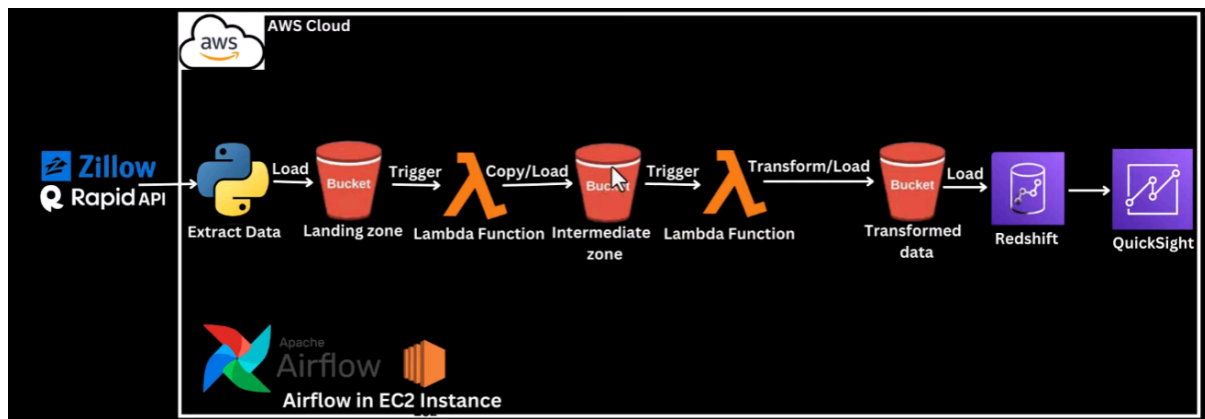
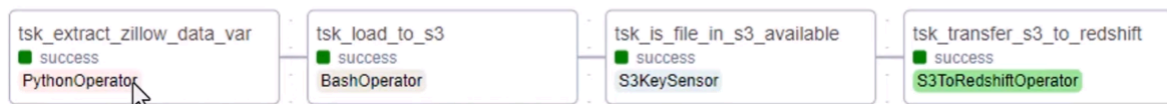


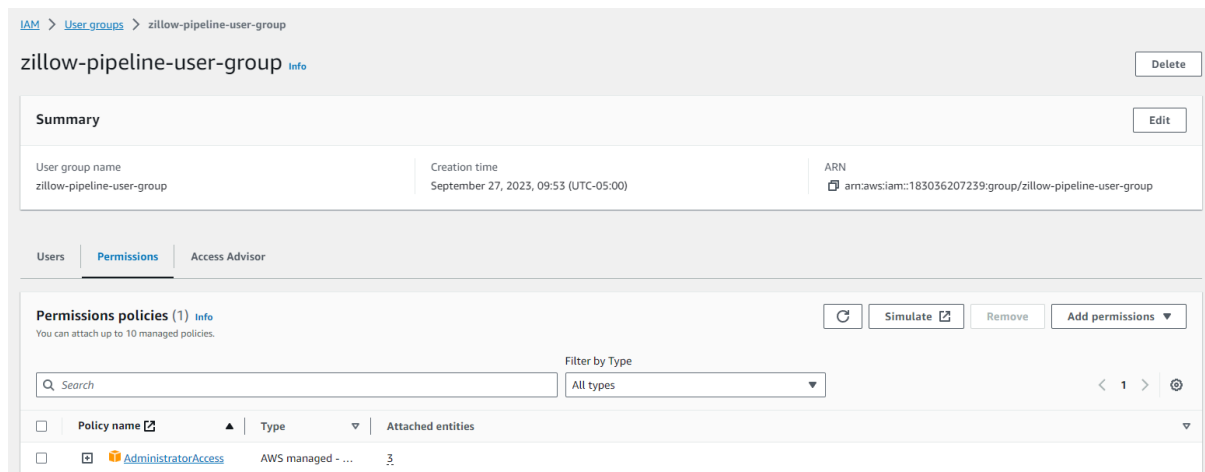
Architecture



DAG TO BUILD



Primero, creamos un User Group para asignar permisos a los servicios que vamos a manejar. Para ello, ingresamos a IAM y creamos un grupo con permisos de Administrator Access.



Ahora debemos crear los usuarios. Para el primer usuario, habilitamos la opción de acceder a la consola de AWS. Le asignamos una contraseña.

User details

User name

zillow-pipeline-user1

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.

Are you providing console access to a person?

User type

☐ Specify a user in Identity Center - Recommended
We recommend that you use Identity Center to provide console access to a person. With Identity Center, you can centrally manage user access to their AWS accounts and cloud applications.

☒ I want to create an IAM user
We recommend that you create IAM users only if you need to enable programmatic access through access keys, service-specific credentials for AWS CodeCommit or Amazon Keyspaces, or a backup credential for emergency account access.

Console password

☐ Autogenerated password
You can view the password after you create the user.

☒ Custom password
Enter a custom password for the user.

ZillowUser#1

Must be at least 8 characters long

Must include at least three of the following mix of character types: uppercase letters (A-Z), lowercase letters (a-z), numbers (0-9), and symbols ! @ # \$ % ^ & * () _ + - (hyphen) = [] { } ' " , . ; : \ | / ~ ` ~`

☒ Show password

☐ Users must create a new password at next sign-in - Recommended
Users automatically get the [IAMUserChangePassword](#) policy to allow them to change their own password.

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

Set permissions

Add user to an existing group or create a new one. Using groups is a best-practice way to manage user's permissions by job functions. [Learn more](#)

Permissions options

☒ Add user to group
Add user to an existing group, or create a new group. We recommend using groups to manage user permissions by job function.

