

**UNIVERSIDAD PRIVADA DE TACNA**



**FACULTAD DE INGENIERIA**

**Escuela Profesional de Ingeniería de Sistema**

**Informe de laboratorio 09: Construyendo un  
ETL Serverless**

**Curso: Inteligencia de negocios**

**DOCENTE: Ing. Patrick Cuadros Quiroga**

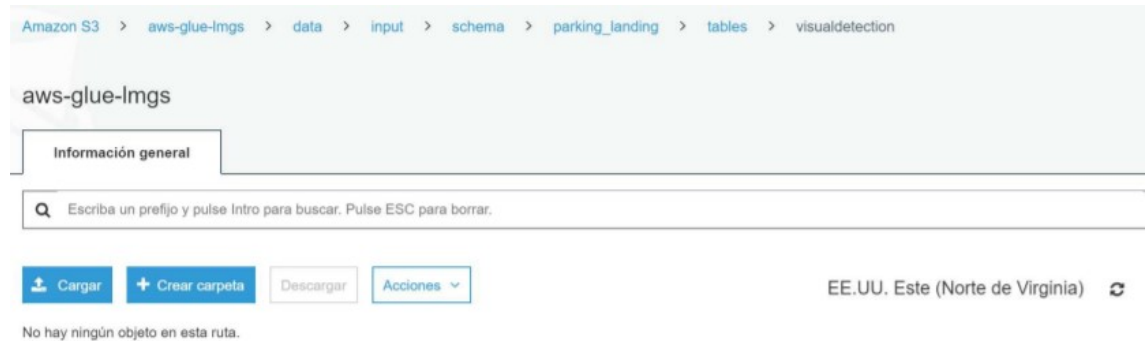
**Alumno: Balcon Coahila, Edwart Juan  
(2013046516)**

**Tacna – Perú**

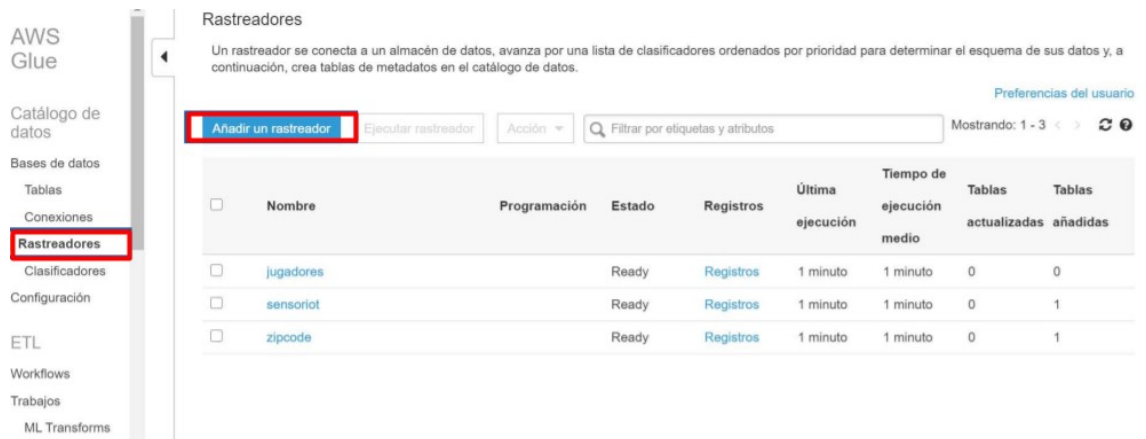
**2021**

# 1. Realizar los siguientes pasos para el laboratorio

## 1.1. ) Entramos a la consola de AWS y accedemos al servicio de S3



## 1.2. Entrar al servicio de Glue



## 1.3. Agregamos el nombre visualdetection landing, clic en Siguiente.

## Añadir un rastreador

☒ Información del rastreador

☐ Crawler source type

☐ Almacén de datos

☐ Rol de IAM

☐ Programación

☐ Salida

☐ Revisar todos los pasos

### Añadir información acerca de su rastreador

Nombre del rastreador

visualdetection\_landing

► Etiquetas, descripción, configuración de seguridad, y clasificadores (opcional)

Siguiente



1.4. Seleccionamos la opción : Data Stores (ya que a partir de un csv se creará una tabla)

ador

### Specify crawler source type


Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores.

