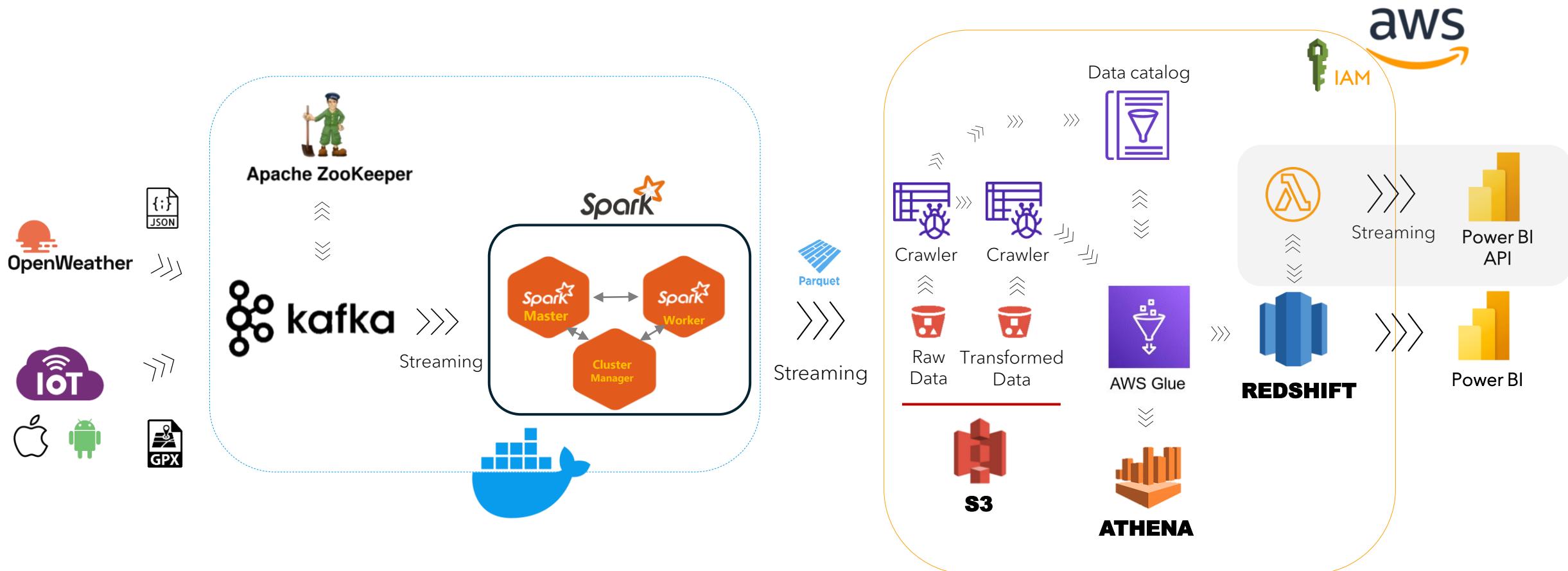


AWS Data Processing End-to-end pipeline | Street Drive lessons realtime monitoring



Project Setup

Docker
Kafka and Spark code and configuration

Creating zookeeper and Broker docker containers

The screenshot shows the PyCharm IDE interface. On the left is the Project tool window displaying a file structure for a project named 'SmartCityRvm' located at 'C:\Users\jrver'. Inside 'SmartCityRvm' is a 'venv' folder containing 'Lib', 'Scripts', '.gitignore', and 'pyvenv.cfg'. The 'docker-compose.yml' file is selected in the Project tool window.

The main editor window displays the contents of the 'docker-compose.yml' file:

```
50     test: [ 'CMD', 'bash', '-c', "nc -z localhost 9092" ]
51
52     interval: 10s
53
54     timeout: 5s
55
56     retries: 5
57
58     networks:
59         - datamasterylabRvm
60
61     spark-master:
62         image: bitnami/spark:latest
63         volumes:
64             - ./jobs:/opt/bitnami/spark/jobs
65         command: bin/spark-class org.apache.spark.deploy.master.Master
66         ports:
67             - "9090:8080"
68             - "7077:7077"
69         networks:
70             - datamasterylabRvm
71
72     networks:
```

The terminal window at the bottom shows the output of running the Docker Compose command, indicating that three services have been created: a Network named 'smartcityrvm_datamasteryLabRvm', a Container named 'zookeeper' (Healthy), and a Container named 'broker' (Started).

```
[+] Running 3/3
✓ Network smartcityrvm_datamasteryLabRvm Created
✓ Container zookeeper Healthy
✓ Container broker Started
(venv) PS C:\Users\jrver\PycharmProjects\Python\SmartCityRvm>
```

Zookeeper, Broker, Spark master and workers docker containers creation

Containers Give feedback								
Container CPU usage ⓘ			Container memory usage ⓘ		Show charts ▾			
Search		☰	<input checked="" type="checkbox"/> Only show running containers					
□	Name	Image		Status	CPU (%)	Port(s)	Last started	Actions
□	smartcityrvm			Running (5/5)	99.52%		10 seconds ago	☰ ⋮ 🗑
□	zookeeper b539fd307c40	confluentinc/cp-zookeeper:7.4.0		Running	0.1%	2181:2181	1 hour ago	☰ ⋮ 🗑
□	broker 90c7f900a36f	confluentinc/cp-server:7.4.0		Running	99.19%	9092:9092 Show all ports (2)	10 seconds ago	☰ ⋮ 🗑
□	spark-master-1 0686e1cf1698	bitnami/spark:latest		Running	0.08%	7077:7077 Show all ports (2)	2 minutes ago	☰ ⋮ 🗑
□	spark-worker-2-1 7a27088f5a39	bitnami/spark:latest		Running	0.07%		2 minutes ago	☰ ⋮ 🗑
□	spark-worker-1-1 abb34dc0ef14	bitnami/spark:latest		Running	0.08%		2 minutes ago	☰ ⋮ 🗑

Checking Spark cluster creation



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 pollution data:

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



Weather in your city

[Guide](#) [API](#) [Dashboard](#) [Marketplace](#) [Pricing](#) [Maps](#) [Our Initiatives](#) [Partners](#) [Blog](#) [For Business](#)[Sign In](#) [Support](#)

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)



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

The depth of historical data have been extended to 100 YEARS.

My services
My API keys
My payments
My profile
Logout

Api Key to call the services from your code

The screenshot shows the OpenWeather API keys management interface. At the top, there's a navigation bar with links like Guide, API, Dashboard, Marketplace, Pricing, Maps, Our Initiatives, Partners, Blog, For Business, Rvm▼, and Sup. Below the navigation bar, there's a search bar with the placeholder "Weather in your city". Underneath the search bar, there's a horizontal menu with links: New Products, Services, API keys (which is underlined, indicating it's the active page), Billing plans, Payments, Block logs, My orders, My profile, and Ask a question.

In the main content area, there's a message: "You can generate as many API keys as needed for your subscription. We accumulate the total load from all of them." Below this message, there's a table with the following columns: Key, Name, Status, Actions, and Create key.

Key	Name	Status	Actions	Create key
c7b4181bea086bc21d2836910f197609	Default	Active	<input checked="" type="checkbox"/> <input type="button" value="Edit"/>	<input type="text" value="SmartCityProject"/> <input type="button" value="Generate"/>

Current weather data

Home / API / Current weather

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}
```

Product concept

Call current weather data

How to make an API call

API response

JSON format API response example

JSON format API response fields

XML format API response example

XML format API response fields

List of weather condition codes

Min/max temperature in current weather

API and forecast API

Bulk downloading

Other features

Geocoding API

Built-in geocoding

Built-in API request by city name

Built-in API request by city ID

Built-in API request by ZIP code

Format

Units of measurement

Multilingual support

Call back function for JavaScript code

Parameters

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

← → ⌂ openweathermap.org/api/air-pollution

Articulos APIs aws BBDD Empleo Forecasting GITHUB Learning My Web NOISE PROJECT OPENAI Python Snowflake Spark Projects TFM Tsystems Utilities

 Weather in your city Guide API Dashboard Marketplace Pricing Maps Our Initiatives Partners Blog For Business

Air Pollution API

[Home](#) / [API](#) / 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 (NO₂), Ozone (O₃), Sulphur dioxide (SO₂), Ammonia (NH₃), and particulates (PM_{2.5} and PM₁₀).

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)

[Air Pollution API concept](#)

[Current air pollution data](#)

[Forecast air pollution data](#)

[Historical air pollution data](#)

[Air Pollution API response](#)

[Example of API response](#)

[Fields in API response](#)

[Air Pollution Index levels scale](#)

OpenWeather API Endpoints Discovery

It takes up to 2 hours been able to get a 200.Ok response.

POSTMAN COLLECTION within my Repository

The screenshot shows a POSTMAN collection interface with the following details:

- Method:** GET
- URL:** <https://api.openweathermap.org/data/3.0/onecall?lat=36.719444&lon=-4.420000&appid=KEY>
- Params:** lat: 36.719444, lon: -4.420000, appid: KEY
- Headers:** (5)
- Body:** (empty)
- Pre-request Script:** (empty)
- Tests:** (empty)
- Settings:** (empty)
- Cookies:** (empty)

The status bar at the bottom indicates:

- 401 Unauthorized
- 246 ms
- 680 B
- Save as example
- ...

The response body is displayed in a modal window:

```
1 {  
2   "cod": 401,  
3   "message": "Please note that using One Call 3.0 requires a separate subscription to the One Call by Call plan. Learn more here https://openweathermap.org/price. If you have a valid subscription to the One Call by Call plan, but still receive this error, then please see https://openweathermap.org/faq#error401 for more info."  
}
```

Frequently Asked Questions

[Home](#) / 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?

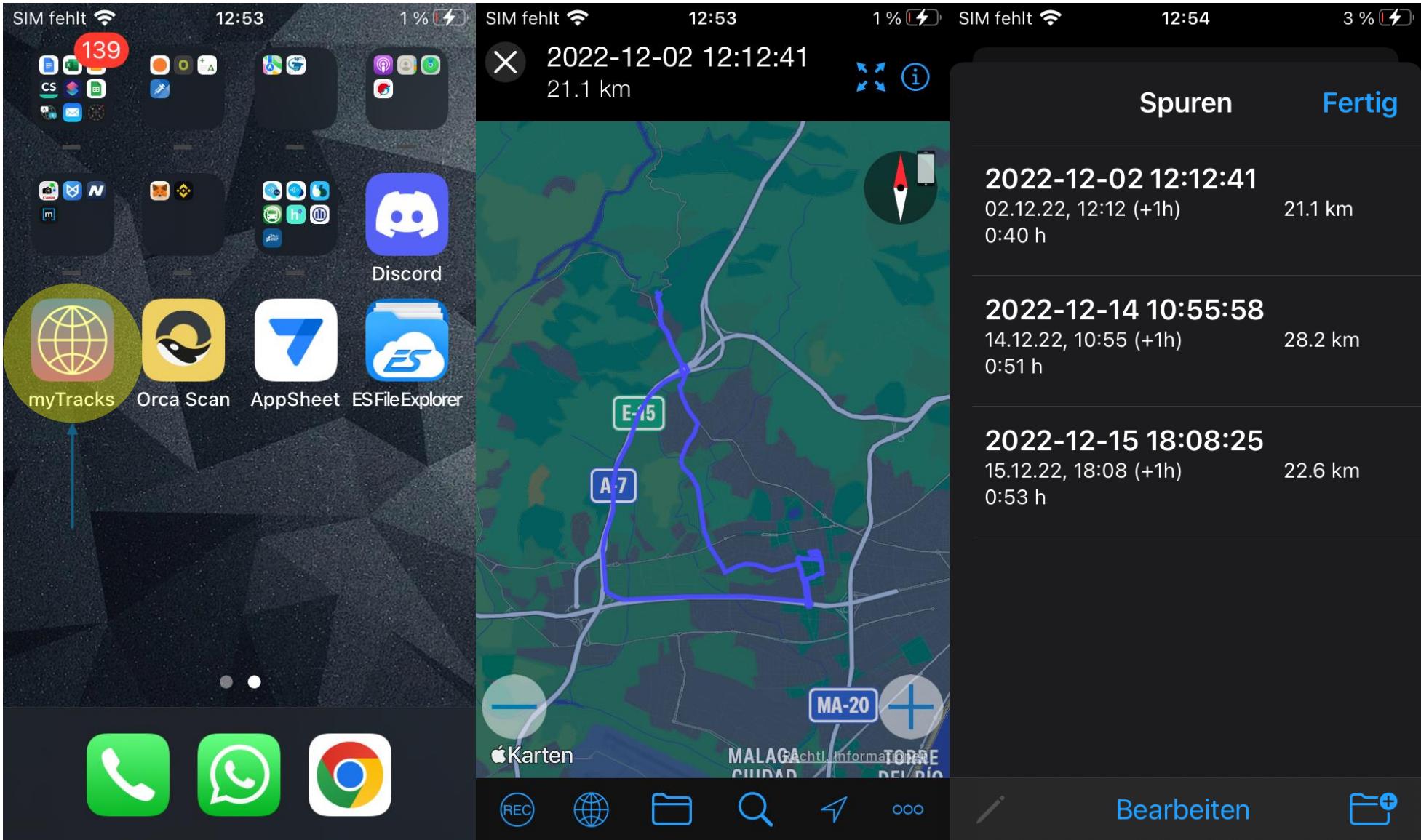


Your API key will be activated automatically, up to 2 hours after your successful registration. We invite you to read the [API documentation](#) that explains how to use our APIs.

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.



AWS

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

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.

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)

Copy ARN Empty Delete Create bucket

✓ Successfully edited Block Public Access settings for this bucket.

X

[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

S | Services | Search [Alt+S]

spark-streaming-data-rvm-bkt

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 that you 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

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](#)

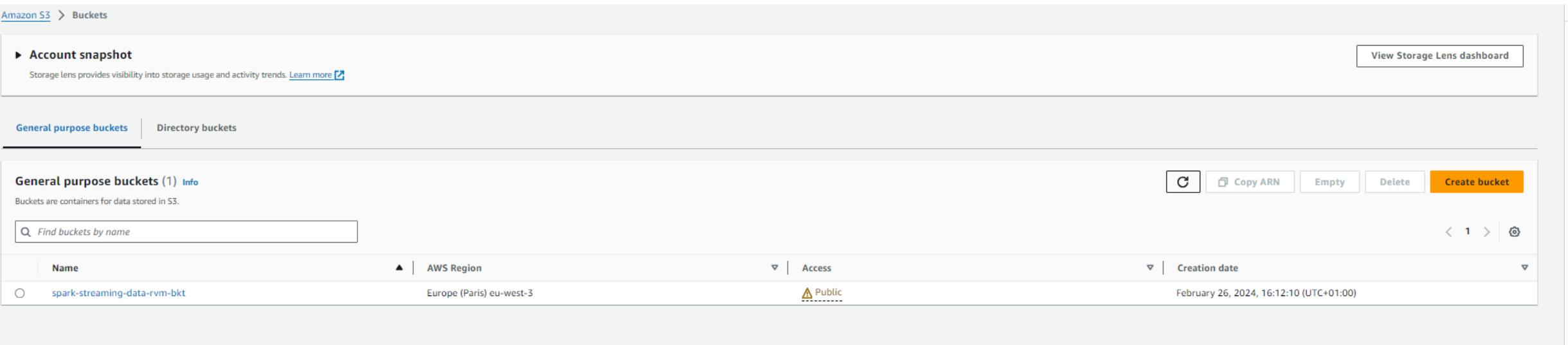
General purpose buckets (1) [Info](#)

Buckets are containers for data stored in S3.

Find buckets by name < 1 > ⌂

Name	AWS Region	Access	Creation date
spark-streaming-data-rvm-bkt	Europe (Paris) eu-west-3	Public	February 26, 2024, 16:12:10 (UTC+01:00)

[C](#) [Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)



Create an User

The screenshot shows the 'Create user' wizard in the AWS IAM console. The title bar says 'Create an User'. The left sidebar shows 'Step 1 Specify user details' is selected. The main area is titled 'Specify user details' and contains a 'User details' section. In the 'User name' field, 'rvm-user' is entered. Below it, a note says 'The user name can have up to 64 characters. Valid characters: A-Z, a-z, 0-9, and + = , . @ _ - (hyphen)'. There is an optional checkbox 'Provide user access to the AWS Management Console - optional' which is unchecked. A note below it says 'If you're providing console access to a person, it's a best practice [to manage their access in IAM Identity Center](#)'. At the bottom right are 'Cancel' and 'Next' buttons.

AWS Services Search [Alt+S] Global ▾

IAM > Users > Create user

Step 1
Specify user details

Step 2
Set permissions

Step 3
Review and create

User details

User name
rvm-user

The user name can have up to 64 characters. Valid characters: A-Z, a-z, 0-9, and + = , . @ _ - (hyphen)

Provide user access to the AWS Management Console - *optional*
If you're providing console access to a person, it's a [best practice](#) to manage their access in IAM Identity Center.

If you are creating programmatic access through access keys or service-specific credentials for AWS CodeCommit or Amazon Keyspaces, you can generate them after you create this IAM user. [Learn more](#)

Cancel Next

Add permissions to the user

The screenshot shows the AWS Identity and Access Management (IAM) service interface. The left sidebar is titled "Identity and Access Management (IAM)" and includes sections for Dashboard, Access management (with "Users" selected), Roles, Policies, Identity providers, Account settings, and Access reports. The main content area is titled "rvm-user" and shows the "Summary" tab. It displays the ARN (arn:aws:iam::851725340236:user/rvm-user), Console access status (Disabled), and Last console sign-in information. Below the summary, the "Permissions" tab is selected, showing "Permissions policies (4)". A table lists four policies: "AdministratorAccess" (AWS managed - job function, Directly), "AmazonS3FullAccess" (AWS managed, Directly), "AWSGlueConsoleFullAccess" (AWS managed, Directly), and "IAMUserChangePassword" (AWS managed, Directly). A "Permissions boundary (not set)" section is also present.

Identity and Access Management (IAM)

IAM > Users > rvm-user

rvm-user [Info](#)

Summary

ARN
arn:aws:iam::851725340236:user/rvm-user

Console access
Disabled

Access key 1
[Create access key](#)

Created
February 26, 2024, 16:49 (UTC+01:00)

Last console sign-in
-

Permissions Groups Tags Security credentials Access Advisor

Permissions policies (4)

Permissions are defined by policies attached to the user directly or through groups.

Filter by Type
All types

Policy name	Type	Attached via
AdministratorAccess	AWS managed - job function	Directly
AmazonS3FullAccess	AWS managed	Directly
AWSGlueConsoleFullAccess	AWS managed	Directly
IAMUserChangePassword	AWS managed	Directly

▶ Permissions boundary (not set)

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 within the code

The screenshot shows the AWS Identity and Access Management (IAM) console. The left sidebar is collapsed, and the main area has a dark header bar with the AWS logo, a search bar, and a date/time indicator (February 26, 2024, 16:49 (UTC+01:00)). The top navigation bar includes tabs for Services, Search, [Alt+S], and Global settings. The main content area is titled "Identity and Access Management (IAM)" and shows the "Security credentials" tab selected. The "Security credentials" section contains three main sections: "Console sign-in", "Multi-factor authentication (MFA)", and "Access keys".

- Console sign-in:** Includes a "Console sign-in link" (https://851725340236.signin.aws.amazon.com/console) and a "Console password" status (Not enabled). A "Enable console access" button is present.
- Multi-factor authentication (MFA):** Shows 0 MFA devices assigned. It includes a "Device type" column, an "Identifier" column, a "Certifications" column, and a "Created on" column. A "Assign MFA device" button is located at the bottom.
- Access keys:** Shows 0 access keys. It includes a "Create access key" button.

The sidebar on the left lists various IAM management options: Dashboard, 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), and Related consoles.

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

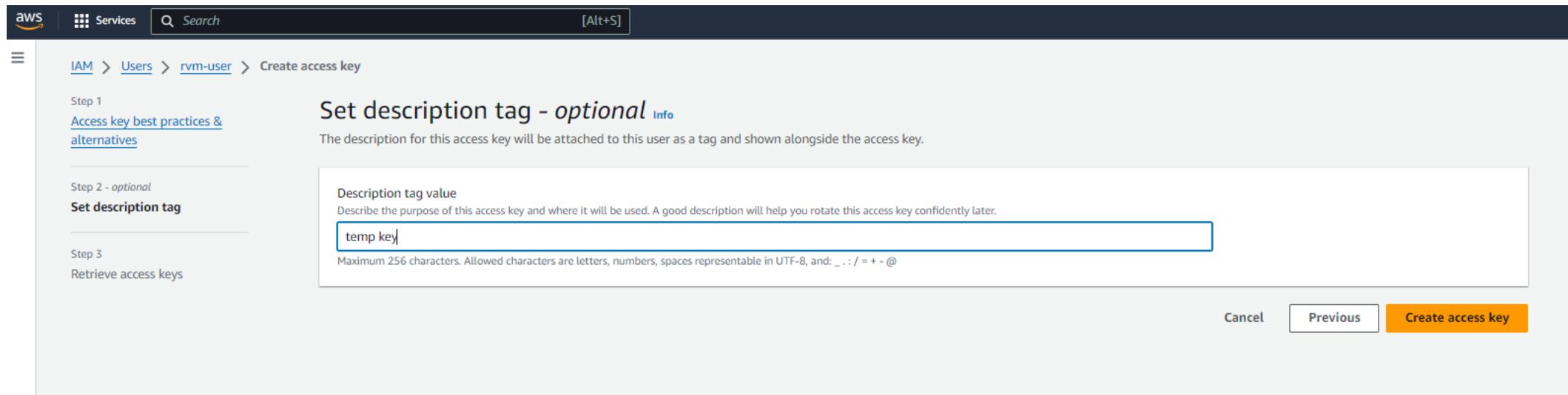
The description for this access key will be attached to this user as a tag and shown alongside the access key.

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
AKIA4MTWJJZGAIMXSUOO	***** 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

The screenshot shows the PyCharm IDE interface with the following details:

- Project:** SmartCityRvm
- File:** docker-compose.yml
- Content:**

```
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
```

The file `docker-compose.yml` is selected in the project tree.

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 ...
```

SC 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

```
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]
-----
|           |       modules      ||  artifacts  |
|   conf     | 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”.

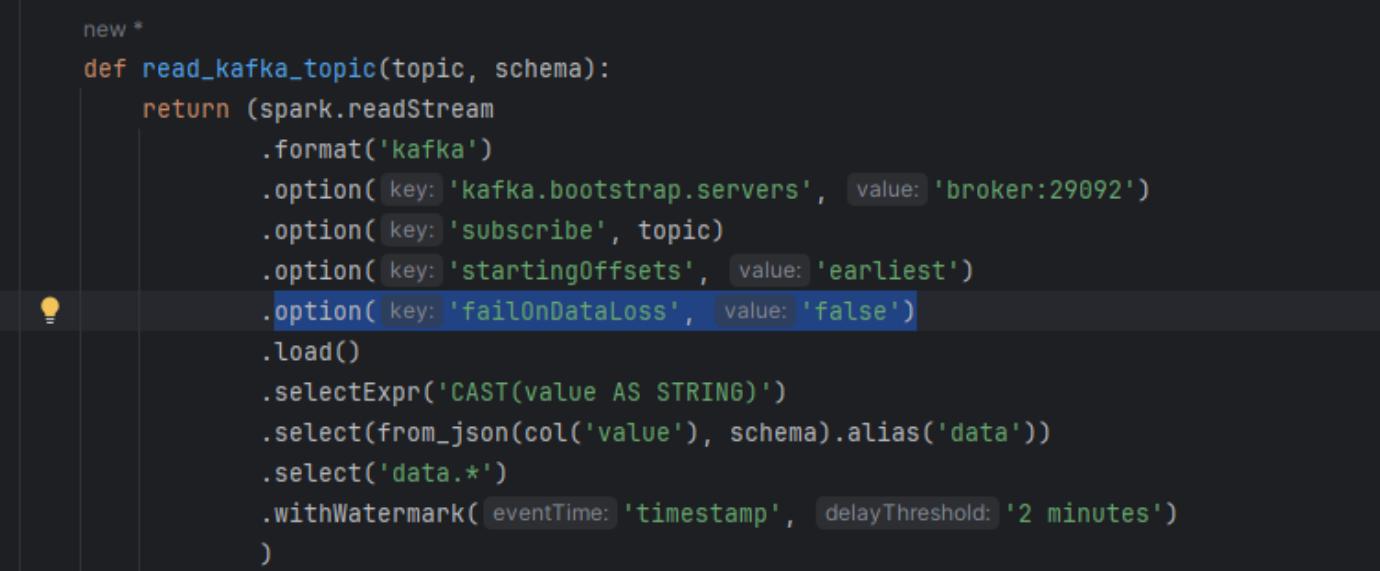
Due to the Broker consistency and connection with Spark, it may occurs that few registers are lost, and this option doesnot allow the process to continue if there are some data missing by default.

“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". “