☐ Copy permissions
Copy all group memberships, attached managed policies, and inline policies from an existing user.

☐ Attach policies directly
Attach a managed policy directly to a user. As a best practice, we recommend attaching policies to a group instead. Then, add the user to the appropriate group.

User groups (1/1)

Search

Create group

☒

Group name

Users

Attached ...

Created

<input checked="" type="checkbox"/>	zillow-pipeline-user-group	0	Administrato...	2023-09-27 (7 minutes ago)
-------------------------------------	----------------------------	---	-----------------	----------------------------

Retrieve password

You can view and download the user's password below or email users instructions for signing in to the AWS Management Console. This is the only time you can view and download this password.

Console sign-in details

Email sign-in instructions

Console sign-in URL

https://183036207239.signin.aws.amazon.com/console

User name

zillow-pipeline-user1

Console password

ZillowUser#1

Hide

Cancel

Download .csv file

Return to users list

Para este usuario creamos una access key que usaremos más adelante. Guardamos la key y el secret.

Use case

☒ Command Line Interface (CLI)
You plan to use this access key to enable the AWS CLI to access your AWS account.

Utilizamos la sign-in URL para ingresar a la cuenta de nuestro user #1.

Sign in as IAM user

Account ID (12 digits) or account alias

183036207239

IAM user name

zillow-pipeline-user1

Password

.....

☐ Remember this account

Sign in

N. Virginia ▾

zillow-pipeline-user1 @ 1830-3620-7239 ▾

Ahora, dentro de esta cuenta creamos una instancia EC2, con los siguientes parámetros.

Quick Start

Amazon Linux
aws


macOS
Mac

Ubuntu
ubuntu

Windows
Microsoft

Red Hat
Red Hat

SUSE Linux
SUS


Browse more AMIs
Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

Ubuntu Server 22.04 LTS (HVM), SSD Volume Type
ami-053b0d53c279acc90 (64-bit (x86)) / ami-0a0c8eebcdd6dcdbd0 (64-bit (Arm))
Virtualization: hvm ENA enabled: true Root device type: ebs

Free tier eligible ▾

▼ Instance type [Info](#)

Instance type

t2.medium
Family: t2 2 vCPU 4 GiB Memory Current generation: true
On-Demand Linux base pricing: 0.0464 USD per Hour
On-Demand RHEL base pricing: 0.1064 USD per Hour
On-Demand Windows base pricing: 0.0644 USD per Hour
On-Demand SUSE base pricing: 0.1464 USD per Hour

▾

☒ All generations

[Compare instance types](#)

[Additional costs apply for AMIs with pre-installed software](#)

Firewall (security groups) [Info](#)

A security group is a set of firewall rules that control the traffic for your instance. Add rules to allow specific traffic to reach your instance.

☒ Create security group

☐ Select existing security group

We'll create a new security group called 'launch-wizard-3' with the following rules:

- ☒ Allow SSH traffic from
Helps you connect to your instance
Anywhere
0.0.0.0/0 ▾
- ☒ Allow HTTPS traffic from the internet
To set up an endpoint, for example when creating a web server
- ☒ Allow HTTP traffic from the internet
To set up an endpoint, for example when creating a web server

Adicionalmente creamos una key pair para conectarnos más adelante. Nuestra instancia ya está lista.

Instances (1) Info				
<input type="text" value="Find instance by attribute or tag (case-sensitive)"/>				
<input type="checkbox"/>	Name	Instance ID	Instance state	Instance type
<input type="checkbox"/>	zillow-pipeline...	i-0839a6926a77989d6	Running	t2.medium

Nos conectamos e instalamos dependencias, creamos un entorno virtual y corremos airflow.

1. `sudo apt update`
2. `sudo apt install python3-pip`
3. `sudo apt install python3.10-venv`
4. `python3 -m venv zillowpipeline_venv`
5. `source zillowpipeline_venv/bin/activate`
 - 5.1. `pip install --upgrade awscli`
 - 5.2. `pip install apache-airflow-providers-amazon`
 - 5.3. `sudo pip install apache-airflow`
 - 5.4. `airflow standalone`

Finalmente iniciamos Airflow con la IP pública de nuestra instancia, a través del puerto 8080. Recordemos que se debe habilitar este puerto en el security group, usando Custom TCP.

Nos dirigimos a RapidAPI y buscamos la API de Zillow.

	Basic	Pro	Recommended Ultra	Mega
Objects	\$0.00 / mo Currently Subscribed Manage And View Usage	\$20.00 / mo Change Plan	\$40.00 / mo Change Plan	\$170.00 / mo Change Plan
Requests	30 / month Hard Limit	10,000 / month + \$0.01 each other	30,000 / month + \$0.005 each other	170,000 / month + \$0.003 each other
Rate Limit	one request per second	2 requests per second	2 requests per second	3 requests per second

Utilizamos el código de request en Python para obtener los datos.

```
Code Snippets Example Responses Results

(Python) Requests Copy Code

import requests

url = "https://zillow56.p.rapidapi.com/search"

querystring = {"location":"houston, tx"}

headers = {
    "X-RapidAPI-Key": "a5526cf9c4mshcb1c83ff9e78685p1e43f3jsna7f2c1a8ab90",
    "X-RapidAPI-Host": "zillow56.p.rapidapi.com"
}

response = requests.get(url, headers=headers, params=querystring)

print(response.json())
```

Creamos el task correspondiente a la extracción de datos provenientes de la API de Zillow. Luego creamos el bucket para almacenarlos.