**Crawler source type**

☒ Data stores

☐ Existing catalog tables

Atrás **Siguiente**



1.5. Elegimos S3, y seleccionamos la carpeta donde se encuentra el archivo csv, clic en Siguiente.

### Añadir un rastreador

- ✓ Información del rastreador
  - visualdetection\_landing
- ✓ Crawler source type
- Data stores
- Almacén de datos
- Rol de IAM
- Programación
- Salida
- Revisar todos los pasos

### Añada un almacén de datos

**Elija un almacén de datos**

S3

**Conexión**

Seleccione una conexión

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any future S3 targets will also use the same connection (or none, if left blank).

[Añadir una conexión](#)


**Rastree datos en**

☒ Ruta especificada en mi cuenta

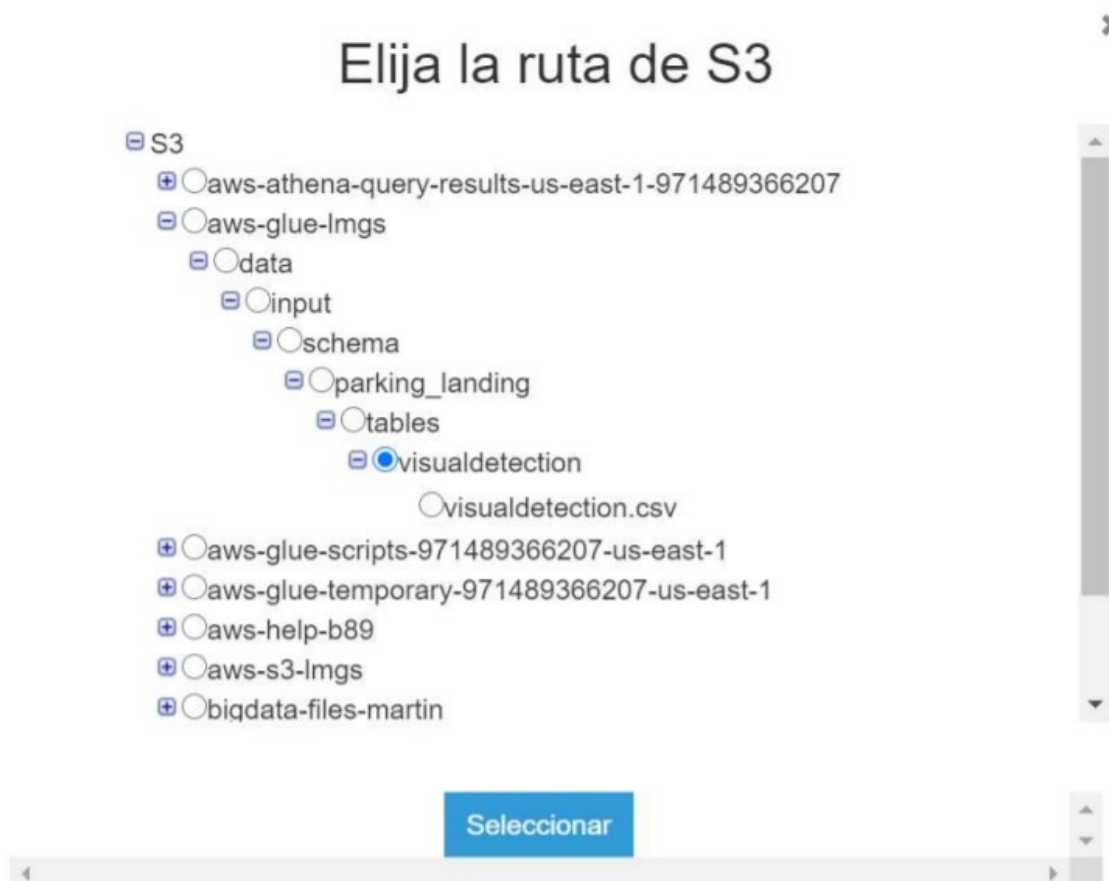
☐ Ruta especificada de otra cuenta

**Ruta de inclusión**

s3://bucket/prefix/object



1.6. Seleccionamos hasta la carpeta visualdeteccion en S3.



1.7. En este caso, solo añadiremos un almacén de datos, elegimos No y clic en Siguiente



The screenshot shows a web interface with a dark header containing the text 'dor'. Below the header, the main content area has a light gray background. At the top right of this area, the title 'Añadir otro almacén de datos' is displayed. Below the title, there are two radio button options: 'Sí' (unselected) and 'No' (selected). At the bottom right, there are two buttons: 'Atrás' (outlined) and 'Siguiente' (solid blue). A blue arrow points from the bottom left towards the 'Siguiente' button.