```
new *
def read_kafka_topic(topic, schema):
    return (spark.readStream
        .format('kafka')
        .option(key: 'kafka.bootstrap.servers', value: 'broker:29092')
        .option(key: 'subscribe', topic)
        .option(key: 'startingOffsets', value: 'earliest')
        .option(key: 'failOnDataLoss', value: 'false')
        .load()
        .selectExpr('CAST(value AS STRING)')
        .select(from_json(col('value'), schema).alias('data'))
        .select('data.*')
        .withWatermark(eventTime: 'timestamp', delayThreshold: '2 minutes')
    )
```

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

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 Run

→ 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

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

Resources in use:

Applications: 1 Running, 0 Completed
Pending: 0 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-2024022711454-0000	(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

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 Comp

Drivers: 0 Run

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

The screenshot shows a PyCharm IDE interface with the following details:

- Project:** SmartCityRvm
- Branch:** master
- Files:** docker-compose.yml, spark-city.py, config.py, .env, main.py, 2022-12-02_12_12_41.gpx
- Code Editor:** The main.py file is open, showing Python code for processing GPX files and publishing to Kafka topics. The code includes logic for reading points from a GPX file, calculating time intervals between points, and sleeping for those intervals. It also includes a final print statement indicating completion.
- Terminal:** The terminal shows the output of running the script. It displays messages being delivered to various Kafka topics (vehicle_data, weather_data, failures_data) and concludes with the message "Simulation completed. All the GPX points have been processed."

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

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	failures_data/	Folder	-	-	-
<input type="checkbox"/>	vehicle_data/	Folder	-	-	-
<input type="checkbox"/>	weather_data/	Folder	-	-	-

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](#) [>](#) [⚙️](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	_spark_metadata/	Folder	-	-	-
<input type="checkbox"/>	part-00000-00109ffb-bea2-4010-8a84-638b74f1a35fc000.snappy.parquet	parquet	February 27, 2024, 12:40:06 (UTC+01:00)	4.1 KB	Standard
<input type="checkbox"/>	part-00000-00251bd4-bbbe-4284-9130-2b6710ff692bc000.snappy.parquet	parquet	February 27, 2024, 12:28:24 (UTC+01:00)	4.1 KB	Standard
<input type="checkbox"/>	part-00000-00accee7-39d7-468c-952d-6c14131abfdac000.snappy.parquet	parquet	February 27, 2024, 12:32:36 (UTC+01:00)	4.1 KB	Standard
<input type="checkbox"/>	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 > Crawlers

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.

Name	State	Schedule	Last run	Last run timestamp	Log
No resources					
No resources to display.					

Filter crawlers

Action ▾ Run Create crawler

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

Table changes from last ...

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

Data Integration and ETL
Legacy pages

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

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

Choose S3 path

X

S3 buckets > spark-streaming-data-rvm-bkt

Objects (1/2)

C

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

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, with 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

_spark_meta/*

**/_spark_metadata

spark_metadata

Cancel

Add an S3 data source

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.

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

Choose an 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.](#)

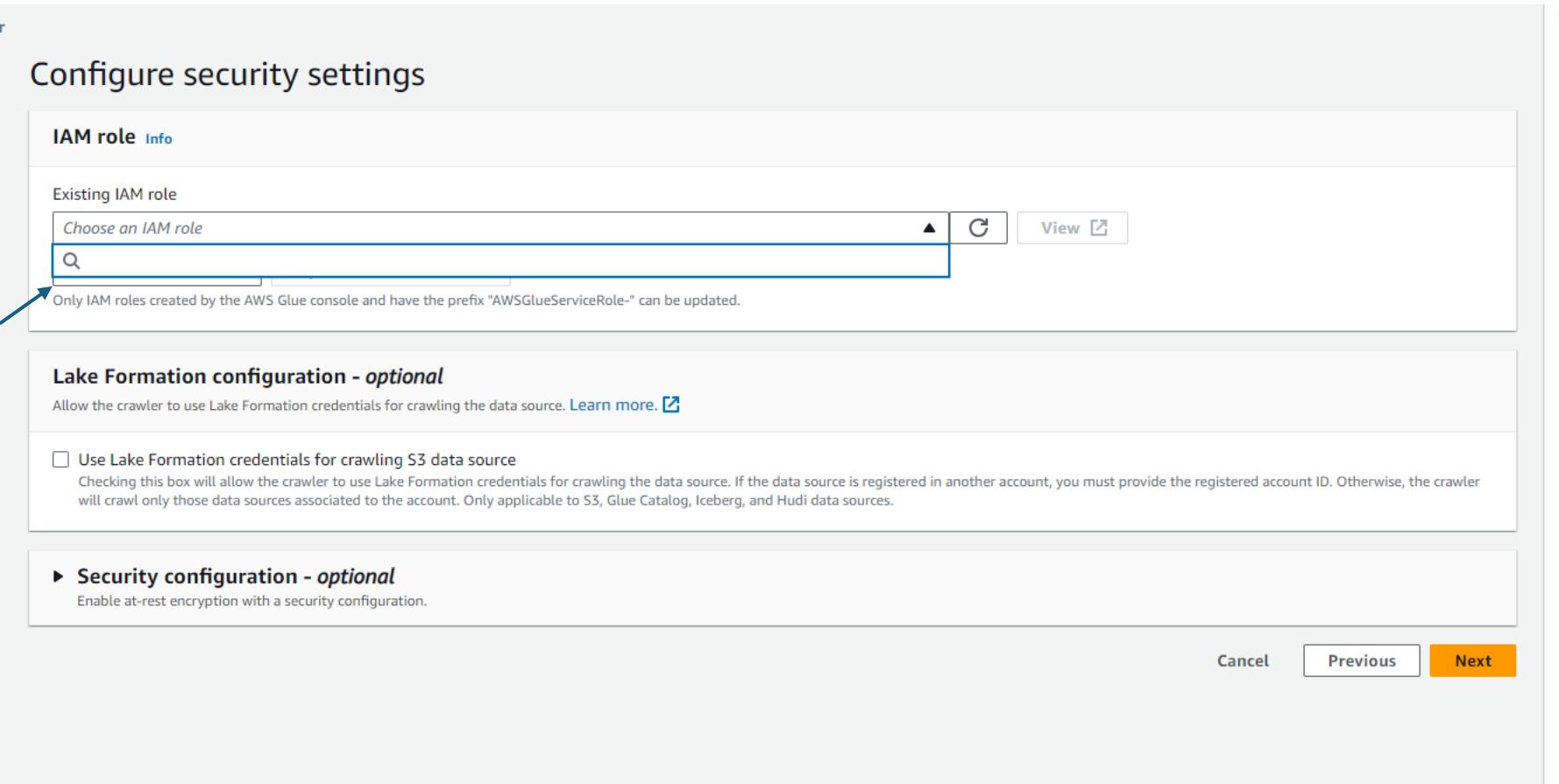
Use Lake Formation credentials for crawling S3 data source

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



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.

Lake F X

Allow the service to use this role to access your data sources.

Use Lambda triggers. Check the box if you want to use Lambda triggers. This will create a Lambda function that triggers the glue job.

Create new IAM role

Enter new IAM role

[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.

X

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 > Crawlers > Add crawler

Step 1 Set crawler properties

Step 2 Choose data sources and classifiers

Step 3 Configure security settings

Step 4 Set output and scheduling

Step 5 Review and create

Set output and scheduling

Output configuration Info

Target database
Choose a database ▾ C
Clear selection Add database ↗
⚠ Target database is required

Table name prefix - optional
Type a prefix added to table names

Maximum table threshold - optional
This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.
Type a number greater than 0

► Advanced options

Crawler schedule
You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)

Frequency
On demand ▾

Cancel Previous Next

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 >

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	smartcitydb	-	-	February 27, 2024 at 12:10:13

Advanced options by default, I didn't touched it

AWS Glue > Crawlers > Add crawler

Step 1
[Set crawler properties](#)

Step 2
[Choose data sources and classifiers](#)

Step 3
[Configure security settings](#)

**Step 4
Set output and scheduling**

Step 5
Review and create

Set output and scheduling

Output configuration Info

Target database

Choose a database

smartcitydb

Table name prefix - *optional*

Type a prefix added to table names

Maximum table threshold - *optional*

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

► Advanced options

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)

Frequency

On demand

[Cancel](#) [Previous](#) [Next](#)

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

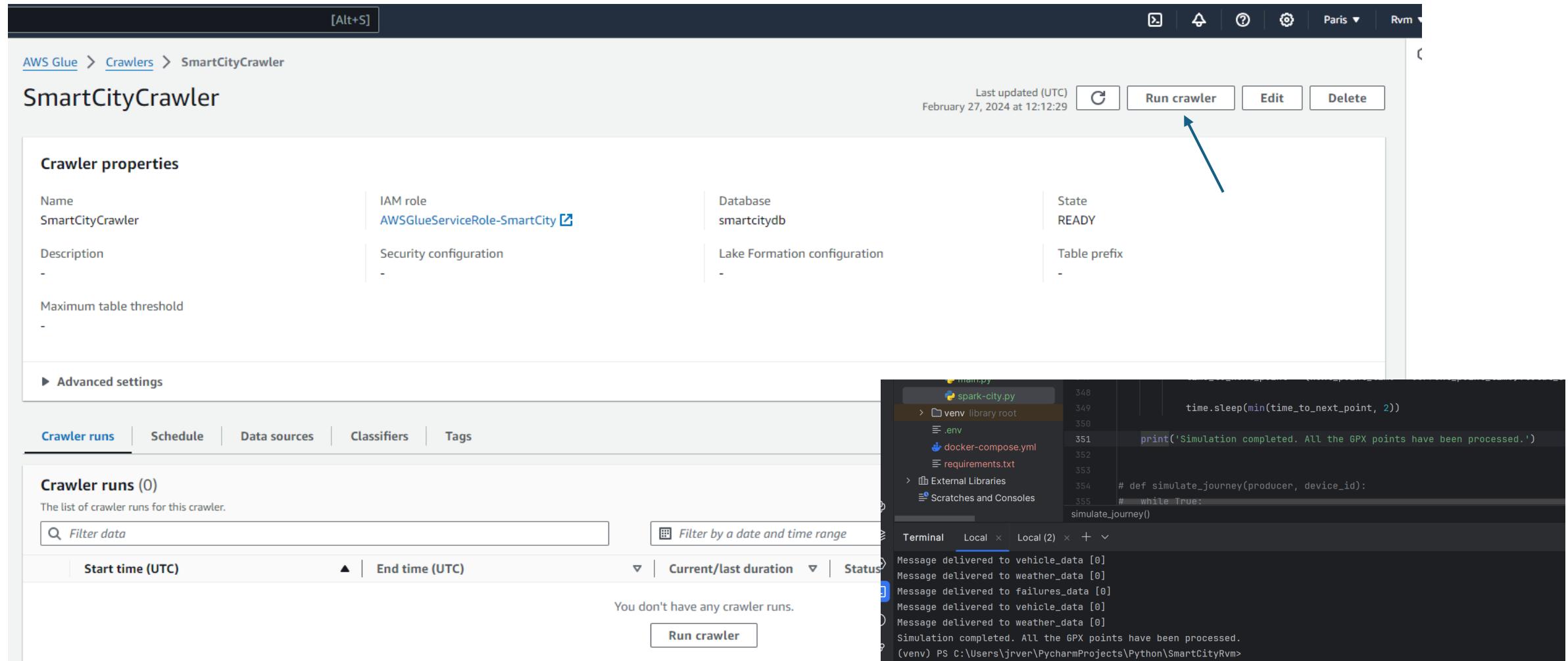
Database	Table prefix - <i>optional</i>	Maximum table threshold - <i>optional</i>	Schedule
smartcitydb	-	-	On demand

[Cancel](#)

[Previous](#)

Create crawler

Run the crawler to start transforming the .parquet files
Run this when the streaming process is completed only!!!



The screenshot shows the AWS Glue Crawler properties page for "SmartCityCrawler". The crawler is in a "READY" state, last updated on February 27, 2024, at 12:12:29. A blue arrow points to the "Run crawler" button. The crawler properties include:

- Name: SmartCityCrawler
- IAM role: AWSGlueServiceRole-SmartCity
- Database: smartcitydb
- Description: -
- Security configuration: -
- Lake Formation configuration: -
- Table prefix: -
- Maximum table threshold: -

The "Advanced settings" section is collapsed. Below the properties, there are tabs for "Crawler runs", "Schedule", "Data sources", "Classifiers", and "Tags". The "Crawler runs" tab shows 0 runs. A "Run crawler" button is located at the bottom.

A terminal window is overlaid on the bottom right, showing the execution of a Python script named "spark-city.py". The script includes code for simulating journeys and delivering messages to various datasets. The terminal output shows the completion of the simulation.