Buckets (1) Info

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

Copy ARN Empty Delete Create bucket

< 1 >

	Name	AWS Region	Access	Creation date
<input type="radio"/>	zillow-pipeline-bucket1	US East (N. Virginia) us-east-1	Bucket and objects not public	September 27, 2023, 14:22:10 (UTC-05:00)

Para conectarnos al S3 desde EC2 creamos un rol con full access.

[IAM](#) > [Roles](#) > zillow_pipeline_ec2_s3_role

zillow_pipeline_ec2_s3_role Info

Allows EC2 instances to call AWS services on your behalf.

Summary

Creation date September 27, 2023, 14:32 (UTC-05:00)	ARN arn:aws:iam:183036207239:role/zillow_pipeline_ec2_s3_role	Instance profile ARN arn:aws:iam:183036207239:instance-profile/zillow_pipeline_ec2_s3_role
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 < 1 >

<input type="checkbox"/>	Policy name	Type	Attached entities
<input type="checkbox"/>	AmazonS3FullAccess	AWS managed	2


Modificamos el rol IAM en el EC2.

EC2 > Instances > i-0839a6926a77989d6 > Modify IAM role

Modify IAM role [Info](#)


Attach an IAM role to your instance.


Instance ID

 i-0839a6926a77989d6 (zillow-pipeline-EC2)

IAM role

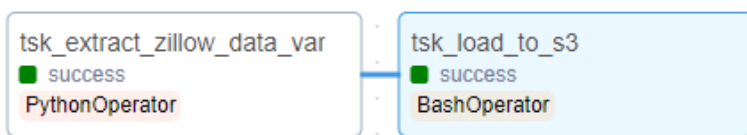
Select an IAM role to attach to your instance or create a new role if you haven't created any. The role you select replaces any roles that are currently attached to your instance.



 [Create new IAM role](#)

Cancel Update IAM role

Hasta ahora el DAG va así.




Creamos un rol para el servicio **Lambda**. Con el permiso `AWSLambdaBasicExecutionRole`, el servicio Lambda puede guardar sus logs en CloudWatch para saber cómo van las ejecuciones.

zillow-pipeline-lambda-role [Info](#) Delete

Allows Lambda functions to call AWS services on your behalf.


Summary Edit



Creation date	September 27, 2023, 14:58 (UTC-05:00)	ARN	 arn:aws:iam::183036207239:role/zillow-pipeline-lambda-role
Last activity	-	Maximum session duration	1 hour

Permissions | Trust relationships | Tags | Access Advisor | Revoke sessions

Permissions policies (2) [Info](#)

You can attach up to 10 managed policies.

Filter by Type All types < 1 > 

<input type="checkbox"/>	Policy name Info	Type	Attached entities
<input type="checkbox"/>	 AmazonS3FullAccess	AWS managed	3
<input type="checkbox"/>	 AWSLambdaBasicExecutionRole	AWS managed	1

Create function [Info](#)

AWS Serverless Application Repository applications have moved to [Create application](#).

☒ Author from scratch
Start with a simple Hello World example.

☐ Use a blueprint
Build a Lambda application from sample code and configuration presets for common use cases.

☐ Container image
Select a container image to deploy for your function.

Basic information


Function name

Enter a name that describes the purpose of your function.

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.



Architecture [Info](#)

Choose the instruction set architecture you want for your function code.

☒ x86_64

☐ arm64

Permissions [Info](#)

By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this default role later when adding triggers.

▼ Change default execution role

Execution role

Choose a role that defines the permissions of your function. To create a custom role, go to the [IAM console](#).


☐ Create a new role with basic Lambda permissions

☒ Use an existing role

☐ Create a new role from AWS policy templates

Existing role

Choose an existing role that you've created to be used with this Lambda function. The role must have permission to upload logs to Amazon CloudWatch Logs.



[View the zillow-pipeline-lambda-role role](#) on the IAM console.

Add trigger

Trigger configuration [Info](#)

 **S3**
aws asynchronous storage

Bucket

Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.

Bucket region: us-east-1

Event types

Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.

All object create events

Prefix - optional

Enter a single optional prefix to limit the notifications to objects with keys that start with matching characters.

Suffix - optional

Enter a single optional suffix to limit the notifications to objects with keys that end with matching characters.

Recursive invocation

If your function writes objects to an S3 bucket, ensure that you are using different S3 buckets for input and output. Writing to the same bucket increases the risk of creating a recursive invocation, which can result in increased Lambda usage and increased costs. [Learn more](#)


☒ I acknowledge that using the same S3 bucket for both input and output is not recommended and that this configuration can cause recursive invocations, increased Lambda usage, and increased costs.


Lambda will add the necessary permissions for AWS S3 to invoke your Lambda function from this trigger. [Learn more](#) about the Lambda permissions model.

[Lambda](#) > [Functions](#) > zillow-pipeline-lambdaFunction

zillow-pipeline-lambdaFunction

Function overview [Info](#)

 **S3**
[+ Add trigger](#)

 **zillow-pipeline-lambdaFunction**
[Layers](#) (0)
[+ Add destination](#)


Probamos este trigger (*guardar cambios con el botón Deploy*) añadiendo un nuevo documento a nuestro S3, con el fin de obtener un log dentro de CloudWatch que nos permita ver ese movimiento.

