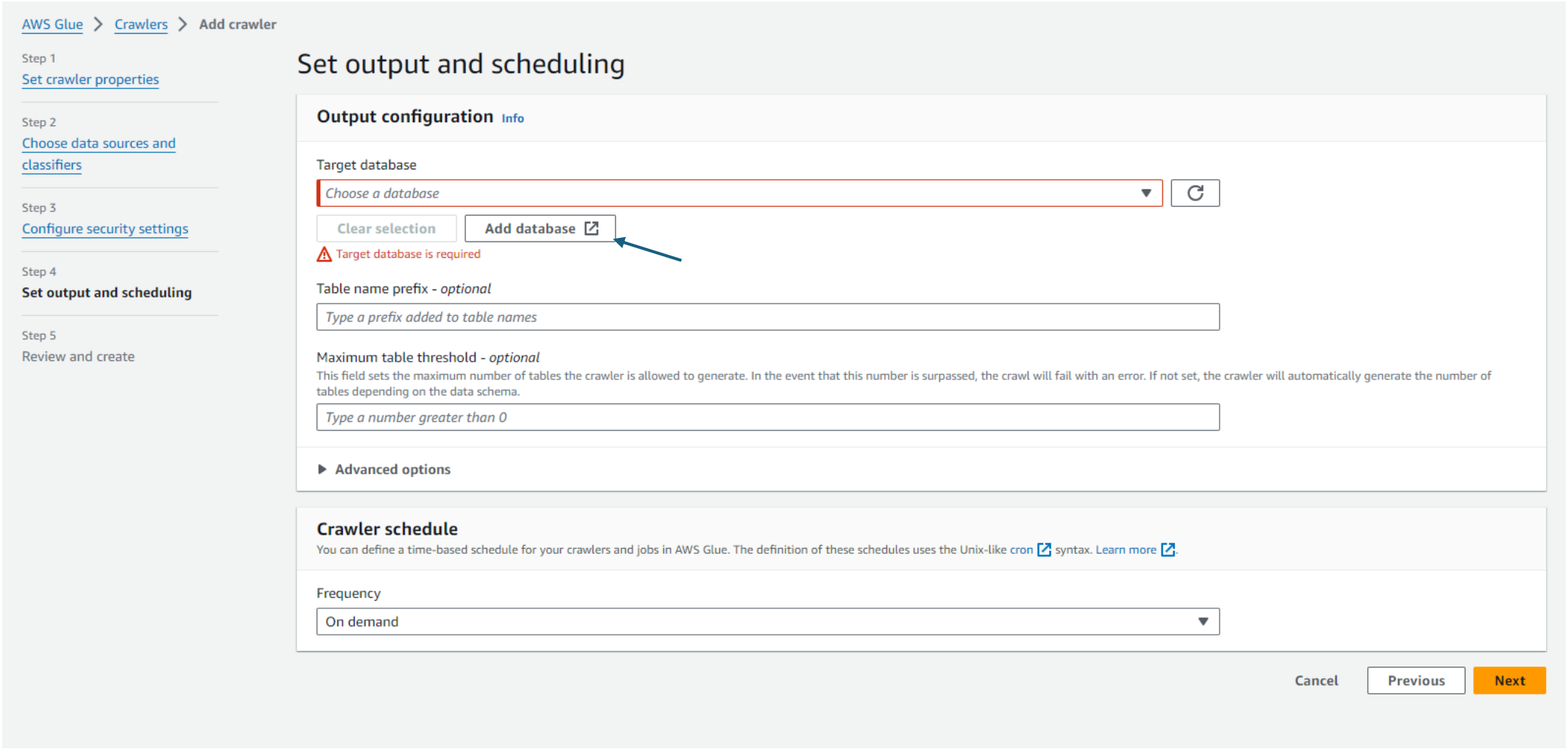
You probably don’t have an IAM role created. Just create a new one.




Configure security settings
IAM role Info
Existing IAM role
Choose an IAM role
IAM role is required
Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.
Create new IAM role
Enter new IAM role
AWSGlueServiceRole-SmartCity
Cancel
Create
Security configuration - optional
Enable at-rest encryption with a security configuration.


IAM Role "AWSGlueServiceRole-SmartCity" successfully created
Successfully created IAM Role "AWSGlueServiceRole-SmartCity". This role trusts AWS Glue and has permissions to access your AWS Glue Crawler targets.
AWS Glue
Crawlers
Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Configure security settings
IAM role Info
Existing IAM role
AWSGlueServiceRole-SmartCity
View
IAM role is required
Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.
Lake Formation configuration - optional
Allow the crawler to use Lake Formation credentials for crawling the data source. Learn more.
Use Lake Formation credentials for crawling S3 data source
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.
Security configuration - optional
Enable at-rest encryption with a security configuration.
Cancel
Previous
Next

You probably don’t have a Database created. Just create a new one.



AWS Glue > Databases > Add database

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name

smartcitydb

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

Location - *optional*

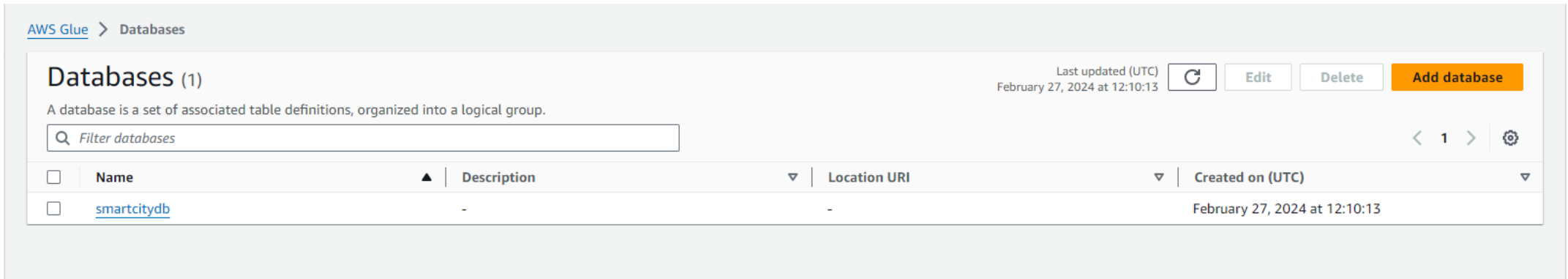
Set the URI location for use by clients of the Data Catalog.

Description - *optional*

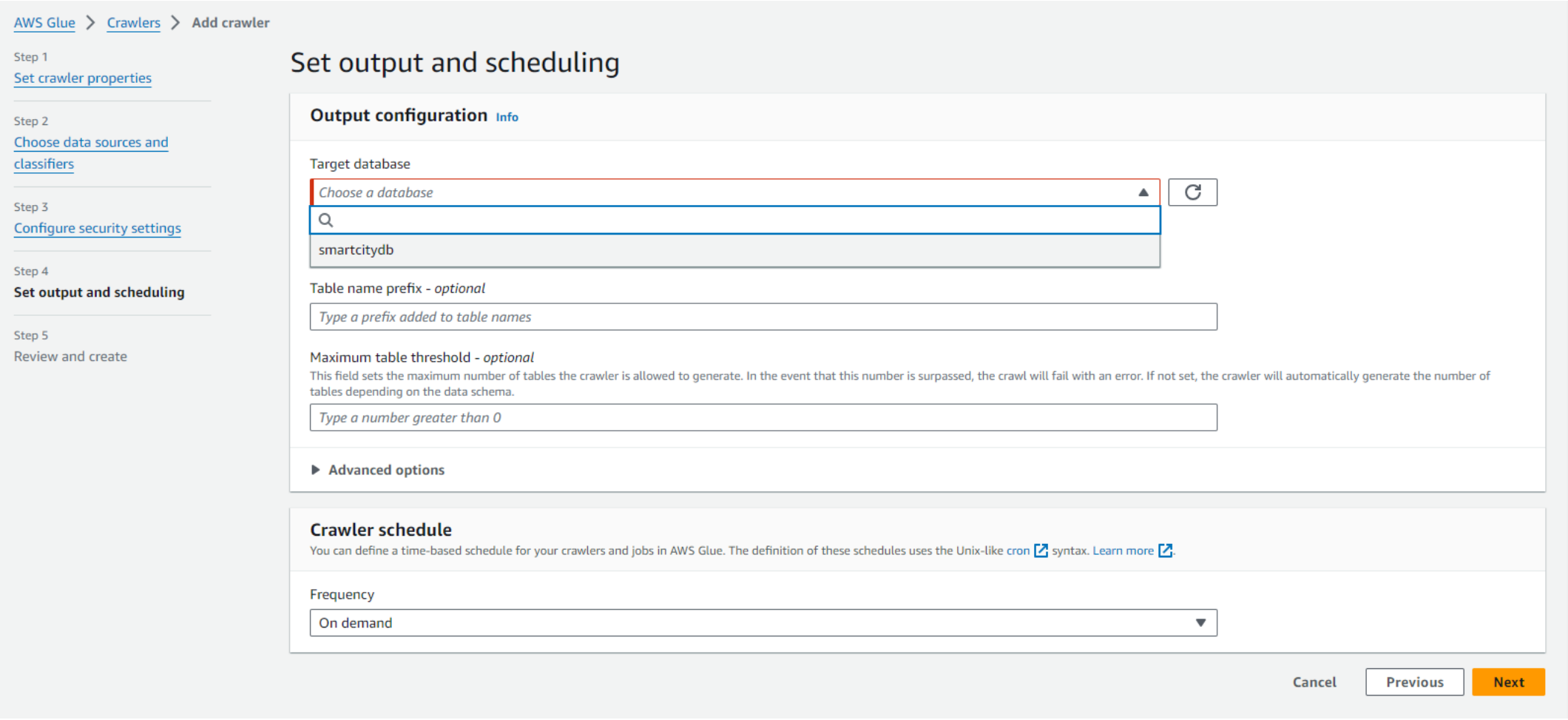
Enter text

Descriptions can be up to 2048 characters long.

Cancel | Create database


AWS Glue
Databases
Databases (1)
A database is a set of associated table definitions, organized into a logical group.
Last updated (UTC)
February 27, 2024 at 12:10:13
Edit
Delete
Add database
Filter databases
1
Name
Description
Location URI
Created on (UTC)
smartcitydb
-
-
February 27, 2024 at 12:10:13

Advanced options by default, I didn’t touched it



Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Review and create

Step 1: Set crawler properties

Edit

Set crawler properties

Name	Description	Tags
SmartCityCrawler	-	-

Step 2: Choose data sources and classifiers

Edit

Data sources (1) Info

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

Type	Data source	Parameters
S3	s3://spark-streaming-data-rvm-bkt/data/	Exclude files, Recrawl all

Step 3: Configure security settings

Edit

Configure security settings

IAM role	Security configuration	Lake Formation configuration
AWSGlueServiceRole-SmartCity	-	-

Step 4: Set output and scheduling

Edit

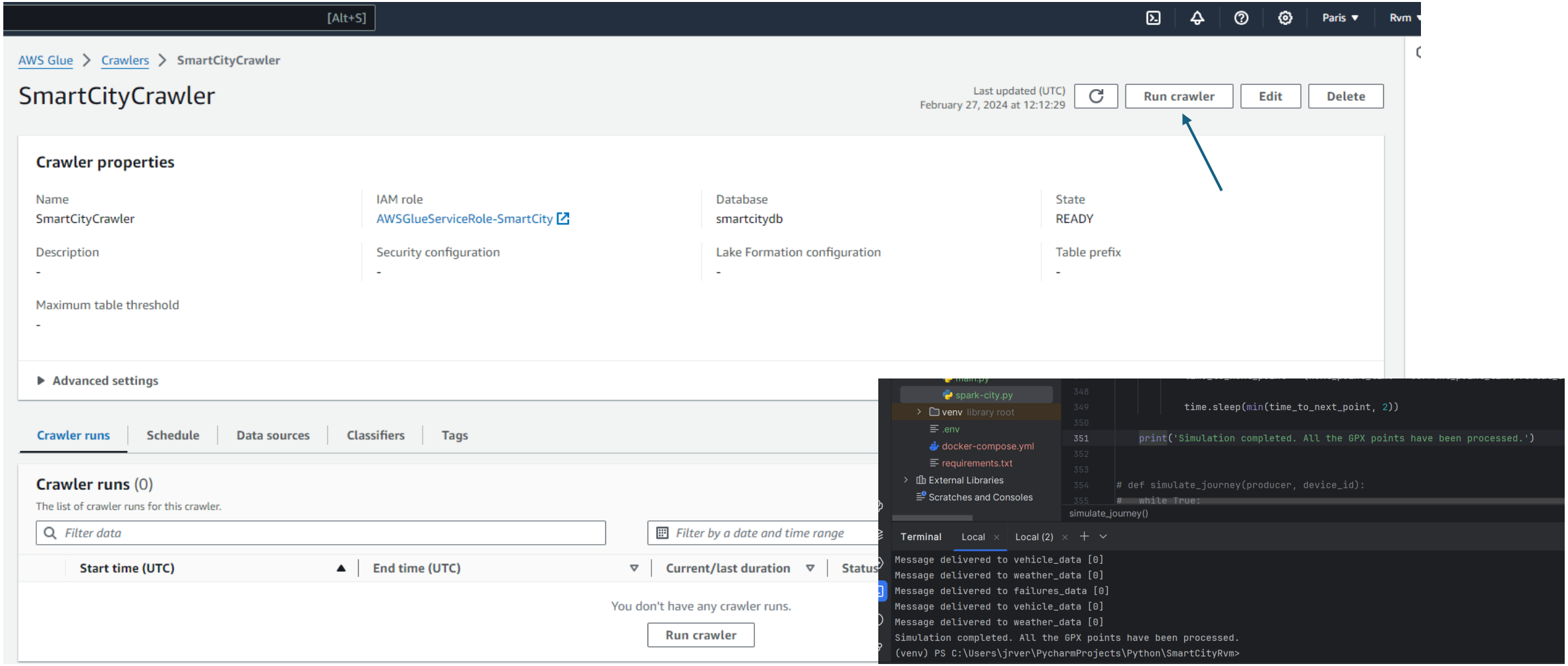
Set output and scheduling

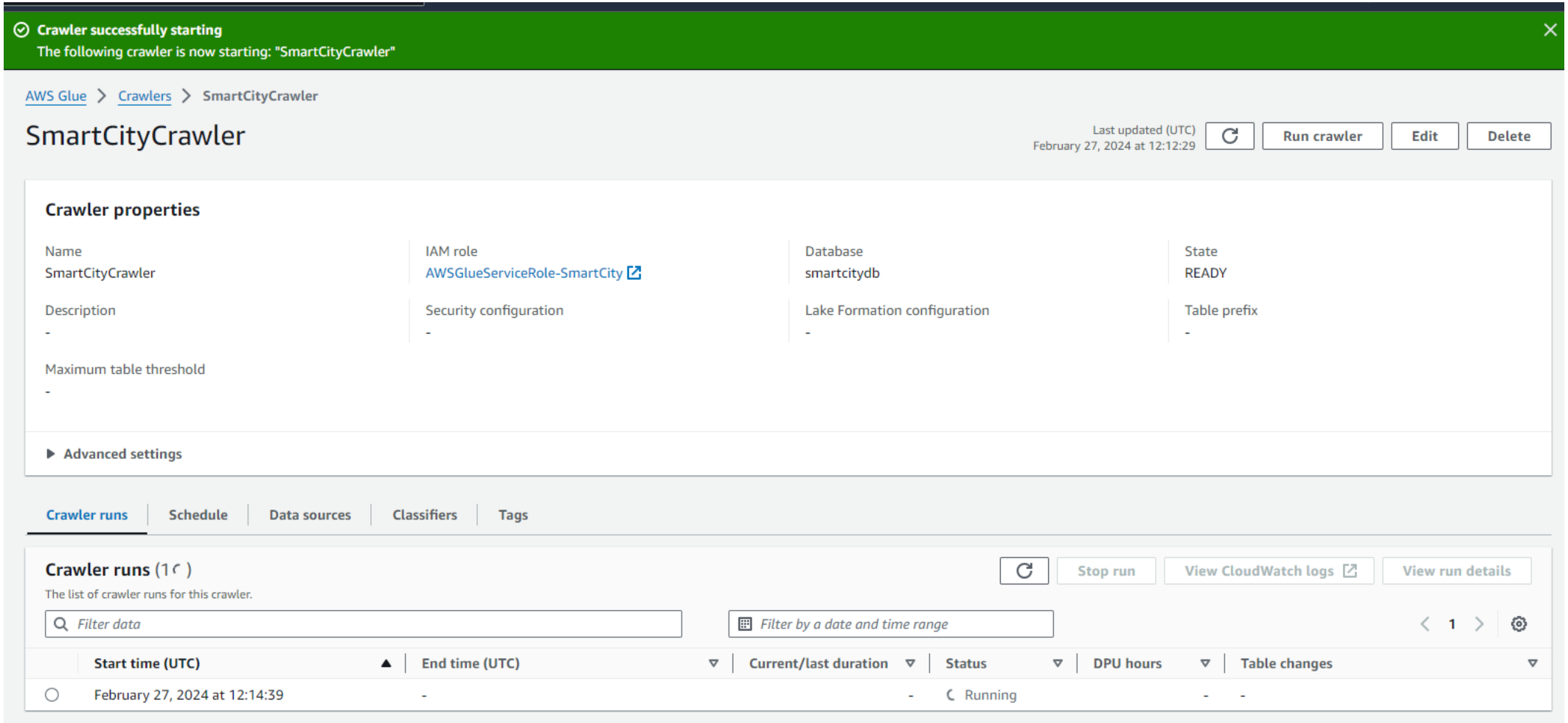
Database	Table prefix - *optional*	Maximum table threshold - *optional*	Schedule
smartcitydb	-	-	On demand

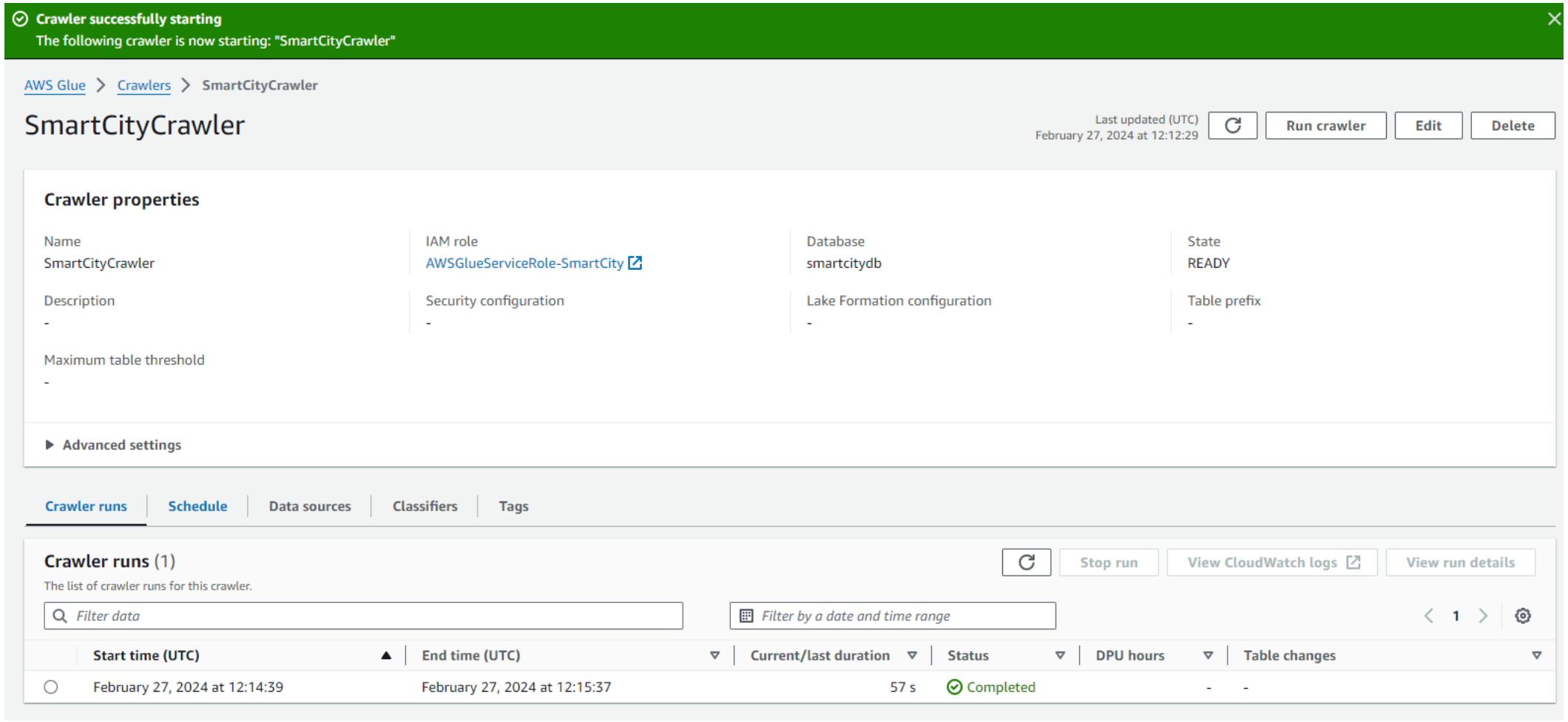
Cancel | Previous | Create crawler

Run the crawler to start transforming the .parquet files

Run this when the streaming process is completed only!!!

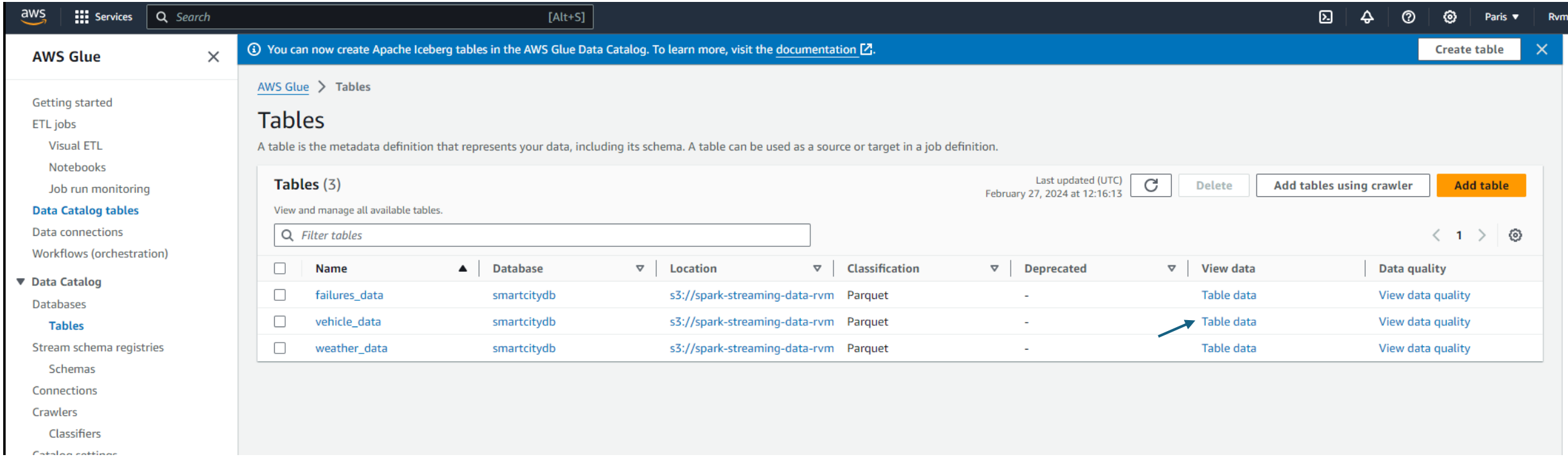


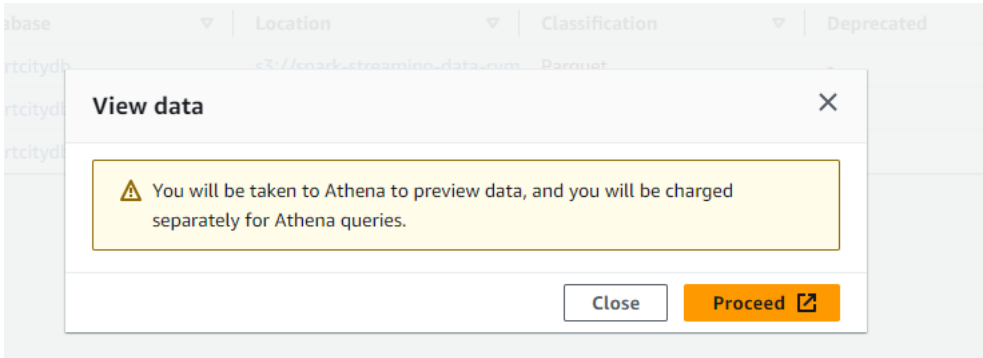

Crawler successfully starting
The following crawler is now starting: "SmartCityCrawler"
AWS Glue > Crawlers > SmartCityCrawler
SmartCityCrawler
Last updated (UTC)
February 27, 2024 at 12:12:29
Run crawler
Edit
Delete
Crawler properties
Name
SmartCityCrawler
IAM role
AWSGlueServiceRole-SmartCity
Database
smartcitydb
State
READY
Description
-
Security configuration
-
Lake Formation configuration
-
Table prefix
-
Maximum table threshold
-
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (1)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes
February 27, 2024 at 12:14:39
-
-
Running
-
-


Crawler successfully starting
The following crawler is now starting: "SmartCityCrawler"
AWS Glue > Crawlers > SmartCityCrawler
SmartCityCrawler
Last updated (UTC)
February 27, 2024 at 12:12:29
Run crawler
Edit
Delete
Crawler properties
Name
SmartCityCrawler
IAM role
AWSGlueServiceRole-SmartCity
Database
smartcitydb
State
READY
Description
-
Security configuration
-
Lake Formation configuration
-
Table prefix
-
Maximum table threshold
-
Advanced settings
Crawler runs
Schedule
Data sources
Classifiers
Tags
Crawler runs (1)
The list of crawler runs for this crawler.
Stop run
View CloudWatch logs
View run details
Filter data
Filter by a date and time range
1
Start time (UTC)
End time (UTC)
Current/last duration
Status
DPU hours
Table changes
February 27, 2024 at 12:14:39
February 27, 2024 at 12:15:37
57 s
Completed
-
-

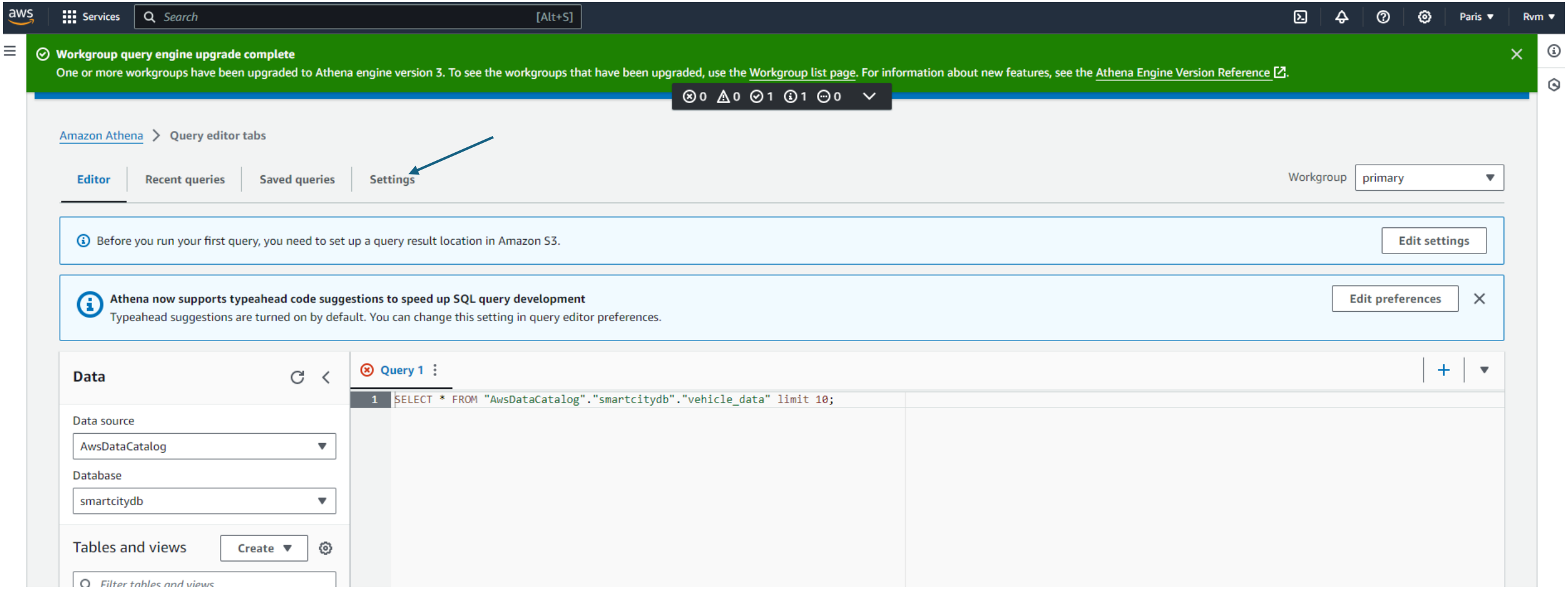
AWS

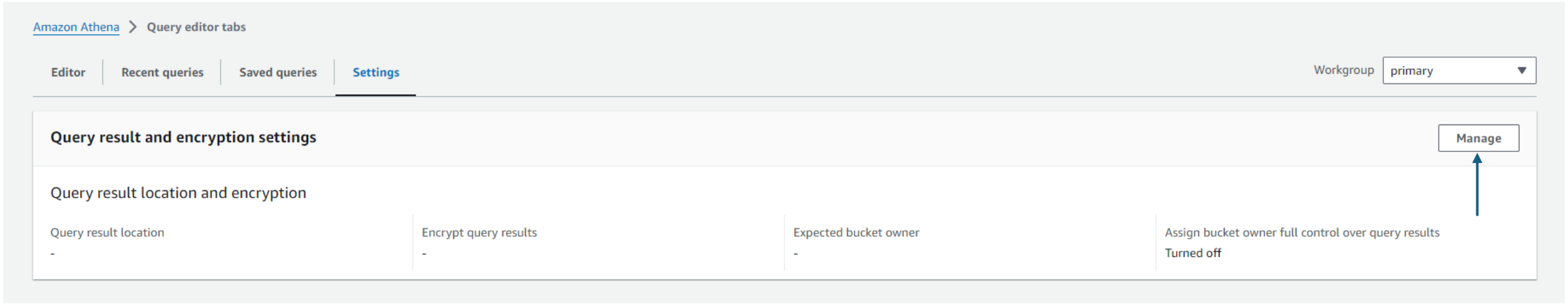
- Access to Glue and explore the transformed tables, available to be queried.
- The queries and exploration are made from ATHENA as you click on “table data”


aws
Services
Search
[Alt+S]
Paris
Rvm
AWS Glue
You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the documentation.
Create table
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
AWS Glue > Tables
Tables
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.
Tables (3)
View and manage all available tables.
Last updated (UTC)
February 27, 2024 at 12:16:13
Delete
Add tables using crawler
Add table
Filter tables
1
Name
Database
Location
Classification
Deprecated
View data
Data quality
failures_data
smartcitydb
s3://spark-streaming-data-rvm
Parquet
-
Table data
View data quality
vehicle_data
smartcitydb
s3://spark-streaming-data-rvm
Parquet
-
Table data
View data quality
weather_data
smartcitydb
s3://spark-streaming-data-rvm
Parquet
-
Table data
View data quality