1.8. Creamos un nuevo rol, que tendrá el permiso de leer el archivo csv de S3 y crear una tabla en Glue. Nombre del rol : AWSGlueServiceRole-Crawler Si ya existe, seleccionar el indicado o concatenarlo con sus iniciales de sus nombres luego se tiene la opción de definir una periodicidad distinta, de acuerdo a la carga de los datos, si por ejemplo sé que todos los días a cierta hora, voy a tener un dataset, podría indicar la hora de ejecución automática de este crawler.

ador

### Elija un rol de IAM

☐ Actualice una política en un rol de IAM  
☐ Elija un rol de IAM existente  
☒ Cree un rol de IAM

**Rol de IAM** ⓘ

AWSGlueServiceRole-

- s3://aws-glue-imgs/data/input/tables/jugadores





ador

Cree una programación para este rastreador

Frecuencia

Ejecutar bajo demanda



Atrás

Siguiente

1.9. Añadimos una nueva base de datos, le ponemos como nombre parking landing, clic en Siguiente. En esta base de datos, se generará de manera automática a partir del dataset que se encuentra en S3, los datos se leerán tal cual están en la fuente de donde se descargó todavía no se ha realizado ninguna transformación.

### Configure la salida del rastreador

**Base de datos ⓘ**

*Elija una base de datos para incluir tablas* ▼

[Añadir una base de datos](#)

**Prefijo añadido a las tablas (opcional) ⓘ**

*Escriba un prefijo añadido a los nombres de las tablas*

- ▶ Agrupación de comportamiento para datos de S3 (opcional)
- ▶ Las opciones de configuración (opcional)

[Atrás](#) [Siguiente](#)

1.10. Asignamos el siguiente nombre a la base de datos: Para no tener inconvenientes, cada uno le pondremos como nombre: parking landing

## Añadir una base de datos

Nombre de la base de datos

► Descripción y ubicación (opcional)

Crear

1.11. Le damos clic en Siguiente.

### Configure la salida del rastreador

**Base de datos ⓘ**

parking\_landing 

[Añadir una base de datos](#)

**Prefijo añadido a las tablas (opcional) ⓘ**

*Escriba un prefijo añadido a los nombres de las tablas*

- Agrupación de comportamiento para datos de S3 (opcional)
- Las opciones de configuración (opcional)

[Atrás](#) [Siguiente](#)