[CloudWatch](#) > [Log groups](#) > /aws/lambda/zillow-pipeline-lambdaFunction

/aws/lambda/zillow-pipeline-lambdaFunction

[Actions](#) [View in Logs Insights](#) [Start tailing](#) [Search log group](#)

Log group details

ARN
 arn:aws:logs:us-east-1:183036207239:log-group:/aws/lambda/zillow-pipeline-lambdaFunction:*

Creation time
16 minutes ago

Retention
Never expire

Stored bytes
-

Metric filters
0

Subscription filters
0

Contributor Insights rules
-

KMS key ID
-

Data protection
-

Sensitive data count
-

[Log streams](#) [Tags](#) [Metric filters](#) [Subscription filters](#) [Contributor Insights](#) [Data protection](#)

Log streams (1)

[Refresh](#) [Delete](#) [Create log stream](#) [Search all log streams](#)

☐ Exact match ☐ Show expired [Info](#)

< 1 > [Refresh](#)

<input type="checkbox"/>	Log stream	Last event time
<input type="checkbox"/>	2023/09/27/[\$LATEST]2d2e5c56cb534b988a1ff572b695f8ae	2023-09-27 17:44:56 (UTC-05:00)

CloudWatch > Log groups > /aws/lambda/zillow-pipeline-lambdaFunction > 2023/09/27/[\$LATEST]2d2e5c56cb534b988a1ff572b695f8ae

Log events
You can use the filter bar below to search for and match terms, phrases, or values in your log events. [Learn more about filter patterns](#)

Filter events

Clear 1m 30m 1h 12h Custom Local Display

Timestamp	Message
No older events at this moment. Retry	
2023-09-27T17:44:55.673-05:00	INIT_START Runtime Version: python:3.10.v13 Runtime Version ARN: arn:aws:lambda:us-east-1::runtime:2d6bcf31c5bdccdc494121a188671b06f079f39a348ad2a53d8864f98f94...
2023-09-27T17:44:56.106-05:00	START RequestId: d57f6b66-5090-43d7-af94-fe6049f4b67e Version: \$LATEST
2023-09-27T17:44:56.107-05:00	zillow-pipeline-bucket1
2023-09-27T17:44:56.107-05:00	zillow-pipeline-user1_accessKeys.csv
2023-09-27T17:44:56.108-05:00	END RequestId: d57f6b66-5090-43d7-af94-fe6049f4b67e
2023-09-27T17:44:56.108-05:00	REPORT RequestId: d57f6b66-5090-43d7-af94-fe6049f4b67e Duration: 1.58 ms Billed Duration: 2 ms Memory Size: 128 MB Max Memory Used: 73 MB Init Duration: 432.07...
No newer events at this moment. Auto retry paused. Resume	

Iniciamos nuevamente el DAG y probamos nuestro trigger de Lambda, en este caso el archivo guardado en el bucket #1 (Landing Zone) debe ser enviado al bucket #2.

Code Test Monitor Configuration Aliases Versions

Code source Info

File Edit Find View Go Tools Window Test Deploy

Go to Anything (Ctrl-P)

Environment

zillow-pipeline-lamb
lambda_function.py

```

1 import boto3
2 import json
3
4 s3_client = boto3.client('s3')
5
6 def lambda_handler(event, context):
7     # TODO implement
8     source_bucket = event['Records'][0]['s3']['bucket']['name']
9     object_key = event['Records'][0]['s3']['object']['key']
10
11     target_bucket = 'copy-of-raw-json-bucket-zillow-1'
12     copy_source = {'Bucket': source_bucket, 'Key': object_key}
13
14     waiter = s3_client.get_waiter('object_exists')
15     waiter.wait(Bucket=source_bucket, Key=object_key)
16     s3_client.copy_object(Bucket=target_bucket, Key=object_key, CopySource=copy_source)
17
18     return {
19         'statusCode': 200,
20         'body': json.dumps('Copy completed successfully')
21     }

```

El archivo se carga correctamente en el bucket #2 correspondiente a Intermediate Zone.

Amazon S3 > Buckets > copy-of-raw-json-bucket-zillow-1

copy-of-raw-json-bucket-zillow-1 Info

Objects Properties Permissions Metrics Management Access Points

Objects (1)
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
response_data_27092023233026.json	json	September 27, 2023, 18:30:36 (UTC-05:00)	56.9 KB	Standard

Creamos una nueva función Lambda para transformar los datos.

Create function [Info](#)

AWS Serverless Application Repository applications have moved to [Create application](#).

☒ **Author from scratch**
Start with a simple Hello World example.

☐ **Use a blueprint**
Build a Lambda application from sample code and configuration presets for common use cases.

☐ **Container image**
Select a container image to deploy for your function.

Basic information

Function name
Enter a name that describes the purpose of your function.
zillow-pipeline-transformation-lambdafunction

Runtime [Info](#)
Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.
Python 3.10

Architecture [Info](#)
Choose the instruction set architecture you want for your function code.
☒ x86_64
☐ arm64

Permissions [Info](#)
By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this default role later when adding triggers.

▼ Change default execution role

Execution role
Choose a role that defines the permissions of your function. To create a custom role, go to the [IAM console](#).

☐ Create a new role with basic Lambda permissions

☒ Use an existing role

☐ Create a new role from AWS policy templates

Existing role
Choose an existing role that you've created to be used with this Lambda function. The role must have permission to upload logs to Amazon CloudWatch Logs.
zillow-pipeline-lambda-role

[View the zillow-pipeline-lambda-role role](#) on the IAM console.

Añadimos un trigger para acceder a nuestro bucket #2.