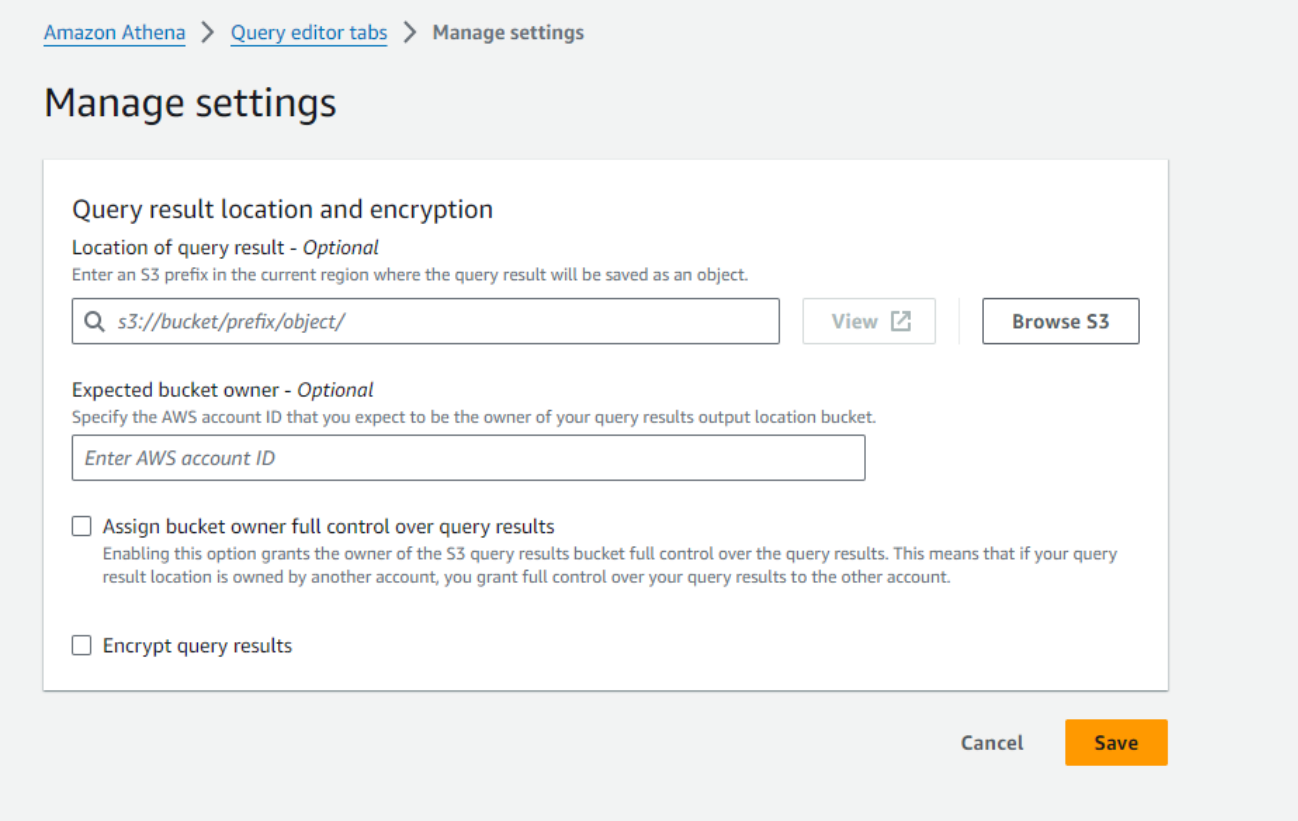

View data
You will be taken to Athena to preview data, and you will be charged separately for Athena queries.
Close
Proceed

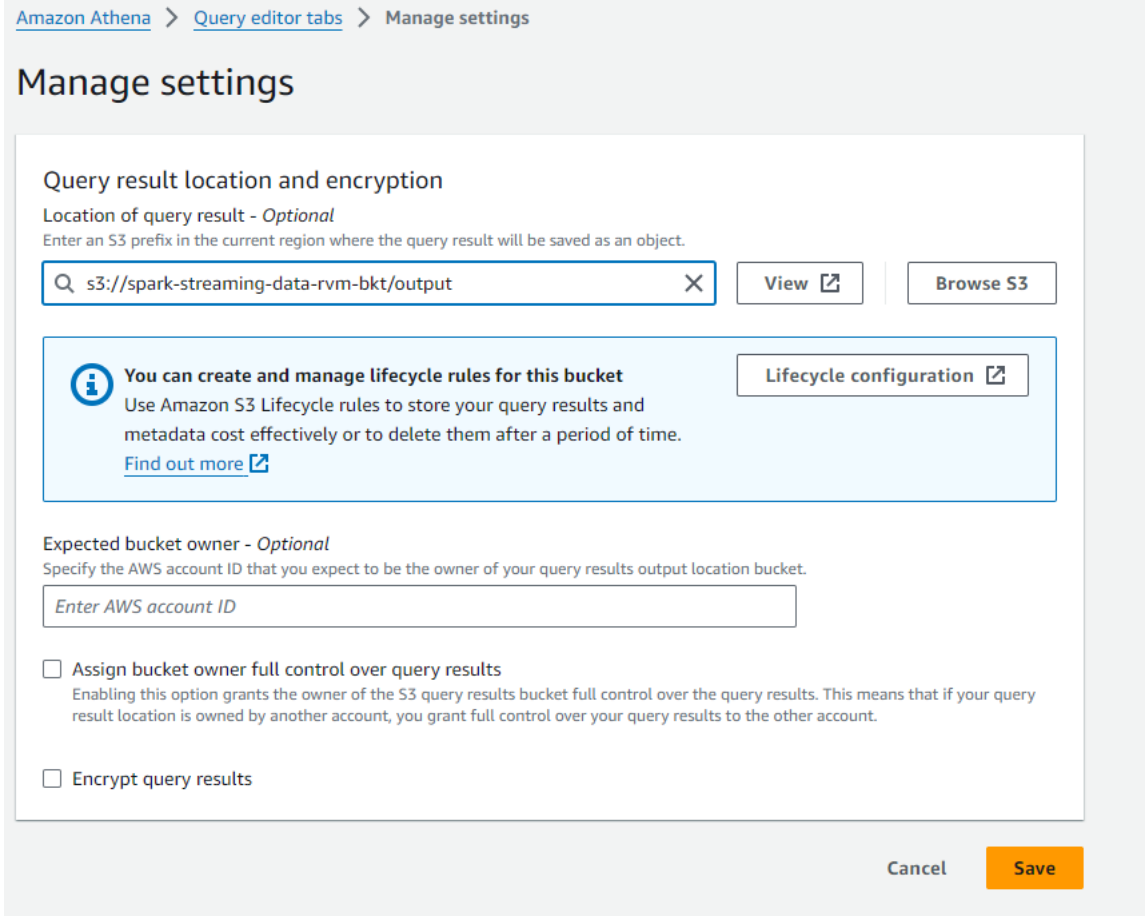
You are not able to run a query against the tables until you set an output directory to the queries



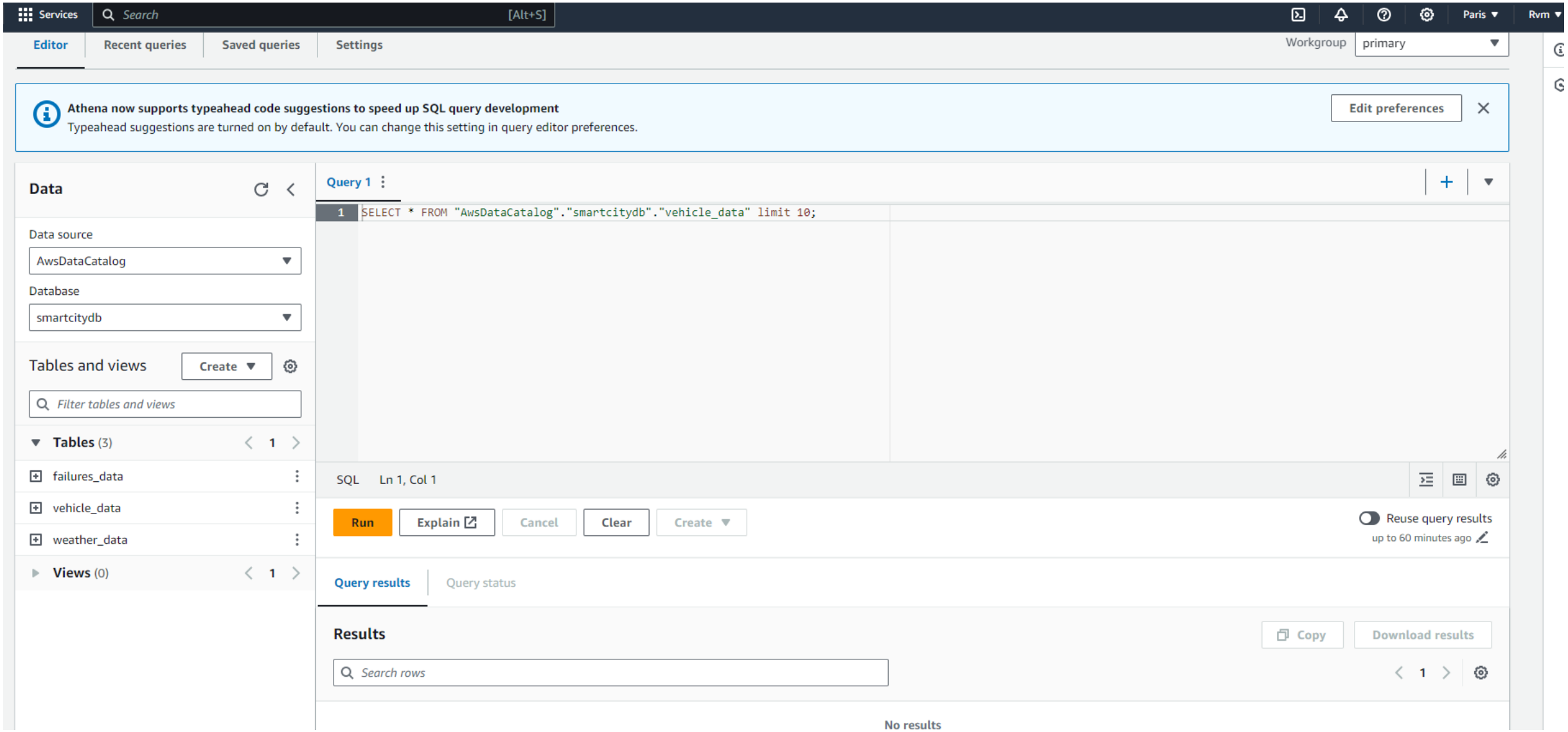

Amazon Athena
Query editor tabs
Editor
Recent queries
Saved queries
Settings
Workgroup
primary
Query result and encryption settings
Manage
Query result location and encryption
Query result location
-
Encrypt query results
-
Expected bucket owner
-
Assign bucket owner full control over query results
Turned off

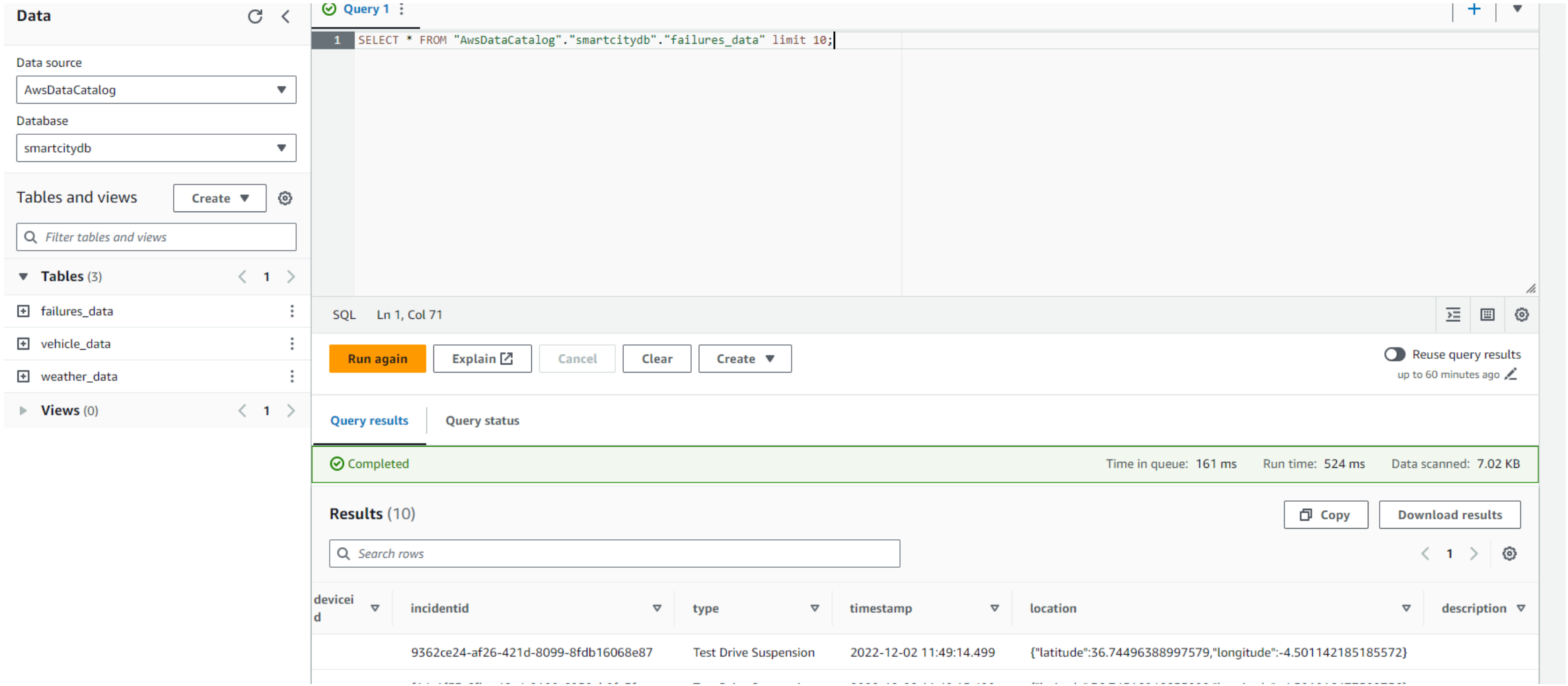
Browse your bucket address and write by hand “/output” at the end, thus a new folder will be created into S3 to save the queries results

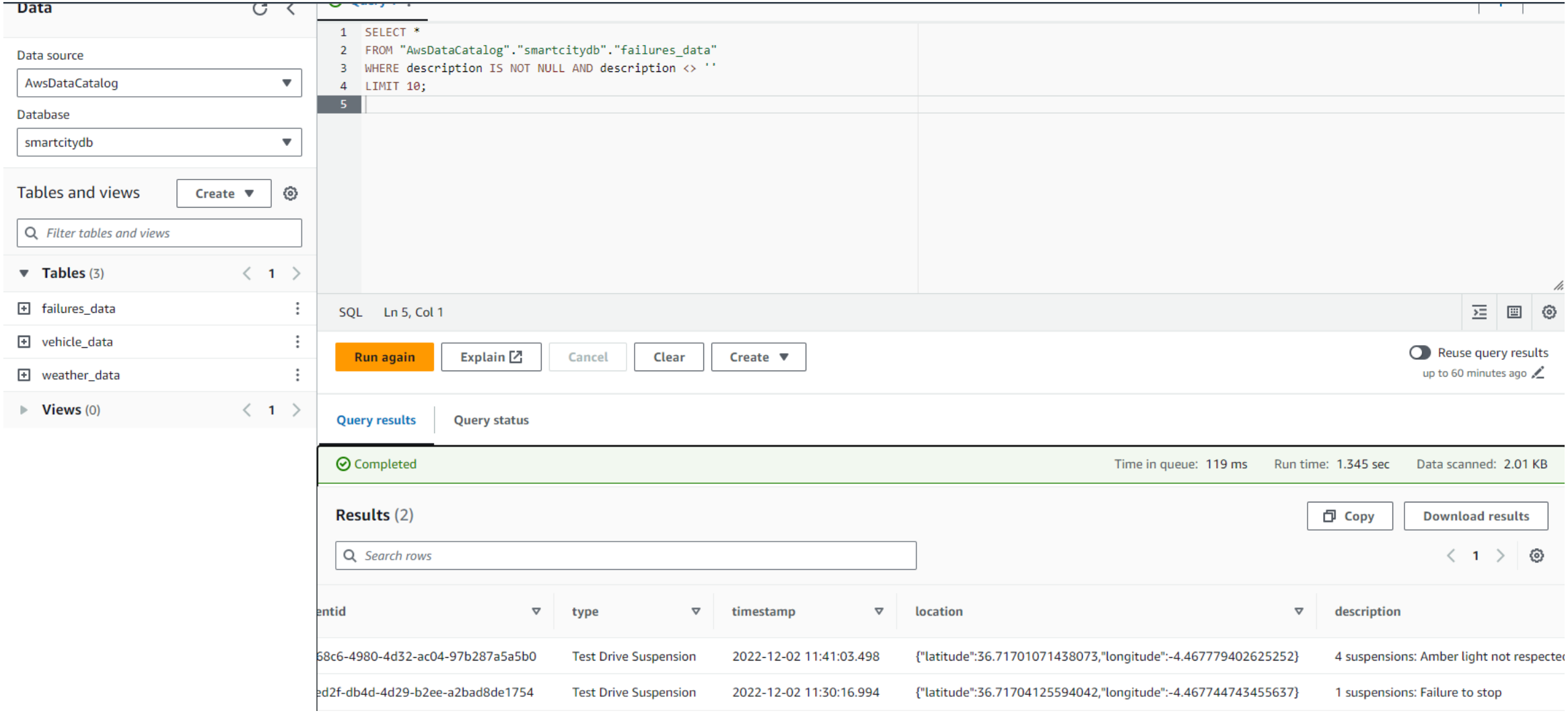


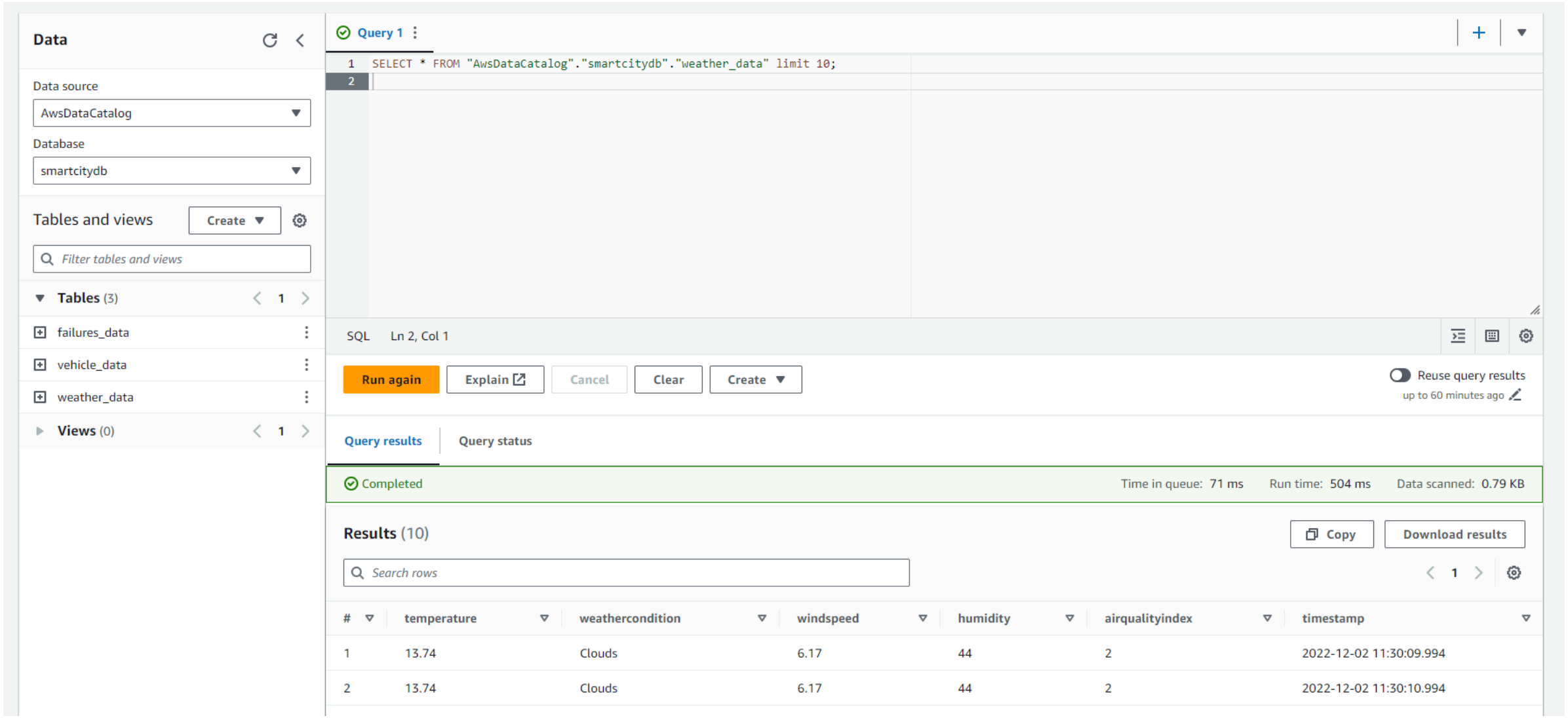


Now you should be able to run an exploratory query to all the tables




Data
Data source
AwsDataCatalog
Database
smartcitydb
Tables and views
Create
Filter tables and views
Tables (3)
failures_data
vehicle_data
weather_data
Views (0)
Query 1
1 SELECT * FROM "AwsDataCatalog"."smartcitydb"."failures_data" limit 10;
SQL Ln 1, Col 71
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query status
Completed
Time in queue: 161 ms
Run time: 524 ms
Data scanned: 7.02 KB
Results (10)
Copy
Download results
Search rows
deviceid
incidentid
type
timestamp
location
description
9362ce24-af26-421d-8099-8fdb16068e87
Test Drive Suspension
2022-12-02 11:49:14.499
{"latitude":36.74496388997579,"longitude":-4.501142185185572}


Data
Data source
AwsDataCatalog
Database
smartcitydb
Tables and views
Create
Filter tables and views
Tables (3)
1
failures_data
vehicle_data
weather_data
Views (0)
1
SELECT *
FROM "AwsDataCatalog"."smartcitydb"."failures_data"
WHERE description IS NOT NULL AND description <> ''
LIMIT 10;
SQL Ln 5, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query status
Completed
Time in queue: 119 ms
Run time: 1.345 sec
Data scanned: 2.01 KB
Results (2)
Copy
Download results
Search rows
1
entid
type
timestamp
location
description
68c6-4980-4d32-ac04-97b287a5a5b0
Test Drive Suspension
2022-12-02 11:41:03.498
{"latitude":36.71701071438073,"longitude":-4.467779402625252}
4 suspensions: Amber light not respecte
ed2f-db4d-4d29-b2ee-a2bad8de1754
Test Drive Suspension
2022-12-02 11:30:16.994
{"latitude":36.71704125594042,"longitude":-4.467744743455637}
1 suspensions: Failure to stop


Data
Data source
AwsDataCatalog
Database
smartcitydb
Tables and views
Create
Filter tables and views
Tables (3)
failures_data
vehicle_data
weather_data
Views (0)
Query 1
1 SELECT * FROM "AwsDataCatalog"."smartcitydb"."weather_data" limit 10;
2
SQL Ln 2, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query status
Completed
Time in queue: 71 ms
Run time: 504 ms
Data scanned: 0.79 KB
Results (10)
Copy
Download results
Search rows
temperature weathercondition windspeed humidity airqualityindex timestamp
1 13.74 Clouds 6.17 44 2 2022-12-02 11:30:09.994
2 13.74 Clouds 6.17 44 2 2022-12-02 11:30:10.994

AWS

Redshift cluster creation, configuration and permissions

Analytics

Amazon Redshift

Fast, fully managed, petabyte-scale cloud data warehouse.

Amazon Redshift makes it easier for you to run and scale analytics without having to manage your data warehouse. Get insights by running real-time and predictive analytics on all of your data, across operational databases, data lake, data warehouse, and thousands of third-party datasets.

Get to powerful insights fast

Get insights from data in seconds without managing data warehouse infrastructure. First-time Redshift Serverless customers receive a $300 credit to use in their account.

Try Redshift Serverless free trial

Getting started

Redshift Serverless overview

Evaluation and POC support

For more granular control

Create, configure, and manage your cluster to control computing resources.

Create cluster

How it works

Amazon Redshift > Clusters > Create cluster

Create cluster Info

Looking for free trial? Try Redshift Serverless. First-time Redshift Serverless customers receive a $300 credit to use in their account.

Launch Redshift Serverless

Cluster configuration

Cluster identifier

This is the unique key that identifies a cluster.

redshift-cluster-1

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

Choose the size of the cluster

- (•) I'll choose
- ( ) Help me choose

Node type Info

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

ra3.4xlarge

AZ configuration Info

Choose if you want to deploy the Redshift cluster in another Availability Zone.

- (•) Single-AZ
  Compute resources are deployed in a single Availability Zone.
- ( ) Multi-AZ - *new*
  Compute resources are deployed in two Availability Zones.

Number of nodes

Enter the number of nodes that you need.

2

Range (2-32)

Cluster configuration

Cluster identifier

This is the unique key that identifies a cluster.

smart-city-redshift-clusterrvm

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

Choose the size of the cluster

- (•) I'll choose
- ( ) Help me choose

Node type | Info

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

dc2.large

Number of nodes

Enter the number of nodes that you need.

1

Range (1-32)

Configuration summary Info

dc2.large | 1 node

$233.60/month

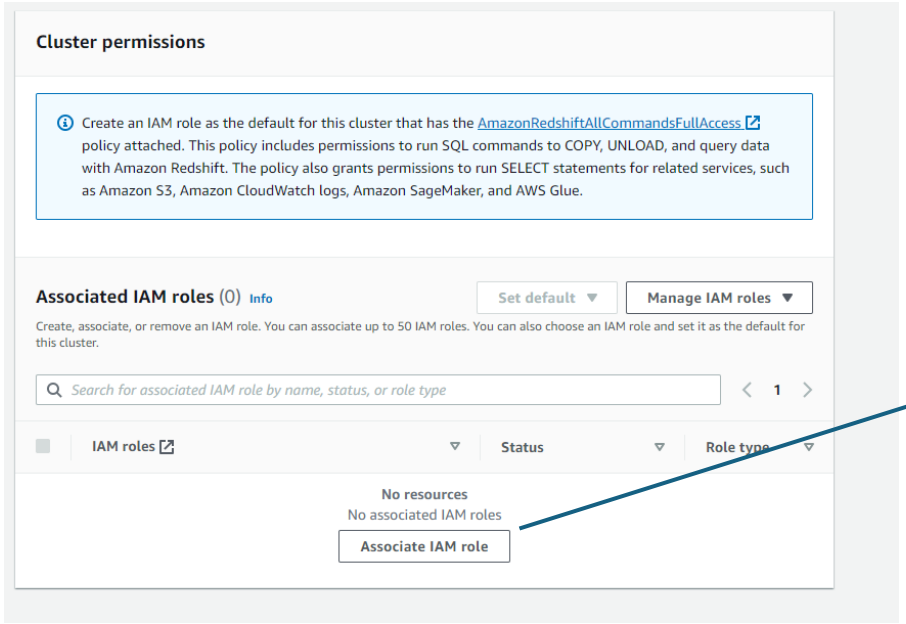
Estimated on-demand compute price

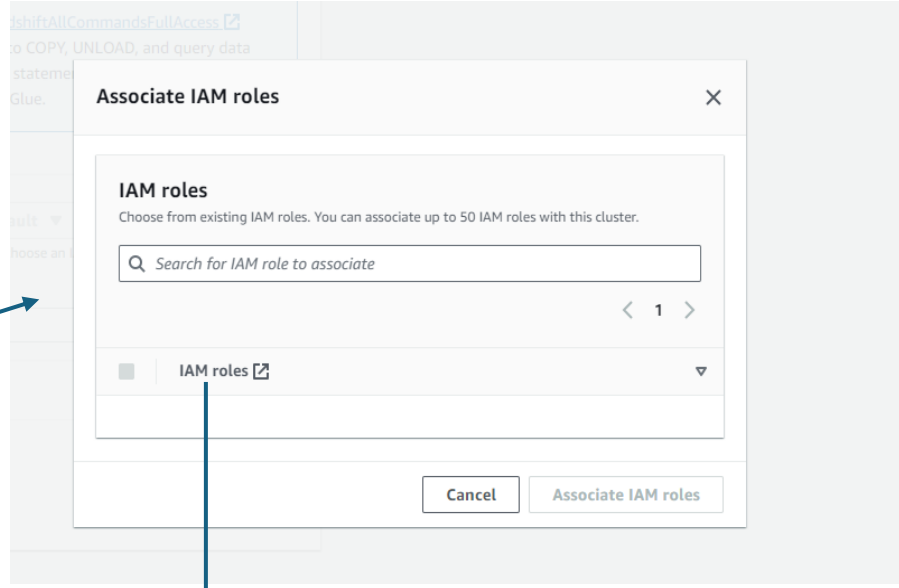
Save more than 60% of your costs by purchasing reserved nodes.
Learn more about pricing

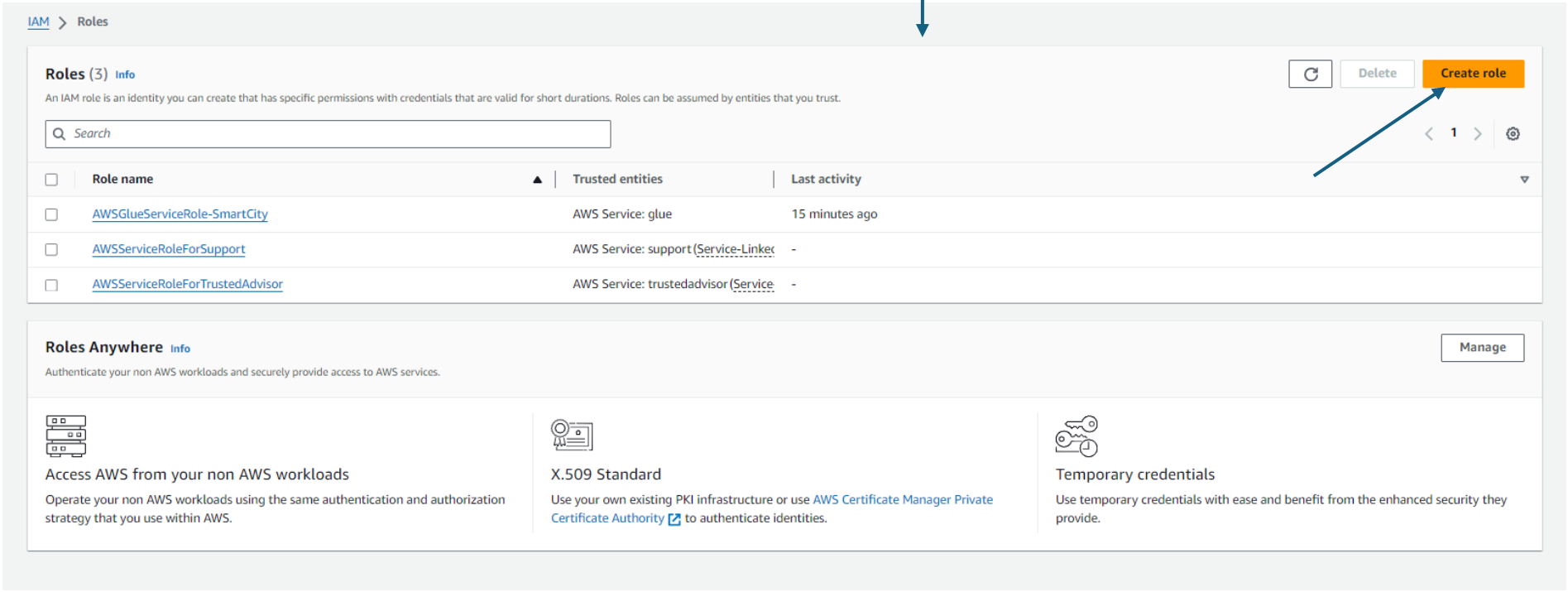
160 GB

Total compressed storage

The total storage capacity for the cluster if you deploy the number of nodes that you chose.


