

CS2032 - Cloud Computing (Ciclo 2024-2)

Data Science

Semana 5 - Taller 1: Data Analytics en S3 - 1 Data Source

ELABORADO POR: GERALDO COLCHADO

Con apoyo de Asistente de Cátedra y Laboratorio:

- Paola Maguiña (paola.maguina@utec.edu.pe)

Contenido

Data Analytics en S3

1. Objetivo del taller 1
2. Ejercicio 1: Datos en S3
3. Ejercicio 2: Consultas con **S3** Select
4. Ejercicio 3: Data Catalog en **Glue**
5. Ejercicio 4: Consultas con **Athena**
6. Ejercicio 5: Ejercicio propuesto
7. Cierre

Objetivo del taller 1:

Data Analytics en S3

- Analizar datos de archivos CSV en S3 con SQL
- Entender y usar Catálogo de Datos

Contenido

Data Analytics en S3

1. Objetivo del taller 1
2. **Ejercicio 1: Datos en S3**
3. Ejercicio 2: Consultas con **S3** Select
4. Ejercicio 3: Data Catalog en **Glue**
5. Ejercicio 4: Consultas con **Athena**
6. Ejercicio 5: Ejercicio propuesto
7. Cierre

Ejercicio 1: Datos en S3

- **Paso 1:**
Crear un bucket en S3 y subir los archivos indicados por el docente

Amazon S3 > Buckets > gcr-univ-data

gcr-univ-data Información

Objetos Propiedades Permisos Métricas

Objetos (2) Información

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Tendrá que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	Tipo
<input type="checkbox"/>	athena/	Carpeta
<input type="checkbox"/>	s3/	Carpeta

Amazon S3 > Buckets > gcr-univ-data > athena/

athena/

Objetos Propiedades

Objetos (9) Información

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Tendrá que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	Tipo
<input type="checkbox"/>	alumno001.csv	CSV
<input type="checkbox"/>	alumno002.csv	CSV
<input type="checkbox"/>	alumno003.csv	CSV
<input type="checkbox"/>	alumno004.csv	CSV
<input type="checkbox"/>	alumno005.csv	CSV
<input type="checkbox"/>	alumno006.csv	CSV
<input type="checkbox"/>	alumno007.csv	CSV
<input type="checkbox"/>	alumno008.csv	CSV
<input type="checkbox"/>	alumno009.csv	CSV

Amazon S3 > Buckets > gcr-univ-data > s3/

s3/

Objetos Propiedades

Objetos (1) Información

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Tendrá que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	Tipo
<input type="checkbox"/>	alumnos.csv	CSV

Contenido

Data Analytics en S3

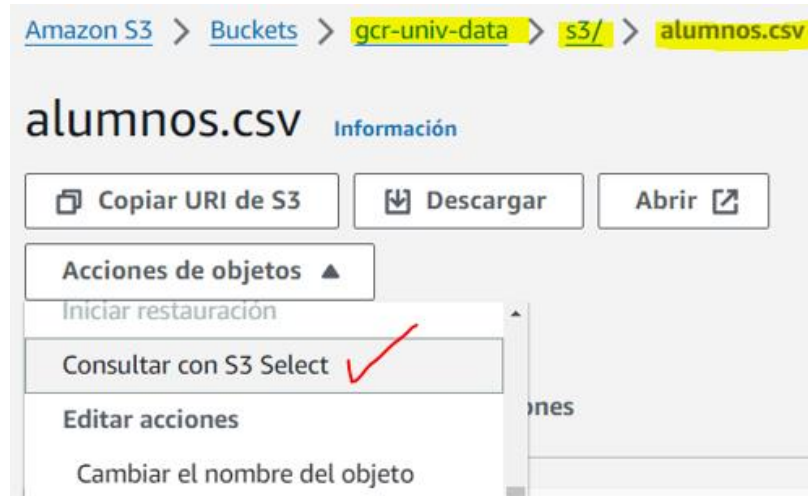
1. Objetivo del taller 1
2. Ejercicio 1: Datos en S3
3. Ejercicio 2: Consultas con S3 Select
4. Ejercicio 3: Data Catalog en Glue
5. Ejercicio 4: Consultas con Athena
6. Ejercicio 5: Ejercicio propuesto
7. Cierre

Ejercicio 2:

Consultas con S3 Select

Tomar en cuenta: S3 Select sólo permite hacer consultas SQL sobre un archivo

- **Paso 1:** Elija “Consultar con S3 Select” y lo indicado.



