

S3, Glue, Athena

Introducción

Este documento servirá como representación de las características generales y consideraciones de la api de nóminas en desarrollo apartir de los conocimientos y aplicación práctica de bases de datos NoSQL, visualizadores de datos y procesamiento de datos.

En la actualidad la cantidad de información en el mundo crece con gran velocidad, es por esto que su almacenamiento se ha visto limitado en cuanto a características como el volumen, velocidad y escalabilidad de la información, siendo esto el motivo principal para problemas relevantes como la pérdida de información y sobrecostos.

Teniendo en cuenta lo anterior, se puede afirmar que la información tiene un enorme poder dentro de muchas áreas, específicamente en la toma de decisiones en cualquier organización, puesto que proporciona una ventaja competitiva o una oportunidad de negocio.

Una posible solución al problema de almacenamiento para grandes volúmenes de datos y su consulta de manera rápida y eficiente es la incorporación de nuevos enfoques tecnológicos en cuanto al almacenamiento, tal enfoque es conocido como bases de datos NoSQL.

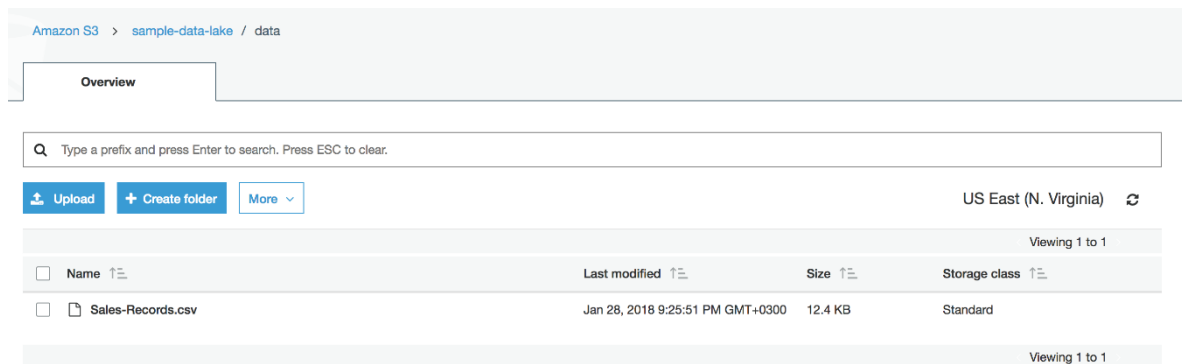
AWS

1. Crear una cuenta de AWS

Se debe tener una cuenta de AWS activa, es fácil de crear.

2. Cargar un archivo csv al S3

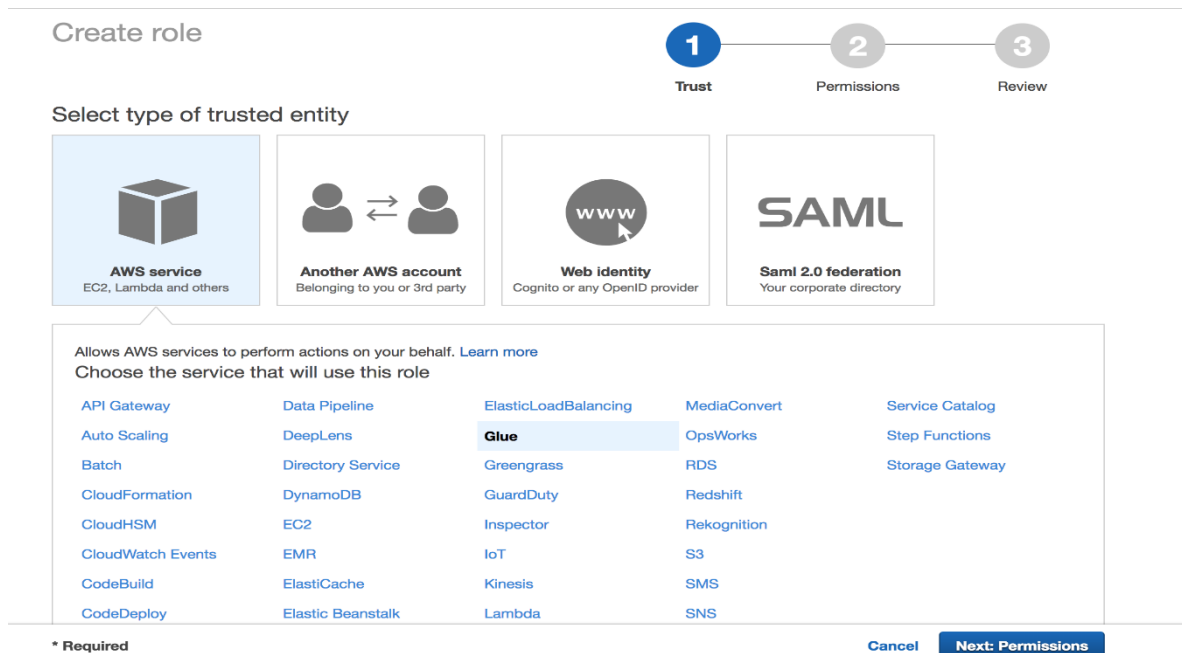
Encontré un csv de registros de ventas, se puede descargar y usar. ([100 registros de ventas](#)). Sube el archivo al S3 y crea un bucket. Presione el botón "Siguierte" para todos los pasos.