Cluster permissions
Create an IAM role as the default for this cluster that has the AmazonRedshiftAllCommandsFullAccess policy attached. This policy includes permissions to run SQL commands to COPY, UNLOAD, and query data with Amazon Redshift. The policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue.
Associated IAM roles (0) Info
Set default
Manage IAM roles
Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.
Search for associated IAM role by name, status, or role type
IAM roles
Status
Role type
No resources
No associated IAM roles
Associate IAM role


Associate IAM roles
IAM roles
Choose from existing IAM roles. You can associate up to 50 IAM roles with this cluster.
Search for IAM role to associate
IAM roles
Cancel
Associate IAM roles


IAM > Roles
Roles (3) Info
An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.
Delete
Create role
Search
Role name
Trusted entities
Last activity
AWSGlueServiceRole-SmartCity
AWS Service: glue
15 minutes ago
AWSServiceRoleForSupport
AWS Service: support (Service-Linke
-
AWSServiceRoleForTrustedAdvisor
AWS Service: trustedadvisor (Service
-
Roles Anywhere Info
Authenticate your non AWS workloads and securely provide access to AWS services.
Manage
Access AWS from your non AWS workloads
Operate your non AWS workloads using the same authentication and authorization strategy that you use within AWS.
X.509 Standard
Use your own existing PKI infrastructure or use AWS Certificate Manager Private Certificate Authority to authenticate identities.
Temporary credentials
Use temporary credentials with ease and benefit from the enhanced security they provide.

Step 1
Select trusted entity

Step 2
Add permissions

Step 3
Name, review, and create

Select trusted entity Info

Trusted entity type

- (●) **AWS service**
  Allow AWS services like EC2, Lambda, or others to perform actions in this account.
- ( ) **AWS account**
  Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.
- ( ) **Web identity**
  Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.
- ( ) **SAML 2.0 federation**
  Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.
- ( ) **Custom trust policy**
  Create a custom trust policy to enable others to perform actions in this account.

Use case

Allow an AWS service like EC2, Lambda, or others to perform actions in this account.

Service or use case

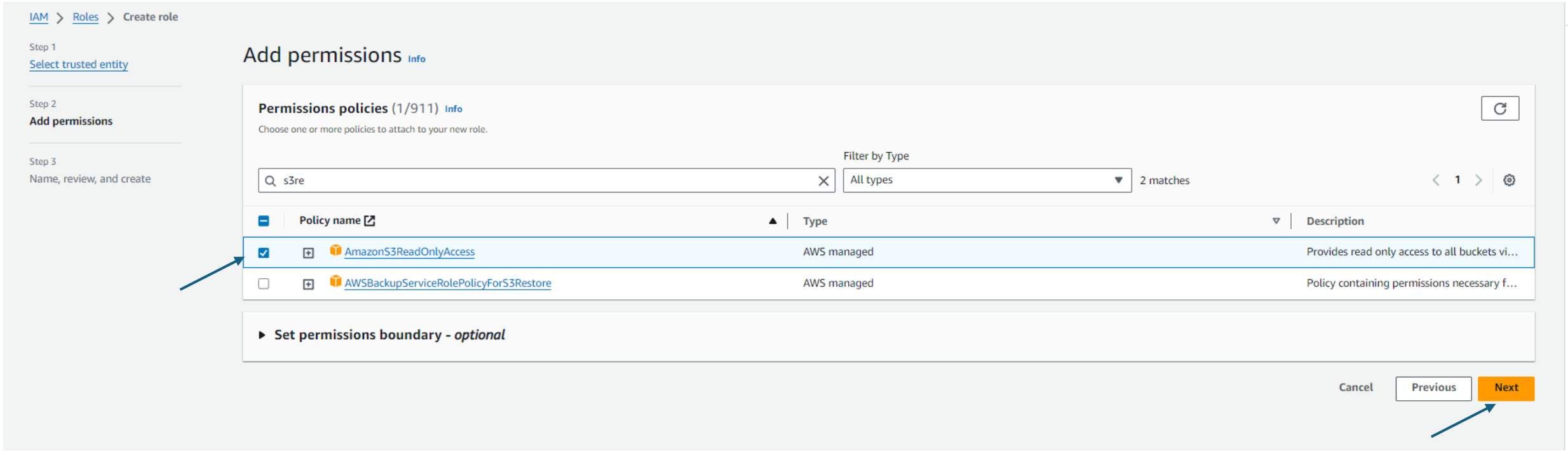
Redshift

Choose a use case for the specified service.

Use case

- (●) **Redshift - Customizable**
  Allows Redshift clusters to call AWS services on your behalf.
- ( ) **Redshift**
  Allows Redshift clusters to call AWS services on your behalf.
- ( ) **Redshift - Scheduler**
  Allow Redshift Scheduler to call Redshift on your behalf.

Cancel Next


IAM > Roles > Create role
Step 1
Select trusted entity
Step 2
Add permissions
Step 3
Name, review, and create
Add permissions Info
Permissions policies (1/911) Info
Choose one or more policies to attach to your new role.
Filter by Type
s3re
All types
2 matches
1
Policy name
Type
Description
AmazonS3ReadOnlyAccess
AWS managed
Provides read only access to all buckets vi...
AWSBackupServiceRolePolicyForS3Restore
AWS managed
Policy containing permissions necessary f...
Set permissions boundary - optional
Cancel
Previous
Next

Step 2
Add permissions

Step 3
Name, review, and create

Role details

Role name
Enter a meaningful name to identify this role.
smart-city-redshit-s3-role
Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

Description
Add a short explanation for this role.
Allows Redshift clusters to call AWS services on your behalf.
Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Step 1: Select trusted entities

Edit

Trust policy

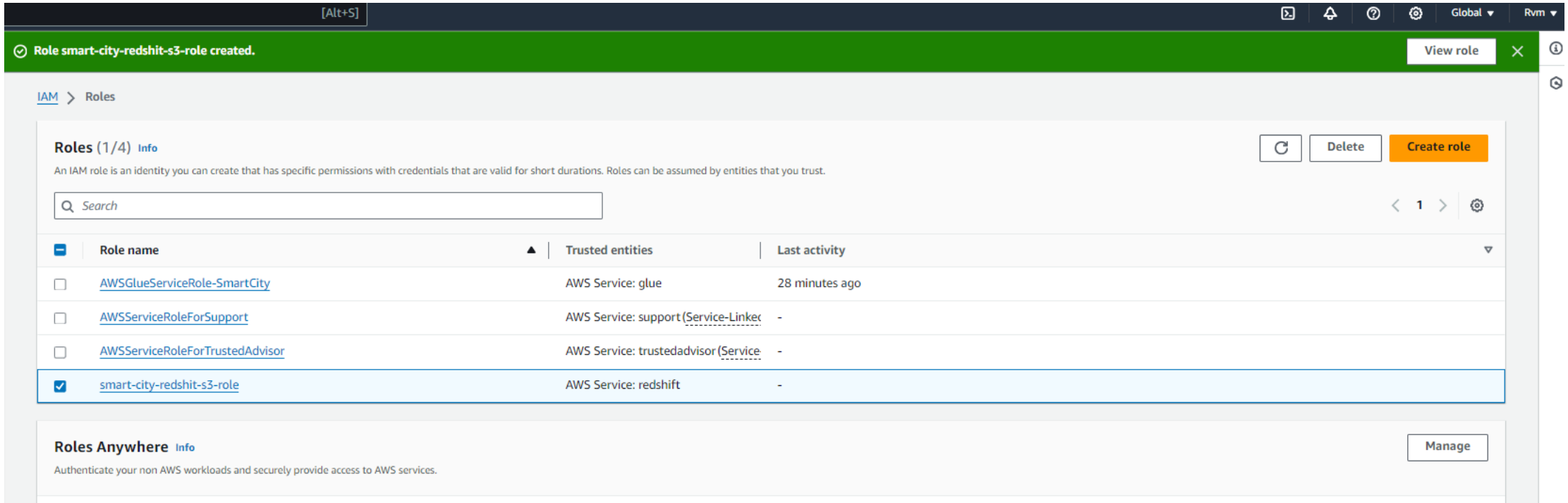
```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "sts:AssumeRole"
            ],
            "Principal": {
                "Service": [
                    "redshift.amazonaws.com"
                ]
            }
        }
    ]
}
```

Step 2: Add permissions

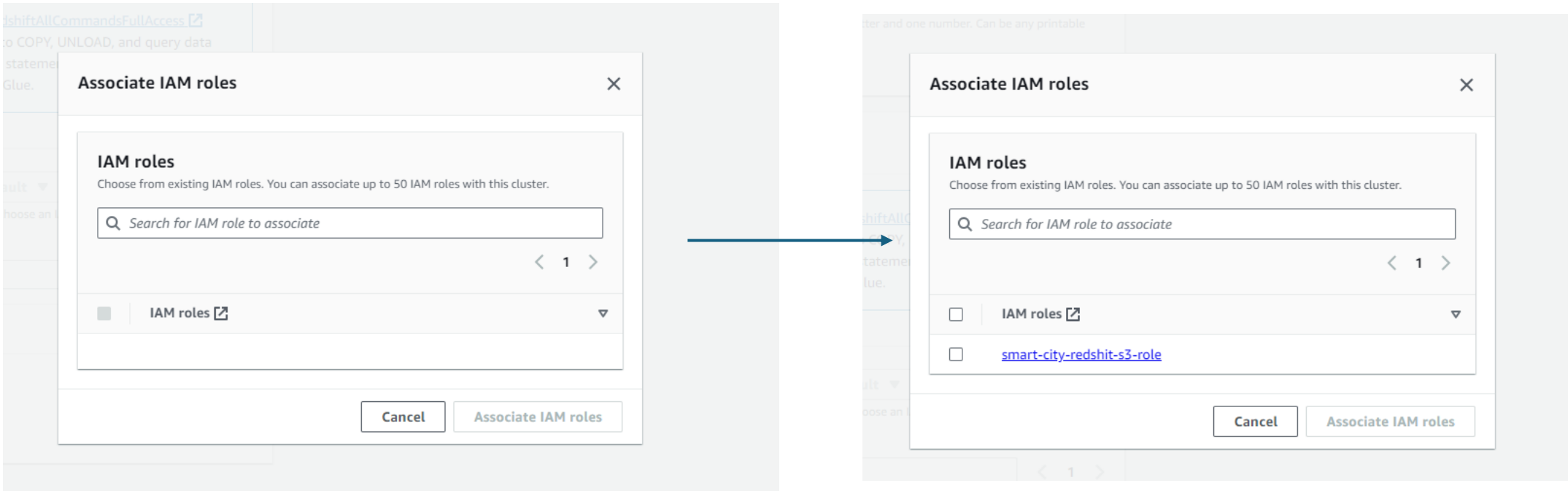
Edit

Permissions policy summary

Policy name	Type	Attached as

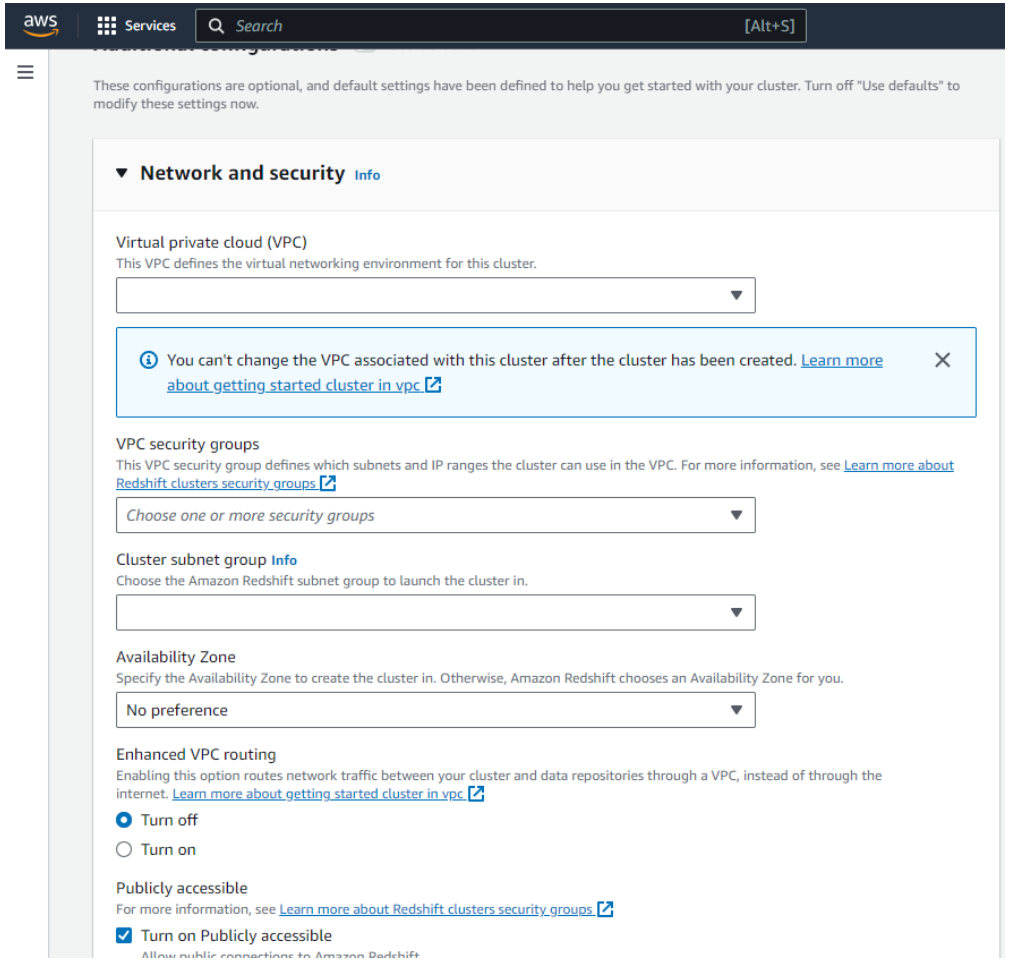

[Alt+S]
Global
Rvm
Role smart-city-redshit-s3-role created.
View role
IAM > Roles
Roles (1/4) Info
An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.
Delete
Create role
Search
1
Role name
Trusted entities
Last activity
AWSGlueServiceRole-SmartCity
AWS Service: glue
28 minutes ago
AWSServiceRoleForSupport
AWS Service: support (Service-Linkec
-
AWSServiceRoleForTrustedAdvisor
AWS Service: trustedadvisor (Service
-
smart-city-redshit-s3-role
AWS Service: redshift
-
Roles Anywhere Info
Authenticate your non AWS workloads and securely provide access to AWS services.
Manage

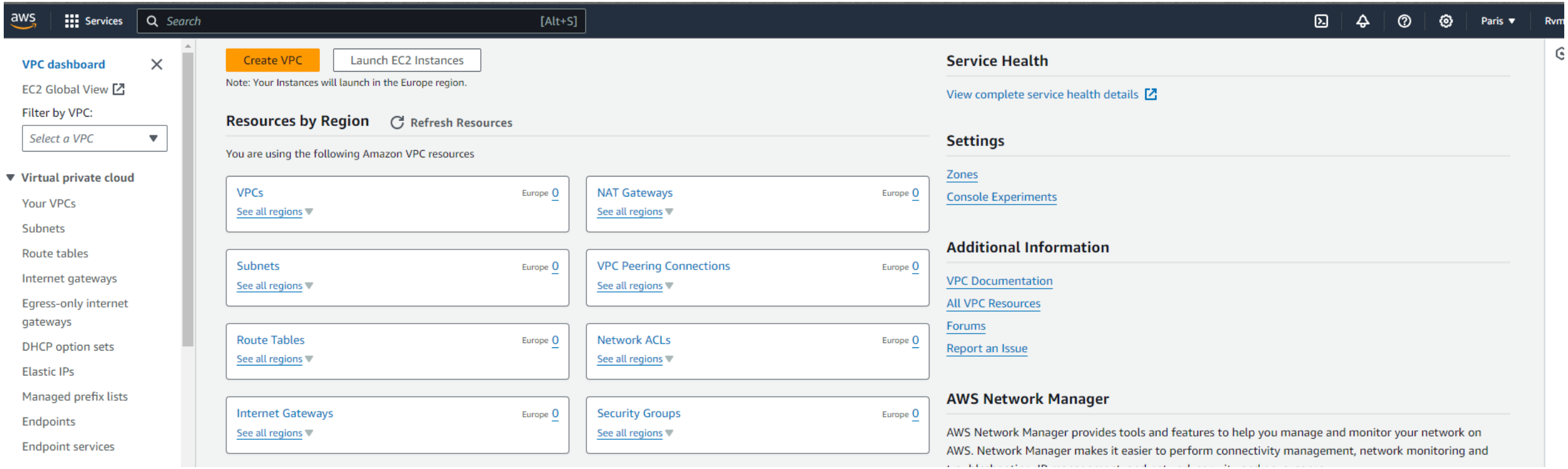
It doesnot refresh automaticaly, so you need to refresh the create cluster page and fill in all the information again.
By the iam roles you will be able to select the recently created role



If you don’t have any VPC. You need to create a new VPC which créate ifself automatically a security group (default).

You need to create a **Cluster subnet group** (this is mandatory to create the redshift cluster). This subnet group Will be assigned to the VPC created before. See next slides




aws
Services
Search
[Alt+S]
Paris
VPC dashboard
EC2 Global View
Filter by VPC:
Select a VPC
Virtual private cloud
Your VPCs
Subnets
Route tables
Internet gateways
Egress-only internet gateways
DHCP option sets
Elastic IPs
Managed prefix lists
Endpoints
Endpoint services
Create VPC
Launch EC2 Instances
Note: Your Instances will launch in the Europe region.
Resources by Region
Refresh Resources
You are using the following Amazon VPC resources
VPCs Europe 0 See all regions
NAT Gateways Europe 0 See all regions
Subnets Europe 0 See all regions
VPC Peering Connections Europe 0 See all regions
Route Tables Europe 0 See all regions
Network ACLs Europe 0 See all regions
Internet Gateways Europe 0 See all regions
Security Groups Europe 0 See all regions
Service Health
View complete service health details
Settings
Zones
Console Experiments
Additional Information
VPC Documentation
All VPC Resources
Forums
Report an Issue
AWS Network Manager
AWS Network Manager provides tools and features to help you manage and monitor your network on AWS. Network Manager makes it easier to perform connectivity management, network monitoring and

Create VPC Info

A VPC is an isolated portion of the AWS Cloud populated by AWS objects, such as Amazon EC2 instances. Mouse over a resource to highlight the related resources.

VPC settings

Resources to create Info
Create only the VPC resource or the VPC and other networking resources.

- ○ VPC only
- ◉ VPC and more

Name tag auto-generation Info
Enter a value for the Name tag. This value will be used to auto-generate Name tags for all resources in the VPC.

☑ Auto-generate

redshift-vpc

IPv4 CIDR block Info
Determine the starting IP and the size of your VPC using CIDR notation.

172.31.0.0/16 — 65,536 IPs

CIDR block size must be between /16 and /28.

IPv6 CIDR block Info

- ◉ No IPv6 CIDR block
- ○ Amazon-provided IPv6 CIDR block

Tenancy Info

Default

Number of Availability Zones (AZs) Info
Choose the number of AZs in which to provision subnets. We recommend at least two AZs for high availability.

1 | 2 | 3

▶ Customize AZs

Number of public subnets Info
The number of public subnets to add to your VPC. Use public subnets for web applications that need to be publicly accessible over the internet.

0 | 1

Number of private subnets Info
The number of private subnets to add to your VPC. Use private subnets to secure backend resources that don't need public access.

0 | 1 | 2

▶ Customize subnets CIDR blocks

NAT gateways ($) Info
Choose the number of Availability Zones (AZs) in which to create NAT gateways. Note that there is a charge for each NAT gateway

None | In 1 AZ | 1 per AZ

VPC endpoints Info
Endpoints can help reduce NAT gateway charges and improve security by accessing S3 directly from the VPC. By default, full access policy is used. You can customize this policy at any time.

None | S3 Gateway

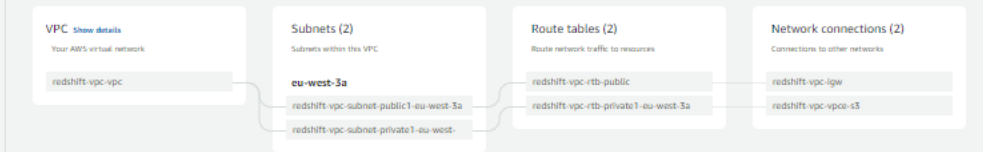
DNS options Info

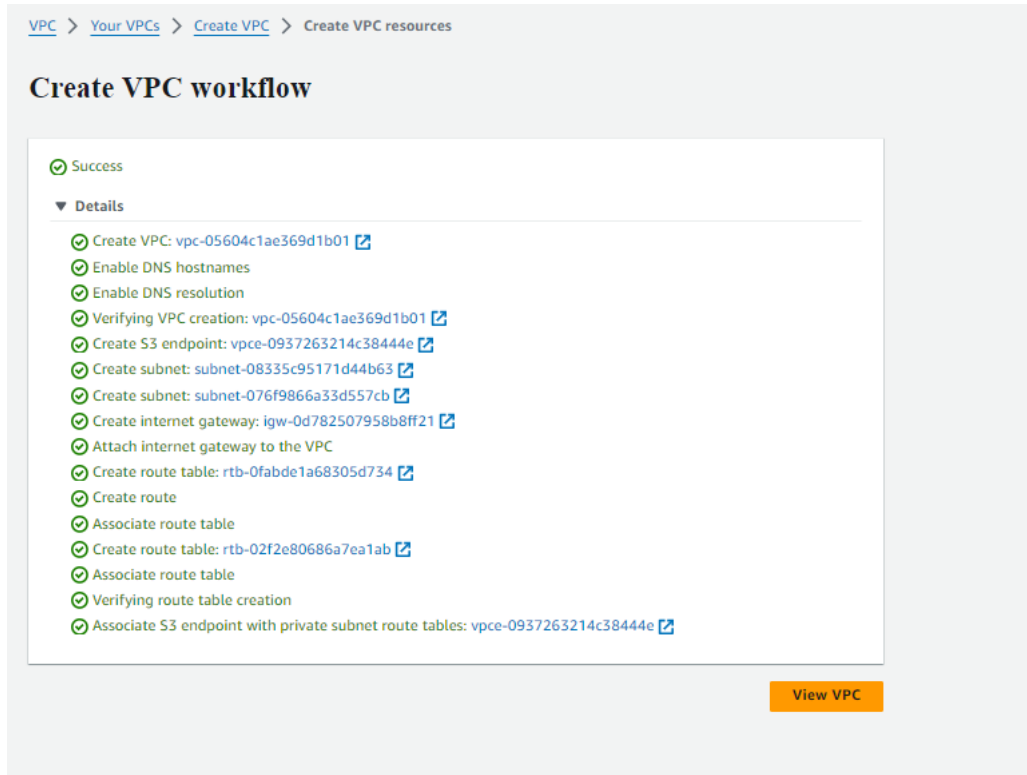
☑ Enable DNS hostnames
☑ Enable DNS resolution

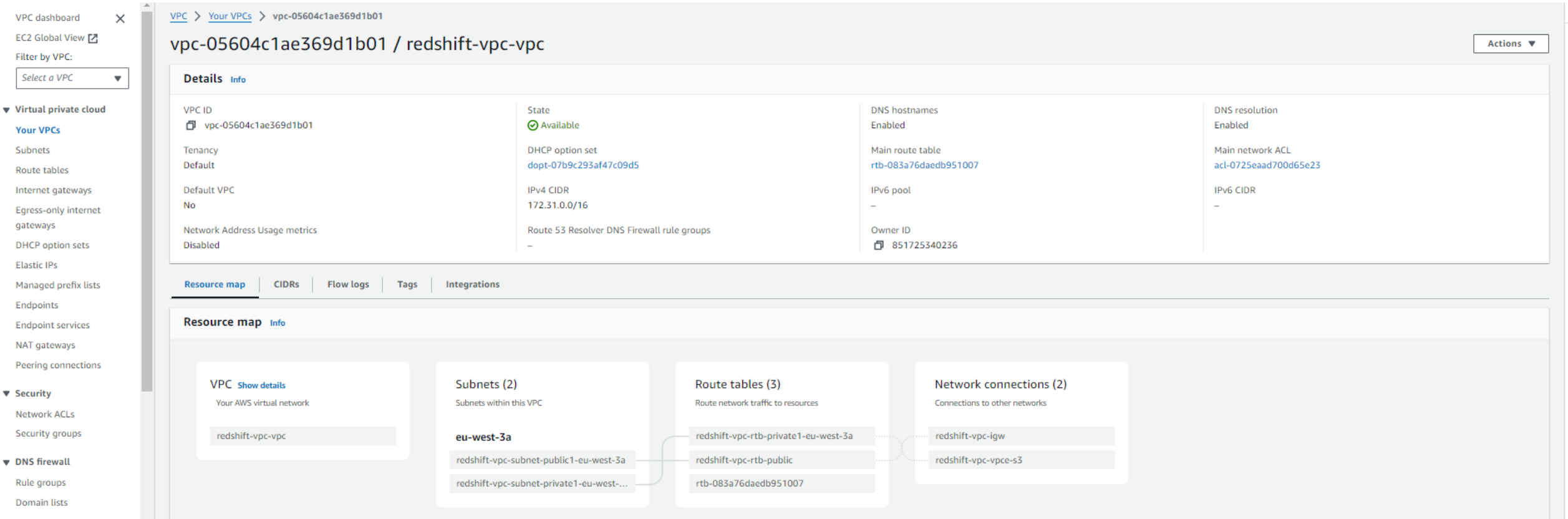
▶ Additional tags

Cancel | Create VPC

Preview



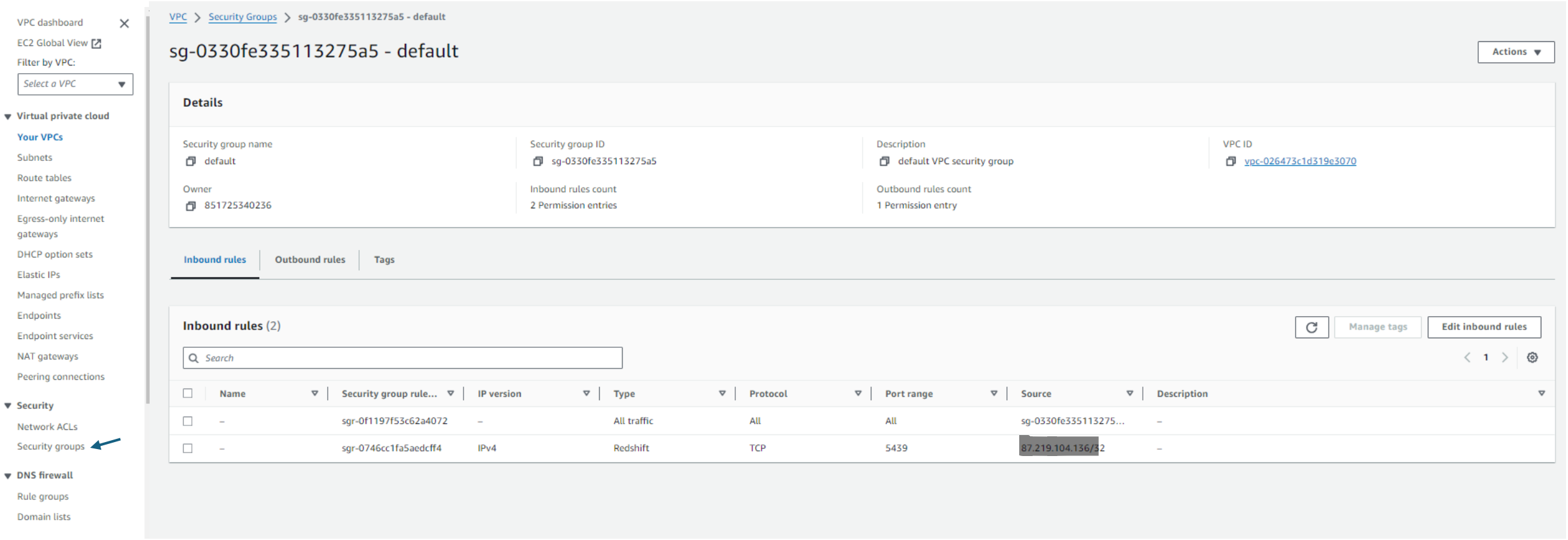

VPC > Your VPCs > Create VPC > Create VPC resources
Create VPC workflow
Success
Details
Create VPC: vpc-05604c1ae369d1b01
Enable DNS hostnames
Enable DNS resolution
Verifying VPC creation: vpc-05604c1ae369d1b01
Create S3 endpoint: vpce-0937263214c38444e
Create subnet: subnet-08335c95171d44b63
Create subnet: subnet-076f9866a33d557cb
Create internet gateway: igw-0d782507958b8ff21
Attach internet gateway to the VPC
Create route table: rtb-0fabde1a68305d734
Create route
Associate route table
Create route table: rtb-02f2e80686a7ea1ab
Associate route table
Verifying route table creation
Associate S3 endpoint with private subnet route tables: vpce-0937263214c38444e
View VPC


VPC dashboard
EC2 Global View
Filter by VPC:
Select a VPC
Virtual private cloud
Your VPCs
Subnets
Route tables
Internet gateways
Egress-only internet gateways
DHCP option sets
Elastic IPs
Managed prefix lists
Endpoints
Endpoint services
NAT gateways
Peering connections
Security
Network ACLs
Security groups
DNS firewall
Rule groups
Domain lists
VPC > Your VPCs > vpc-05604c1ae369d1b01
vpc-05604c1ae369d1b01 / redshift-vpc-vpc
Actions
Details Info
VPC ID
vpc-05604c1ae369d1b01
State
Available
DNS hostnames
Enabled
DNS resolution
Enabled
Tenancy
Default
DHCP option set
dopt-07b9c293af47c09d5
Main route table
rtb-083a76daedb951007
Main network ACL
acl-0725eaad700d65e23
Default VPC
No
IPv4 CIDR
172.31.0.0/16
IPv6 pool
–
IPv6 CIDR
–
Network Address Usage metrics
Disabled
Route 53 Resolver DNS Firewall rule groups
–
Owner ID
851725340236
Resource map
CIDRs
Flow logs
Tags
Integrations
Resource map Info
VPC Show details
Your AWS virtual network
redshift-vpc-vpc
Subnets (2)
Subnets within this VPC
eu-west-3a
redshift-vpc-subnet-public1-eu-west-3a
redshift-vpc-subnet-private1-eu-west-...
Route tables (3)
Route network traffic to resources
redshift-vpc-rtb-private1-eu-west-3a
redshift-vpc-rtb-public
rtb-083a76daedb951007
Network connections (2)
Connections to other networks
redshift-vpc-igw
redshift-vpc-vpce-s3

Important:
Go to VPC console > Security Groups > security group (default)

There add an **inbound rule**:
Type:Custom port: 5439 source: My Ip

This allow the traffic from your IP to readshift. Without this you won't be able to connect redshift from your pc



Network and security Info

Virtual private cloud (VPC)
This VPC defines the virtual networking environment for this cluster.

redshift-vpc-vpc
vpc-05604c1ae369d1b01

You can't change the VPC associated with this cluster after the cluster has been created. Learn more about getting started cluster in vpc

VPC security groups
This VPC security group defines which subnets and IP ranges the cluster can use in the VPC. For more information, see Learn more about Redshift clusters security groups

Choose one or more security groups

default
sg-0e14a3b2c904bdccd

Cluster subnet group Info
Choose the Amazon Redshift subnet group to launch the cluster in.