1.12. Clic en Finalizar



## Rastreadores

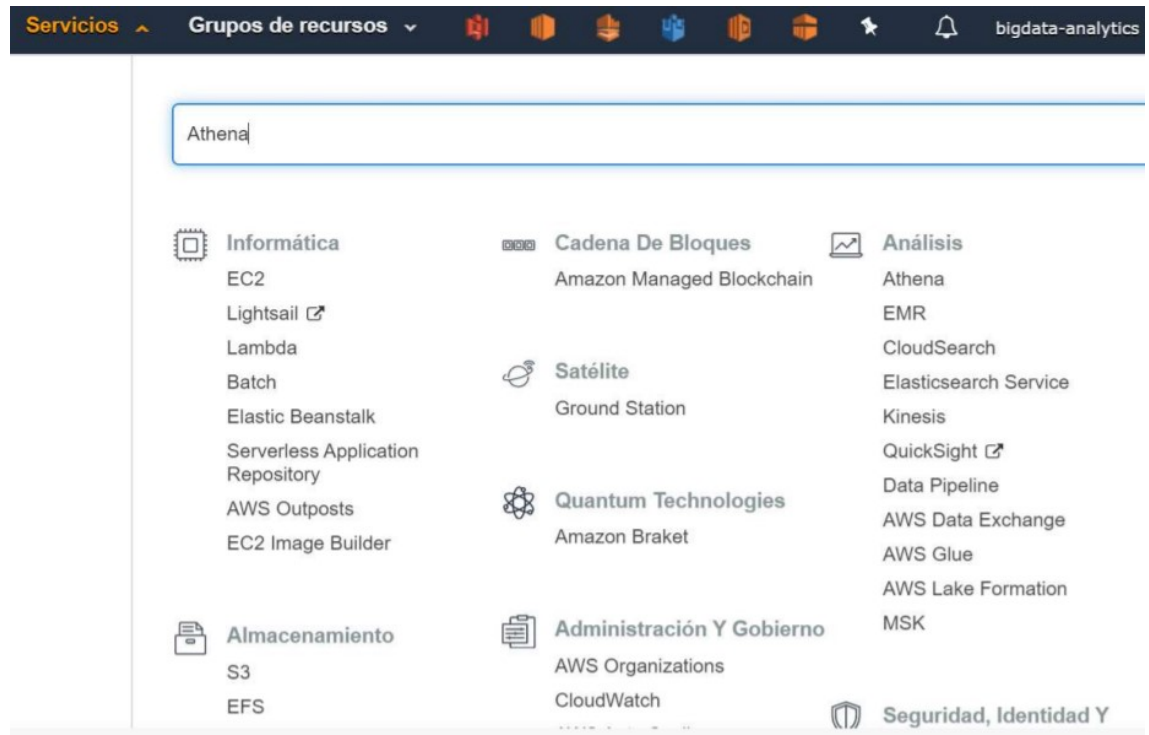
Un rastreador se conecta a un almacén de datos, avanza por una lista de clasificadores ordenados por prioridad para determinar el esquema de sus datos y, a continuación, crea tablas de metadatos en el catálogo de datos.

El rastreador "visualdetection\_landing" se ha completado y ha realizado los siguientes cambios: 1 tablas creadas, 0 tablas actualizadas. Consulte las tablas creadas en la base de datos [parking\\_landing](#).

[Preferencias del usuario](#)

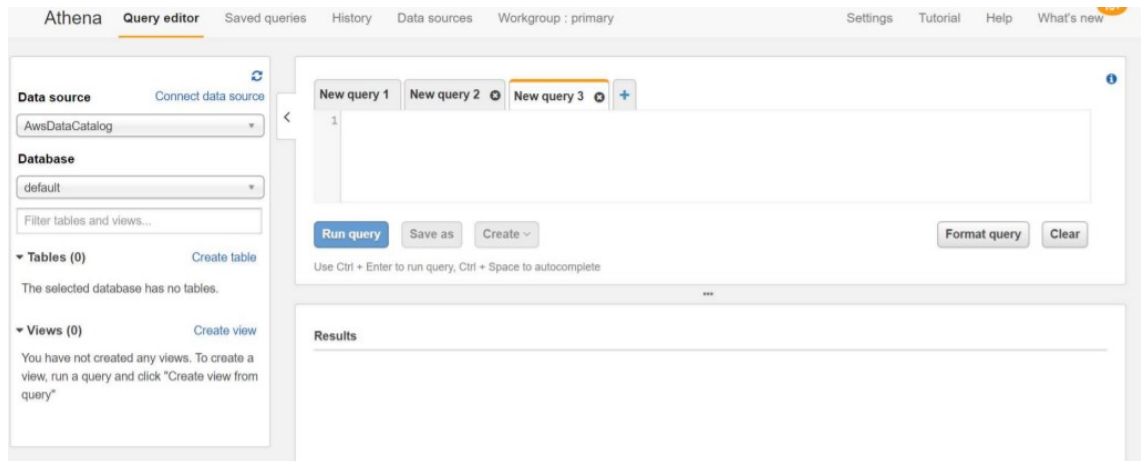
<input type="checkbox"/>	Nombre	Programación	Estado	Registros	Última ejecución	Tiempo de ejecución medio	Tablas actualizadas	Tablas añadidas
<input type="checkbox"/>	<a href="#">jugadores</a>		Ready	<a href="#">Registros</a>	53 segundos	53 segundos	0	1
<input type="checkbox"/>	<a href="#">visualdetection_landing</a>		Ready	<a href="#">Registros</a>	40 segundos	40 segundos	0	1

1.15. Ahora para poder realizar una consulta SQL en el servicio de Athena, entramos al servicio.



1.16. Antes de realizar una consulta SQL en Athena, debemos

realizar la siguiente configuración. Clic en Settings.