```
main.py
spark-city.py
venv library root
.env
docker-compose.yml
requirements.txt
External Libraries
Scratches and Consoles

348
349
350
351 print('Simulation completed. All the GPX points have been processed.')
352
353
354 # def simulate_journey(producer, device_id):
355 #     while True:
#         simulate_journey()

Terminal Local Local (2) +
Message delivered to vehicle_data [0]
Message delivered to weather_data [0]
Message delivered to failures_data [0]
Message delivered to vehicle_data [0]
Message delivered to weather_data [0]
Simulation completed. All the GPX points have been processed.
(venv) PS C:\Users\jrver\PycharmProjects\Python\SmartCityRvm>
```

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 | C Run crawler Edit Delete

Crawler properties

Name	IAM role	Database	State
SmartCityCrawler	AWSGlueServiceRole-SmartCity	smartcitydb	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.

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	-	-

View CloudWatch logs View run details

Crawler successfully starting

The following crawler is now starting: "SmartCityCrawler"

X

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

< 1 > |

Filter data

Filter by a date and time range

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 X You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the documentation. Create table X

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

AWS Glue > Tables

Tables

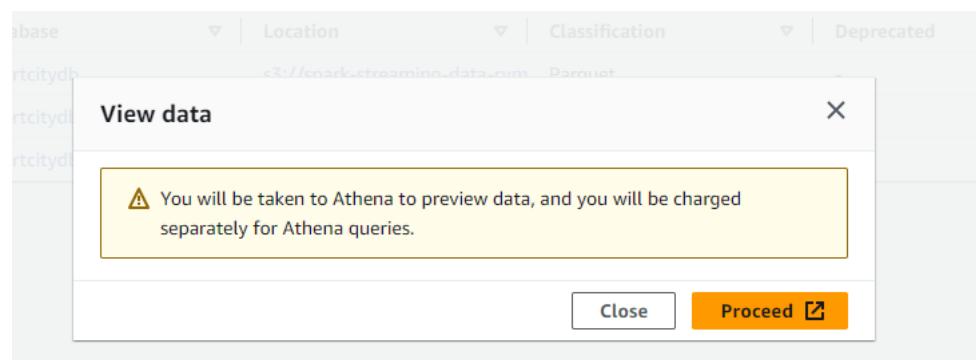
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (3) Last updated (UTC)
February 27, 2024 at 12:16:13 C Delete Add tables using crawler Add table

View and manage all available tables.

	Name	Database	Location	Classification	Deprecated	View data	Data quality
<input type="checkbox"/>	failures_data	smartcitydb	s3://spark-streaming-data-rvm	Parquet	-	Table data	View data quality
<input type="checkbox"/>	vehicle_data	smartcitydb	s3://spark-streaming-data-rvm	Parquet	-	Table data	View data quality
<input type="checkbox"/>	weather_data	smartcitydb	s3://spark-streaming-data-rvm	Parquet	-	Table data	View data quality

Filter tables < 1 > ⚙



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



AWS Services Search [Alt+S] Workgroup query engine upgrade complete One or more workgroups have been upgraded to Athena engine version 3. To see the workgroups that have been upgraded, use the [Workgroup list page](#). For information about new features, see the [Athena Engine Version Reference](#).

Amazon Athena > Query editor tabs

Editor Recent queries Saved queries **Settings** Workgroup primary

Before you run your first query, you need to set up a query result location in Amazon S3. Edit settings

Athena now supports typeahead code suggestions to speed up SQL query development Typeahead suggestions are turned on by default. You can change this setting in query editor preferences. Edit preferences X

Data Data source: AwsDataCatalog Database: smartcitydb Tables and views Create Filter tables and views

Query 1 : 1 SELECT * FROM "AwsDataCatalog"."smartcitydb"."vehicle_data" limit 10;

[Editor](#) | [Recent queries](#) | [Saved queries](#) | [Settings](#)Workgroup [primary](#) ▾**Query result and encryption settings****Query result location and encryption**

Query result location

-

Encrypt query results

-

Expected bucket owner

-

Assign bucket owner full control over query results

Turned off

[Manage](#)

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

Amazon Athena > Query editor tabs > Manage settings

Manage settings

Query result location and encryption

Location of query result - *Optional*
Enter an S3 prefix in the current region where the query result will be saved as an object.

[View](#) [Browse S3](#)

Expected bucket owner - *Optional*
Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

Assign bucket owner full control over query results
Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

Encrypt query results

[Cancel](#) [Save](#)

Amazon Athena > Query editor tabs > Manage settings

Manage settings

Query result location and encryption

Location of query result - *Optional*
Enter an S3 prefix in the current region where the query result will be saved as an object.

[View](#) [Browse S3](#)

You can create and manage lifecycle rules for this bucket [Lifecycle configuration](#)

Use Amazon S3 Lifecycle rules to store your query results and metadata cost effectively or to delete them after a period of time.
[Find out more](#)

Expected bucket owner - *Optional*
Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

Assign bucket owner full control over query results
Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

Encrypt query results

[Cancel](#) [Save](#)

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

The screenshot shows the AWS Athena Query Editor interface. The top navigation bar includes 'Services', a search bar, and various icons for account management and permissions. The main area is titled 'Editor' and shows a message about typeahead code suggestions. A sidebar on the left is titled 'Data' and contains sections for 'Data source' (set to 'AwsDataCatalog'), 'Database' (set to 'smartcitydb'), and 'Tables and views'. It lists three tables: 'failures_data', 'vehicle_data', and 'weather_data'. Below this is a section for 'Views (0)'. The central workspace is titled 'Query 1 :'. It contains a single SQL query: 'SELECT * FROM "AwsDataCatalog"."smartcitydb"."vehicle_data" limit 10;'. Below the query is a toolbar with buttons for 'Run' (highlighted in orange), 'Explain', 'Cancel', 'Clear', and 'Create'. To the right of the toolbar is a switch for 'Reuse query results' which is set to 'up to 60 minutes ago'. The bottom section is titled 'Results' and shows a search bar with 'Search rows' placeholder text. The status message 'No results' is displayed.

Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

Editor Recent queries Saved queries Settings Workgroup primary

Query 1 :

```
1 | SELECT * FROM "AwsDataCatalog"."smartcitydb"."vehicle_data" limit 10;
```

Data source: AwsDataCatalog
Database: smartcitydb
Tables and views: Create, Filter tables and views
Tables (3): failures_data, vehicle_data, weather_data
Views (0)

SQL Ln 1, Col 1

Run Explain Cancel Clear Create Reuse query results up to 60 minutes ago

Query results | Query status

Results

No results

Data C <  **Query 1 :**

1 `SELECT * FROM "AwsDataCatalog"."smartcitydb"."failures_data" limit 10;`

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 >

SQL Ln 1, Col 71   

[Run again](#) [Explain](#)  [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)  [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): failures_data, vehicle_data, weather_data

Views (0)

SQL: SELECT * FROM "AwsDataCatalog"."smartcitydb"."failures_data" WHERE description IS NOT NULL AND description <> '' LIMIT 10;

Run again, Explain, Cancel, Clear, Create, Reuse query results up to 60 minutes ago

Query results: Completed, Time in queue: 119 ms, Run time: 1.345 sec, Data scanned: 2.01 KB

Results (2):

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 respected
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

Copy, Download results, Search rows, <, 1, >,

Data  

Data source: AwsDataCatalog

Database: smartcitydb

Tables and views:  

Filter tables and views:  Filter tables and views

Tables (3)  1 

- + failures_data 
- + vehicle_data 
- + weather_data 

Views (0)  1 

Query 1 :

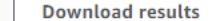
```
1 SELECT * FROM "AwsDataCatalog"."smartcitydb"."weather_data" limit 10;
```

SQL Ln 2, Col 1   

     Reuse query results up to 60 minutes ago 

Query results 

Completed Time in queue: 71 ms Run time: 504 ms Data scanned: 0.79 KB

Results (10)  

Search rows  1  

#	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



Services

Search

[Alt+S]

Paris ▾ Rvm ▾

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.

How it works



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](#)

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

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](#)

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

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

```
graph TD; A[Cluster permissions] --> B[Associate IAM roles]; C[Roles] --> D[Create role]
```

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.

Search

Role name	Trusted entities	Last activity
AWSGlueServiceRole-SmartCity	AWS Service: glue	15 minutes ago
AWSServiceRoleForSupport	AWS Service: support (Service-Linked)	-
AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked)	-

[Create role](#)

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.

```
graph TD; A[Roles] --> B[Create role]; C[Create role]
```

Step 1

Select trusted entity Info

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

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

All types

2 matches

 s3re

<

1

>



Policy name

- | Policy name | Type | Description |
|---|-------------|--|
| <input checked="" type="checkbox"/> AmazonS3ReadOnlyAccess | AWS managed | Provides read only access to all buckets vi... |
| <input type="checkbox"/> AWSBackupServiceRolePolicyForS3Restore | AWS managed | Policy containing permissions necessary f... |

► Set permissions boundary - optional

[Cancel](#)[Previous](#)[Next](#)

AWS Services Search [Alt+S] Global ▾ Rvm ▾

Step 2 Add permissions

Step 3 Name, review, and create

Role details

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

Description
Add a short explanation for this role.
 Maximum 1000 characters. Use alphanumeric and '+,=,@-_ characters.

Step 1: Select trusted entities

Edit

Trust policy

```
1: {
2:     "Version": "2012-10-17",
3:     "Statement": [
4:         {
5:             "Effect": "Allow",
6:             "Action": [
7:                 "sts:AssumeRole"
8:             ],
9:             "Principal": {
10:                 "Service": [
11:                     "redshift.amazonaws.com"
12:                 ]
13:             }
14:         }
15:     ]
16: }
```

Step 2: Add permissions

Edit

Permissions policy summary

Policy name	Type	Attached as

[Alt+S]

Role smart-city-redshift-s3-role created.

View role X ⓘ

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.

Search < 1 > ⓘ

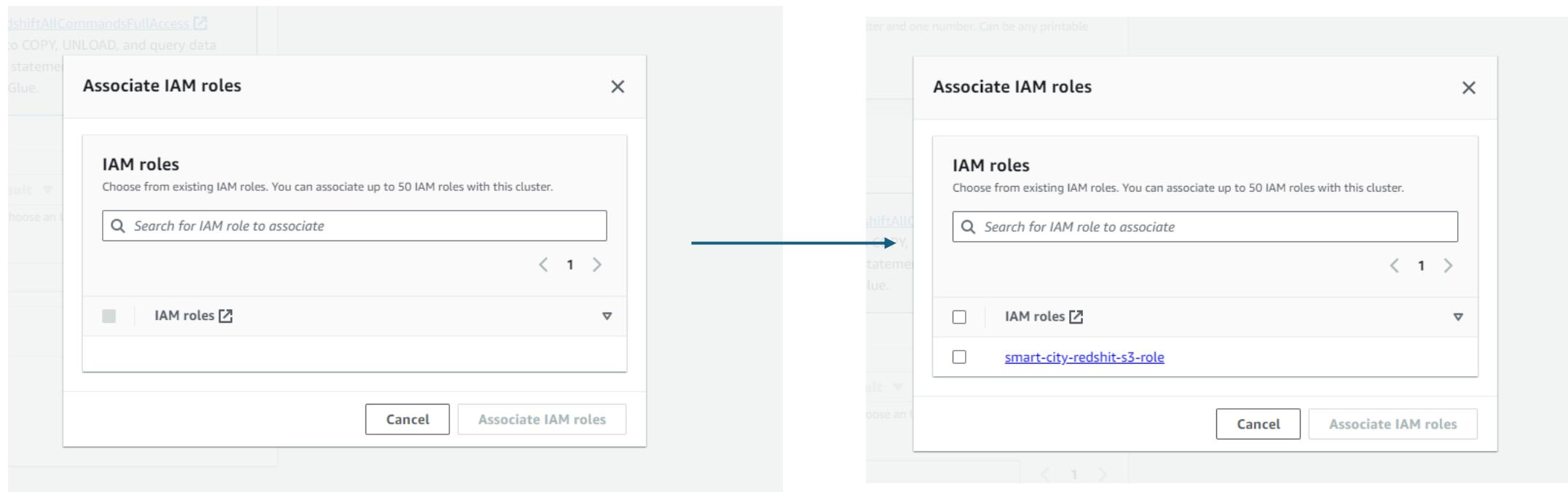
Role name	Trusted entities	Last activity
AWSGlueServiceRole-SmartCity	AWS Service: glue	28 minutes ago
AWSServiceRoleForSupport	AWS Service: support (Service-Linked)	-
AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked)	-
smart-city-redshift-s3-role	AWS Service: redshift	-

Roles Anywhere ⓘ Manage

Authenticate your non AWS workloads and securely provide access to AWS services.

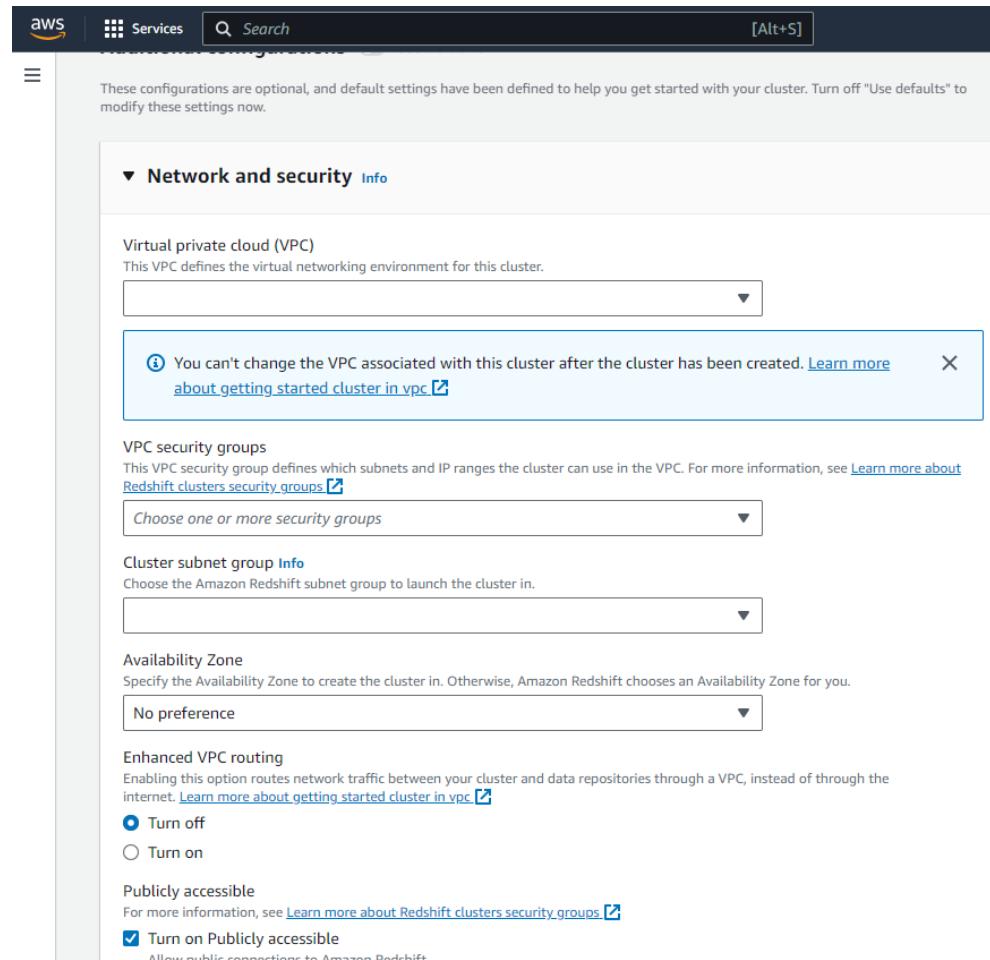
It does not refresh automatically, so you need to refresh the create cluster page and fill in all the information again.

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 creates itself 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 Rvm

VPC dashboard X Create VPC Launch EC2 Instances Note: Your Instances will launch in the Europe region.

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

Resources by Region Refresh Resources You are using the following Amazon VPC resources

VPCs	Europe 0
Subnets	Europe 0
Route Tables	Europe 0
Internet Gateways	Europe 0
NAT Gateways	Europe 0
VPC Peering Connections	Europe 0
Network ACLs	Europe 0
Security Groups	Europe 0

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 /24.

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 one AZ 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** **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](#)

Preview