Add trigger

Trigger configuration [Info](#)

S3
aws asynchronous storage

Bucket
Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.
s3/copy-of-raw-json-bucket-zillow-1

Bucket region: us-east-1

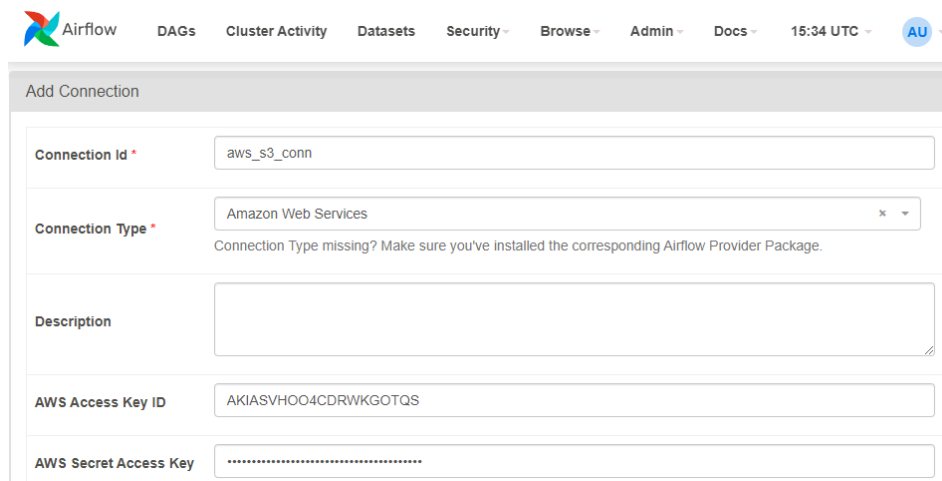
Event types
Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.

All object create events

Escribimos nuestro código para tomar los registros en formato json, seleccionar algunas columnas y subir el archivo csv al bucket #3 creado. A su vez, creamos un sensor de S3 para corroborar que nuestro archivo quedó cargado correctamente.

```
is_file_in_s3_available = S3KeySensor(  
    task_id = 'tsk_is_file_in_s3_available',  
    bucket_key = '{{ti.xcom_pull("tsk_extract_zillow_data_var")[1]}}',  
    bucket_name = s3_bucket,  
    aws_conn_id = 'aws_s3_conn',  
    wildcard_match = False, # Set this to True if you want to use wildcards in the  
    timeout = 60, # Optional: Timeout for the sensor (in seconds)  
    poke_interval = 5, # Optional: Time interval between S3 checks (in seconds)  
)
```

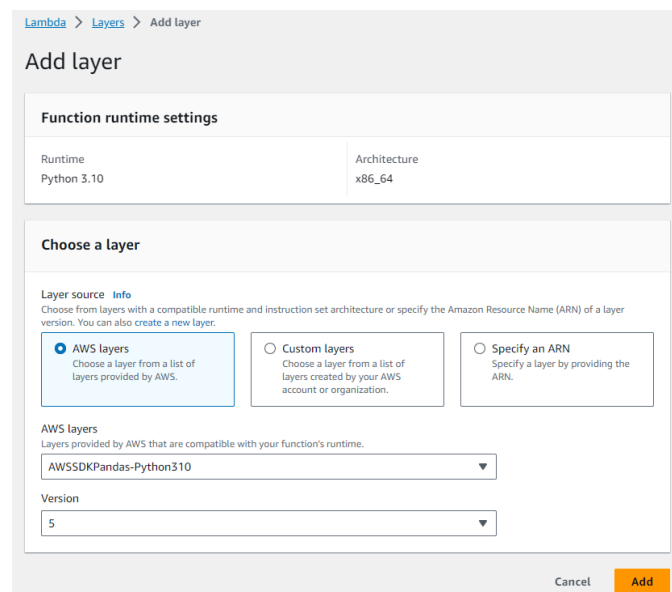
Creamos la conexión al servicio S3 en Airflow. Utilizamos la key y secret del user #1 que creamos anteriormente.



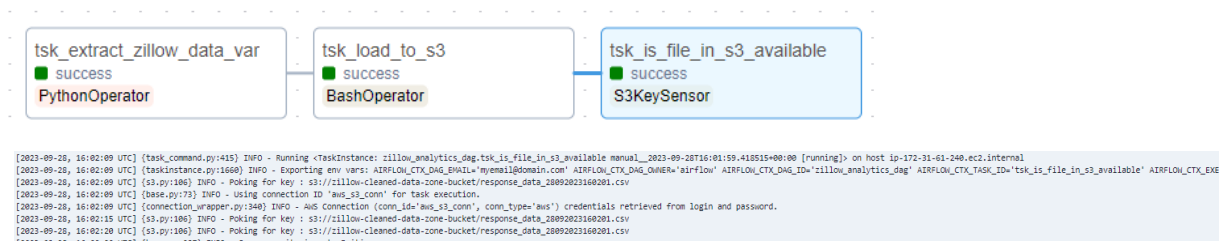
Ingresamos a las funciones Lambda y editamos el timeout hasta 5 min. En la sección Configuration.



Ahora, en nuestra última función Lambda debemos crear un Layer para poder usar la dependencia Pandas, si no arrojará error.



Corremos el DAG y se ejecuta correctamente.