Availability Zone
Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

Enhanced VPC routing
Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. Learn more about getting started cluster in vpc

- (•) Turn off
- ( ) Turn on

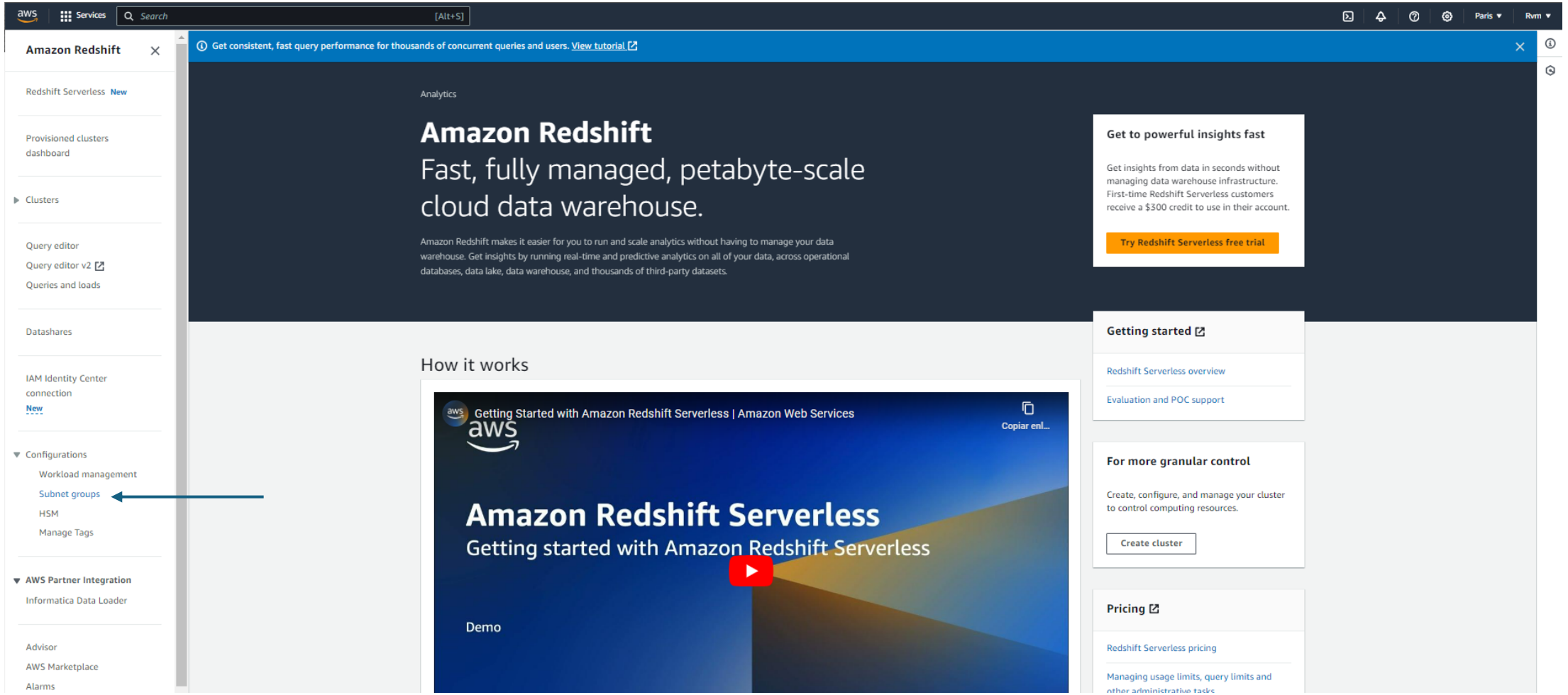
Publicly accessible
For more information, see Learn more about Redshift clusters security groups

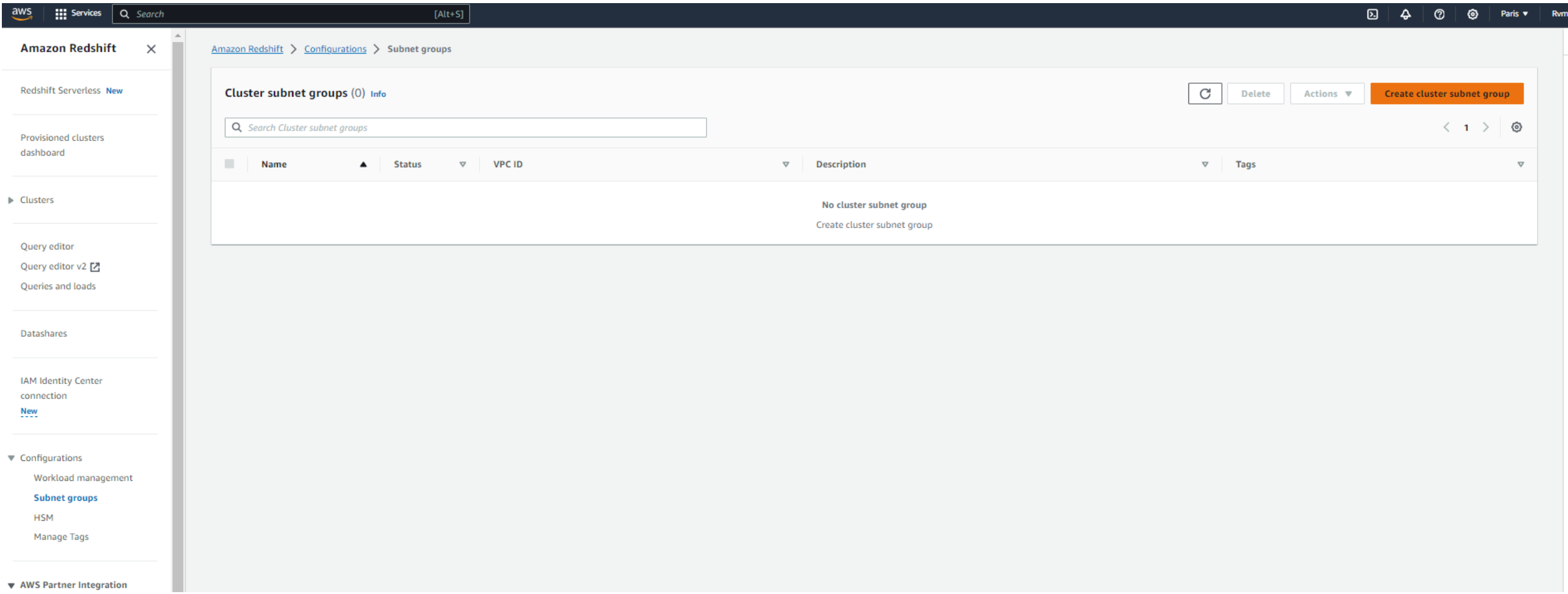
- [x] Turn on Publicly accessible
  Allow public connections to Amazon Redshift.

Elastic IP address
Select the Elastic IP address for connecting to the cluster.

It can take about ten minutes for the setting to change and connections to succeed.

create a **Cluster subnet group** (this is mandatory to create the redshift cluster). This subnet group Will be assigned to the VPC created before. See next slides




aws
Services
Search
[Alt+S]
Paris
Rvm
Amazon Redshift
Redshift Serverless New
Provisioned clusters dashboard
Clusters
Query editor
Query editor v2
Queries and loads
Datashares
IAM Identity Center connection
New
Configurations
Workload management
Subnet groups
HSM
Manage Tags
AWS Partner Integration
Amazon Redshift > Configurations > Subnet groups
Cluster subnet groups (0) Info
Delete
Actions
Create cluster subnet group
Search Cluster subnet groups
1
Name
Status
VPC ID
Description
Tags
No cluster subnet group
Create cluster subnet group

Create cluster subnet group Info

Cluster subnet group details

Name
You can't modify the name after your subnet group has been created.
cluster-subnet-group-1
The name must be 1-255 characters. Valid characters are A-Z, a-z, 0-9, space, hyphen (-), underscore (_), and period (.).

Description
redshift-cluster-subnet-group

Add subnets

VPC
Choose the VPC that contains the subnets that you want to include in your cluster subnet group.
vpc-05604c1ae369d1b01

Add all the subnets for this VPC

Availability Zone
Choose an Availability Zone

Subnet

Add subnet

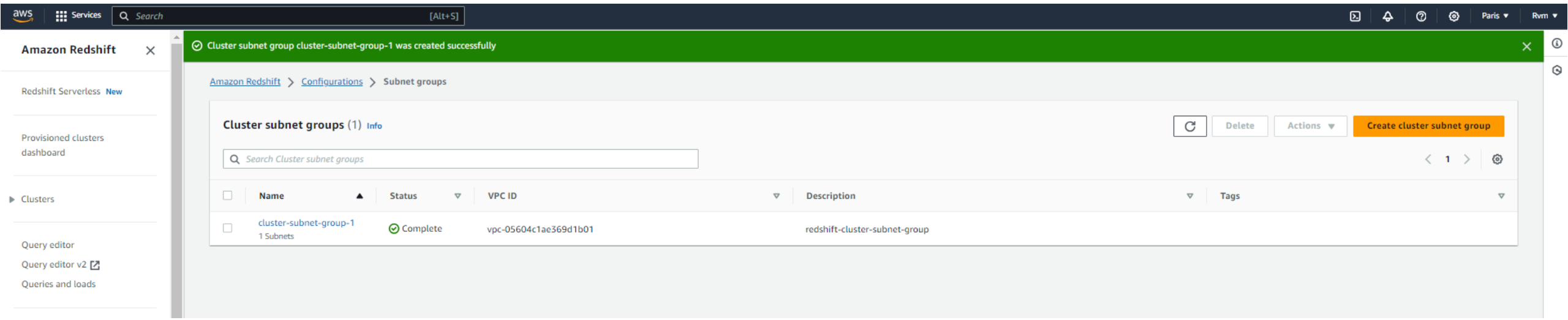
Subnets in this cluster subnet group (1)

Remove all

Availability Zone	Subnet ID	CIDR block	IPv6 CIDR block	Action
eu-west-3a	subnet-076f98...	172.31.128.0/20	-	Remove

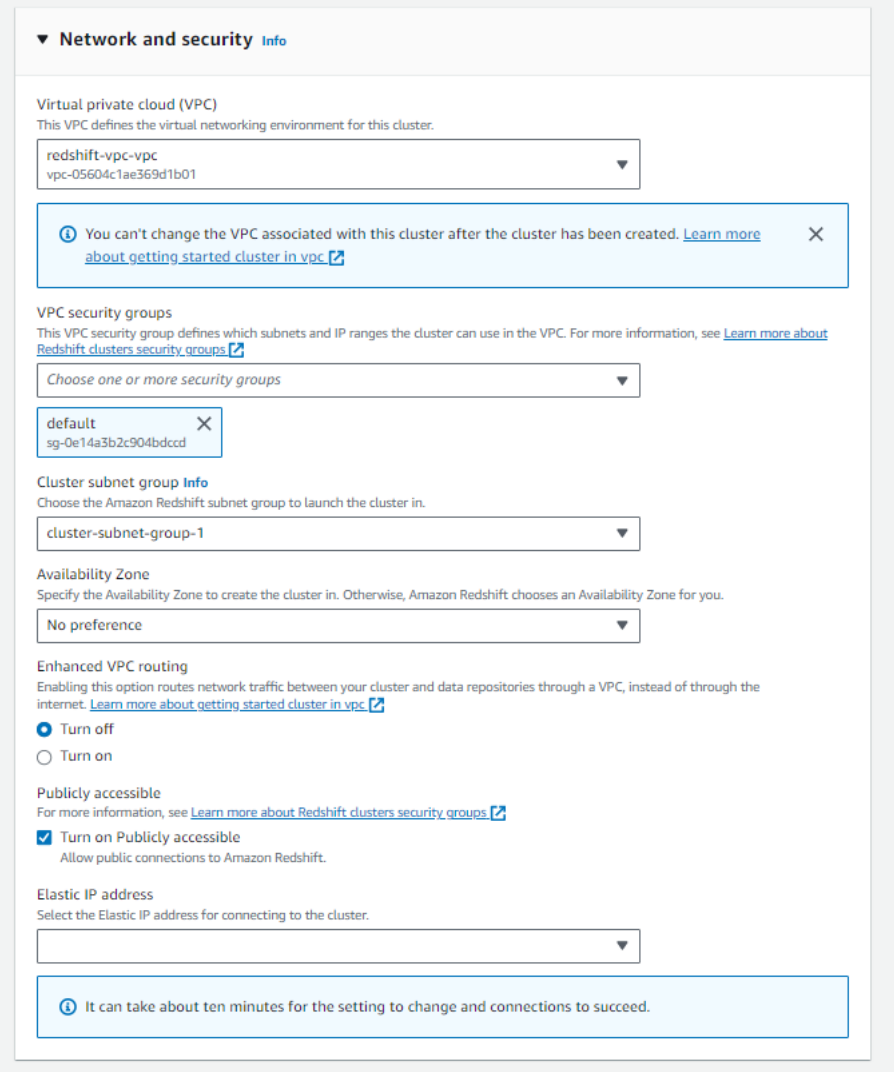
Cancel

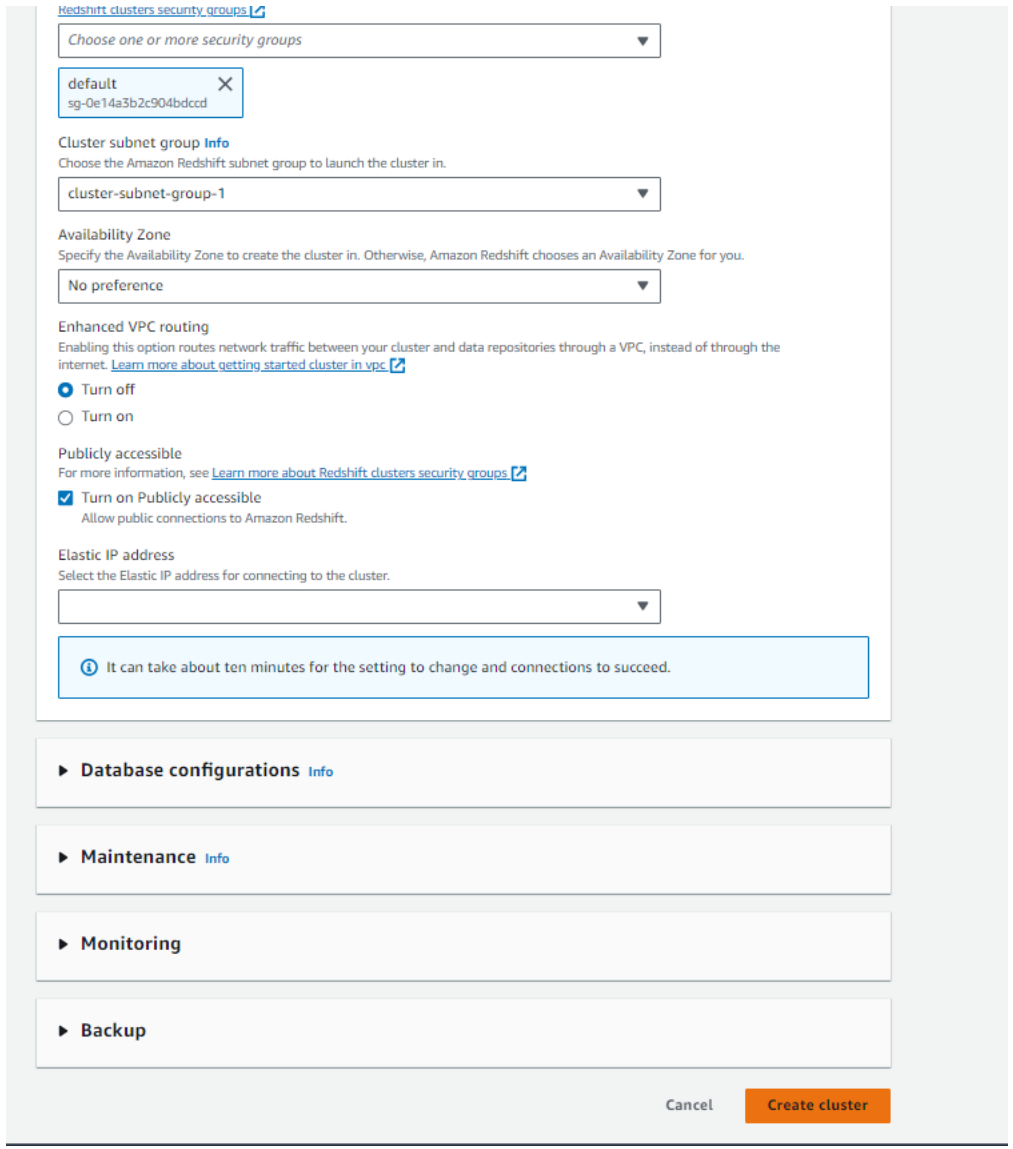
Create cluster subnet group

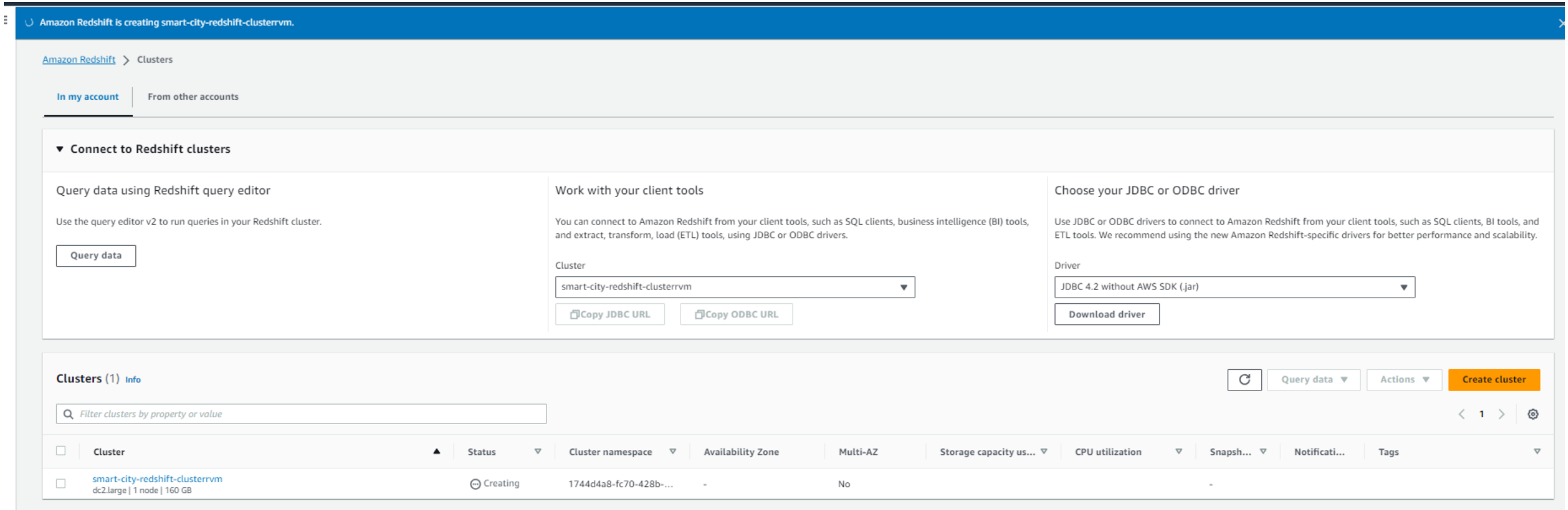

aws
Services
Search
[Alt+S]
Paris
Rvm
Amazon Redshift
Redshift Serverless New
Provisioned clusters dashboard
Clusters
Query editor
Query editor v2
Queries and loads
Cluster subnet group cluster-subnet-group-1 was created successfully
Amazon Redshift > Configurations > Subnet groups
Cluster subnet groups (1) Info
Delete
Actions
Create cluster subnet group
Search Cluster subnet groups
1
Name
Status
VPC ID
Description
Tags
cluster-subnet-group-1
1 Subnets
Complete
vpc-05604c1ae369d1b01
redshift-cluster-subnet-group

After resfreshing, again fill in the the fields...

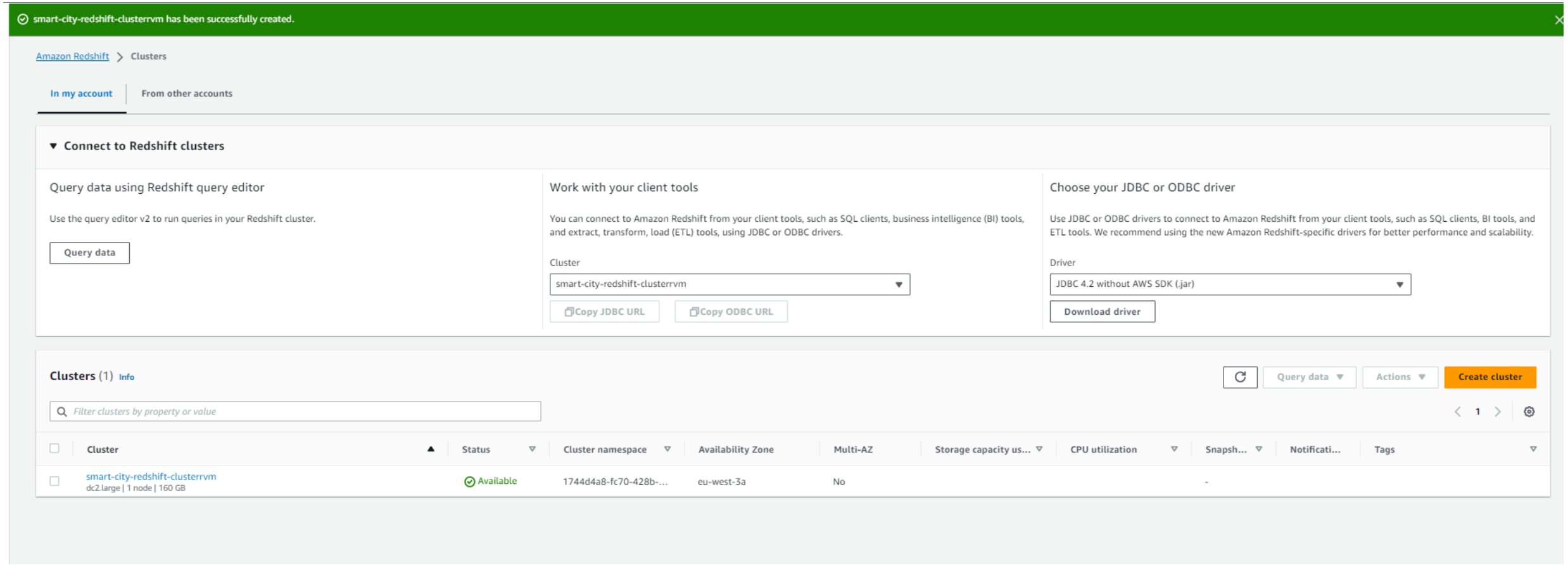
If you have read this documentation previously, you can create all the needed VPC and cluster subnet group in advance to ease the process




Redshift clusters security groups
Choose one or more security groups
default
sg-0e14a3b2c904bdccd
Cluster subnet group Info
Choose the Amazon Redshift subnet group to launch the cluster in.
cluster-subnet-group-1
Availability Zone
Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.
No preference
Enhanced VPC routing
Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. Learn more about getting started cluster in vpc
Turn off
Turn on
Publicly accessible
For more information, see Learn more about Redshift clusters security groups
Turn on Publicly accessible
Allow public connections to Amazon Redshift.
Elastic IP address
Select the Elastic IP address for connecting to the cluster.
It can take about ten minutes for the setting to change and connections to succeed.
Database configurations Info
Maintenance Info
Monitoring
Backup
Cancel
Create cluster

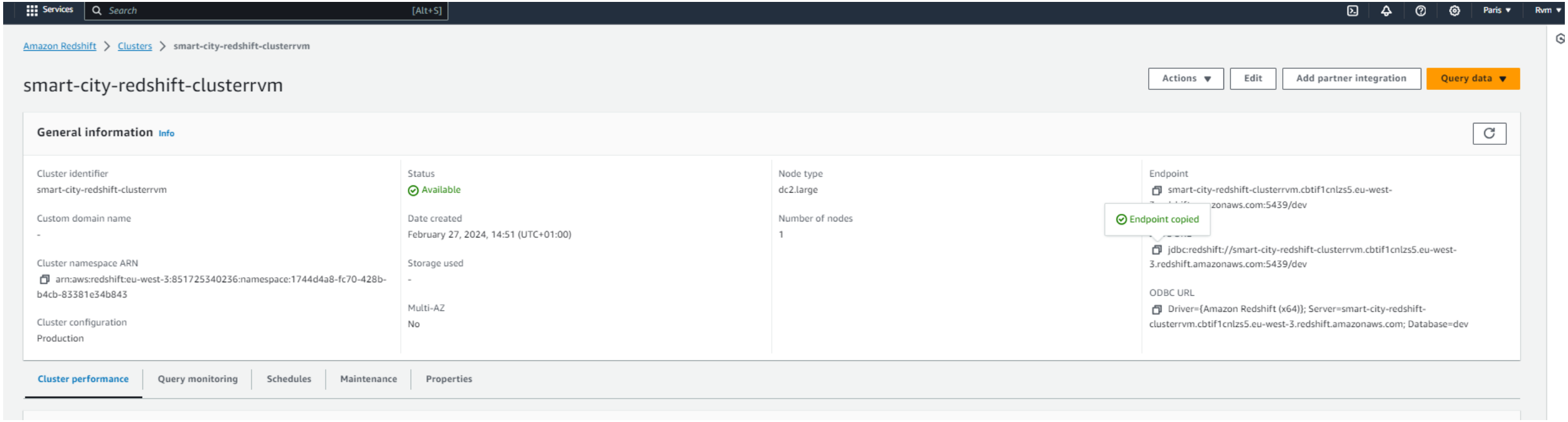

Amazon Redshift is creating smart-city-redshift-clusterrvm.
Amazon Redshift > Clusters
In my account
From other accounts
Connect to Redshift clusters
Query data using Redshift query editor
Use the query editor v2 to run queries in your Redshift cluster.
Query data
Work with your client tools
You can connect to Amazon Redshift from your client tools, such as SQL clients, business intelligence (BI) tools, and extract, transform, load (ETL) tools, using JDBC or ODBC drivers.
Cluster
smart-city-redshift-clusterrvm
Copy JDBC URL
Copy ODBC URL
Choose your JDBC or ODBC driver
Use JDBC or ODBC drivers to connect to Amazon Redshift from your client tools, such as SQL clients, BI tools, and ETL tools. We recommend using the new Amazon Redshift-specific drivers for better performance and scalability.
Driver
JDBC 4.2 without AWS SDK (.jar)
Download driver
Clusters (1) Info
Query data
Actions
Create cluster
Filter clusters by property or value
1
Cluster
Status
Cluster namespace
Availability Zone
Multi-AZ
Storage capacity us...
CPU utilization
Snapsh...
Notificati...
Tags
smart-city-redshift-clusterrvm
dc2.large | 1 node | 160 GB
Creating
1744d4a8-fc70-428b-...
-
No
-