```

graph LR
    VPC[redshift-vpc] --- euwest3a[eu-west-3a]
    VPC --- euwest3b[eu-west-3b]
    euwest3a --- rtbpublic[redshift-vpc-rtb-public]
    euwest3b --- rtbprivate[redshift-vpc-rtb-private1-eu-west-3a]
    rtbpublic --- lgw[redshift-vpc-lgw]
    rtbprivate --- lgw
    lgw --- s3[redshift-vpc-s3]
  
```

[Cancel](#) [Create VPC](#)

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 X

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-0725ead700d65e23
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

```
graph LR; V[VPC] --> S1[Subnets (2)]; V --> RT1[Route tables (3)]; V --> NC1[Network connections (2)]; S1 --> RT1; RT1 --> NC1;
```

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-3a

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

The screenshot shows the AWS VPC dashboard with the 'Security groups' section selected. A blue arrow points to the 'Security groups' link in the left sidebar. The main view displays the details for the 'sg-0330fe335113275a5 - default' security group. The 'Inbound rules' tab is active, showing two entries:

Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description
-	sgr-0f1197f53c62a4072	-	All traffic	All	All	sg-0330fe335113275...	-
-	sgr-0746cc1fa5aedcff4	IPv4	Redshift	TCP	5439	87.219.104.136/32	-

▼ 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 [X](#)
sg-0e14a3b2c904bdccc

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](#)

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

The screenshot shows the AWS Amazon Redshift Serverless console. On the left, a sidebar menu lists various options: 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 (Informatica Data Loader), Advisor, AWS Marketplace, and Alarms. A blue arrow points to the 'Subnet groups' link under 'Configurations'. The main content area features a large banner for 'Amazon Redshift' with the subtext 'Fast, fully managed, petabyte-scale cloud data warehouse.' Below this is a 'How it works' section with a video thumbnail titled 'Getting Started with Amazon Redshift Serverless | Amazon Web Services'. To the right of the video are several callout boxes: 'Get to powerful insights fast' (with a 'Try Redshift Serverless free trial' button), 'Getting started' (with links to 'Redshift Serverless overview' and 'Evaluation and POC support'), 'For more granular control' (with a 'Create cluster' button), and 'Pricing' (with links to 'Redshift Serverless pricing' and 'Managing usage limits, query limits and other administrative tasks').

AWS Services Search [Alt+S] Paris Rvm

Amazon Redshift X

Redshift Serverless New

Provisioned clusters dashboard

Clusters

Query editor

Query editor v2

Queries and loads

Datasources

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

Search Cluster subnet groups

Name Status VPC ID Description Tags

No cluster subnet group
Create cluster subnet group

Actions Delete Create cluster subnet group

< 1 > ⚙️

Create cluster subnet group Info

Cluster subnet group details

Name

You can't modify the name after your subnet group has been created.

The name must be 1-255 characters. Valid characters are A-Z, a-z, 0-9, space, hyphen (-), underscore (_), and period (.).

Description

Add subnets

VPC

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

Availability Zone

Subnet

Subnets in this cluster subnet group (1)

Availability Zone	Subnet ID	CIDR block	IPv6 CIDR block	Action
eu-west-3a	subnet-076f98...	172.31.128.0/20	-	<input type="button" value="Remove"/>

aws Services Search [Alt+S] Paris Rvm

Amazon Redshift X Cluster subnet group cluster-subnet-group-1 was created successfully

Amazon Redshift > Configurations > Subnet groups

Cluster subnet groups (1) Info

Search Cluster subnet groups

Create cluster subnet group

Name	Status	VPC ID	Description	Tags
cluster-subnet-group-1 1 Subnets	Complete	vpc-05604c1ae369d1b01	redshift-cluster-subnet-group	

Clusters

Query editor

Query editor v2

Queries and loads

The screenshot shows the AWS Amazon Redshift Subnet groups page. A green success message at the top states "Cluster subnet group cluster-subnet-group-1 was created successfully". Below this, the breadcrumb navigation shows "Amazon Redshift > Configurations > Subnet groups". The main title is "Cluster subnet groups (1) Info". A search bar is present above the table. The table has columns: Name, Status, VPC ID, Description, and Tags. One row is listed: "cluster-subnet-group-1" (with "1 Subnets" link), "Complete" status, "vpc-05604c1ae369d1b01" VPC ID, "redshift-cluster-subnet-group" description, and no tags. Action buttons include "Create cluster subnet group" (orange), "Delete", and "Actions". The left sidebar includes links for "Redshift Serverless New", "Provisioned clusters dashboard", "Clusters", "Query editor", "Query editor v2", and "Queries and loads".

After refreshing, 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

▼ 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](#) X

VPC security groups
This VPC's 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 X
sg-0e14a3b2c904bdcccd

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.

[Redshift clusters security groups](#)

Choose one or more security groups

default [X](#)
sg-0e1a3b2c904bdcc

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)

Query data

Actions

Create cluster

Filter clusters by property or value

< 1 >

<input type="checkbox"/>	Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ	Storage capacity us...	CPU utilization	Snapsh...	Notificati...	Tags	
<input type="checkbox"/>	smart-city-redshift-clusterrvm dc2.large 1 node 160 GB	Creating	1744d4a8-fc70-428b-...	-	No	-	-	-	-	-	

It takes around 15 minutes to be created

smart-city-redshift-clusterrvm has been successfully created.

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

Filter clusters by property or value

C Query data Actions Create cluster

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	Available	1744d4a8-fc70-428b-b...	eu-west-3a	No	-	-	-	-	-

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.

The screenshot shows the AWS Redshift Cluster details page for the cluster 'smart-city-redshift-clusterrvm'. The 'General information' section displays various cluster details. On the right side, under the 'Endpoint' section, there is a copied URL: 'jdbc:redshift://smart-city-redshift-clusterrvm.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com:5439/dev'. A green status bubble indicates that the endpoint has been copied. Below this, the ODBC URL is also shown: 'Driver={Amazon Redshift (x64)}; Server=smart-city-redshift-clusterrvm.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com; Database=dev'.

Services Search [Alt+S] Paris ▾ Rvm ▾

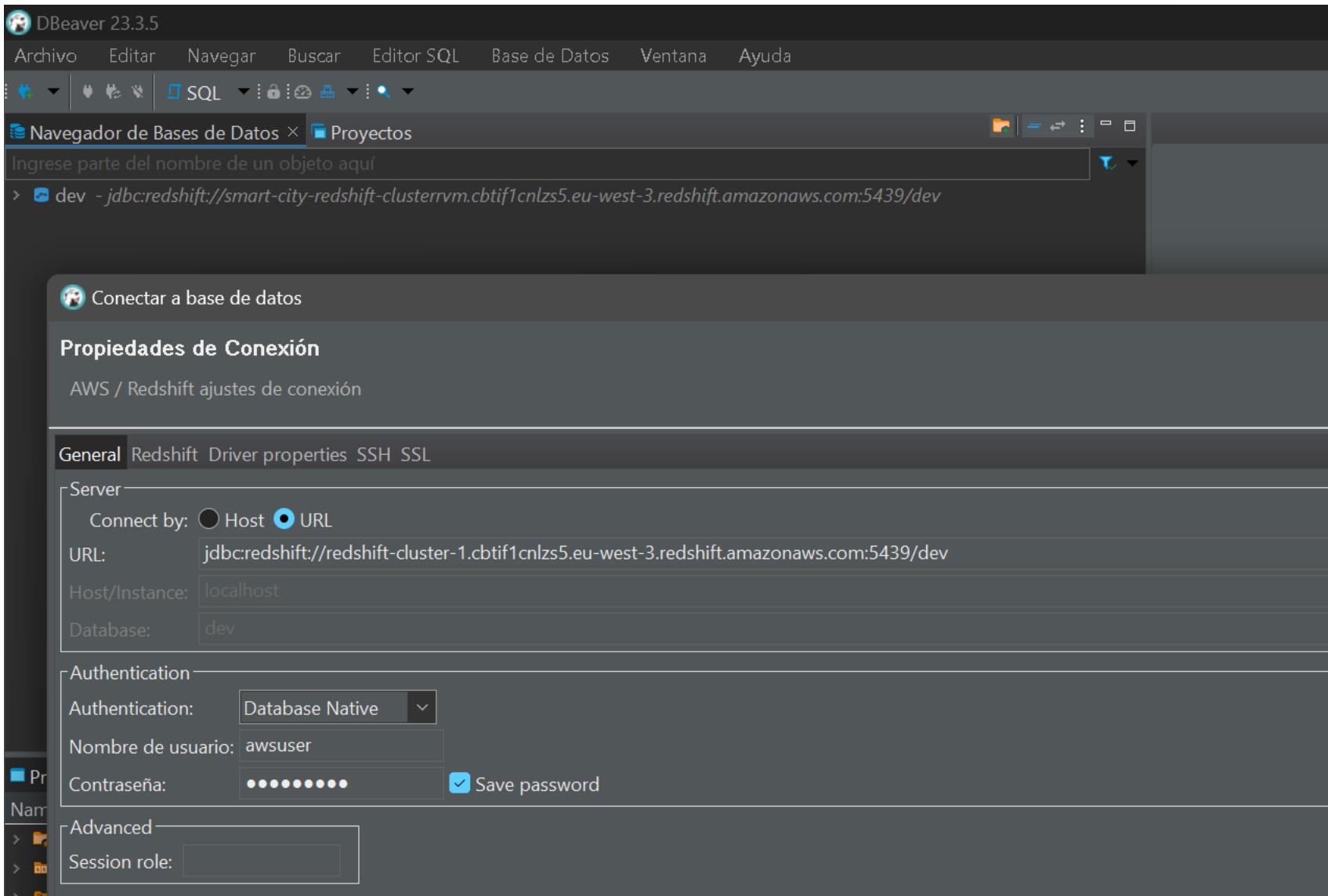
Amazon Redshift > Clusters > smart-city-redshift-clusterrvm

smart-city-redshift-clusterrvm Actions ▾ Edit Add partner integration Query data ▾

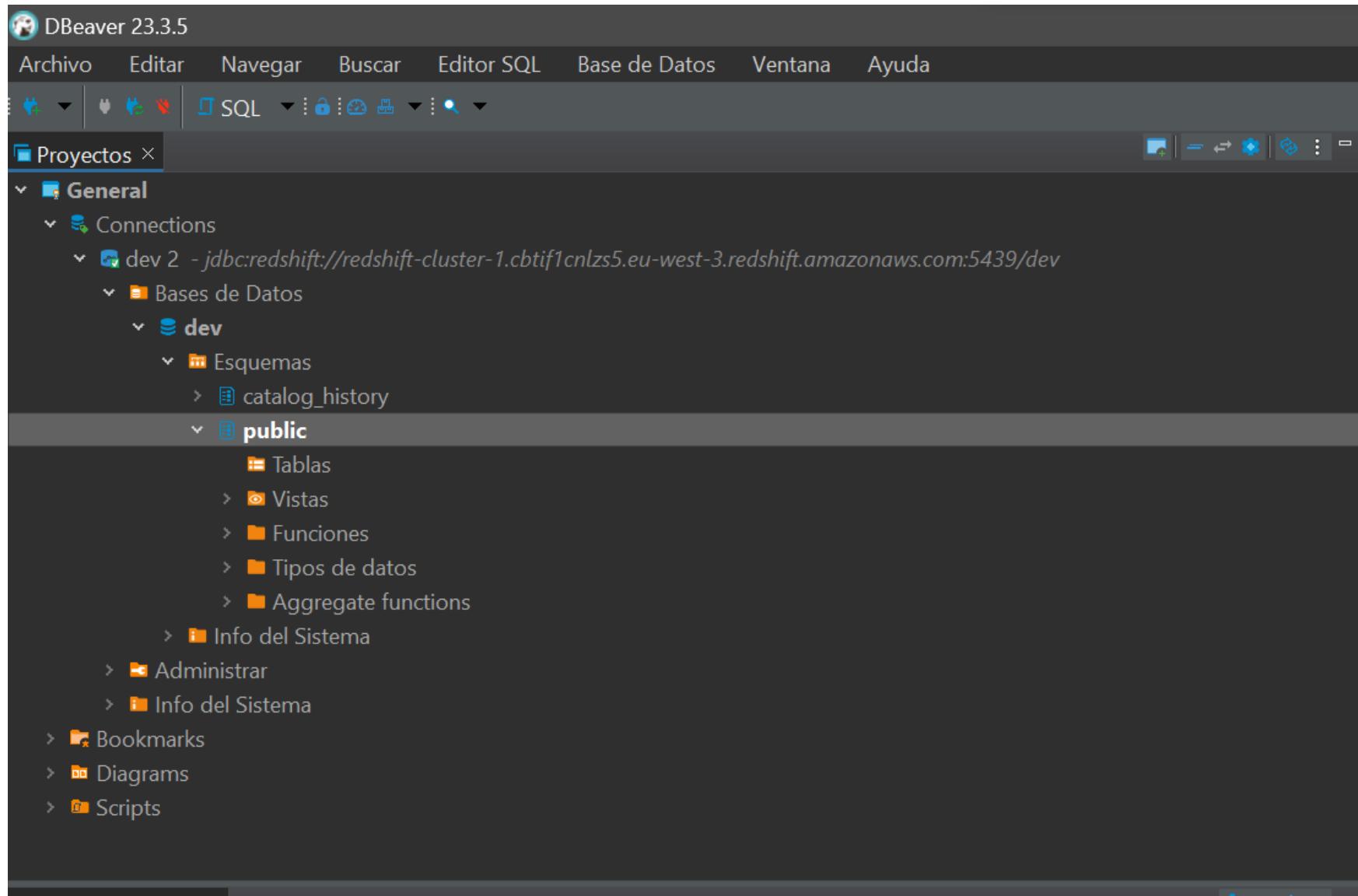
General information Info C

Cluster identifier	Status	Node type	Endpoint
smart-city-redshift-clusterrvm	Available	dc2.large	smart-city-redshift-clusterrvm.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com:5439/dev
Custom domain name	Date created	Number of nodes	Endpoint copied
-	February 27, 2024, 14:51 (UTC+01:00)	1	jdbc:redshift://smart-city-redshift-clusterrvm.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com:5439/dev
Cluster namespace ARN	Storage used	ODBC URL	
arn:aws:redshift:eu-west-3:851725340236:namespace:1744d4a8-fc70-428b-b4cb-83381e34b843	-	Driver={Amazon Redshift (x64)}; Server=smart-city-redshift-clusterrvm.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com; Database=dev	
Cluster configuration	Multi-AZ		
Production	No		

Cluster performance Query monitoring Schedules Maintenance Properties



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.



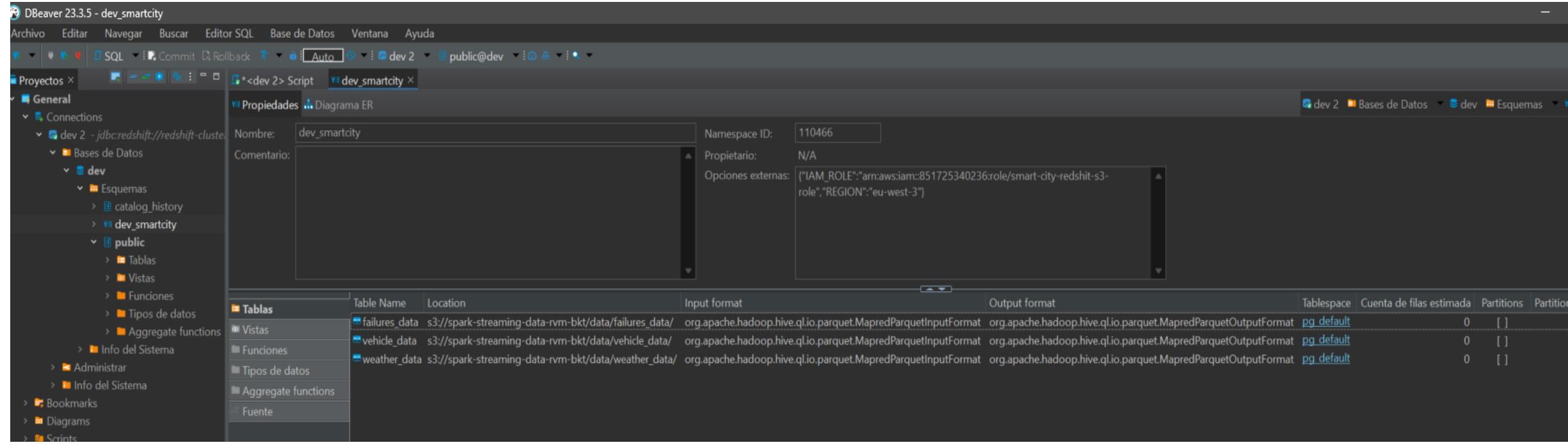
Create an external schema pointing to the Glue catalog and created database.
Pass your IAM role – go to IAM and select the role created before and copy the arn.
This schema allows to query against redshift and retrieving the data from Glue

The screenshot shows the DBeaver interface with the following details:

- Toolbar:** Archivo, Editar, Navegar, Buscar, Editor SQL, Base de Datos, Ventana, Ayuda.
- Connections:** dev 2 - jdbc:redshift://redshift-cluster-1.cbtif1cnlzs5.eu-west-3.redshift.amazonaws.com:5439/dev?user=public&password=public
- Script Editor:** *<dev 2> Script contains the SQL command:

```
create external schema dev_smartcity
from data catalog
database smartcitydb
iam_role 'arn:aws:iam::851725340236:role/smart-city-redshift-s3-role'
region 'eu-west-3';
```
- Project Explorer:** Shows the project structure under General > Connections > dev 2 > Bases de Datos > dev > Esquemas, listing catalog_history and dev_smartcity.
- Statistics:** Estadísticas 1 table showing the execution details of the query:

Name	Value
Updated Rows	0
Query	create external schema dev_smartcity from data catalog database smartcitydb iam_role 'arn:aws:iam::851725340236:role/smart-city-redshift-s3-role' region 'eu-west-3'
Start time	Tue Feb 27 19:31:26 CET 2024
Finish time	Tue Feb 27 19:31:26 CET 2024
- Project - General:** Shows the project structure under General > Scripts.



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-cluster.00000.eu-west-1.amazonaws.com:5439/dev?ssl=true&user=public&password=smartcity123&db=smartcitydb
- 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
- Administrador
- 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-redshift-s3-role'
    region 'eu-west-3';

select * from dev_smartcity.vehicle_data
```

dev_smartcity.vehicle_data_612614a9b105b 1

Enter a SQL expression to filter results (use Ctrl+Space)

	iceid	timestamp	location	speed	length	elevation	nic	make	model	year	nic	fueltype
1	e-Rvm-001	2022-12-02 11:31:37.992	{"latitude":36.71255645159326,"longitude":-4.466778561477C}	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":-4.466452421624E}	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":-4.487963737924123}	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":-4.487738214437E}	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	64,263201599	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":-4.4970218762791}	24,530598223	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":-4.497029637922C}	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":-4.497044272725C}	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":-4.497060215104E}	22,42299788068	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":-4.494401626293C}	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":-4.494482746351C}	58,430515878	0,0157289196	141,2803611755	Yamaha	MT-07	2.024	Gasoline		

Beaver 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-cluster

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

```
*<dev 2> Script > dev_smartcity
•create external schema dev_smartcity
    from data catalog
    database smartcitydb
    iam_role 'arn:aws:iam::851725340236:role/smart-city-redshift-s3-role'
    region 'eu-west-3';

    select * from dev_smartcity.failures_data
```

dev_dev_smartcity_failures_data_612614d9aa06f 1 >

```
select * from dev_smartcity.failures_data
```

Enter a SQL expression to filter results (use Ctrl+Space)

Grilla		deviceid	incidentid	type	timestamp	location	description
Grilla	1	'2a4' [NULL]	bba42109-795e-4006-81a5-806ebceee7e2	Test Drive Suspension	2022-12-02 11:42:39.331	{"latitude":36.71334440357677,"longitude":-4.4669327046763 [NULL]}	
Texto	2	8bd4 [NULL]	b1370abb-33a4-45a9-b047-6b1f5b025cee	Test Drive Suspension	2022-12-02 11:42:40.998	{"latitude":36.71331671535663,"longitude":-4.4669873267453 [NULL]}	
Grilla	3	2d52 [NULL]	7abce15a-1596-45e6-b851-23375444090e	Test Drive Suspension	2022-12-02 11:42:41.997	{"latitude":36.71331026129118,"longitude":-4.4670130316135 [NULL]}	
Texto	4	2a4a [NULL]	36eb9580-bd6d-4646-a48c-3bb9b4602856	Test Drive Suspension	2022-12-02 11:42:42.997	{"latitude":36.71330455321512,"longitude":-4.4671360664111 [NULL]}	
Grilla	5	e6d0e [NULL]	eaf61adf-eae8-42f9-8e6f-a7c10124a6b6	Test Drive Suspension	2022-12-02 11:16:10.993	{"latitude":36.75938627034125,"longitude":-4.4970360249322 [NULL]}	
Texto	6	62fa [NULL]	018b0538-a33c-4693-9527-4a23eb52761	Test Drive Suspension	2022-12-02 11:16:11.993	{"latitude":36.75928593057838,"longitude":-4.4970146846068 [NULL]}	
Grilla	7	4328a [NULL]	44e3e0df-0e4e-4de4-8c66-ad2d973ea0eb	Test Drive Suspension	2022-12-02 11:16:12.993	{"latitude":36.75920151643154,"longitude":-4.4969787262422 [NULL]}	
Texto	8	f5d52 [NULL]	b72a546b-4c27-4cf7-a6f2-63ae52e0d30e	Test Drive Suspension	2022-12-02 11:16:13.993	{"latitude":36.75910529218313,"longitude":-4.4969448633533 [NULL]}	
Grilla	9	bf96 [NULL]	0fc36221-0694-45c0-a497-510623a9dd66	Test Drive Suspension	2022-12-02 11:16:14.993	{"latitude":36.75900778550353,"longitude":-4.496910698716C [NULL]}	

Beaver 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-cluster

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

```
*<dev 2> Script > dev_smartcity
•create external schema dev_smartcity
    from data catalog
    database smartcitydb
    iam_role 'arn:aws:iam::851725340236:role/smart-city-redshift-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 >

```
select * from dev_smartcity.failures_data WHERE description IS NOT NULL AND description <> ''
```

Enter a SQL expression to filter results (use Ctrl+Space)

Grilla		id	deviceid	incidentid	type	timestamp	location	description
Grilla	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":-4.46777940262524 suspensions: Amber light not respected}	
Texto	2	8ebc7ff0-202b-425b-a7f8-960b54eefb73	[NULL]	8b6ced2f-db4d-4d29-b2ee-a2bad8de1754	Test Drive Suspension	2022-12-02 11:30:16.994	{"latitude":36.71704125594042,"longitude":-4.46774474345561 suspensions: Failure to stop}	

Proyectos x

General

Connections

dev 2 - jdbc:redshift://redshift-cluster-1.ckjwv3qyv3o.us-west-2.redshift.amazonaws.com:5439/dev?ssl=true&user=smartcityuser&password=SmartCity123456789&db=smartcitydb

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 x #dev_smarty