El archivo transformado en formato .csv queda almacenado en el bucket #3.

zillow-cleaned-data-zone-bucket

Info

Objects

Properties

Permissions

Metrics

Management

Access Points

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Refresh

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

Name

Type

Last modified

Size

response_data_28092023160201.csv

csv

September 28, 2023, 11:02:18 (UTC-05:00)

	A	B	C	D	E	F	G	H	I	
1	bathrooms	bedrooms	city	homeStatus	homeType	livingArea	price	rentZestimate	zipcode	
2	3	3	Houston	FOR_SALE	TOWNHOUSE	2073	358000	2907	77004	
3	3	4	Houston	FOR_SALE	SINGLE_FAMILY	2402	319000	2500	77089	
4	2	3	Houston	FOR_SALE	SINGLE_FAMILY	1608	277000	1794	77064	
5	3	4	Houston	FOR_SALE	SINGLE_FAMILY	2172	270000	2328	77082	
6	3	3	Houston	FOR_SALE	TOWNHOUSE	1756	330000	2564	77054	
7	2	4	Houston	FOR_SALE	SINGLE_FAMILY	2174	450000	2470	77080	
8	0		Houston	FOR_SALE	LOT		450000	1114	77025	
9	4	3	Houston	FOR_SALE	TOWNHOUSE	3309	945000	5209	77030	
10	3	4	Houston	FOR_SALE	SINGLE_FAMILY	2605	429000	2600	77077	
11	7	6	Houston	FOR_SALE	SINGLE_FAMILY	6638	1799000	10775	77024	
12	4	3	Houston	FOR_SALE	SINGLE_FAMILY	2674	925000	4911	77008	

Como siguiente paso, creamos nuestro cluster dentro de Amazon Redshift usando la siguiente configuraci3n. **Se debe tener cuidado con el cobro por usar este servicio.**

Amazon Redshift

>

Clusters

>

Create cluster

Create cluster

Info

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.xlplus

Number of nodes

Enter the number of nodes that you need.

1

Range (1-32)

Configuration summary

Info

ra3.xlplus | 1 node

\$792.78/month

Estimated on-demand compute price

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

[Learn more about pricing](#)

4 TB

Max compressed storage

RA3 stores data in Redshift managed storage. Each RA3.xlplus node gets up to 32 TB of compressed data capacity in managed storage to ensure optimal query performance.

\$0.024/GB/month

Estimated storage price

Pay only for the amount of data you store in managed storage when running an RA3 cluster.

Database configurations

Admin user name
Enter a login ID for the admin user of your DB instance.

The name must be 1-128 alphanumeric characters, and it can't be a [reserved word](#).

☐ **Auto generate password**
Amazon Redshift can generate a password for you, or you can specify your own password.

Admin user password

Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except `"/", "'", "`, or `"@"`.

☐ **Show password**

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.

Nuestro cluster se crea correctamente.

Amazon Redshift

>

Clusters

>

redshift-cluster-1

redshift-cluster-1

Actions

Edit

Add partner integration

Query data

General information

Cluster identifier	redshift-cluster-1	Status	Available	Node type	ra3.xplus	Endpoint	redshift-cluster-1.cxvlgag1ej3.us-east-1.redshift.amazonaws.com:5439/dev
Custom domain name	- new	Date created	September 28, 2023, 12:04 (UTC-05:00)	Number of nodes	1	JDBC URL	jdbc:redshift://redshift-cluster-1.cxvlgag1ej3.us-east-1.redshift.amazonaws.com:5439/dev
Cluster ARN	arn:aws:redshift:us-east-1:183036207239:namespace:Sacb25f7-b3e2-45eb-abc6-486fac684b6a	Storage used	0.00% (0.00 of 4 TB used)			ODBC URL	Driver={Amazon Redshift (x64)}; Server=redshift-cluster-1.cxvlgag1ej3.us-east-1.redshift.amazonaws.com; Database=dev
Cluster configuration	Production	Multi-AZ	No				

Dentro de las herramientas de Amazon Redshift tenemos el *Query editor v2*. Lo seleccionamos e ingresamos al cluster con nuestras credenciales.

Redshift query editor v2

+

Untitled 1

x

Editor

Create

Load data

Filter resources

redshift-cluster-1

Connect to redshift-cluster-1

x

Authentication

Federated user

Temporary credentials

Database user name and password

AWS Secrets Manager

Database

dev

User name

awsuserzillow

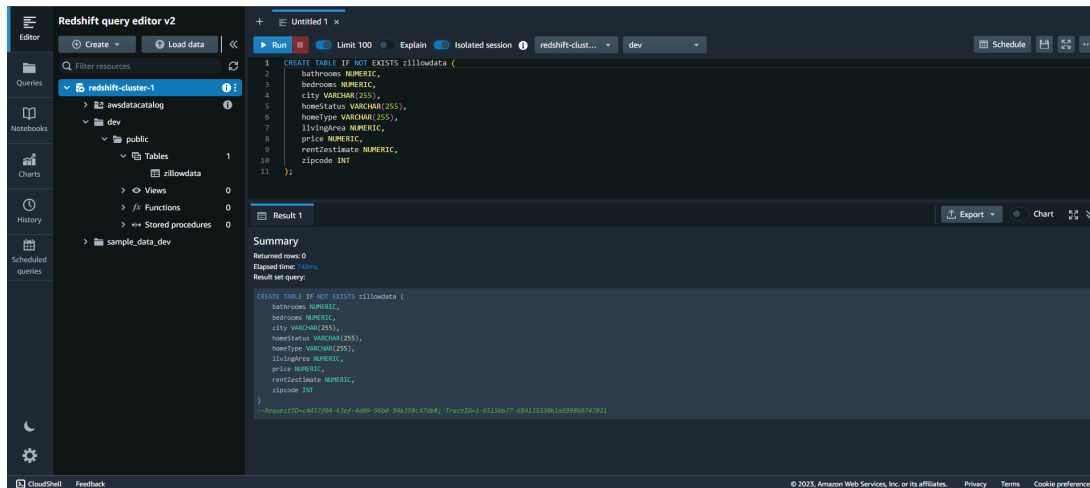
Password

☐ Show password

Cancel

Create connection

Se crea la tabla correspondiente, en la base de datos Dev en Redshift.