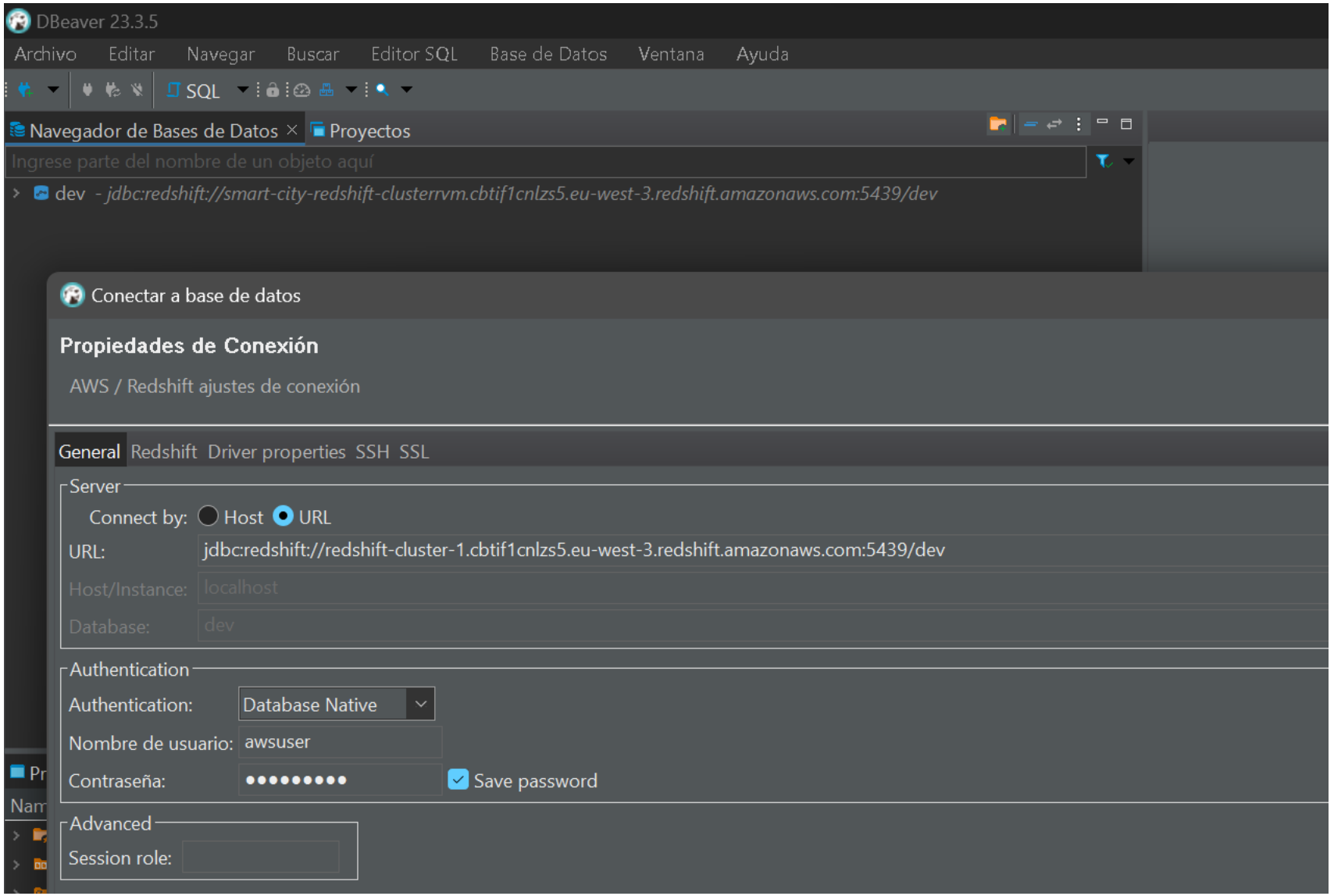
It takes around 15 minutes to be created



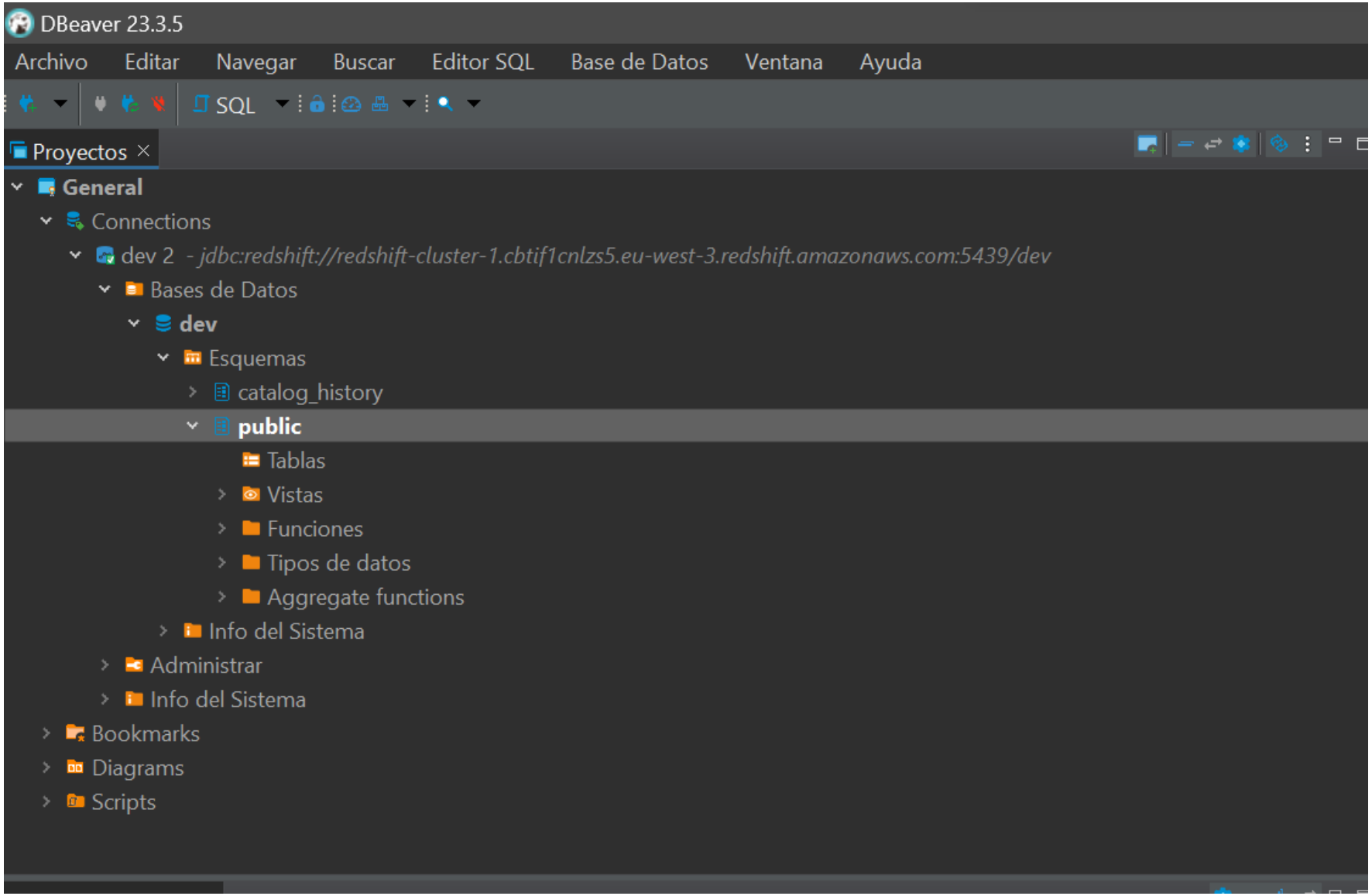
The Redshift cluster is already created, then lets explore the content and create an external schema that points to the tables on Glue. In this way you will be able to query Redshift database directly and retrieve the transformed data persisted in Glue.

To do that, just copy the jdbc connection chain and create a server connection in your computer using Dbeaver.

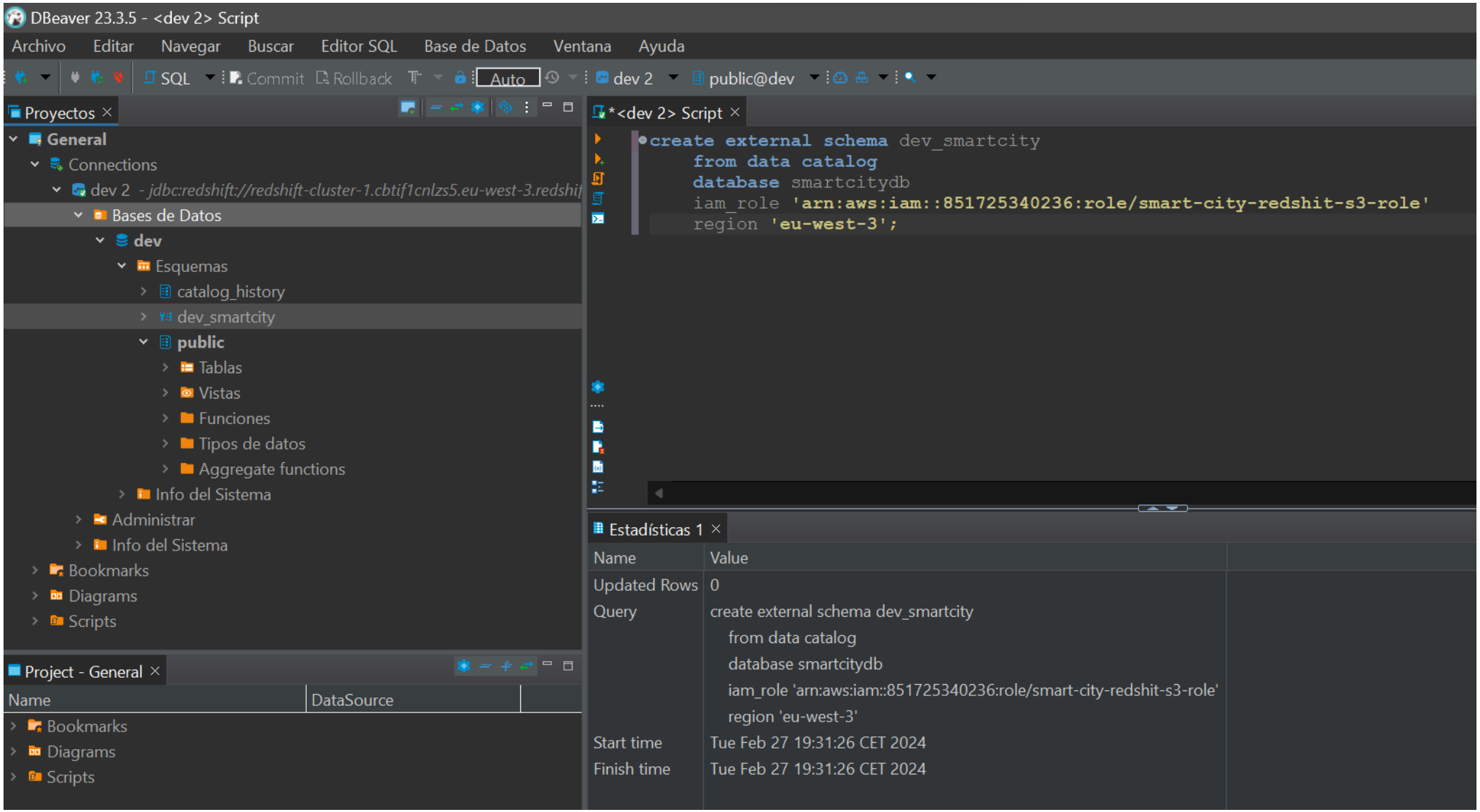


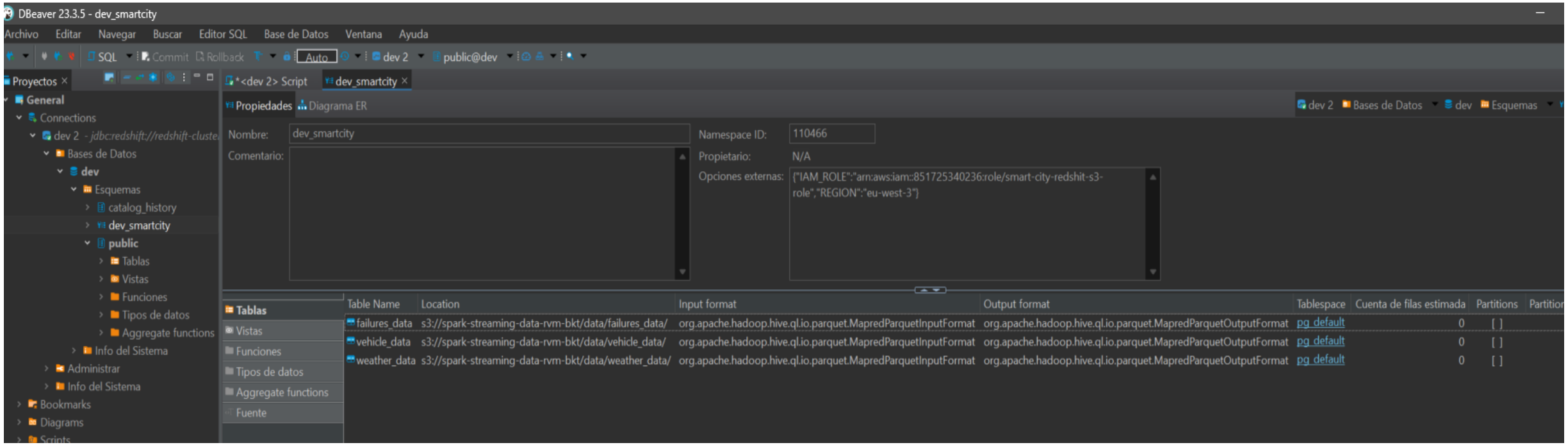

DBeaver 23.3.5
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
SQL
Navegador de Bases de Datos
Proyectos
Ingrese parte del nombre de un objeto aquí
dev - jdbc:redshift://smart-city-redshift-clusterrvm.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com:5439/dev
Conectar a base de datos
Propiedades de Conexión
AWS / Redshift ajustes de conexión
General Redshift Driver properties SSH SSL
Server
Connect by: Host URL
URL: jdbc:redshift://redshift-cluster-1.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com:5439/dev
Host/Instance: localhost
Database: dev
Authentication
Authentication: Database Native
Nombre de usuario: awsuser
Contraseña: ●●●●●●●●●
Save password
Advanced
Session role:

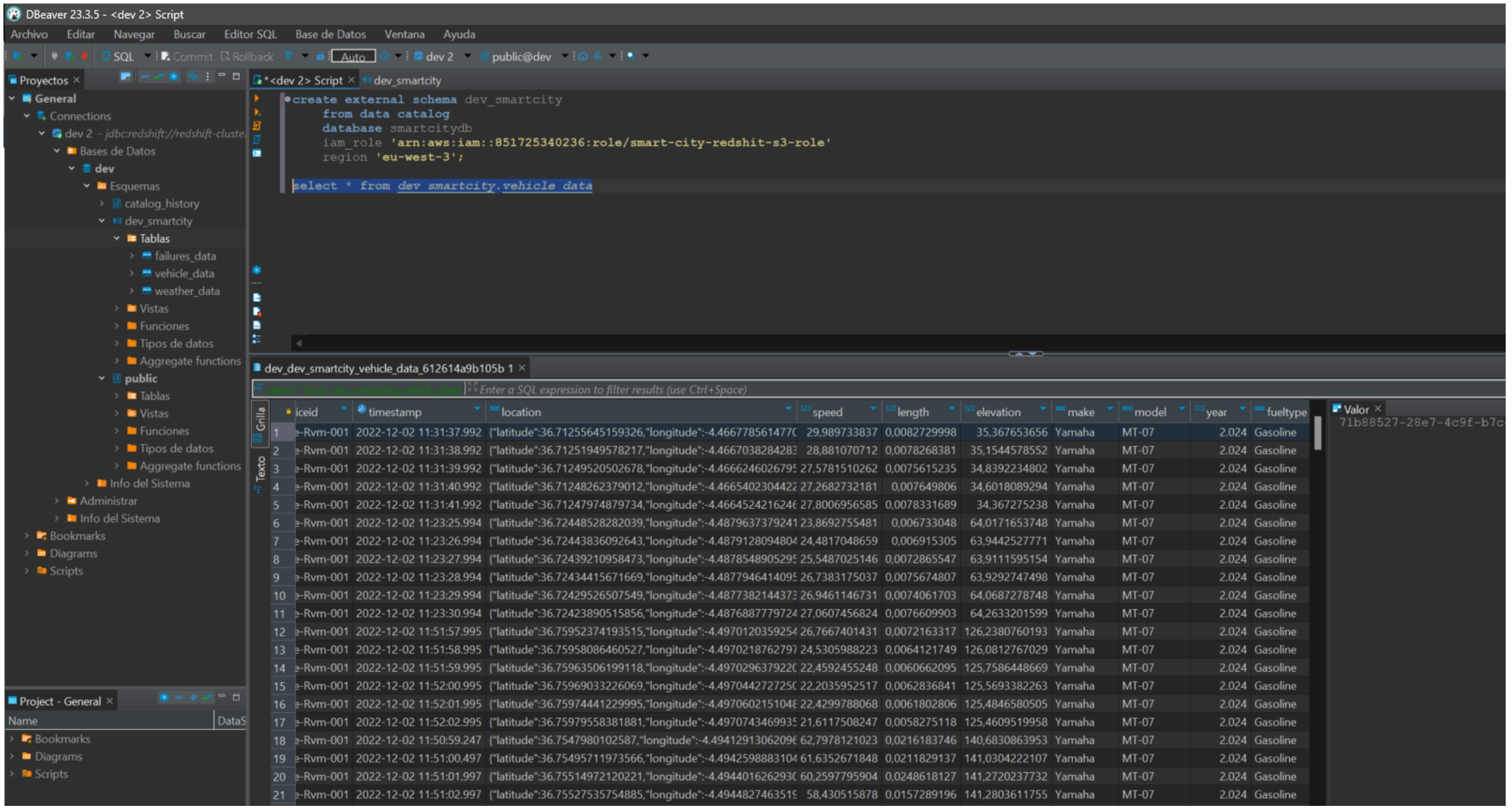
Once connected to redshift you can explore the existing schemas.
In this case, public is where we want to work in, but there aren't tables yet.

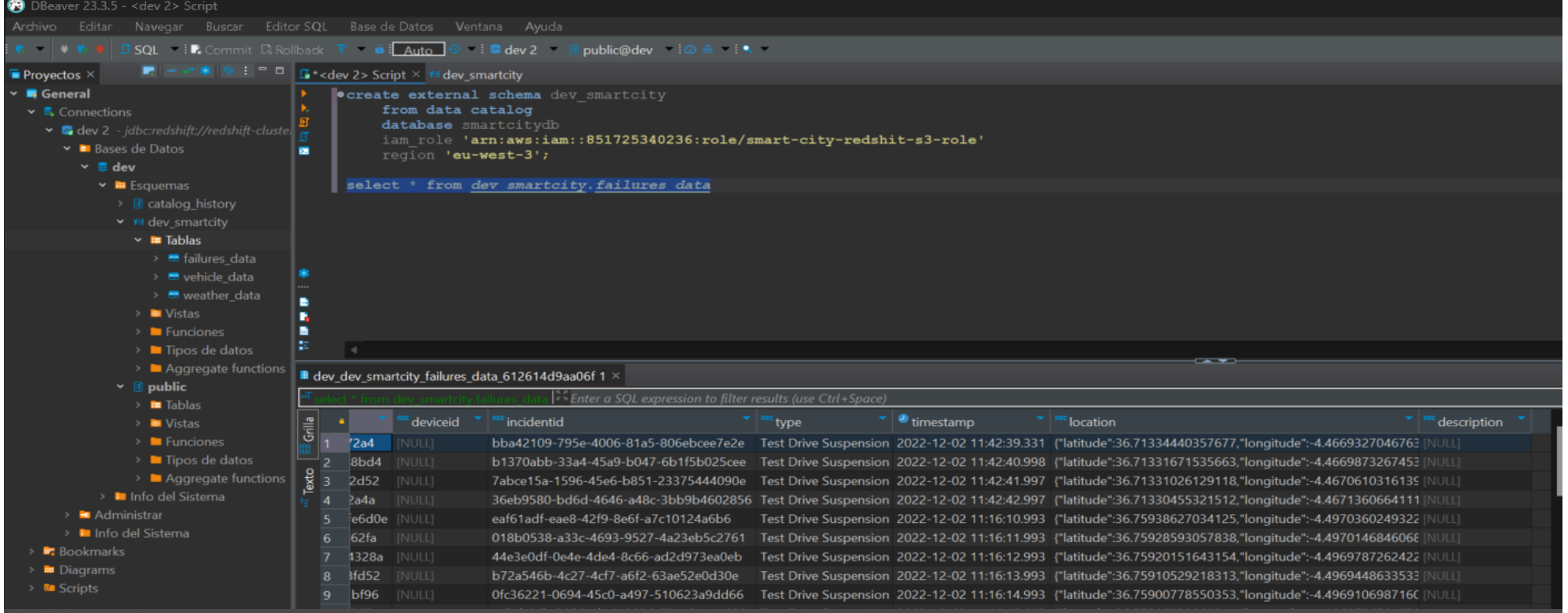


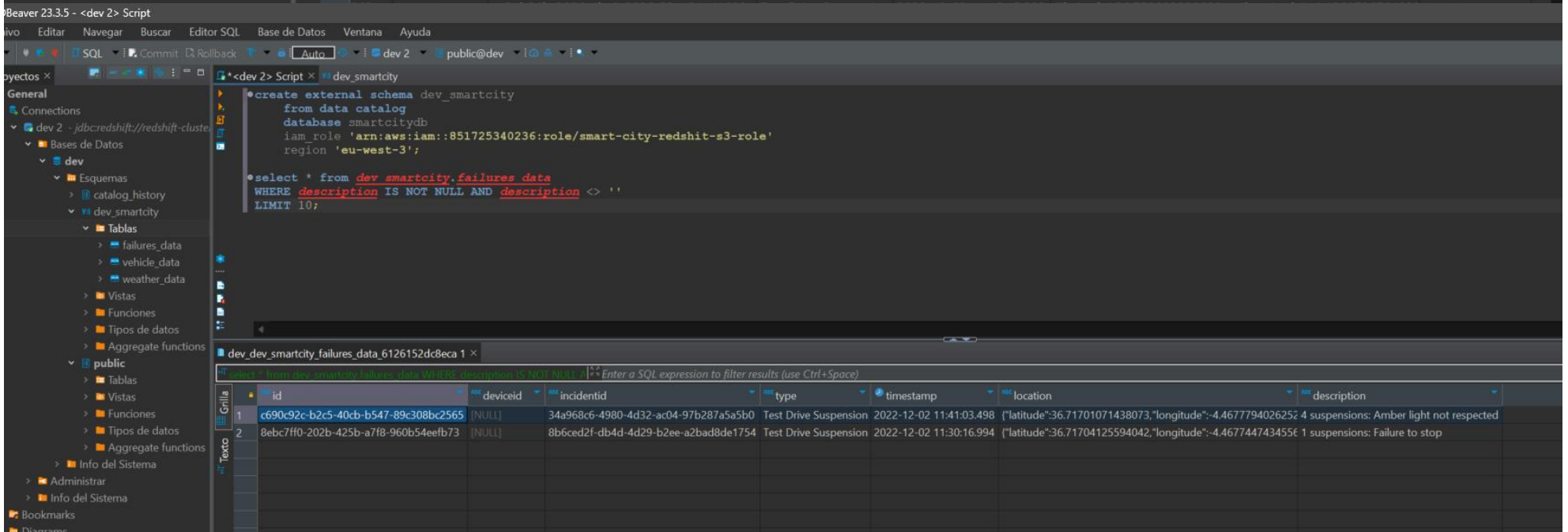
Create an external schema pointing to the Glue catalog and created database.
Pass your IAM role – go to IAM and select the role created before and copy the arn.
This schema allows to query against redshift and retrieving the data from Glue

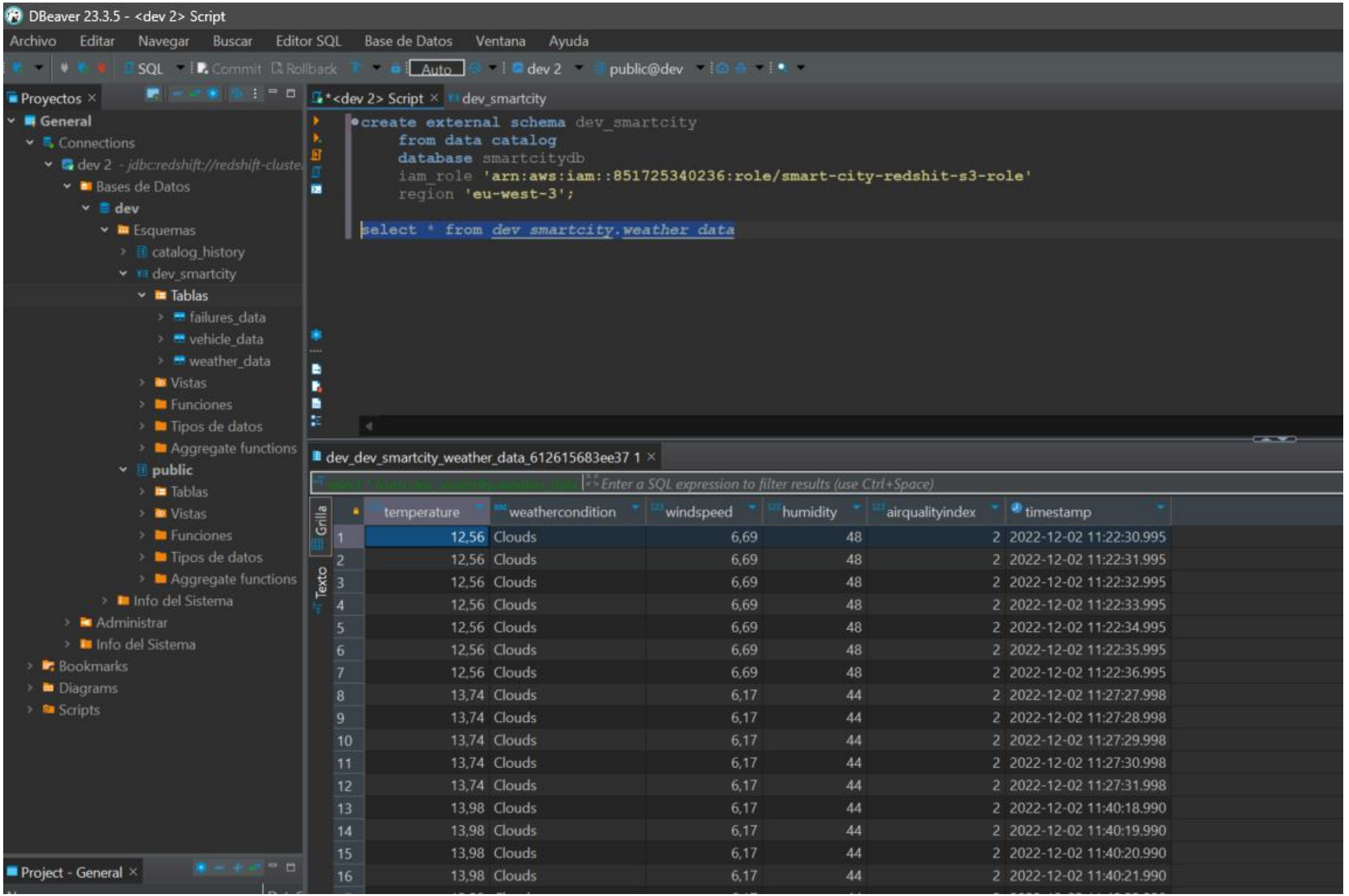



DBeaver 23.3.5 - dev_smartcity
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
SQL Commit Rollback Auto dev 2 public@dev
Proyectos
General
Connections
dev 2 - jdbc:redshift://redshift-cluste
Bases de Datos
dev
Esquemas
catalog_history
dev_smartcity
public
Tablas
Vistas
Funciones
Tipos de datos
Aggregate functions
Info del Sistema
Administrar
Info del Sistema
Bookmarks
Diagrams
Scripts
*<dev 2> Script
dev_smartcity
Propiedades
Diagrama ER
dev 2 Bases de Datos dev Esquemas
Nombre: dev_smartcity
Comentario:
Namespace ID: 110466
Propietario: N/A
Opciones externas: {"IAM_ROLE":"arn:aws:iam::851725340236:role/smart-city-redshit-s3-role","REGION":"eu-west-3"}
Tablas
Vistas
Funciones
Tipos de datos
Aggregate functions
Fuente
Table Name Location Input format Output format Tablespace Cuenta de filas estimada Partitions Partitior
failures_data s3://spark-streaming-data-rvm-bkt/data/failures_data/ org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat pg_default 0 [ ]
vehicle_data s3://spark-streaming-data-rvm-bkt/data/vehicle_data/ org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat pg_default 0 [ ]
weather_data s3://spark-streaming-data-rvm-bkt/data/weather_data/ org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat pg_default 0 [ ]