1.17. Indicamos el bucket que hemos creado, y le agregamos /data/output/, en esta ruta se guardarán 2 archivos (1 el resultado de los queries y el segundo con la metadata).

## Settings

Settings apply by default to all new queries. [Learn more](#)

Workgroup: **primary**

Query result location

s3://aws-glue-lmgs/data/output/

Example: s3://query-results-bucket/folder/

Encrypt query results



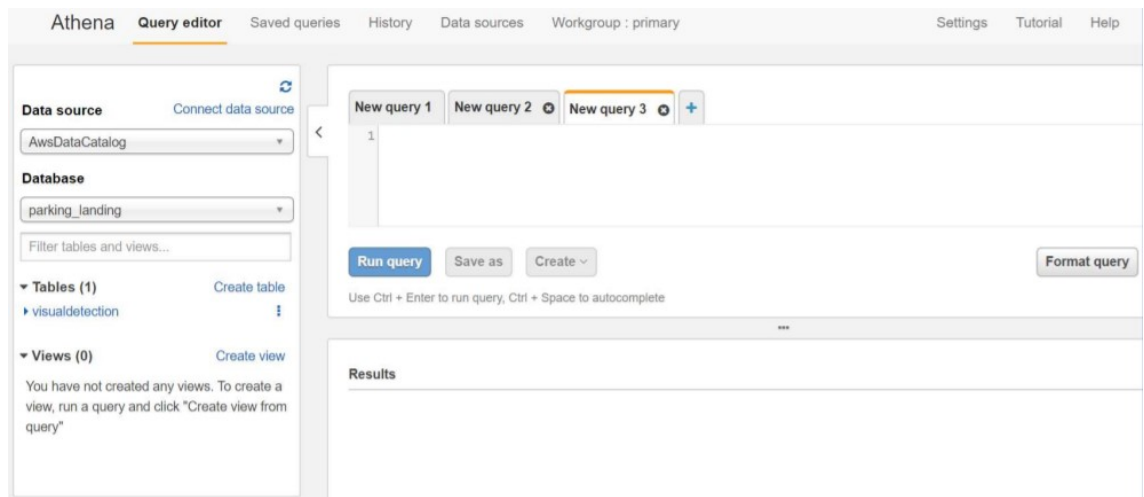
Autocomplete



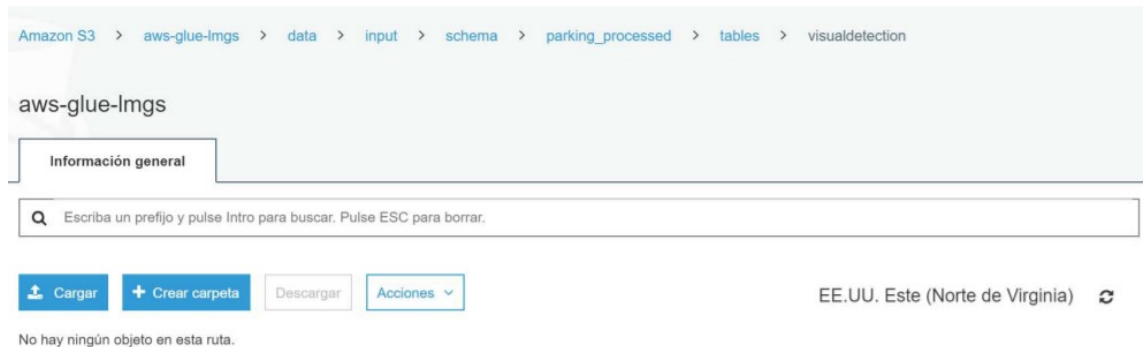
Cancel

Save

1.18. Ahora podremos ejecutar consultas SQL en Athena. Clic en Settings.

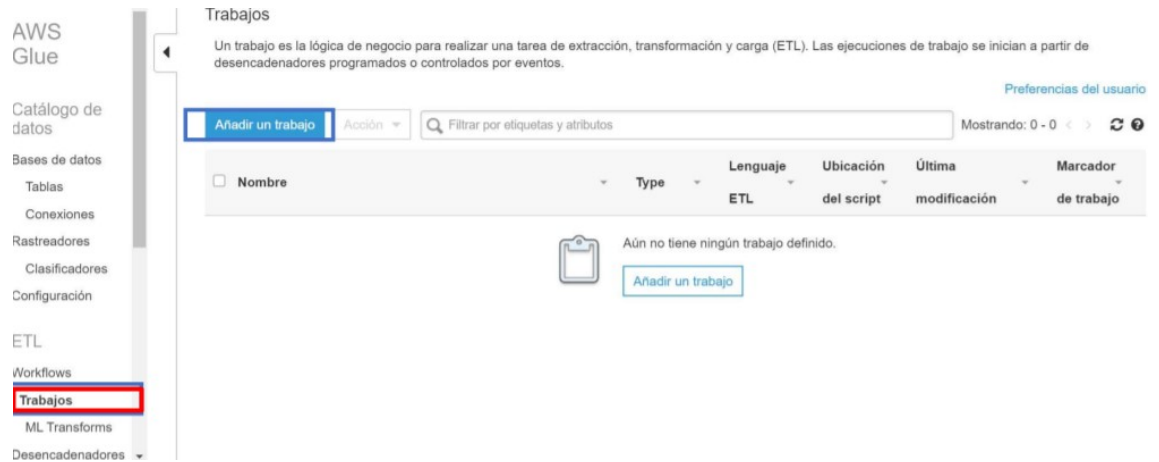


1.19. AEjecutamos algunos queries para probar Athena: **SELECT \* FROM parking\_landing.visual detection limit 10; SELECT \* FROM parking\_landing.visual detection where camera = 'A';** Clic en Settings.



1.20. Luego, nos vamos a Glue para crear el job de transformación. Clic en Settings.





1.21. Definir como nombre del job : visualdeteccion processed, clic en siguiente. Clic en Settings.

The screenshot shows the 'Añadir un trabajo' (Add Job) form in the AWS Glue console. The form is titled 'Propiedades del trabajo' and includes several fields: 'Nombre' (Name) with the value 'visualdeteccion\_processed', 'Rol de IAM' (IAM Role) with the value 'AWSGlueServiceRole-Crawler', 'Type' with the value 'Spark', and 'Glue version' with the value 'Spark 2.4, Python 3 with improved job startup times (Glue Version 2.0)'. There are also radio buttons for 'Este trabajo ejecuta' (This job runs) with options for 'Un script recomendado generado por AWS Glue' (Recommended script generated by AWS Glue) and 'Un script existente proporcionado por usted' (Existing script provided by you).