Luego, creamos la tarea (task) en nuestro DAG para poder tomar el archivo de nuestro bucket #3 y cargar esa información en la tabla de Redshift. Dentro de esta tarea debe haber una conexión establecida en Airflow. Usamos el endpoint del cluster y las credenciales.

Add Connection

Connection Id *

Connection Type *

Description

Host

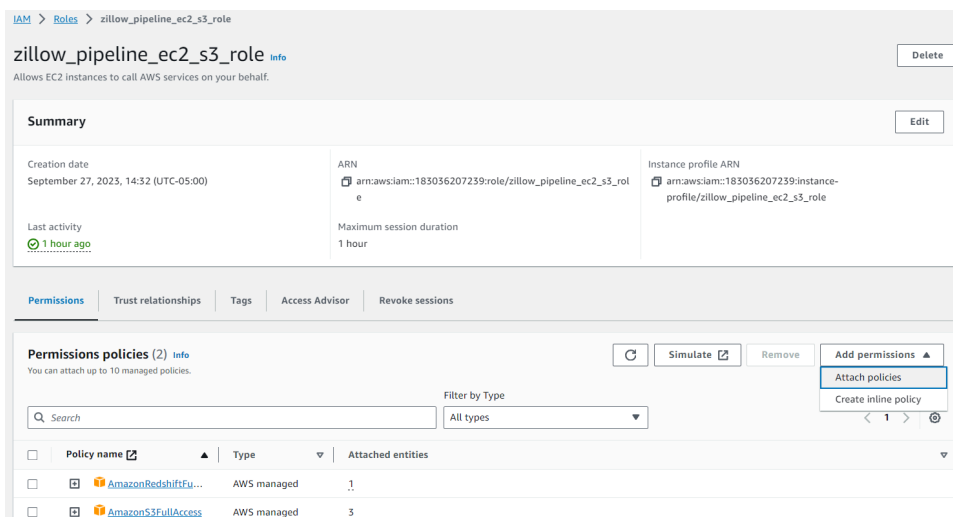
Database

User

Password

Port

Debemos añadir el permiso de acceso a Redshift en nuestro rol de EC2.



Para lograr una conexión exitosa con el servicio de Redshift debemos crear una *inbound rule* permitiendo el tráfico desde cualquier punto. Podemos limitar el acceso al permitir solo nuestra IP, pero en este caso no lo haremos así.

VPC security group

Specify which instances and devices can connect to the cluster.

[sg-01beca6514d318e67](#)

All traffic

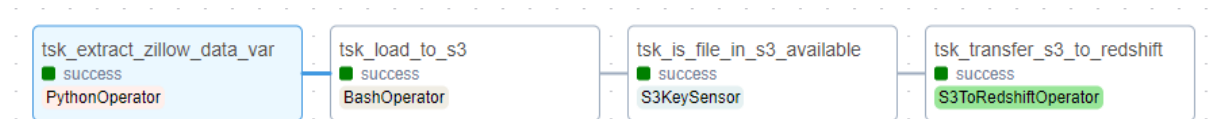
All

All

Custom

0.0.0.0/0

Corremos nuestro DAG y vemos que se ejecuta correctamente.



Consultamos la tabla en Redshift y vemos que ya se han cargado los datos del archivo .csv.

13 SELECT * FROM zillowdata;								
Result 1 (41)								
	bathrooms	bedrooms	city	homestatus	hometype	livingarea	price	re
<input type="checkbox"/>	0	NULL	Houston	FOR_SALE	LOT	NULL	148998	1
<input type="checkbox"/>	4	5	Houston	FOR_SALE	SINGLE_FAMILY	3360	875000	5
<input type="checkbox"/>	2	4	Houston	FOR_SALE	SINGLE_FAMILY	1955	214900	1
<input type="checkbox"/>	4	4	Houston	FOR_SALE	SINGLE_FAMILY	3264	1025000	6
<input type="checkbox"/>	2	3	Houston	FOR_SALE	SINGLE_FAMILY	1515	261990	1
<input type="checkbox"/>	4	4	Houston	FOR_SALE	SINGLE_FAMILY	3489	559900	3
<input type="checkbox"/>	0	NULL	Houston	FOR_SALE	LOT	NULL	95000	1
<input type="checkbox"/>	2	3	Houston	FOR_SALE	SINGLE_FAMILY	1388	190000	1
<input type="checkbox"/>	3	4	Houston	FOR_SALE	SINGLE_FAMILY	2255	500000	2
<input type="checkbox"/>	2	4	Houston	FOR_SALE	SINGLE_FAMILY	1860	310000	2
<input type="checkbox"/>	2	3	Houston	FOR_SALE	SINGLE_FAMILY	1669	263000	1
<input type="checkbox"/>	2	2	Houston	FOR_SALE	SINGLE_FAMILY	1074	243000	1
<input type="checkbox"/>	3	3	Houston	FOR_SALE	TOWNHOUSE	2226	360000	3
<input type="checkbox"/>	2	3	Houston	FOR_SALE	SINGLE_FAMILY	1466	242000	1

En caso de correr nuevamente nuestro DAG, se cargaran los datos nuevamente a nuestra tabla zillowdata, sin importar que estén duplicados.

Finalmente utilizamos el servicio QuickSight para visualizar nuestros datos. Elegimos la siguiente opción al momento de abrir el servicio.