DBeaver 23.3.5 - <dev 2> Script
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
Proyectos
General
Connections
dev 2 - jdbc:redshift://redshift-cluste
Bases de Datos
dev
Esquemas
catalog_history
dev_smartcity
Tablas
failures_data
vehicle_data
weather_data
Vistas
Funciones
Tipos de datos
Aggregate functions
public
Tablas
Vistas
Funciones
Tipos de datos
Aggregate functions
Info del Sistema
Administrar
Info del Sistema
Bookmarks
Diagrams
Scripts
Project - General
Name
Bookmarks
Diagrams
Scripts
*<dev 2> Script
dev_smartcity
create external schema dev_smartcity
from data catalog
database smartcitydb
iam_role 'arn:aws:iam::851725340236:role/smart-city-redshit-s3-role'
region 'eu-west-3';
select * from dev_smartcity.vehicle_data
dev_dev_smartcity_vehicle_data_612614a9b105b 1
Enter a SQL expression to filter results (use Ctrl+Space)
Grilla
Texto
iceid timestamp location speed length elevation make model year fueltype
1 e-Rvm-001 2022-12-02 11:31:37.992 {"latitude":36.71255645159326,"longitude":-4.466778561477(29,989733837 0,0082729998 35,367653656 Yamaha MT-07 2.024 Gasoline
2 e-Rvm-001 2022-12-02 11:31:38.992 {"latitude":36.71251949578217,"longitude":-4.4667038284283 28,881070712 0,0078268381 35,1544578552 Yamaha MT-07 2.024 Gasoline
3 e-Rvm-001 2022-12-02 11:31:39.992 {"latitude":36.71249520502678,"longitude":-4.4666246026795 27,5781510262 0,0075615235 34,8392234802 Yamaha MT-07 2.024 Gasoline
4 e-Rvm-001 2022-12-02 11:31:40.992 {"latitude":36.71248262379012,"longitude":-4.4665402304422 27,2682732181 0,007649806 34,6018089294 Yamaha MT-07 2.024 Gasoline
5 e-Rvm-001 2022-12-02 11:31:41.992 {"latitude":36.71247974879734,"longitude":-4.4664524216246 27,8006956585 0,0078331689 34,367275238 Yamaha MT-07 2.024 Gasoline
6 e-Rvm-001 2022-12-02 11:23:25.994 {"latitude":36.72448528282039,"longitude":-4.4879637379241 23,8692755481 0,006733048 64,0171653748 Yamaha MT-07 2.024 Gasoline
7 e-Rvm-001 2022-12-02 11:23:26.994 {"latitude":36.72443836092643,"longitude":-4.4879128094804 24,4817048659 0,006915305 63,9442527771 Yamaha MT-07 2.024 Gasoline
8 e-Rvm-001 2022-12-02 11:23:27.994 {"latitude":36.72439210958473,"longitude":-4.4878548905295 25,5487025146 0,0072865547 63,9111595154 Yamaha MT-07 2.024 Gasoline
9 e-Rvm-001 2022-12-02 11:23:28.994 {"latitude":36.72434415671669,"longitude":-4.4877946414095 26,7383175037 0,0075674807 63,9292747498 Yamaha MT-07 2.024 Gasoline
10 e-Rvm-001 2022-12-02 11:23:29.994 {"latitude":36.72429526507549,"longitude":-4.4877382144373 26,9461146731 0,0074061703 64,0687278748 Yamaha MT-07 2.024 Gasoline
11 e-Rvm-001 2022-12-02 11:23:30.994 {"latitude":36.72423890515856,"longitude":-4.4876887779724 27,0607456824 0,0076609903 64,2633201599 Yamaha MT-07 2.024 Gasoline
12 e-Rvm-001 2022-12-02 11:51:57.995 {"latitude":36.75952374193515,"longitude":-4.4970120359254 26,7667401431 0,0072163317 126,2380760193 Yamaha MT-07 2.024 Gasoline
13 e-Rvm-001 2022-12-02 11:51:58.995 {"latitude":36.75958086460527,"longitude":-4.4970218762797 24,5305988223 0,0064121749 126,0812767029 Yamaha MT-07 2.024 Gasoline
14 e-Rvm-001 2022-12-02 11:51:59.995 {"latitude":36.75963506199118,"longitude":-4.4970296379220 22,4592455248 0,0060662095 125,7586448669 Yamaha MT-07 2.024 Gasoline
15 e-Rvm-001 2022-12-02 11:52:00.995 {"latitude":36.75969033226069,"longitude":-4.4970442727250 22,2035952517 0,0062836841 125,5693382263 Yamaha MT-07 2.024 Gasoline
16 e-Rvm-001 2022-12-02 11:52:01.995 {"latitude":36.75974441229995,"longitude":-4.4970602151048 22,4299788068 0,0061802806 125,4846580505 Yamaha MT-07 2.024 Gasoline
17 e-Rvm-001 2022-12-02 11:52:02.995 {"latitude":36.75979558381881,"longitude":-4.4970743469935 21,6117508247 0,0058275118 125,4609519958 Yamaha MT-07 2.024 Gasoline
18 e-Rvm-001 2022-12-02 11:50:59.247 {"latitude":36.7547980102587,"longitude":-4.49412913062096 62,7978121023 0,0216183746 140,6830863953 Yamaha MT-07 2.024 Gasoline
19 e-Rvm-001 2022-12-02 11:51:00.497 {"latitude":36.75495711973566,"longitude":-4.4942598883104 61,6352671848 0,0211829137 141,0304222107 Yamaha MT-07 2.024 Gasoline
20 e-Rvm-001 2022-12-02 11:51:01.997 {"latitude":36.75514972120221,"longitude":-4.4944016262930 60,2597795904 0,0248618127 141,2720237732 Yamaha MT-07 2.024 Gasoline
21 e-Rvm-001 2022-12-02 11:51:02.997 {"latitude":36.75527535754885,"longitude":-4.4944827463519 58,430515878 0,0157289196 141,2803611755 Yamaha MT-07 2.024 Gasoline
Valor
71b88527-28e7-4c9f-b7c


DBeaver 23.3.5 - <dev 2> Script
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
SQL Commit Rollback Auto dev 2 public@dev
Proyectos
General
Connections
dev 2 - jdbc:redshift://redshift-cluste
Bases de Datos
dev
Esquemas
catalog_history
dev_smartcity
Tablas
failures_data
vehicle_data
weather_data
Vistas
Funciones
Tipos de datos
Aggregate functions
public
Tablas
Vistas
Funciones
Tipos de datos
Aggregate functions
Info del Sistema
Administrar
Info del Sistema
Bookmarks
Diagrams
Scripts
*<dev 2> Script
dev_smartcity
create external schema dev_smartcity
from data catalog
database smartcitydb
iam_role 'arn:aws:iam::851725340236:role/smart-city-redshit-s3-role'
region 'eu-west-3';
select * from dev_smartcity.failures_data
dev_dev_smartcity_failures_data_612614d9aa06f 1
Enter a SQL expression to filter results (use Ctrl+Space)
Grilla
Texto
deviceid incidentid type timestamp location description
1 [NULL] bba42109-795e-4006-81a5-806ebcee7e2e Test Drive Suspension 2022-12-02 11:42:39.331 {"latitude":36.71334440357677,"longitude":-4.4669327046763 [NULL]
2 [NULL] b1370abb-33a4-45a9-b047-6b1f5b025cee Test Drive Suspension 2022-12-02 11:42:40.998 {"latitude":36.71331671535663,"longitude":-4.4669873267453 [NULL]
3 [NULL] 7abce15a-1596-45e6-b851-23375444090e Test Drive Suspension 2022-12-02 11:42:41.997 {"latitude":36.71331026129118,"longitude":-4.4670610316139 [NULL]
4 [NULL] 36eb9580-bd6d-4646-a48c-3bb9b4602856 Test Drive Suspension 2022-12-02 11:42:42.997 {"latitude":36.71330455321512,"longitude":-4.4671360664111 [NULL]
5 [NULL] eaf61adf-eae8-42f9-8e6f-a7c10124a6b6 Test Drive Suspension 2022-12-02 11:16:10.993 {"latitude":36.75938627034125,"longitude":-4.4970360249322 [NULL]
6 [NULL] 018b0538-a33c-4693-9527-4a23eb5c2761 Test Drive Suspension 2022-12-02 11:16:11.993 {"latitude":36.7592859305783 8,"longitude":-4.4970146846068 [NULL]
7 [NULL] 44e3e0df-0e4e-4de4-8c66-ad2d973ea0eb Test Drive Suspension 2022-12-02 11:16:12.993 {"latitude":36.75920151643154,"longitude":-4.4969787262422 [NULL]
8 [NULL] b72a546b-4c27-4cf7-a6f2-63ae52e0d30e Test Drive Suspension 2022-12-02 11:16:13.993 {"latitude":36.75910529218313,"longitude":-4.4969448633533 [NULL]
9 [NULL] 0fc36221-0694-45c0-a497-510623a9dd66 Test Drive Suspension 2022-12-02 11:16:14.993 {"latitude":36.75900778550353,"longitude":-4.4969106987160 [NULL]

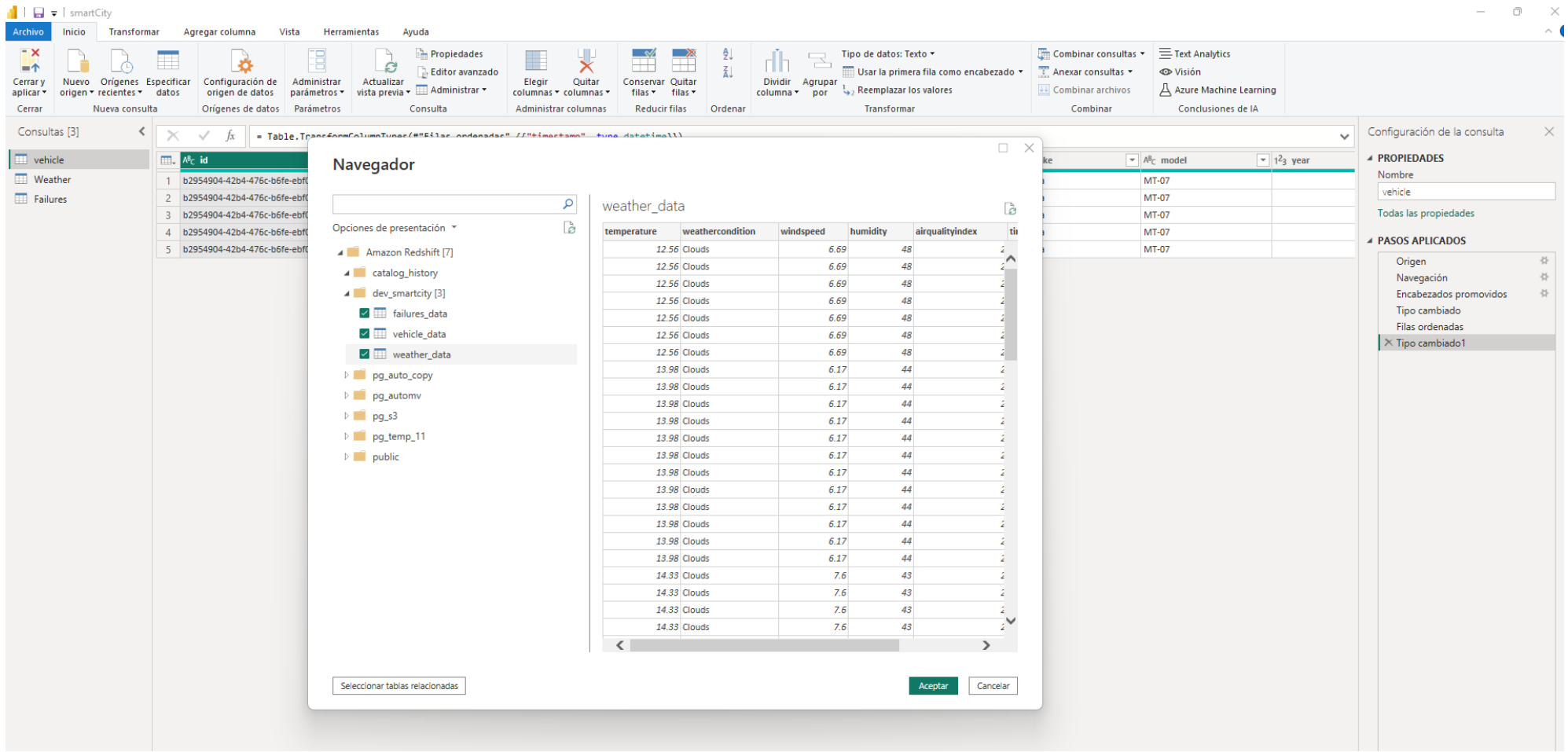

DBeaver 23.3.5 - <dev 2> Script
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
SQL Commit Rollback Auto dev 2 public@dev
Proyectos
General
Connections
dev 2 - jdbc:redshift://redshift-cluste
Bases de Datos
dev
Esquemas
catalog_history
dev_smartcity
Tablas
failures_data
vehicle_data
weather_data
Vistas
Funciones
Tipos de datos
Aggregate functions
public
Tablas
Vistas
Funciones
Tipos de datos
Aggregate functions
Info del Sistema
Administrar
Info del Sistema
Bookmarks
Diagrams
*<dev 2> Script
dev_smartcity
create external schema dev_smartcity
from data catalog
database smartcitydb
iam_role 'arn:aws:iam::851725340236:role/smart-city-redshit-s3-role'
region 'eu-west-3';
select * from dev_smartcity.failures_data
WHERE description IS NOT NULL AND description <> ''
LIMIT 10;
dev_dev_smartcity_failures_data_6126152dc8eca 1
Enter a SQL expression to filter results (use Ctrl+Space)
Grilla
Texto
id deviceid incidentid type timestamp location description
1 c690c92c-b2c5-40cb-b547-89c308bc2565 [NULL] 34a968c6-4980-4d32-ac04-97b287a5a5b0 Test Drive Suspension 2022-12-02 11:41:03.498 {"latitude":36.71701071438073,"longitude":-4.4677794026252 4 suspensions: Amber light not respected
2 8ebc7ff0-202b-425b-a7f8-960b54eefb73 [NULL] 8b6ced2f-db4d-4d29-b2ee-a2bad8de1754 Test Drive Suspension 2022-12-02 11:30:16.994 {"latitude":36.7170412559404 2,"longitude":-4.467744743455 1 suspensions: Failure to stop


DBeaver 23.3.5 - <dev 2> Script
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
SQL Commit Rollback Auto dev 2 public@dev
Proyectos
General
Connections
dev 2 - jdbc:redshift://redshift-cluste
Bases de Datos
dev
Esquemas
catalog_history
dev_smartcity
Tablas
failures_data
vehicle_data
weather_data
Vistas
Funciones
Tipos de datos
Aggregate functions
public
Tablas
Vistas
Funciones
Tipos de datos
Aggregate functions
Info del Sistema
Administrar
Info del Sistema
Bookmarks
Diagrams
Scripts
Project - General
*<dev 2> Script
dev_smartcity
create external schema dev_smartcity
from data catalog
database smartcitydb
iam_role 'arn:aws:iam::851725340236:role/smart-city-redshit-s3-role'
region 'eu-west-3';
select * from dev_smartcity.weather_data
dev_dev_smartcity_weather_data_612615683ee37 1
Enter a SQL expression to filter results (use Ctrl+Space)
Grilla
Texto
temperature weathercondition windspeed humidity airqualityindex timestamp
1 12,56 Clouds 6,69 48 2 2022-12-02 11:22:30.995
2 12,56 Clouds 6,69 48 2 2022-12-02 11:22:31.995
3 12,56 Clouds 6,69 48 2 2022-12-02 11:22:32.995
4 12,56 Clouds 6,69 48 2 2022-12-02 11:22:33.995
5 12,56 Clouds 6,69 48 2 2022-12-02 11:22:34.995
6 12,56 Clouds 6,69 48 2 2022-12-02 11:22:35.995
7 12,56 Clouds 6,69 48 2 2022-12-02 11:22:36.995
8 13,74 Clouds 6,17 44 2 2022-12-02 11:27:27.998
9 13,74 Clouds 6,17 44 2 2022-12-02 11:27:28.998
10 13,74 Clouds 6,17 44 2 2022-12-02 11:27:29.998
11 13,74 Clouds 6,17 44 2 2022-12-02 11:27:30.998
12 13,74 Clouds 6,17 44 2 2022-12-02 11:27:31.998
13 13,98 Clouds 6,17 44 2 2022-12-02 11:40:18.990
14 13,98 Clouds 6,17 44 2 2022-12-02 11:40:19.990
15 13,98 Clouds 6,17 44 2 2022-12-02 11:40:20.990
16 13,98 Clouds 6,17 44 2 2022-12-02 11:40:21.990

POWER BI

- DirectQuery or import the data from redshift
- With a student License / Premium / PRO you are able to use the redshift-powerBI connector

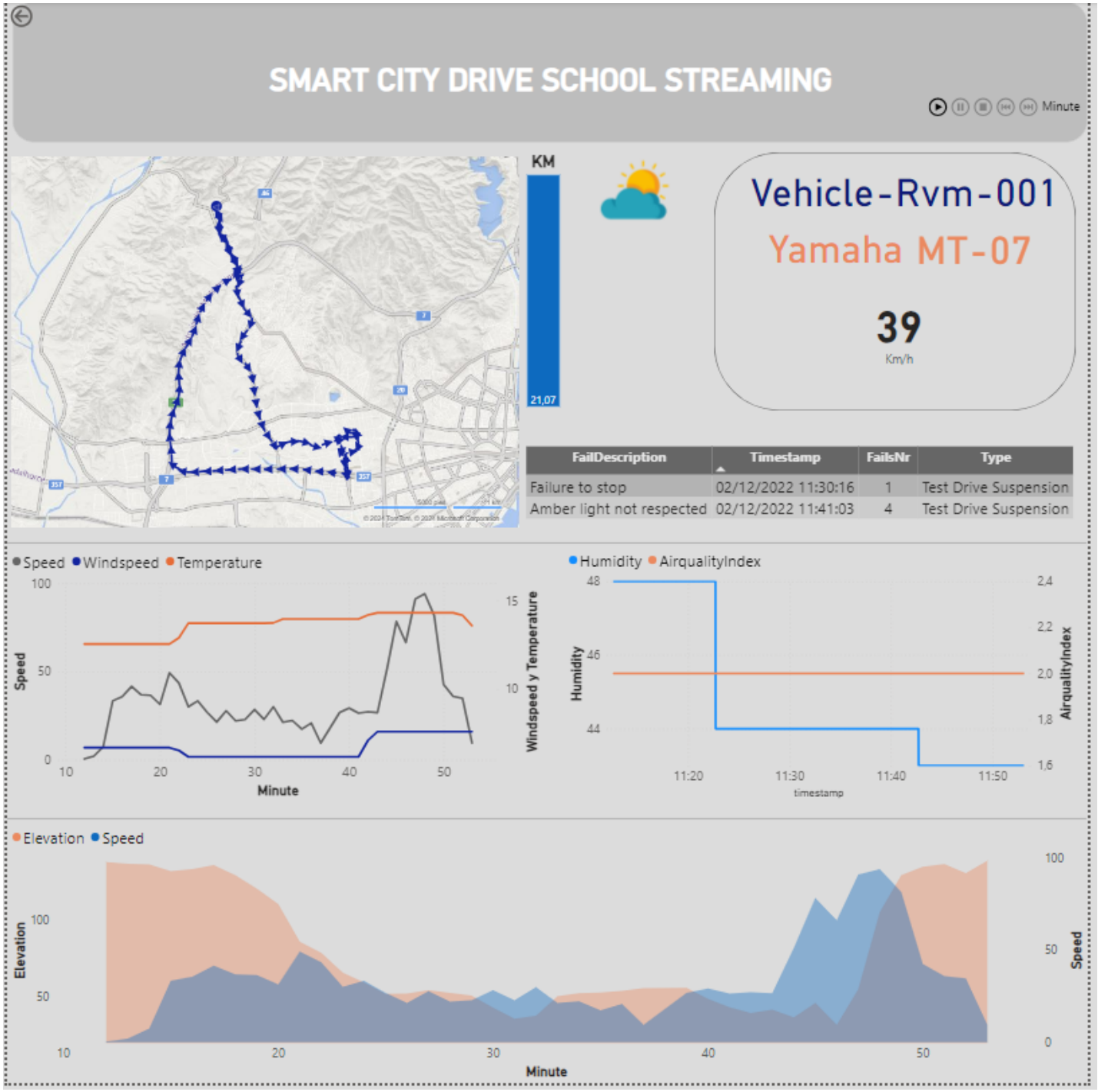
On the backgroung you can see other tables. This was the test I made before in order to prepare the dashboard a bit, to avoid overcosts for having the aws services deployed. In this way, It was just conecting, cleaning some columns and visualize.



It is not a streaming intake from redshift.

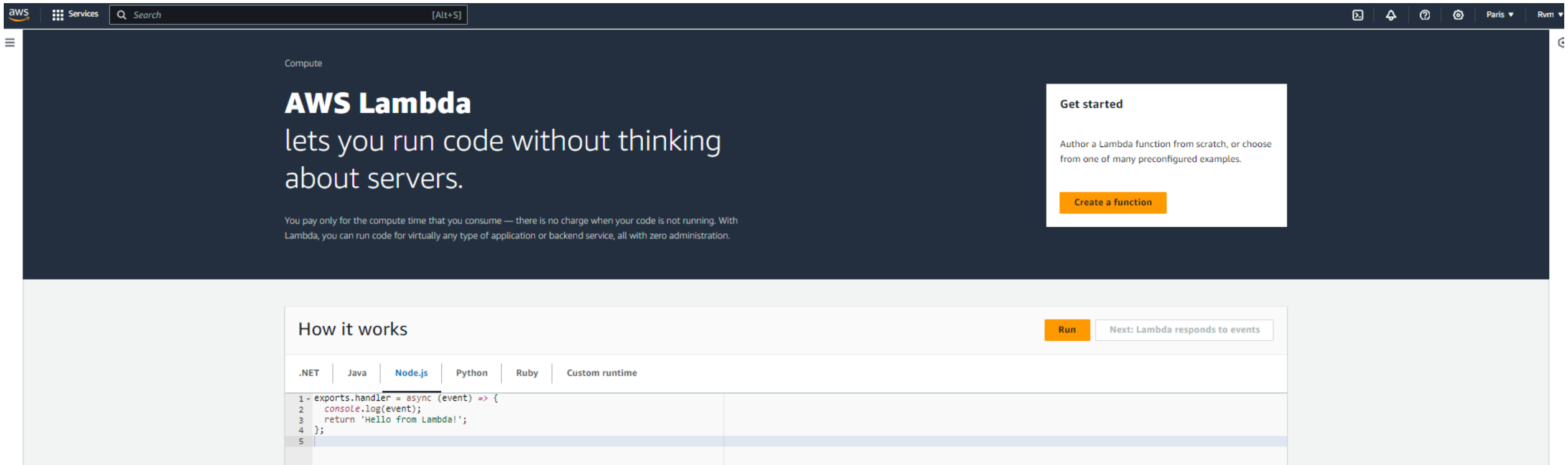
The best way I found to simulate a streaming was implementing the tool “play axis” in powerbi.

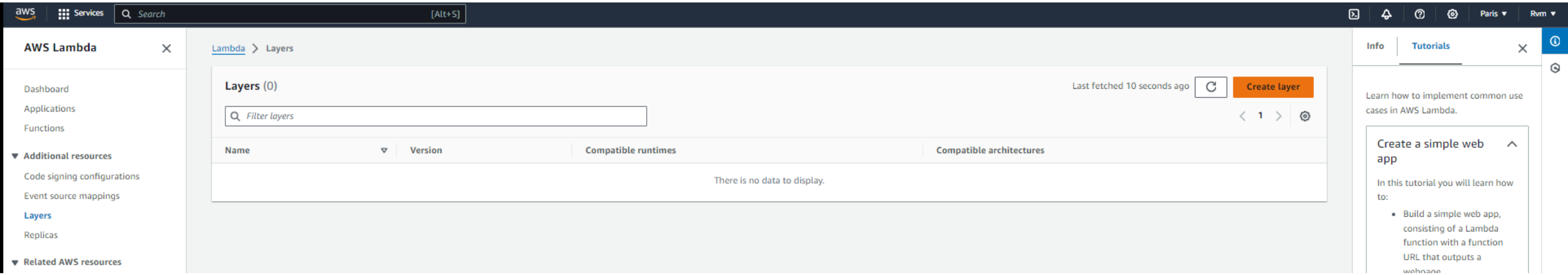
This allows to represent the data sequentially.



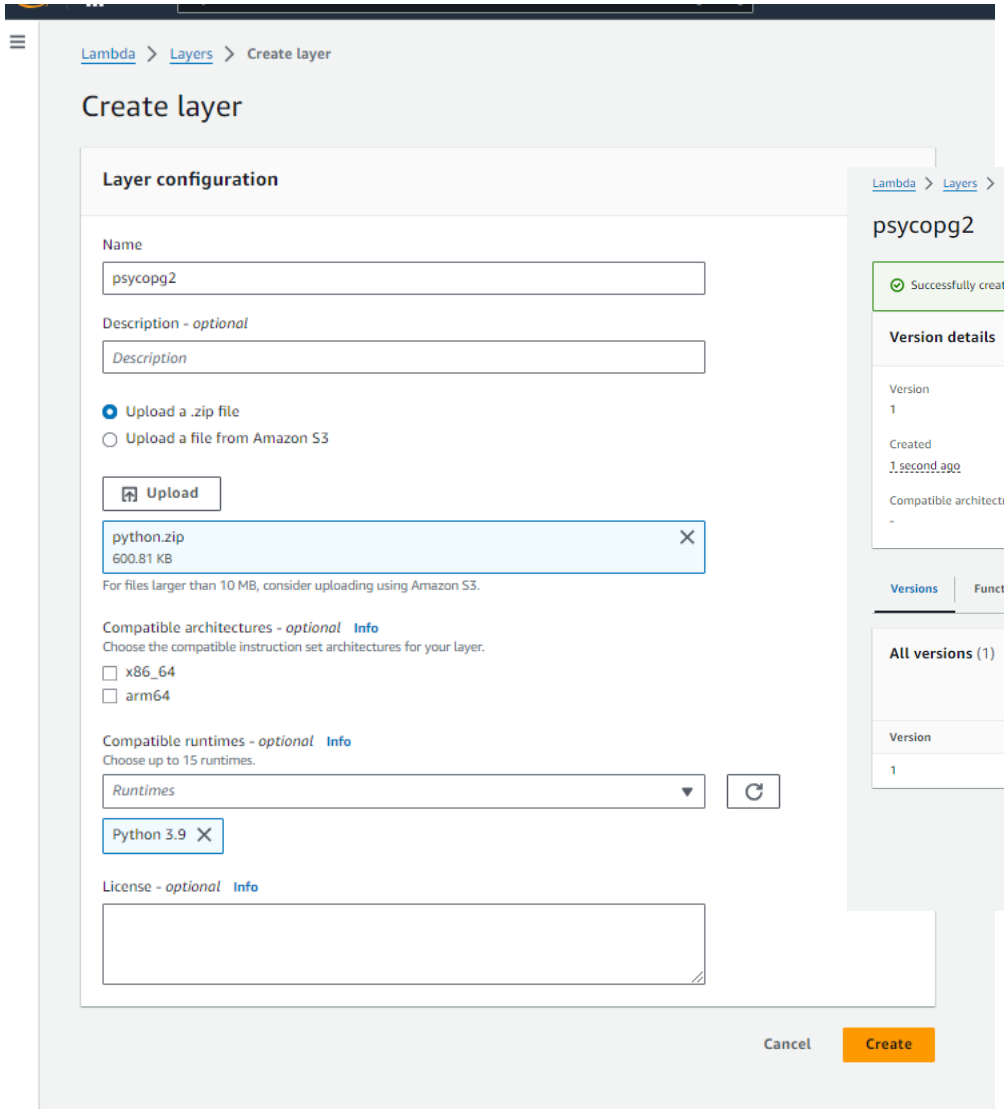
AWS LAMBDA

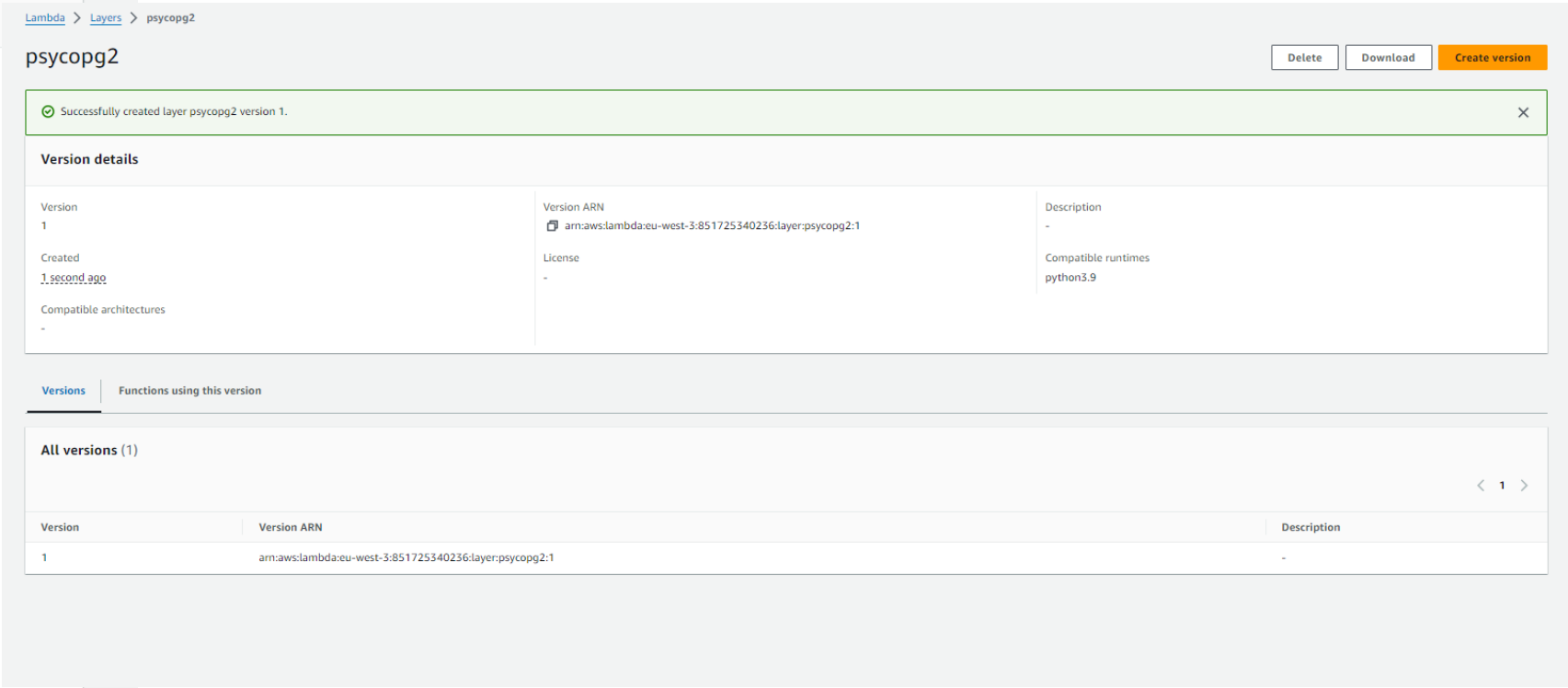
Create a Lambda function to simulate a streaming leverage. The function queries against redshift, getting the vehicle_data table, and sending the file in little batchs to Power BI API.