3. Crea un rol

Ingresa a <https://console.aws.amazon.com/iam/>, click en "Roles" y luego en el botón azul "Crear rol".

Elige "Glue" y luego.



Se debe dar permisos a Glue y S3.

Create role



Attach permissions policies

Choose one or more policies to attach to your new role.

Create policy Refresh

Filter: Policy type		Q glue	Showing 6 results	
	Policy name	Attachments	Description	
<input type="checkbox"/>	AWSGlueConsoleFullAccess	0	Provides full access to AWS Glue via the AWS Management ...	
<input type="checkbox"/>	AWSGlueServiceNotebookRole	0	Policy for AWS Glue service role which allows customer to m...	
<input checked="" type="checkbox"/>	AWSGlueServiceRole	4	Policy for AWS Glue service role which allows access to relat...	
<input type="checkbox"/>	AWSGlueServiceRole-sample	1	This policy will be used for Glue Crawler and Job execution. ...	
<input type="checkbox"/>	AWSGlueServiceRole-sample2	1	This policy will be used for Glue Crawler and Job execution. ...	
<input type="checkbox"/>	AWSGlueServiceRole-sample3	1	This policy will be used for Glue Crawler and Job execution. ...	

Create role



Attach permissions policies

Choose one or more policies to attach to your new role.

Create policy Refresh

Filter: Policy type		Q s3	Showing 4 results	
	Policy name	Attachments	Description	
<input type="checkbox"/>	AmazonDMSRedshiftS3Role	0	Provides access to manage S3 settings for Redshift endpoint...	
<input checked="" type="checkbox"/>	AmazonS3FullAccess	1	Provides full access to all buckets via the AWS Management...	
<input type="checkbox"/>	AmazonS3ReadOnlyAccess	2	Provides read only access to all buckets via the AWS Manag...	
<input type="checkbox"/>	QuickSightAccessForS3StorageManagement...	0	Policy used by QuickSight team to access customer data pr...	

Create role



Attach permissions policies

Choose one or more policies to attach to your new role.

Create policy Refresh

Filter: Policy type Q glue Showing 6 results			
	Policy name	Attachments	Description
<input type="checkbox"/>	AWSGlueConsoleFullAccess	0	Provides full access to AWS Glue via the AWS Management ...
<input type="checkbox"/>	AWSGlueServiceNotebookRole	0	Policy for AWS Glue service role which allows customer to m...
<input checked="" type="checkbox"/>	AWSGlueServiceRole	4	Policy for AWS Glue service role which allows access to relat...
<input type="checkbox"/>	AWSGlueServiceRole-sample	1	This policy will be used for Glue Crawler and Job execution. ...
<input type="checkbox"/>	AWSGlueServiceRole-sample2	1	This policy will be used for Glue Crawler and Job execution. ...
<input type="checkbox"/>	AWSGlueServiceRole-sample3	1	This policy will be used for Glue Crawler and Job execution. ...

Create role



Attach permissions policies

Choose one or more policies to attach to your new role.

Create policy Refresh

Filter: Policy type Q s3 Showing 4 results			
	Policy name	Attachments	Description
<input type="checkbox"/>	AmazonDMSRedshiftS3Role	0	Provides access to manage S3 settings for Redshift endpoint...
<input checked="" type="checkbox"/>	AmazonS3FullAccess	1	Provides full access to all buckets via the AWS Management...
<input type="checkbox"/>	AmazonS3ReadOnlyAccess	2	Provides read only access to all buckets via the AWS Manag...
<input type="checkbox"/>	QuickSightAccessForS3StorageManagement...	0	Policy used by QuickSight team to access customer data pr...

Ingrese un nombre de rol (AWSGlueServiceRoleDefault) y haz click en el botón "Crear rol".

4. Funcionamiento de Glue

Agregar una nueva base de datos (sampledb)

Agregar una nueva tabla mediante la opción "Agregar tablas mediante un rastreador"

Ingrese el nombre del rastreador (rastreador de ventas) y haga clic en el botón Siguiente

Add crawler

Crawler info

Data store

IAM Role

Schedule

Output

Review all steps

Add information about your crawler

Crawler name

salescrawler

▶ Description and classifiers (optional)

Next

Ingresa la ruta del archivo csv que se almacenó en S3 y click en el botón Siguiente

Add crawler

✓ Crawler info

salescrawler

Data store

IAM Role

Schedule

Output

Review all steps

Add a data store

Data store

S3

Crawl data in

☒ Specified path in my account

☐ Specified path in another account

Include path

s3://sample-data-lake/data

All folders and files contained in the include path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

▶ Exclude patterns (optional)

Back

Next

No necesitamos agregar un nuevo almacén de datos, selecciona "No", click en Siguiente

Add crawler

✓ Crawler info

salescrawler

Data store

S3: s3://sample-dat...