Configuración de entrada

Ruta
s3://gcr-univ-data/s3/alumnos.csv

Tamaño
703.0 B

Formato
☒ CSV
☐ JSON
☐ Apache Parquet

CSV delimitador
☒ Coma
☐ Tabulador
☐ Personalizado

☒ Excluir la primera línea de CSV datos
Habilite esta configuración si CSV contiene

Compresión
☒ Ninguno

Configuración de salida

Formato
☒ CSV
☐ JSON

CSV delimitador
☒ Coma
☐ Tabulador
☐ Personalizado

Tomar en cuenta: S3 Select sólo permite hacer consultas SQL sobre un archivo

Ejercicio 2:

Consultas con S3 Select

- **Paso 2:** Realice la siguiente consulta SQL

SELECT * FROM s3object s

Estado

✔ Se han devuelto correctamente 9 registros en 1515 ms

Bytes devueltos: 631 B

Sin procesar

Formateado

								< 1 >	
1	Alejandro	Rosalez	2000/12/12	123 Main St.	Any Town	MD	301-555-0158		
2	Jane	Doe	2004/10/05	456 State St.	Anywhere	WA	360-555-0163		
3	John	Stiles	2006/09/20	1980 8th St.	Nowhere	NY	914-555-0122		
4	Li	Juan	2001/06/29	1323 22nd Ave.	Anytown	NY	914-555-0149		
5	Jeanette	Lawrence	2001/02/23	3467 Clark Estate	Michaelberg	MA	437-104-7046		
6	David	Lyons	2005/06/23	69512 King Road	West Kellytown	CA	001-066-838-4285		
7	Nicolas	Garcia	2002/04/24	025 Jason Valley	Mcclureview	NM	629-221-9150		
8	Carlos	Sparks	2005/11/18	3208 Jason Corner	Burnettport	RI	579-4467		
9	Marta	Obrien	2004/04/28	224 Keith Roads	South Matthewfort	OH	138-298-4220		

Tomar en cuenta: S3 Select sólo permite hacer consultas SQL sobre un archivo

Ejercicio 2:

Consultas con S3 Select

- **Paso 3:** Realice las siguientes consultas SQL

```
SELECT * FROM s3object s WHERE s.alumno_id = '2'
```

```
SELECT * FROM s3object s WHERE s.birthday >= '2004'
```

```
SELECT * FROM s3object s WHERE s.state = 'NY'
```