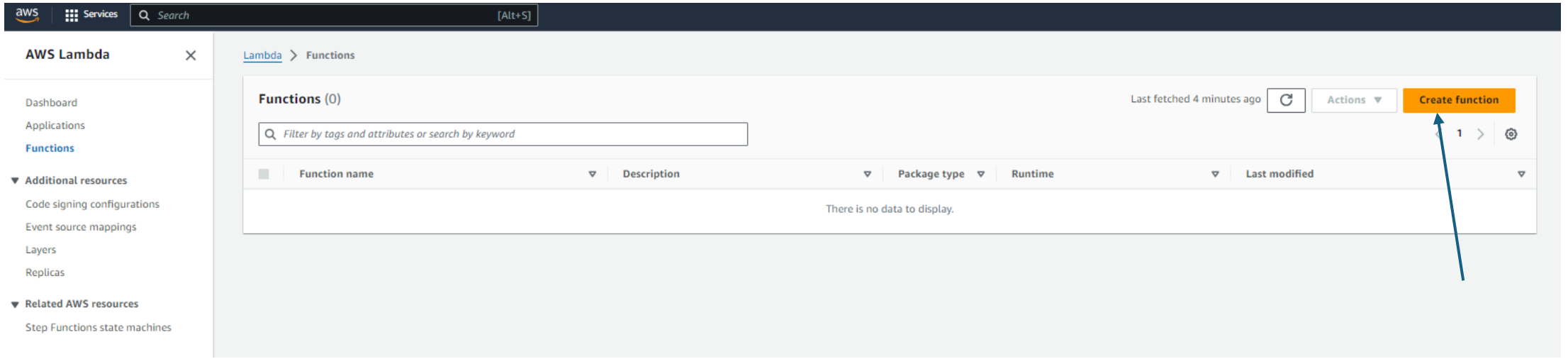

aws
Services
Search
[Alt+S]
Paris
Rvm
Compute
AWS Lambda
lets you run code without thinking about servers.
You pay only for the compute time that you consume — there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service, all with zero administration.
Get started
Author a Lambda function from scratch, or choose from one of many preconfigured examples.
Create a function
How it works
Run
Next: Lambda responds to events
.NET
Java
Node.js
Python
Ruby
Custom runtime
exports.handler = async (event) => {
console.log(event);
return 'Hello from Lambda!';
};

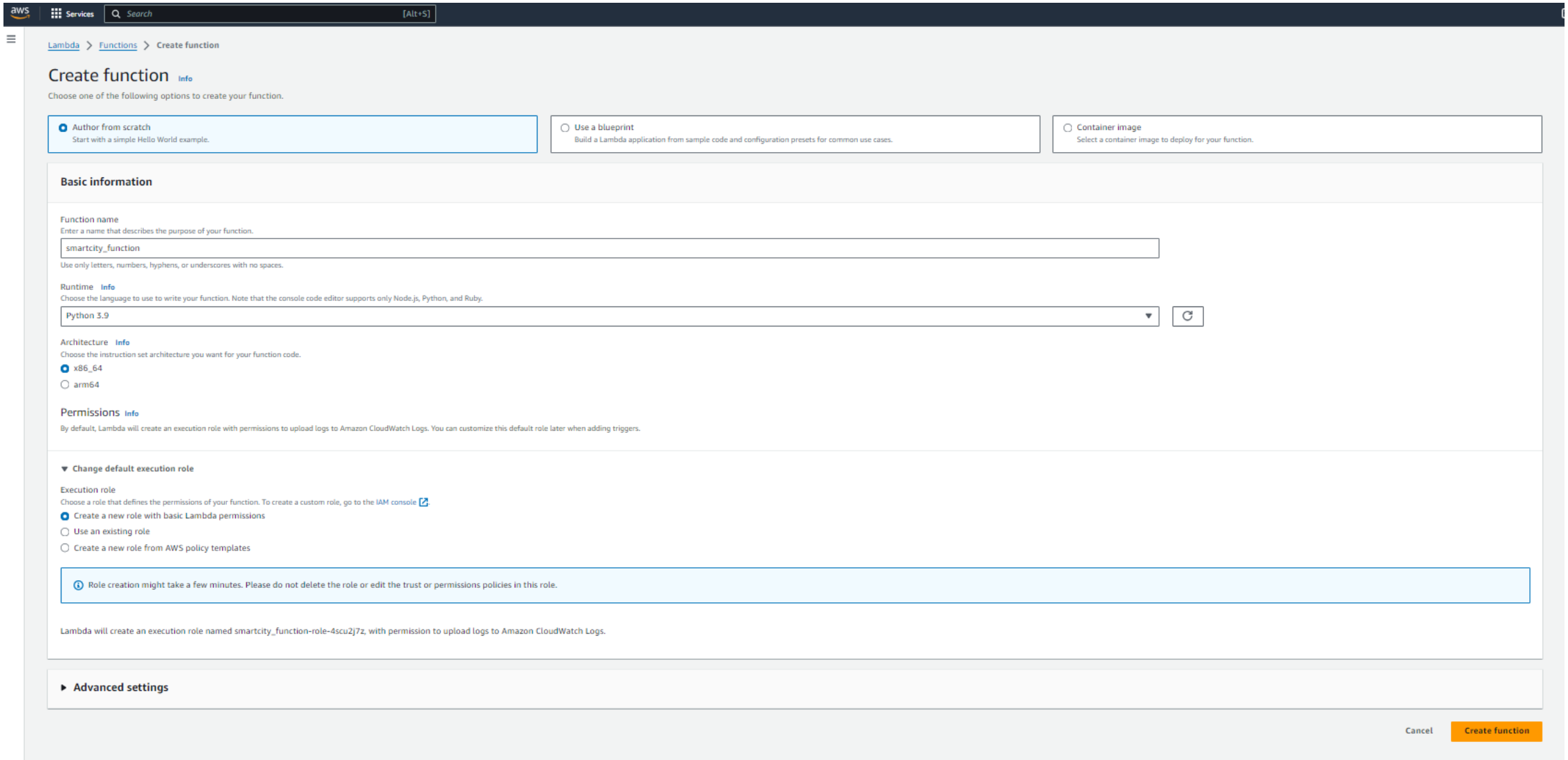

aws
Services
Search
[Alt+S]
Paris
Rvm
AWS Lambda
Lambda
Layers
Dashboard
Applications
Functions
Additional resources
Code signing configurations
Event source mappings
Layers
Replicas
Related AWS resources
Layers (0)
Last fetched 10 seconds ago
Create layer
Filter layers
1
Name
Version
Compatible runtimes
Compatible architectures
There is no data to display.
Info
Tutorials
Learn how to implement common use cases in AWS Lambda.
Create a simple web app
In this tutorial you will learn how to:
Build a simple web app, consisting of a Lambda function with a function URL that outputs a

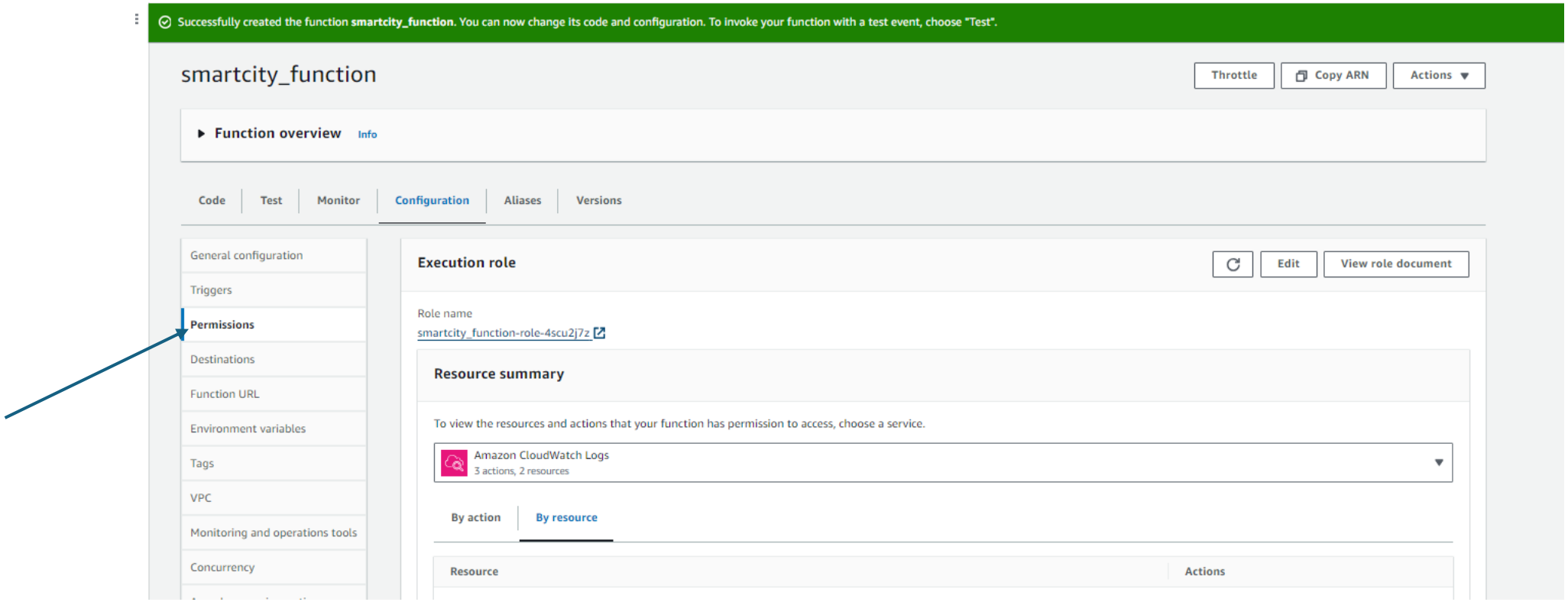
La Lambda’s layer allows to upload the libraries needed to run the function properly. This aws environment does not have many Python libraries installed by default.
Check the Commands_and_Comments.txt file in my Repo to know hoy to create this python.zip.

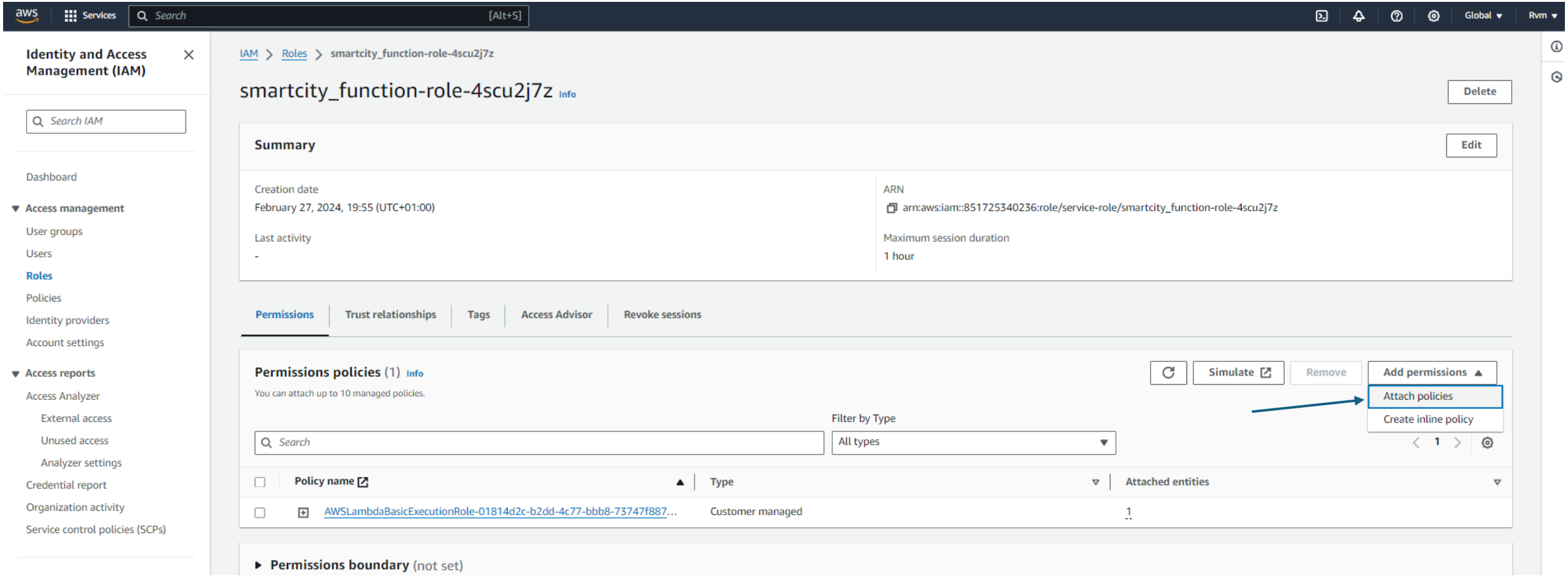





aws
Services
Search
[Alt+S]
AWS Lambda
Dashboard
Applications
Functions
Additional resources
Code signing configurations
Event source mappings
Layers
Replicas
Related AWS resources
Step Functions state machines
Lambda
Functions
Functions (0)
Last fetched 4 minutes ago
Actions
Create function
Filter by tags and attributes or search by keyword
1
Function name
Description
Package type
Runtime
Last modified
There is no data to display.


aws
Services
Search
[Alt+S]
Lambda > Functions > Create function
Create function Info
Choose one of the following options to create your function.
Author from scratch
Start with a simple Hello World example.
Use a blueprint
Build a Lambda application from sample code and configuration presets for common use cases.
Container image
Select a container image to deploy for your function.
Basic information
Function name
Enter a name that describes the purpose of your function.
smartcity_function
Use only letters, numbers, hyphens, or underscores with no spaces.
Runtime Info
Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.
Python 3.9
Architecture Info
Choose the instruction set architecture you want for your function code.
x86_64
arm64
Permissions Info
By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this default role later when adding triggers.
Change default execution role
Execution role
Choose a role that defines the permissions of your function. To create a custom role, go to the IAM console.
Create a new role with basic Lambda permissions
Use an existing role
Create a new role from AWS policy templates
Role creation might take a few minutes. Please do not delete the role or edit the trust or permissions policies in this role.
Lambda will create an execution role named smartcity_function-role-4scu2j7z, with permission to upload logs to Amazon CloudWatch Logs.
Advanced settings
Cancel
Create function


Successfully created the function smartcity_function. You can now change its code and configuration. To invoke your function with a test event, choose "Test".
smartcity_function
Throttle
Copy ARN
Actions
Function overview
Info
Code
Test
Monitor
Configuration
Aliases
Versions
General configuration
Triggers
Permissions
Destinations
Function URL
Environment variables
Tags
VPC
Monitoring and operations tools
Concurrency
Execution role
Edit
View role document
Role name
smartcity_function-role-4scu2j7z
Resource summary
To view the resources and actions that your function has permission to access, choose a service.
Amazon CloudWatch Logs
3 actions, 2 resources
By action
By resource
Resource
Actions


aws
Services
Search
[Alt+S]
Global
Rvm
Identity and Access Management (IAM)
Search IAM
Dashboard
Access management
User groups
Users
Roles
Policies
Identity providers
Account settings
Access reports
Access Analyzer
External access
Unused access
Analyzer settings
Credential report
Organization activity
Service control policies (SCPs)
IAM > Roles > smartcity_function-role-4scu2j7z
smartcity_function-role-4scu2j7z Info
Delete
Summary
Edit
Creation date
February 27, 2024, 19:55 (UTC+01:00)
Last activity
-
ARN
arn:aws:iam::851725340236:role/service-role/smartcity_function-role-4scu2j7z
Maximum session duration
1 hour
Permissions
Trust relationships
Tags
Access Advisor
Revoke sessions
Permissions policies (1) Info
You can attach up to 10 managed policies.
Simulate
Remove
Add permissions
Attach policies
Create inline policy
Filter by Type
Search
All types
1
Policy name
Type
Attached entities
AWSLambdaBasicExecutionRole-01814d2c-b2dd-4c77-bbb8-73747f887...
Customer managed
1
Permissions boundary (not set)

IAM > Roles > smartcity_function-role-4scu2j7z > Add permissions

Attach policy to smartcity_function-role-4scu2j7z

▶ **Current permissions policies** (1)

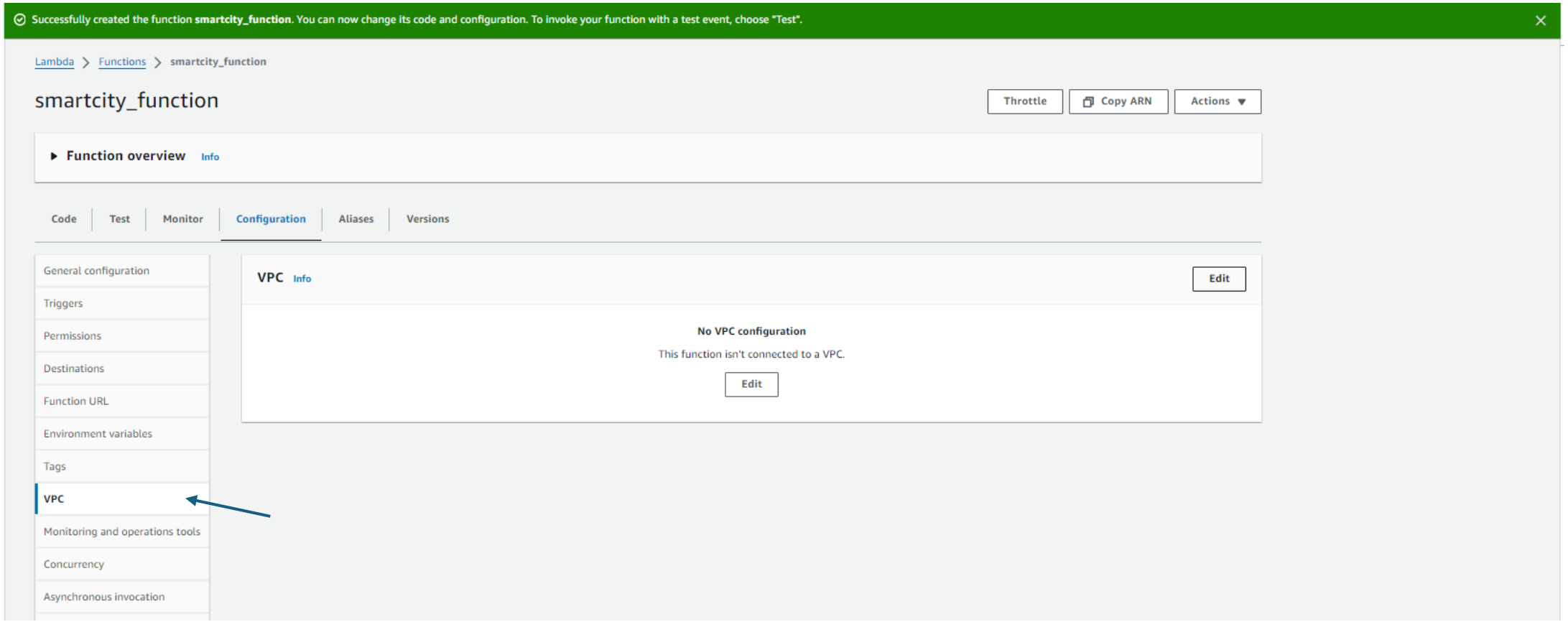
Other permissions policies (1/913)

Filter by Type

lambdav | All types | 1 match | 1

Policy name	Type	Description
AWSLambdaVPCAccessExecutionRole	AWS managed	Provides minimum permissions for a Lam…

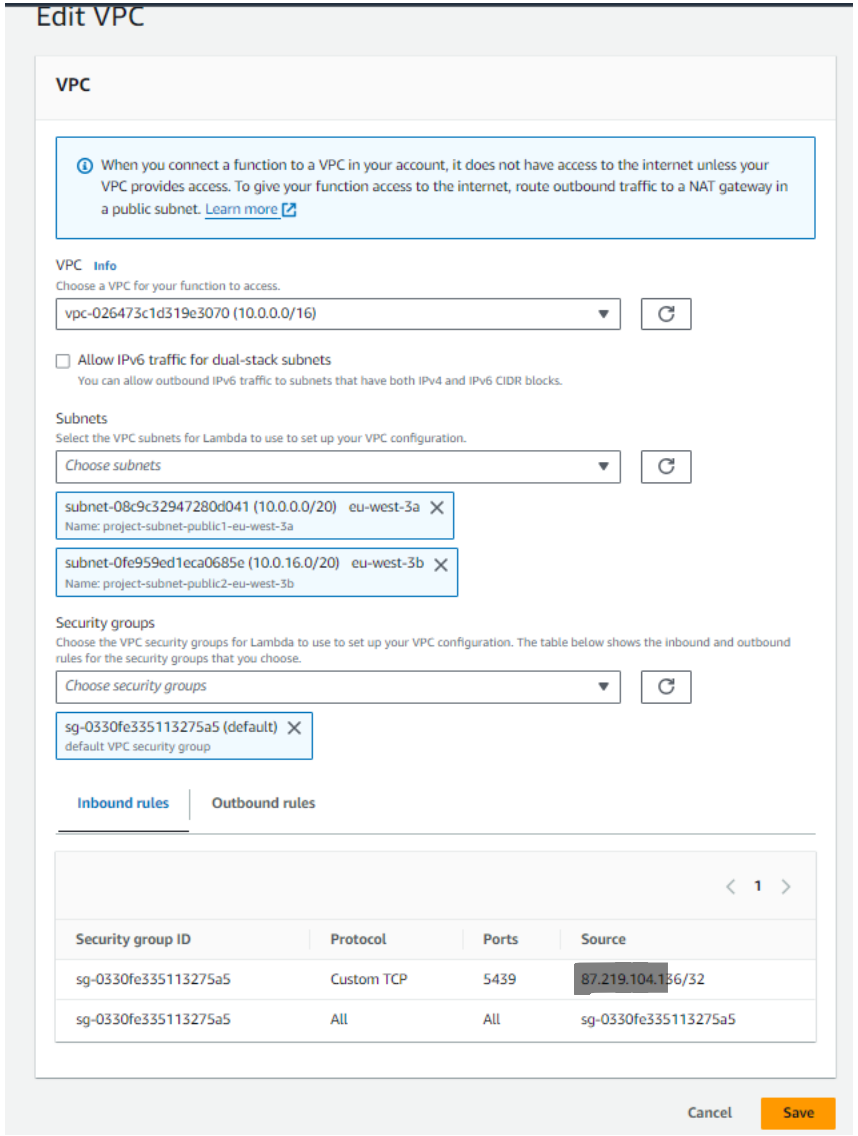
Cancel | Add permissions

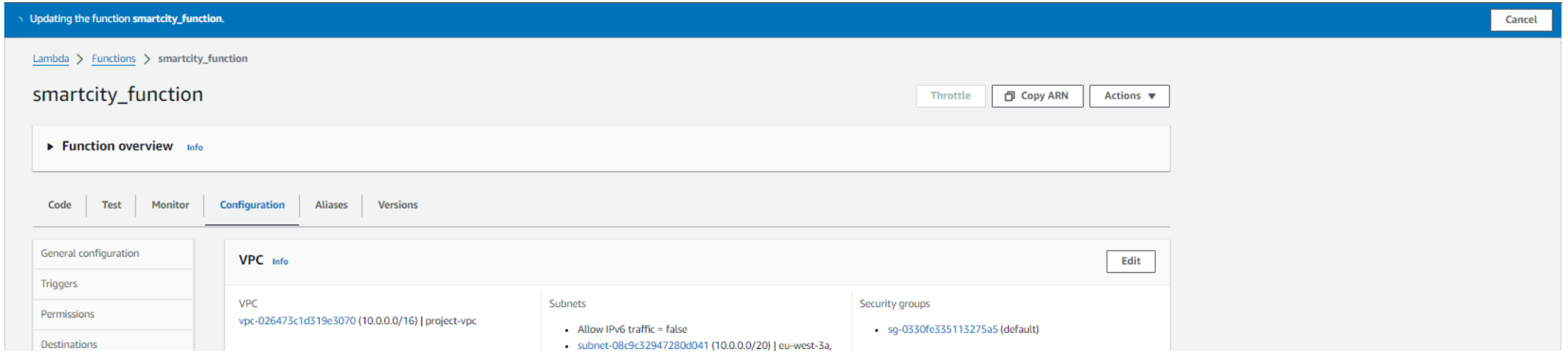

Successfully created the function smartcity_function. You can now change its code and configuration. To invoke your function with a test event, choose "Test".
Lambda > Functions > smartcity_function
smartcity_function
Throttle
Copy ARN
Actions
Function overview Info
Code
Test
Monitor
Configuration
Aliases
Versions
General configuration
Triggers
Permissions
Destinations
Function URL
Environment variables
Tags
VPC
Monitoring and operations tools
Concurrency
Asynchronous invocation
VPC Info
Edit
No VPC configuration
This function isn't connected to a VPC.
Edit

The same VPC as for Redshift.

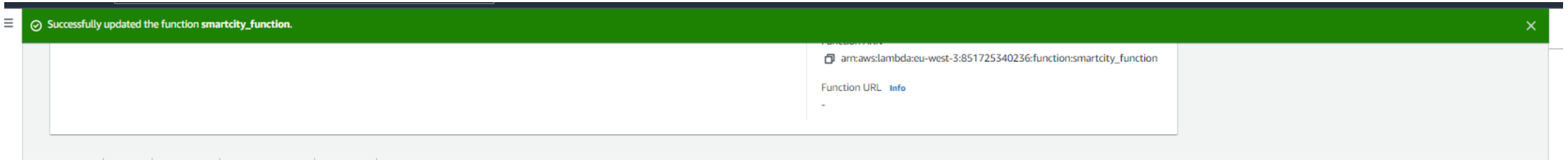
Important:
Go to
VPC console > Security Groups > security group (default)