1.22. Seleccionamos la tabla visualdeteccion que está dentro de la base de datos que hemos creado cada uno (es la que se encuentra en formato csv), clic en Siguiente. Clic en Settings.

Añadir un trabajo

Propiedades del trabajo

visualdeteccion\_proces

sed

Origen de datos

Transform type

Destino de datos

Esquema

Elija los orígenes de datos

Filtrar por atributos o buscar por palabra clave

Mostrando: 1 - 2

Nombre	Base de datos	Ubicación	Clasificación
<input type="radio"/> csv	flight	s3://crawler-public-us-east-1/fligh...	csv
<input checked="" type="radio"/> visualdeteccion	parking_landing	s3://aws-glue-imgs/data/input/sch...	csv

Atrás

Siguiente

1.23. Elegimos cambiar esquema y siguiente. Clic en Settings.

Choose a transform type

☒ Change schema  
Change schema of your source data and create a new target dataset

☐ Find matching records  
Use machine learning to find matching records within your source data

Atrás

Siguiente

1.24. Modificar el nombre de bucket por el nuestro.

17

## Añadir un trabajo

✓ Propiedades del trabajo

visualdetection

✓ Origen de datos

visualdetection

✓ Transform type

Change schema

○ Destino de datos

○ Esquema

### Elija los destinos de datos

☒ Crear tablas en el destino de datos

☐ Usar las tablas del catálogo de datos y actualizar el destino de datos

#### Almacén de datos

Amazon S3

#### Formato

Parquet

#### Conexión

- Seleccione una opción -

Añadir una conexión

#### Ruta de destino

s3://aws-glue-lmgs/data/input/schema/parking\_processed/