```
create external schema dev_smarty
    from data catalog
    database smartcitydb
    iam role 'arn:aws:iam::851725340236:role/smart-city-redshift-s3-role'
    region 'eu-west-3';

select * from dev_smarty.weather_data
```

dev_dev_smarty_weather_data_612615683ee37 1 x

Enter a SQL expression to filter results (use Ctrl+Space)

Grilla	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

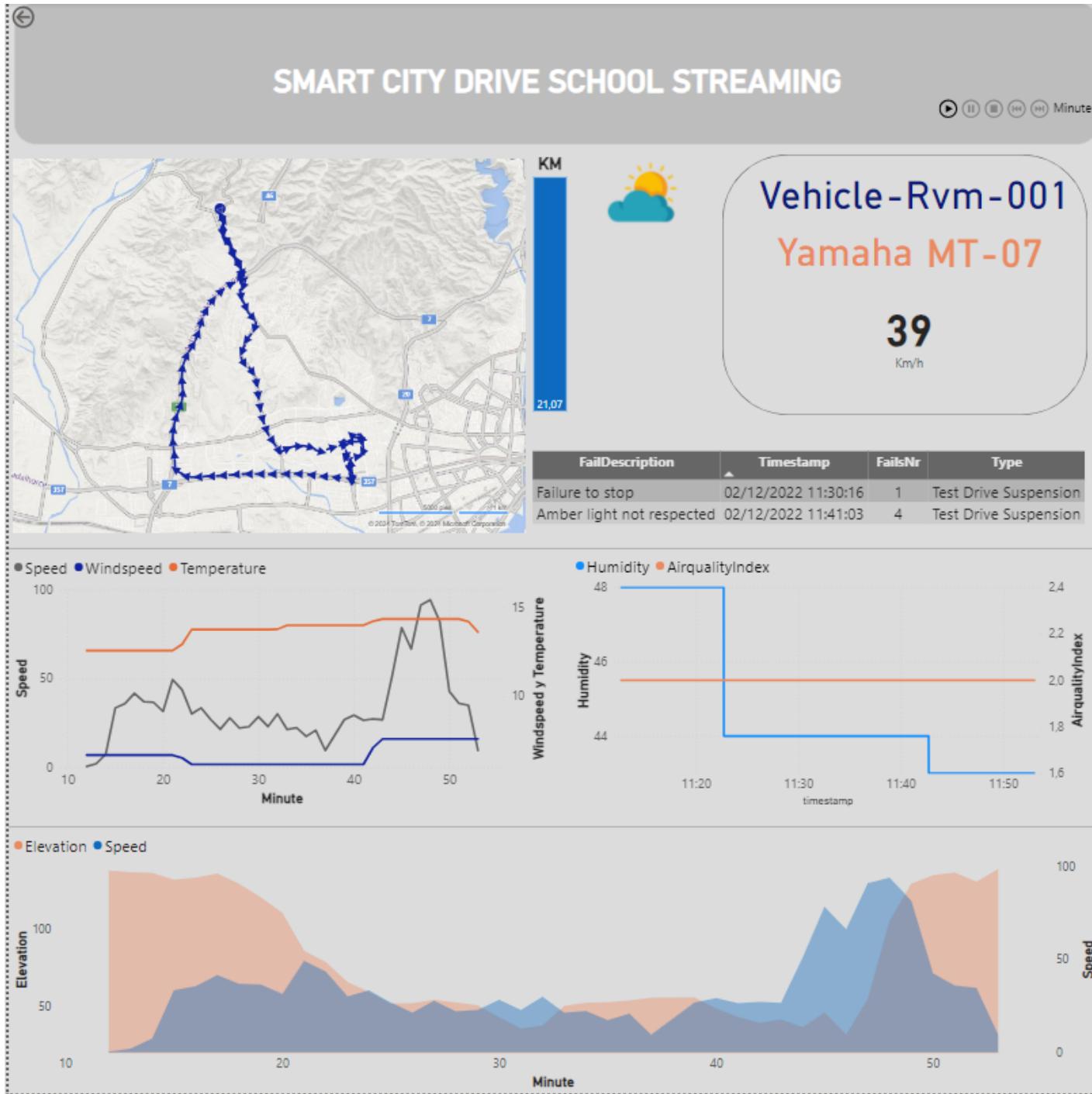
On the background you can see other tables. This was the test I made before in order to prepare the dashboard a bit, to avoid overcosts for having the aws services deployed. In this way, It was just conecting, cleaning some columns and visualize.

The screenshot shows the Microsoft Power BI desktop interface. The top ribbon includes 'Archivo', 'Inicio', 'Transformar', 'Agregar columna', 'Vista', 'Herramientas', and 'Ayuda'. Below the ribbon are various icons for file operations like 'Nuevo' (New), 'Orígenes de datos' (Data Sources), 'Propiedades' (Properties), 'Editor avanzado' (Advanced Editor), 'Consultas' (Queries), and 'Transformar' (Transform). A 'Navegador' (Navigator) pane on the left lists three queries: 'vehicle', 'Weather', and 'Failures'. The main area displays two tables: 'vehicle' and 'weather_data'. The 'weather_data' table has columns: 'temperature', 'weathercondition', 'windspeed', 'humidity', 'airqualityindex', and 'time'. The data shows various weather conditions like 'Clouds' with temperatures ranging from 12.56 to 14.33. The 'vehicle' table has columns: 'id', 'model', and 'year'. The data shows vehicle models like 'MT-07' across different years. On the right, a 'Configuración de la consulta' (Query Configuration) pane is open, showing 'PROPIEDADES' (Properties) with 'Nombre' (Name) set to 'vehículo' and 'Todas las propiedades' (All properties) selected. Under 'PASOS APLICADOS' (Applied steps), 'Tipo cambiado1' is listed. At the bottom, there are buttons for 'Seleccionar tablas relacionadas' (Select related tables), 'Aceptar' (Accept), and 'Cancelar' (Cancel).

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 Q 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

.NET Java Node.js Python Ruby Custom runtime

```
1 - exports.handler = async (event) => {
2   console.log(event);
3   return 'Hello from Lambda!';
4 };
5 
```

Run Next: Lambda responds to events

AWS Services Search [Alt+S] Paris Rvm

AWS Lambda X

Lambda > Layers

Last fetched 10 seconds ago C Create layer

Layers (0)

Filter layers

Name Version Compatible runtimes Compatible architectures

There is no data to display.

Additional resources

- Code signing configurations
- Event source mappings
- Layers **Layers**
- Replicas

Related AWS resources

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 webpage

The screenshot shows the AWS Lambda service interface. On the left, there's a sidebar with links like Dashboard, Applications, Functions, Additional resources (Code signing configurations, Event source mappings, Layers, Replicas), and Related AWS resources. The main area is titled 'Layers' and shows a message 'There is no data to display.' There's a search bar labeled 'Filter layers' and a 'Create layer' button. A navigation bar at the top right includes 'Info' and 'Tutorials' tabs, along with other standard AWS navigation icons.

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 how to create this python.zip.

The image shows two screenshots of the AWS Lambda console. The left screenshot is titled 'Create layer' and shows the 'Layer configuration' section. It includes fields for 'Name' (psycopg2), 'Description - optional' (empty), 'Upload a .zip file' (selected), 'Upload a file from Amazon S3' (unchecked), a 'Upload' button, and a preview of a 'python.zip' file (600.81 KB). Below these are sections for 'Compatible architectures - optional' (x86_64, arm64 selected) and 'Compatible runtimes - optional' (Python 3.9 selected). The right screenshot shows the 'psycopg2' layer details page, which displays a success message ('Successfully created layer psycopg2 version 1.'), version details (Version 1, Created 1 second ago, Compatible architectures empty), and a 'Versions' tab showing one version (All versions (1)).

Create layer

Layer configuration

Name: psycopg2

Description - optional:

Upload a .zip file

Upload a file from Amazon S3

Upload

python.zip
600.81 KB

For files larger than 10 MB, consider uploading using Amazon S3.

Compatible architectures - optional Info
Choose the compatible instruction set architectures for your layer.

x86_64

arm64

Compatible runtimes - optional Info
Choose up to 15 runtimes.

Runtimes: Python 3.9

License - optional Info

Create

psycopg2

Successfully created layer psycopg2 version 1.

Version details

Version	Version ARN	Description
1	arn:aws:lambda:eu-west-3:851725340236:layer:psycopg2:1	- Compatible runtimes python3.9

Versions Functions using this version

All versions (1)

Version	Version ARN	Description
1	arn:aws:lambda:eu-west-3:851725340236:layer:psycopg2:1	-

A screenshot of the AWS Lambda Functions page. The left sidebar shows navigation options: Dashboard, Applications, Functions (selected), Additional resources (Code signing configurations, Event source mappings, Layers, Replicas), and Related AWS resources (Step Functions state machines). The main content area is titled "Functions (0)" and includes a search bar with placeholder text "Filter by tags and attributes or search by keyword". A table header with columns "Function name", "Description", "Package type", "Runtime", and "Last modified" is present, followed by a message "There is no data to display." At the top right of the main area, there are buttons for "Actions" and "Create function" (which is highlighted with a blue arrow). The status bar at the bottom indicates "Last fetched 4 minutes ago".

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 Info
Enter a name that describes the purpose of your 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.
 C

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 [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

▶ Function overview Info

Code | Test | Monitor | Configuration **Configuration** | Aliases | Versions

General configuration

Triggers

Permissions

Destinations

Function URL

Environment variables

Tags

VPC

Monitoring and operations tools

Concurrency

Execution role

Role name
`smartcity_function-role-4scu2j7z`

Resource summary

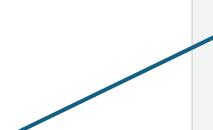
To view the resources and actions that your function has permission to access, choose a service.

Amazon CloudWatch Logs
3 actions, 2 resources

By action | By resource

Resource	Actions

Throttle | Copy ARN | Actions ▾



AWS Services Search [Alt+S] IAM > Roles > smartcity_function-role-4scu2j7z Global Rvm

smartcity_function-role-4scu2j7z Info

Summary Edit

Creation date: February 27, 2024, 19:55 (UTC+01:00) ARN: arn:aws:iam::851725340236:role/service-role/smartcity_function-role-4scu2j7z

Last activity: - 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.

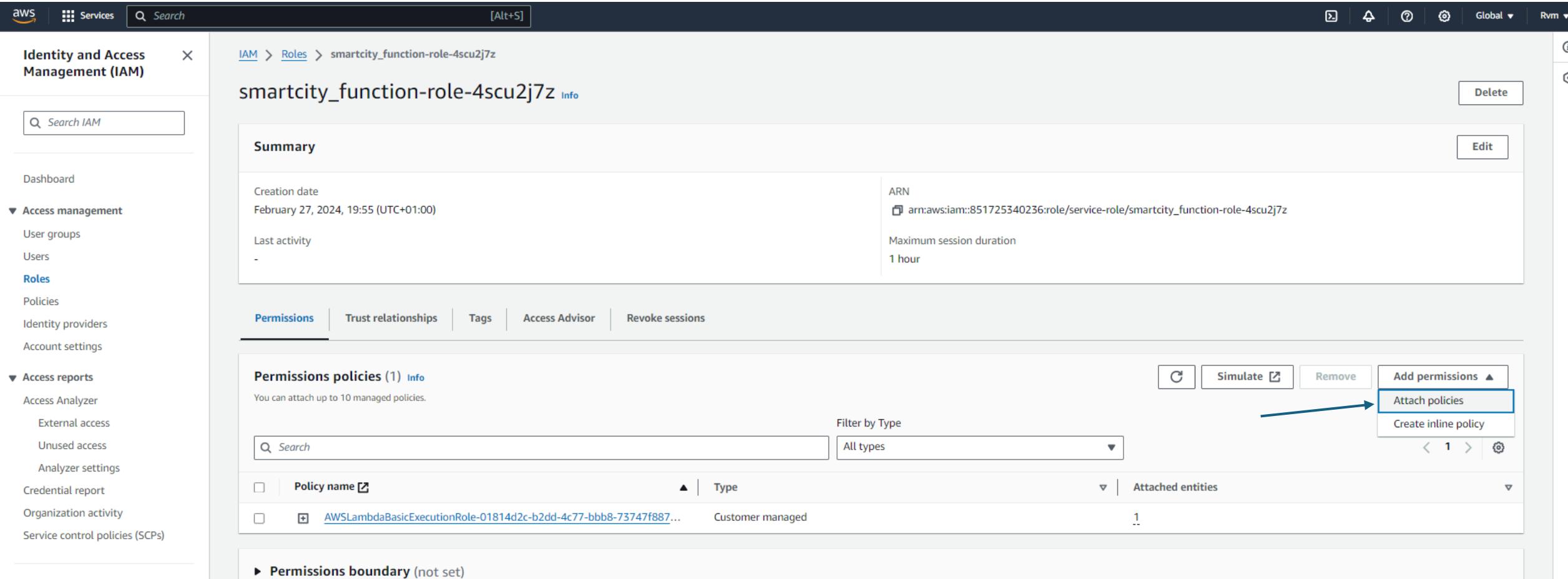
Filter by Type: All types

Attached entities: 1

Actions: C Simulate Remove Add permissions ▲ Create inline policy

Policy name: AWSLambdaBasicExecutionRole-01814d2c-b2dd-4c77-bbb8-73747f887... **Type:** Customer managed

Permissions boundary: (not set)



Attach policy to smartcity_function-role-4scu2j7z

▶ Current permissions policies (1)

Other permissions policies (1/913)

Filter by Type

<input checked="" type="checkbox"/>	Policy name	Type	Description
<input checked="" type="checkbox"/>	AWSLambdaVPCAccessExecutionRole	AWS managed	Provides minimum permissions for a Lam...

⌚ Successfully created the function **smartcity_function**. You can now change its code and configuration. To invoke your function with a test event, choose "Test". X

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 Edit ←

No VPC configuration

This function isn't connected to a VPC.

Edit

Monitoring and operations tools

Concurrency

Asynchronous invocation

The screenshot shows the AWS Lambda Function Configuration page for a function named 'smartcity_function'. The 'Configuration' tab is active. On the left, a sidebar lists various configuration sections: General configuration, Triggers, Permissions, Destinations, Function URL, Environment variables, Tags, and VPC. The 'VPC' section is currently selected and highlighted with a blue border. An arrow points from the text '←' to the 'Edit' button within this section. At the top of the page, a green banner displays a success message: '⌚ Successfully created the function smartcity_function. You can now change its code and configuration. To invoke your function with a test event, choose "Test".'. Other visible buttons include 'Throttle', 'Copy ARN', and 'Actions ▾'.

The same VPC as for Redshift.

Important:

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

There add an inbound rule

Edit VPC

VPC

When you connect a function to a VPC in your account, it does not have access to the internet unless your VPC provides access. To give your function access to the internet, route outbound traffic to a NAT gateway in a public subnet. [Learn more](#)

VPC [Info](#)
Choose a VPC for your function to access.
vpc-026473c1d319e3070 (10.0.0.0/16) [C](#)

Allow IPv6 traffic for dual-stack subnets
You can allow outbound IPv6 traffic to subnets that have both IPv4 and IPv6 CIDR blocks.

Subnets
Select the VPC subnets for Lambda to use to set up your VPC configuration.
[Choose subnets](#) [C](#)

subnet-08c9c32947280d041 (10.0.0.0/20) eu-west-3a [X](#)
Name: project-subnet-public1-eu-west-3a

subnet-0fe959ed1eca0685e (10.0.16.0/20) eu-west-3b [X](#)
Name: project-subnet-public2-eu-west-3b

Security groups
Choose the VPC security groups for Lambda to use to set up your VPC configuration. The table below shows the inbound and outbound rules for the security groups that you choose.
[Choose security groups](#) [C](#)

sg-0330fe335113275a5 (default) [X](#)
default VPC security group

Inbound rules [Outbound rules](#)

Security group ID	Protocol	Ports	Source
sg-0330fe335113275a5	Custom TCP	5439	87.219.104.136/32
sg-0330fe335113275a5	All	All	sg-0330fe335113275a5

[Cancel](#) [Save](#)

Updating the function smartcity_function.

Lambda > Functions > smartcity_function

smartcity_function

Throttle Copy ARN Actions ▾

▶ Function overview Info

Code Test Monitor Configuration Aliases Versions

General configuration Triggers Permissions Destinations

VPC Info

VPC vpc-026473c1d319e3070 (10.0.0.0/16) | project-vpc

Subnets

- Allow IPv6 traffic = false
- subnet-08c9c32947280d041 (10.0.0.0/20) | eu-west-3a,

Security groups

- sg-0330fe335113275a5 (default)

Edit

The screenshot shows the AWS Lambda function configuration page for 'smartcity_function'. The 'Configuration' tab is selected. Under the 'VPC' section, it shows a single VPC named 'project-vpc' with one subnet. The subnet has its 'Allow IPv6 traffic' setting disabled. A security group named 'sg-0330fe335113275a5' is assigned. There are tabs for 'General configuration', 'Triggers', 'Permissions', and 'Destinations' on the left.

Wait until updating succes

Successfully updated the function smartcity_function.

arn:aws:lambda:eu-west-3:851725340236:function:smartcity_function

Function URL Info

The screenshot shows a success message at the top: 'Successfully updated the function smartcity_function.' Below it, the ARN of the function is displayed: 'arn:aws:lambda:eu-west-3:851725340236:function:smartcity_function'. A 'Function URL' button is also visible.

And proceed adding a layer (which contains the Python libraries)

Layers Info

Edit Add a layer

Merge order	Name	Layer version	Compatible runtimes	Compatible architectures	Version ARN
There is no data to display.					

The screenshot shows the 'Layers' configuration page. It has a header 'Layers Info' and buttons for 'Edit' and 'Add a layer'. A table lists 'Merge order', 'Name', 'Layer version', 'Compatible runtimes', 'Compatible architectures', and 'Version ARN'. A note at the bottom states 'There is no data to display.'

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

+ 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)

Environment smartcity_function

lambda_function