There add an inbound rule

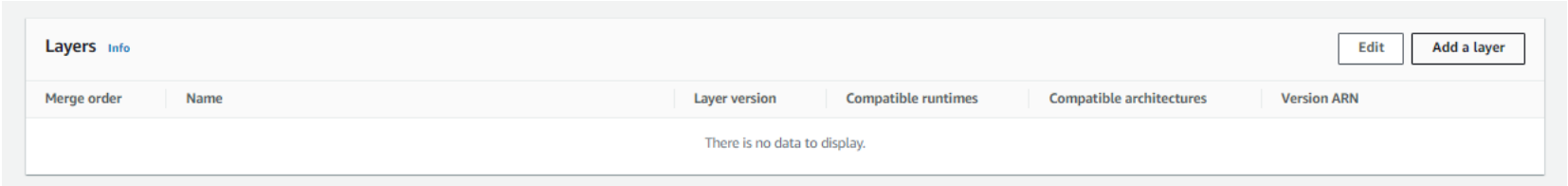


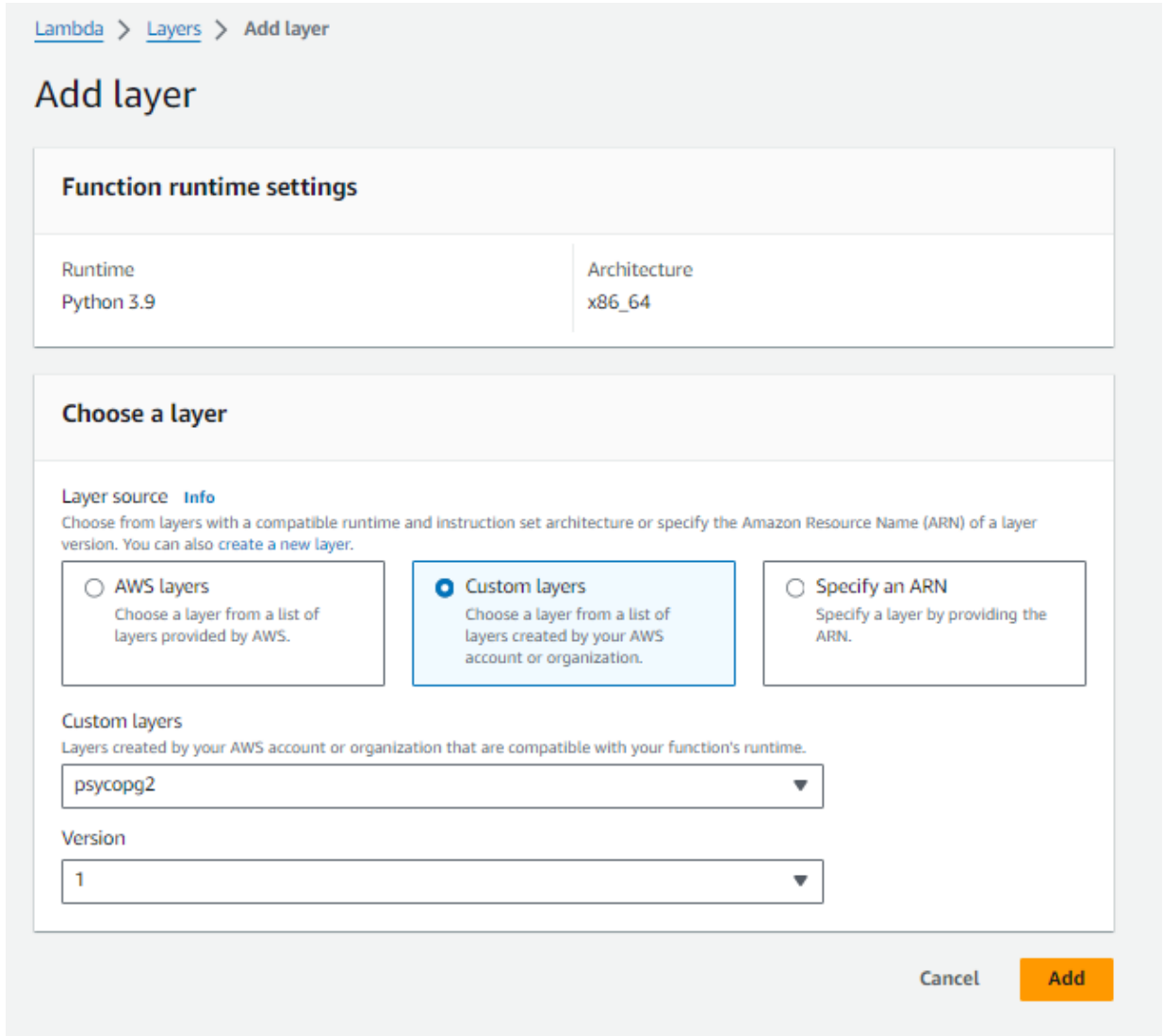


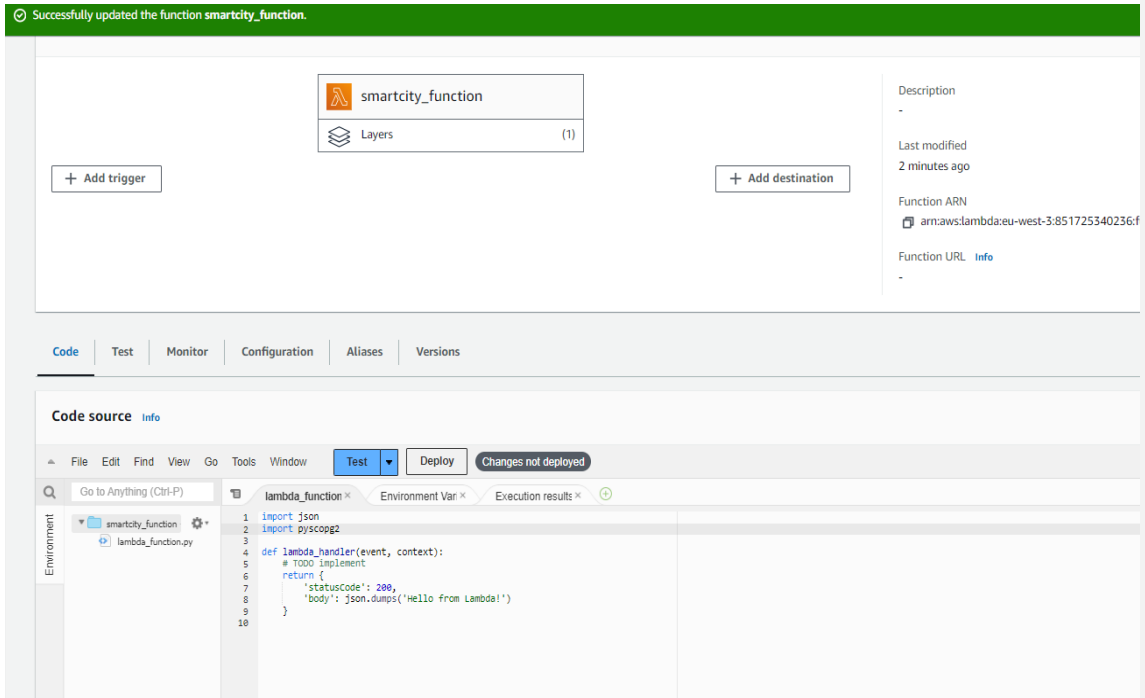
Wait until updating succes

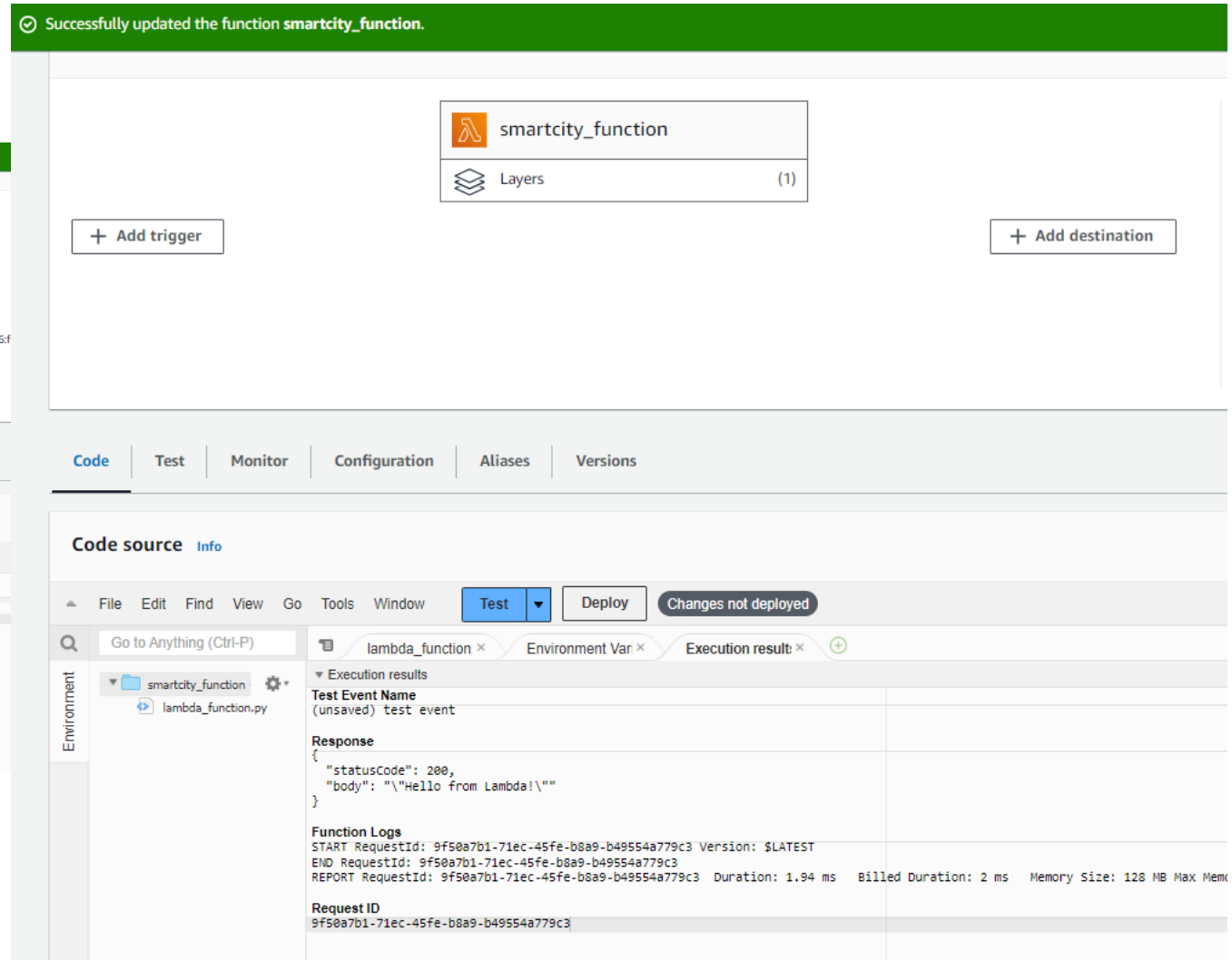


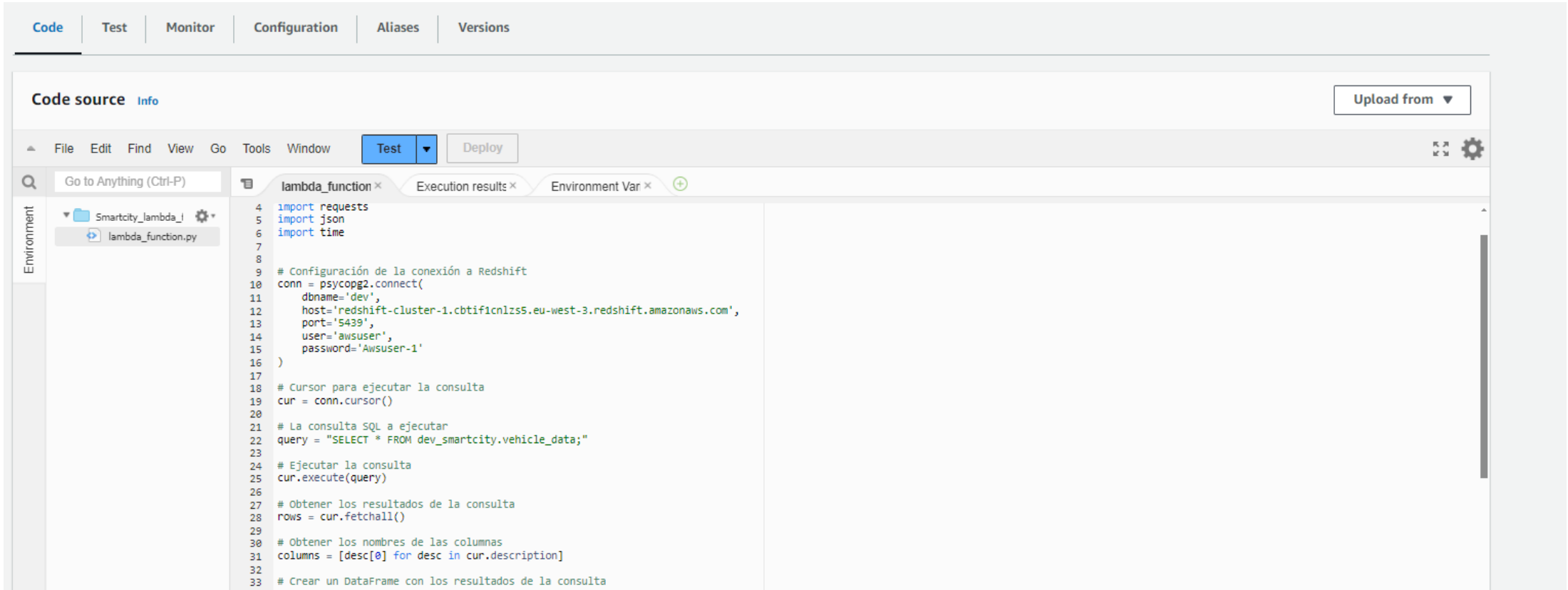
And proceed adding a layer (which contains the Python libraries




Lambda > Layers > Add layer
Add layer
Function runtime settings
Runtime
Python 3.9
Architecture
x86_64
Choose a layer
Layer source Info
Choose from layers with a compatible runtime and instruction set architecture or specify the Amazon Resource Name (ARN) of a layer version. You can also create a new layer.
AWS layers
Choose a layer from a list of layers provided by AWS.
Custom layers
Choose a layer from a list of layers created by your AWS account or organization.
Specify an ARN
Specify a layer by providing the ARN.
Custom layers
Layers created by your AWS account or organization that are compatible with your function's runtime.
psycopg2
Version
1
Cancel
Add


Successfully updated the function smartcity_function.
smartcity_function
Layers (1)
Add trigger
Add destination
Description
-
Last modified
2 minutes ago
Function ARN
arn:aws:lambda:eu-west-3:851725340236:f
Function URL Info
-
Code
Test
Monitor
Configuration
Aliases
Versions
Code source Info
File Edit Find View Go Tools Window
Test
Deploy
Changes not deployed
Go to Anything (Ctrl-P)
lambda_function
Environment Var
Execution results
Environment
smartcity_function
lambda_function.py
import json
import pyscopg2
def lambda_handler(event, context):
# TODO implement
return {
'statusCode': 200,
'body': json.dumps('Hello from Lambda!')
}


Successfully updated the function smartcity_function.
smartcity_function
Layers (1)
Add trigger
Add destination
Code
Test
Monitor
Configuration
Aliases
Versions
Code source Info
File Edit Find View Go Tools Window
Test
Deploy
Changes not deployed
Go to Anything (Ctrl-P)
lambda_function
Environment Var
Execution result
Environment
smartcity_function
lambda_function.py
Execution results
Test Event Name
(unsaved) test event
Response
{
"statusCode": 200,
"body": "\"Hello from Lambda!\""
}
Function Logs
START RequestId: 9f50a7b1-71ec-45fe-b8a9-b49554a779c3 Version: $LATEST
END RequestId: 9f50a7b1-71ec-45fe-b8a9-b49554a779c3
REPORT RequestId: 9f50a7b1-71ec-45fe-b8a9-b49554a779c3 Duration: 1.94 ms Billed Duration: 2 ms Memory Size: 128 MB Max Mem
Request ID
9f50a7b1-71ec-45fe-b8a9-b49554a779c3

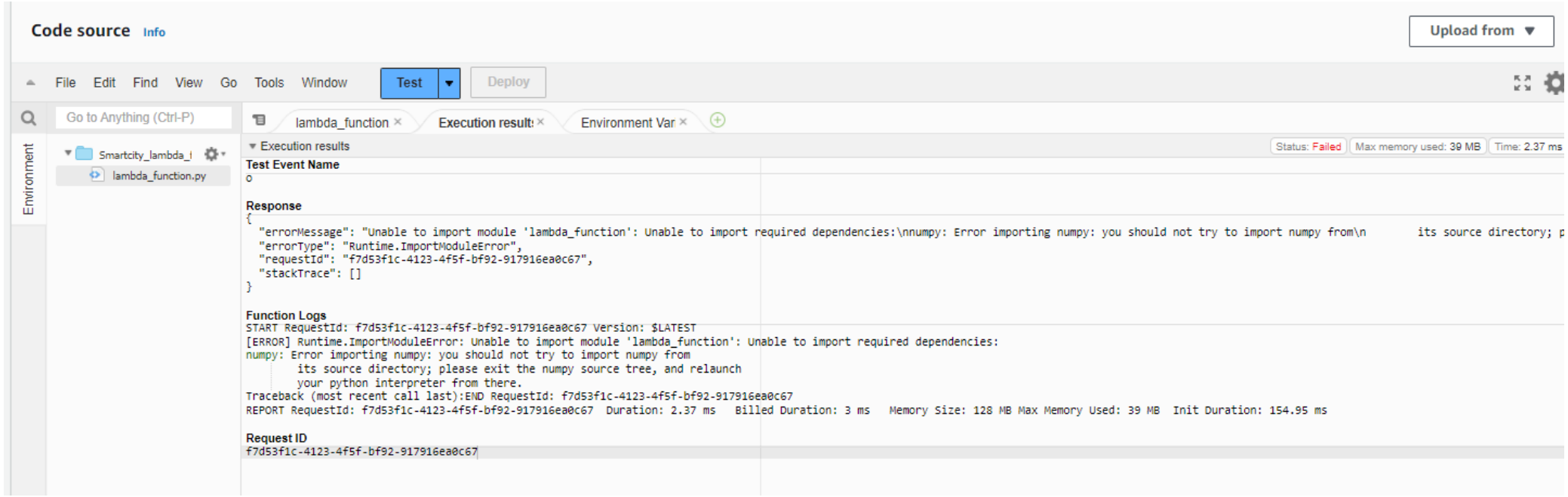

Code
Test
Monitor
Configuration
Aliases
Versions
Code source Info
Upload from
File Edit Find View Go Tools Window
Test
Deploy
Go to Anything (Ctrl-P)
Environment
Smartcity_lambda_f
lambda_function.py
lambda_function
Execution results
Environment Var
import requests
import json
import time
# Configuración de la conexión a Redshift
conn = psycopg2.connect(
dbname='dev',
host='redshift-cluster-1.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com',
port='5439',
user='awsuser',
password='Awsuser-1'
)
# Cursor para ejecutar la consulta
cur = conn.cursor()
# La consulta SQL a ejecutar
query = "SELECT * FROM dev_smartcity.vehicle_data;"
# Ejecutar la consulta
cur.execute(query)
# Obtener los resultados de la consulta
rows = cur.fetchall()
# Obtener los nombres de las columnas
columns = [desc[0] for desc in cur.description]
# Crear un DataFrame con los resultados de la consulta

Although I was able to query redshift from Lambda, unfortunately I pretended to save the table vehicle_data into a dataframe and transform it, in order to pass the data as Json to PowerBI API correctly.

That’s why the location columna in my code, retrieves a json containing latitude and longitud as key value in json format for this columna only.

the libraries needed to transform a dataframe with panda aren’t to much, the problem I found is the bunch of libraries that work behind and ables the main library runs correctly.

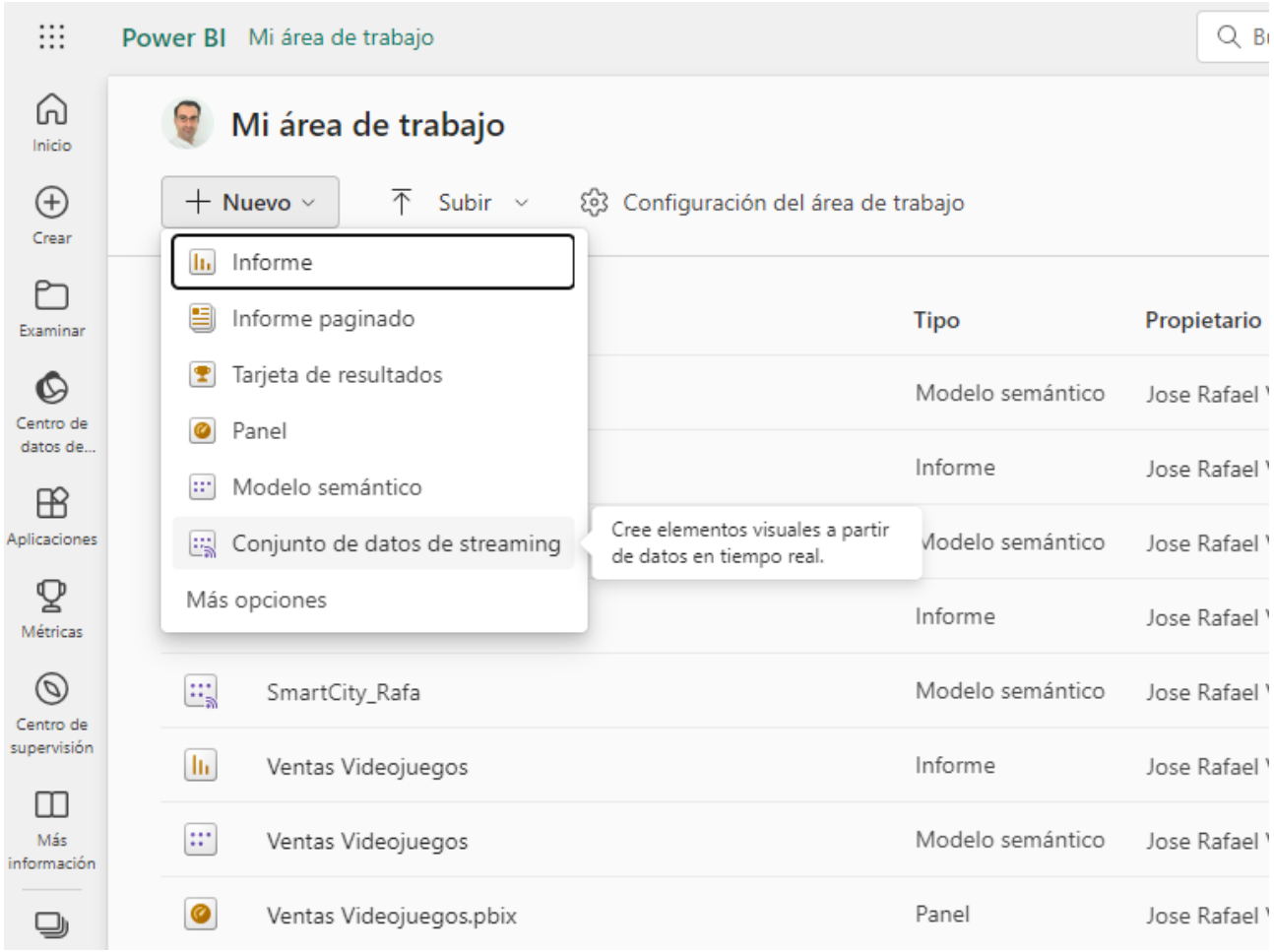
I felt in an endless Flow of errors adding to the Layer all the libraries inside the Python.zip, until I reached the error below and I decided to stop and trying such streaming simulation with lambda in the next Project.

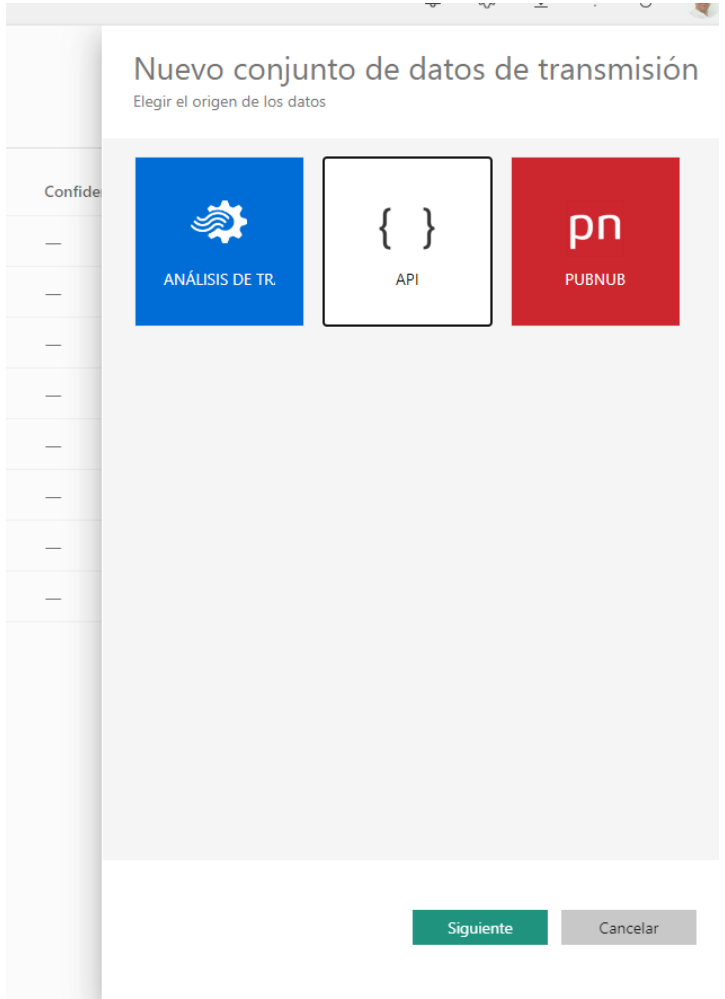


PowerBI API

https://app.powerbi.com/groups/me/list?experience=power-bi

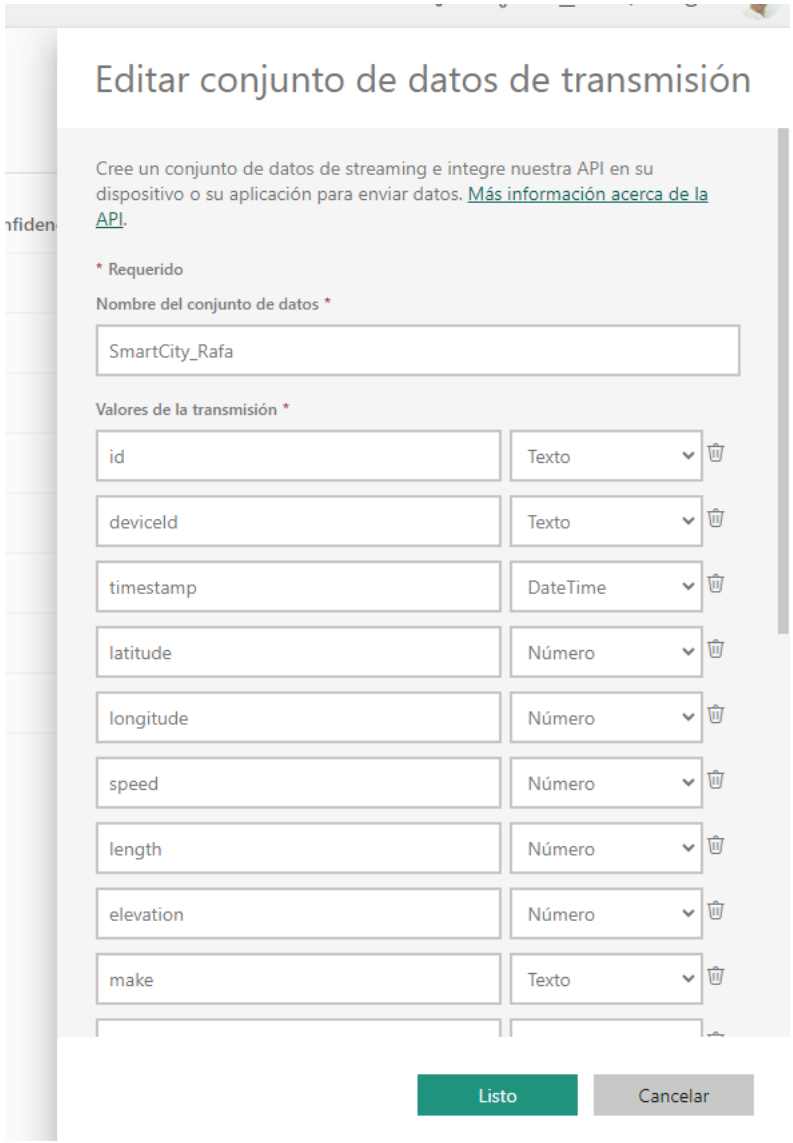
- Create a new streaming data set
- Select API
- Give it a name and entry the fields (next slide)

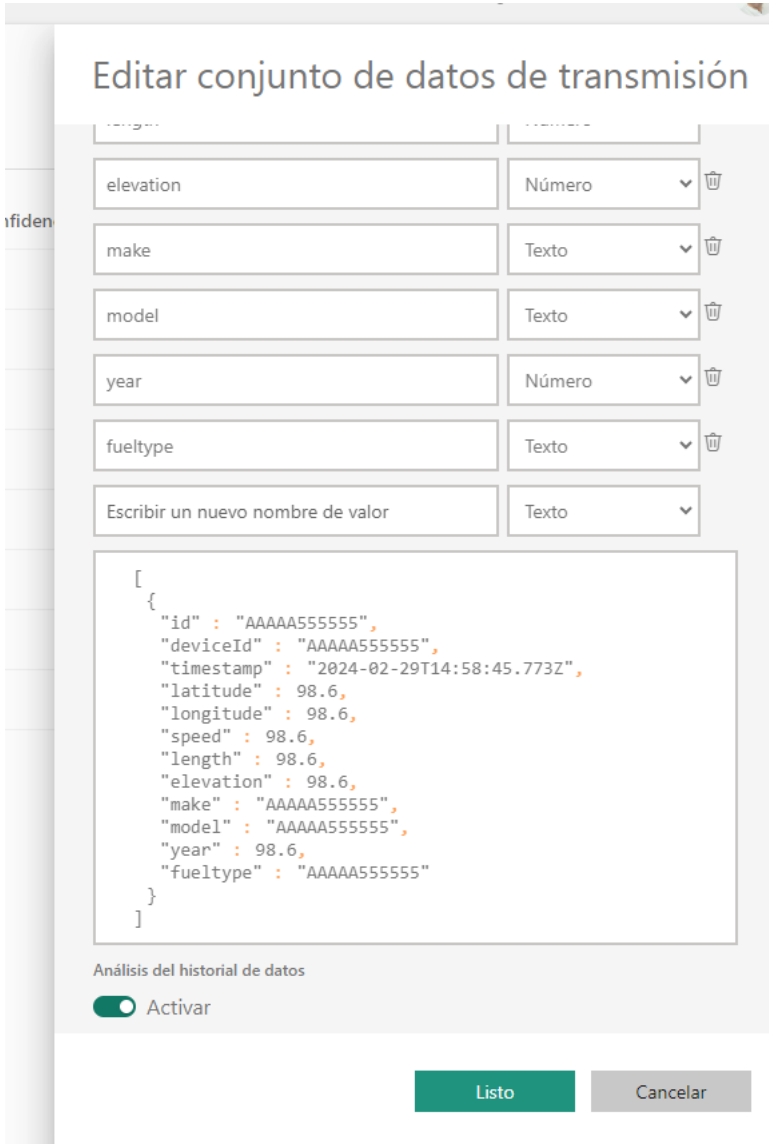




PowerBI API

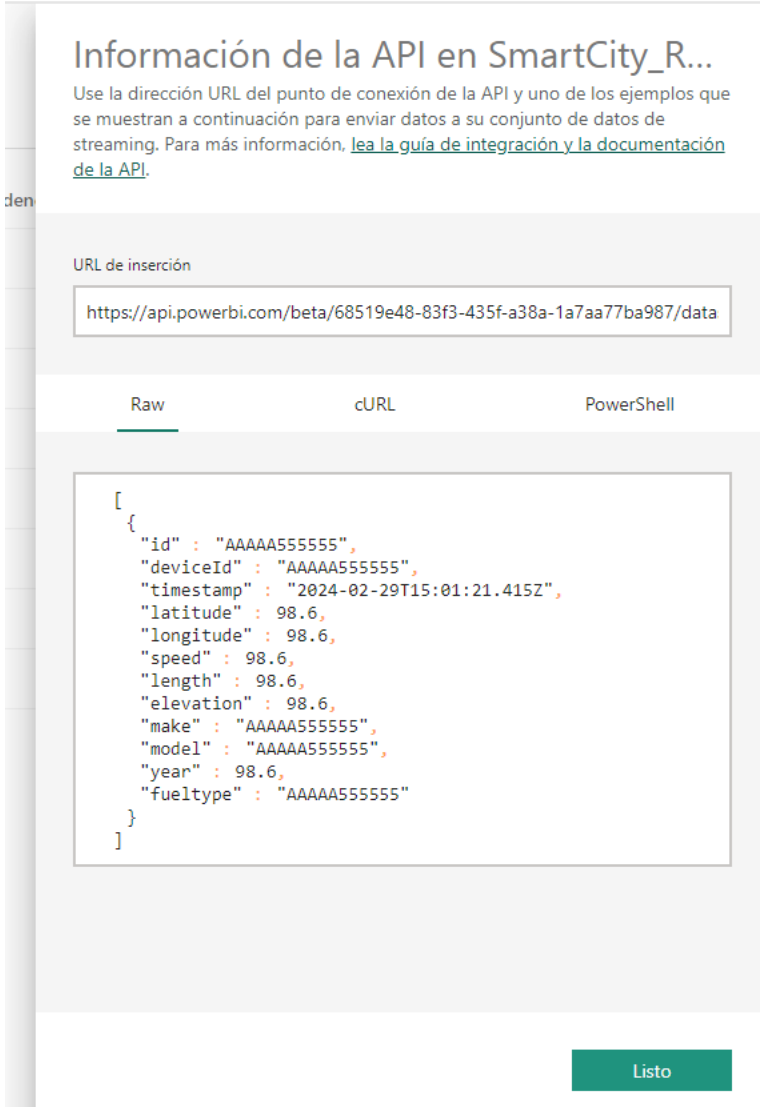
Once you add all the fields, PBI shows an example JSON schema which you must use in order to send the data to the endpoint.





PowerBI API

Finally you get the Endpoint URL and json schema to be used



PowerBI API
Create a dashboard
In my case I’m not receiving data because of the issue with the Lambda Layer and the library uploading