IAM Role

Schedule

Output

Review all steps

Add another data store

☐ Yes

☒ No

Back

Next

Selecciona el rol que se creó antes, click en Siguiente.

Add crawler

✓ Crawler info

salescrawler

✓ Data store

S3: s3://sample-dat...

✓ IAM Role

○ Schedule

○ Output

○ Review all steps

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. [Learn more](#)

☐ Update a policy in an IAM role

☒ Choose an existing IAM role

☐ Create an IAM role

IAM role ⓘ

AWSGlueServiceRoleDefault

↕

This role must provide permissions similar to the AWS managed policy, **AWSGlueServiceRole**, plus access to your Amazon S3 data stores.

- s3://sample-data-lake/data

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

Back

Next

Ejecutar bajo demanda, Siguiente.

Add crawler

✓ Crawler info

salescrawler

✓ Data store

S3: s3://sample-dat...

✓ IAM Role

arn:aws:iam::962386333785:role/AWSGlueServiceRoleDefault

○ Schedule

○ Output

○ Review all steps

Create a schedule for this crawler

Frequency

Run on demand

▼

Back

Next

Selecciona "sampledb" y luego click en Siguiente

Add crawler

✓ Crawler info

salescrawler

✓ Data store

S3: s3://sample-dat...

✓ IAM Role

arn:aws:iam::962386333785:role/AWSGlueServiceRoleDefault

✓ Schedule

Run on demand

○ Output

○ Review all steps

Configure the crawler's output

Database ⓘ

sampledb

▼

Add database

Prefix added to tables (optional) ⓘ

Type a prefix added to table names

► Configuration options (optional)

Back

Next

Add crawler

- ✓ **Crawler info**
salescrawler
- ✓ **Data store**
S3: s3://sample-dat...
- ✓ **IAM Role**
arn:aws:iam::962386333785:role/AWSGlueServiceRoleDefault
- ✓ **Schedule**
Run on demand
- ✓ **Output**
sampledb
- **Review all steps**

Data stores

Data store S3
Include path s3://sample-data-lake/data
Exclude patterns

IAM role

IAM role arn:aws:iam::962386333785:role/AWSGlueServiceRoleDefault

Schedule

Schedule Run on demand

Output

Database sampledb
Prefix added to tables (optional)
► Schema change policy

[Back](#) [Finish](#)

Se creó el crawler. Posteriormente se debe dar click en “Ejecutar ahora”. Al ejecutar el crawler, incorporaremos la información de Sales-Records.csv a Glue.

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

ETL

Jobs

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.

Crawler **salescrawler** was created to run on demand. [Run it now?](#)

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

Action ▾

<input type="checkbox"/>	Name	Schedule	Status	Logs
<input type="checkbox"/>	salescrawler		Ready	

Una vez completado el crawler, en la columna "Tablas agregadas". Sales-Records.csv ya está en la plataforma Glue.

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

ETL

Jobs

Crawlers

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

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

Action ▾

Showing: 1 - 1 < > ↺

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	salescrawler		Stopping		17 secs	17 secs	0	1

Puedes ver la tabla de "datos" en la sección Tablas.

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

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.

Add tables

Action

Filter by attributes or search by keyword

Save view

Showing: 1 - 1

<input type="checkbox"/>	Name	Database	Location	Classification	Last updated	De
<input type="checkbox"/>	data	sampledb	s3://sample-data-lake/data/	csv	30 January 2018 11:2...	

5. Athena

Podemos consultar este csv usando Athena.

Con este ejemplo se listan 10 registros usando la opción "Vista previa de la tabla".

Athena

Query Editor

Saved Queries

History

AWS Glue Data Catalog

Settings

Tutorial

Help

What's new

DATABASE

sampledb

TABLES

Filter Tables...

data

Create table

1 SELECT * FROM "sampledb"."data" limit 10;

Preview table

Show properties

Delete table

Generate Create Table DDL

Run Query

Save As

Format query

New Query

(Run time: 1.39 seconds, Data scanned: 12.45KB)

Results

	region	country	item type	sales channel	order priority	order date	order id	ship date	units sold	unit p
1	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669185933	6/27/2010	9925	255.21
2	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.7
3	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.2
4	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33
5	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.2

Conclusión

Podemos consultar un archivo csv usando S3, Glue y Athena. Debido a la arquitectura escalable de AWS, los costos de consulta son muy bajos.

Como cualquier desarrollo, este tipo de arquitectura y diseño tiene sus beneficios e inconvenientes, pero podría destacar:

1. No es necesario realizar mantenimiento de los servidores donde se tienen instalados programas y aplicaciones. El código se ejecuta en un contenedor temporal por lo que ya no es necesario instalar software, gestionar puertos de acceso o estar pendiente de las actualizaciones.
2. Se puede realizar un escalamiento de manera horizontal tanto como se requiera. Es posible añadir todos los clusters, balanceo de cargas etc, conforme se necesite.
3. Relacionado a los costos, solamente se va a pagar por el tiempo que se estén utilizando los procesos.
4. Las funciones que se utilicen, es posible integrarlas con el resto de servicios que ofrece la plataforma, como son logging, virtualización o endpoints.