```
1 import json
2 import psycopg2
3
4 def lambda_handler(event, context):
5     # TODO implement
6     return {
7         'statusCode': 200,
8         'body': json.dumps('Hello from Lambda!')
9     }
10
```

Successfully updated the function smartcity_function.

smartcity_function

+ Add trigger

+ Add destination

Layers (1)

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)

Environment smartcity_function

lambda_function

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_I lambda_function.py

lambda_function Execution results Environment Var +

```
4 import requests
5 import json
6 import time
7
8
9 # Configuración de la conexión a Redshift
10 conn = psycopg2.connect(
11     dbname='dev',
12     host='redshift-cluster-1.cbtif1cnlz5.eu-west-3.redshift.amazonaws.com',
13     port='5439',
14     user='awsuser',
15     password='Awsuser-1'
16 )
17
18 # Cursor para ejecutar la consulta
19 cur = conn.cursor()
20
21 # La consulta SQL a ejecutar
22 query = "SELECT * FROM dev_smartcity.vehicle_data;"
23
24 # Ejecutar la consulta
25 cur.execute(query)
26
27 # Obtener los resultados de la consulta
28 rows = cur.fetchall()
29
30 # Obtener los nombres de las columnas
31 columns = [desc[0] for desc in cur.description]
32
33 # 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 column in my code, retrieves a json containing latitude and longitude as key value in json format for this column only.

the libraries needed to transform a dataframe with pandas aren't too much, the problem I found is the bunch of libraries that work behind and makes the main library run 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 try such streaming simulation with lambda in the next Project.

The screenshot shows the AWS Lambda Test console interface. The top navigation bar includes 'Code source' and 'Info' tabs, and a 'Test' button which is currently selected. On the right, there is an 'Upload from' button. Below the navigation, there are tabs for 'Execution result' and 'Environment Var'. The 'Execution results' tab is active, showing a 'Test Event Name' input field with '0' and a 'Response' JSON object. The 'Response' object contains an error message about failing to import numpy. The 'Function Logs' section displays the full error stacktrace, including the error message and the Python interpreter's response. At the bottom, the 'Request ID' is listed as 'f7d53f1c-4123-4f5f-bf92-917916ea0c67'.

```
Code source Info
File Edit Find View Go Tools Window Test Deploy
Upload from
Go to Anything (Ctrl-P) lambda_function Execution result Environment Var
Smartcity_lambda_1 lambda_function.py
Execution results
Test Event Name 0
Status: Failed Max memory used: 39 MB Time: 2.37 ms
Response
{
  "errorMessage": "Unable to import module 'lambda_function': Unable to import required dependencies:\nnumpy: Error importing numpy: you should not try to import numpy from\nits source directory; p",
  "errorType": "Runtime.ImportModuleError",
  "requestId": "f7d53f1c-4123-4f5f-bf92-917916ea0c67",
  "stackTrace": []
}
Function Logs
START RequestId: f7d53f1c-4123-4f5f-bf92-917916ea0c67 Version: $LATEST
[ERROR] Runtime.ImportModuleError: Unable to import module 'lambda_function': Unable to import required dependencies:
numpy: Error importing numpy: you should not try to import numpy from
its source directory; please exit the numpy source tree, and relaunch
your python interpreter from there.
Traceback (most recent call last):END RequestId: f7d53f1c-4123-4f5f-bf92-917916ea0c67
REPORT RequestId: f7d53f1c-4123-4f5f-bf92-917916ea0c67 Duration: 2.37 ms Billed Duration: 3 ms Memory Size: 128 MB Max Memory Used: 39 MB Init Duration: 154.95 ms
Request ID
f7d53f1c-4123-4f5f-bf92-917916ea0c67
```

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)

Power BI Mi área de trabajo

Inicio Crear Examinar Centro de datos de... Aplicaciones Métricas Centro de supervisión Más información

+ Nuevo Subir Configuración del área de trabajo

Informe

Informe paginado Tarjeta de resultados Panel Modelo semántico Conjunto de datos de streaming Más opciones

SmartCity_Rafa Ventas Videojuegos Ventas Videojuegos Ventas Videojuegos.pbix

Cree elementos visuales a partir de datos en tiempo real.

Tipo	Propietario
Modelo semántico	Jose Rafael
Informe	Jose Rafael
Modelo semántico	Jose Rafael
Informe	Jose Rafael
Modelo semántico	Jose Rafael
Informe	Jose Rafael
Modelo semántico	Jose Rafael
Panel	Jose Rafael

Nuevo conjunto de datos de transmisión

Elegir el origen de los datos

ANÁLISIS DE TR API PUBNUB

Siguiente Cancelar

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.

Editar conjunto de datos de transmisión

Cree un conjunto de datos de streaming e integre nuestra API en su dispositivo o su aplicación para enviar datos. [Más información acerca de la API](#).

* Requerido

Nombre del conjunto de datos *

SmartCity_Rafa

Valores de la transmisión *

id	Texto	eliminar
deviceid	Texto	eliminar
timestamp	DateTime	eliminar
latitude	Número	eliminar
longitude	Número	eliminar
speed	Número	eliminar
length	Número	eliminar
elevation	Número	eliminar
make	Texto	eliminar

Listo Cancelar

Editar conjunto de datos de transmisión

elevation	Número	eliminar
make	Texto	eliminar
model	Texto	eliminar
year	Número	eliminar
fueltype	Texto	eliminar
Escribir un nuevo nombre de valor	Texto	eliminar

```
[{"id": "AAAAA555555", "deviceId": "AAAAA555555", "timestamp": "2024-02-29T14:58:45.773Z", "latitude": 98.6, "longitude": 98.6, "speed": 98.6, "length": 98.6, "elevation": 98.6, "make": "AAAAA555555", "model": "AAAAA555555", "year": 98.6, "fueltype": "AAAAA555555"}]
```

Análisis del historial de datos

Activar

Listo Cancelar

PowerBI API

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

Información de la API en SmartCity_R...

Use la dirección URL del punto de conexión de la API y uno de los ejemplos que se muestran a continuación para enviar datos a su conjunto de datos de streaming. Para más información, [lea la guía de integración y la documentación de la API](#).

den

URL de inserción

```
https://api.powerbi.com/beta/68519e48-83f3-435f-a38a-1a7aa77ba987/data
```

Raw cURL PowerShell

```
[  
  {  
    "id" : "AAAAAA555555",  
    "deviceId" : "AAAAAA555555",  
    "timestamp" : "2024-02-29T15:01:21.415Z",  
    "latitude" : 98.6,  
    "longitude" : 98.6,  
    "speed" : 98.6,  
    "length" : 98.6,  
    "elevation" : 98.6,  
    "make" : "AAAAAA555555",  
    "model" : "AAAAAA555555",  
    "year" : 98.6,  
    "fueltype" : "AAAAAA555555"  
  }  
]
```

Listo

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

The screenshot illustrates the PowerBI API process for creating a dashboard. It consists of three main sections:

- Left Panel:** A list of datasets and reports:
 - SmartCity_Rafa
 - Ventas Videojuegos
 - Ventas Videojuegos
- Middle Panel:** A context menu for the "SmartCity_Rafa" item, with options "Crear informe" (Create Report) and "Eliminar" (Delete). A blue arrow points from the "Crear informe" option to the "Informe" section in the bottom panel.
- Bottom Panel:** The main workspace showing a world map titled "SmartCity". The map displays oceans and continents with labels for the Arctic, Pacific, Atlantic, and Indian Oceans. Below the map, the text "Suma de elevation, Suma de latitude, Suma de longitude, Suma de elevation y Suma de speed por timestamp" is visible. The left sidebar contains navigation links like "Inicio", "Crear", "Borrar", "Serie de datos de...", "Aplicaciones", "Métricas", "Serie de operación", "Área de trabajo", "El Área de trabajo", and "smartCity".