Sign up for Standard Edition [here](#).

Standard

[Back](#)

Authentication method

- ☒ Use IAM federated identities & QuickSight-managed users
Authenticate with single sign-on (SAML or OpenID Connect), AWS IAM credentials, or QuickSight credentials
- ☐ Use IAM federated identities only
Authenticate with single sign-on (SAML or OpenID Connect) or AWS IAM credentials

QuickSight region

Select a region

US East (N. Virginia) ▼

Account info

QuickSight account name

You will need this for you and others to sign in

zillow-quicksight-1

Notification email address

For QuickSight to send important notifications

jhonbuesaquillo1403@gmail.com

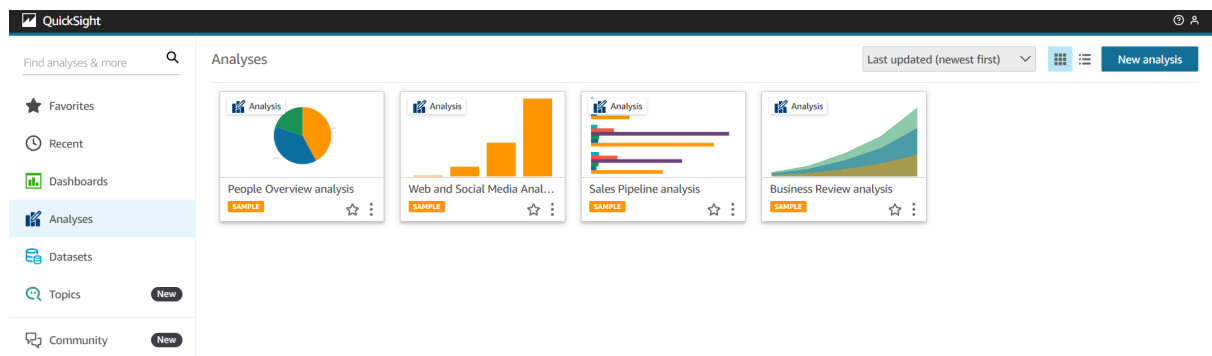
QuickSight access to AWS services

Make your existing AWS data and users available in QuickSight. [Learn more](#)

Allow access and autodiscovery for these resources

- ☒ Amazon Redshift
- ☒ Amazon RDS
- ☒ IAM
- ☐ Amazon S3
[Select S3 buckets](#)

Creamos nuestra cuenta en QuickSight y nos ofrece la siguiente interfaz.



Vamos a la opción Datasets y creamos uno nuevo. Nuestra source es Redshift (Auto-discovered).

New Redshift data source

Data source name

zillowdataset

Instance ID

redshift-cluster-zillow1 ▼

Connection type

Choose a VPC connection ▼

Database name

dev

Username

awsuserzillow

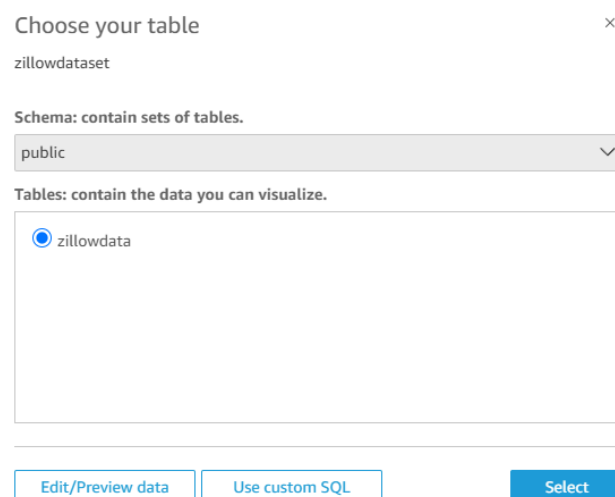
Password

Validate connection

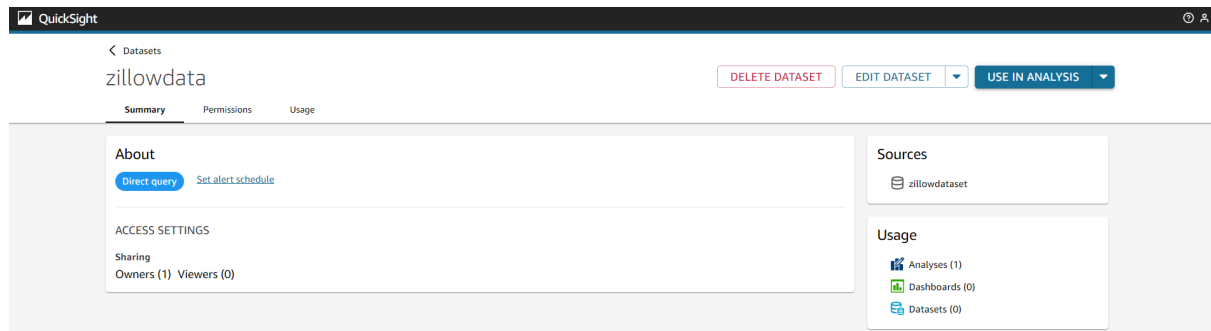
SSL is enabled

Create data source

Seleccionamos nuestra tabla y nos vamos a Edit/Preview data.



Dentro de los datasets del servicio ya podemos ver el que hemos creado anteriormente. Ingresamos a *Use in Analysis*.



Dentro de esta herramienta podemos explorar las diferentes opciones que nos ofrece. Para este caso solo planteamos algunas gráficas sencillas para validar los datos de nuestra tabla zillowdata.