Contenido

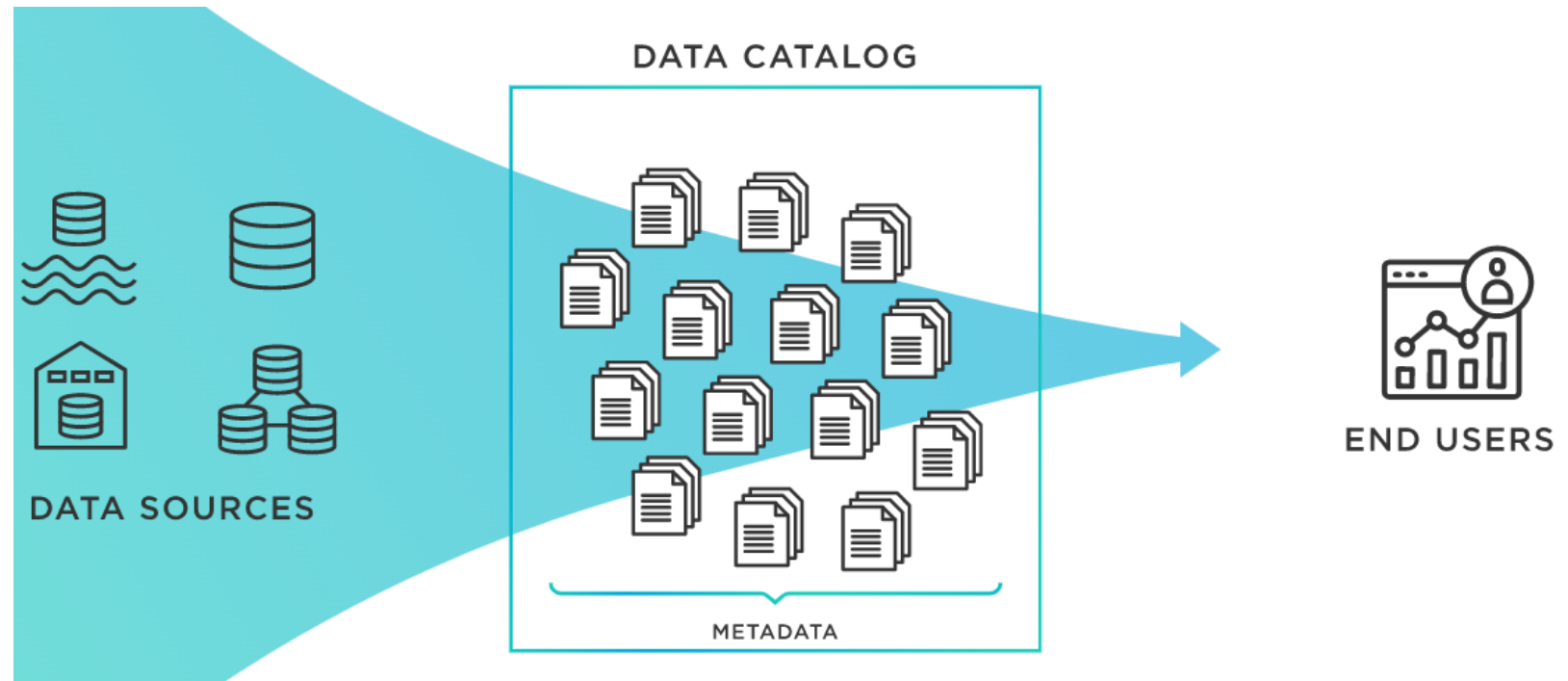
Data Analytics en S3

1. Objetivo del taller 1
2. Ejercicio 1: Datos en S3
3. Ejercicio 2: Consultas con **S3** Select
4. **Ejercicio 3: Data Catalog en **Glue****
5. Ejercicio 4: Consultas con **Athena**
6. Ejercicio 5: Ejercicio propuesto
7. Cierre

Ejercicio 3:

¿ Qué es un Data Catalog?

*“Un catálogo de datos es un **inventario** completo de todos los **conjuntos de datos** que una **organización** **tiene** y pone a disposición para su **uso**”*



Ejercicio 3:

Data Catalog en Glue

- **Paso 1:** Cree una Base de Datos “universidad”

[AWS Glue](#) > [Databases](#) > [Add database](#)

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name

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

Description - *optional*

Descriptions can be up to 2048 characters long.

Database settings

Location - *optional*

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

[Cancel](#) [✓ Create database](#)

Ejercicio 3:

Data Catalog en Glue

- **Paso 2:** Cree una tabla “alumnos”

[AWS Glue](#) > [Tables](#) > [Add table](#)

Step 1
Set table properties

Step 2
Choose or define schema

Step 3
Review and create



Set table properties

Table details

Name

If you plan to access the table from Amazon Athena, then the name should be under 256 characters and contain only lowercase letters (a-z), numbers (0-9), and underscore (_). For more information, see [Athena names](#).

Database

  [Create database](#)

Description - optional

Descriptions can be up to 2048 characters long.

Ejercicio 3:

Data Catalog en Glue

- **Paso 2:** Cree una tabla “alumnos”

Table format

Data Catalog managed tables support data compaction for Apache Iceberg table

☒ **Standard AWS Glue table (default)**
Create a standard AWS Glue table.

☐ Apache Iceberg table
Create an Apache Iceberg table.

Data store

Select the type of source

☒ **S3**
☐ Kinesis
☐ Kafka

Data location is specified in

☒ my account
☐ another account

Include path

Path must be in the form s3://bucket/prefix/. It must end with a slash (/) and not contain any spaces.

Data format

Classification

Choose the format of the data in your table.

☐ Avro
☒ **CSV**
☐ JSON
☐ XML
☐ Parquet
☐ ORC

Delimiter

Ejercicio 3:

Data Catalog en Glue

- **Paso 2:** Cree una tabla "alumnos"

Step 2/3

Choose or define schema

Schema

☒ Define or upload schema
Manually define schema

☐ Choose from Glue Schema Registry
Select existing schema from your Glue Schema Registry.

Schema (0)

Edit schema as JSON ✓

Delete

Edit schema as JSON

```
1 [
2   {
3     "Name": "alumno_id",
4     "Type": "int"
5   },
6   {
7     "Name": "first_name",
8     "Type": "string"
9   },
10  {
11    "Name": "last_name",
12    "Type": "string"
13  },
14  {
15    "Name": "birthday",
16    "Type": "string"
17  },
18  {
19    "Name": "street_address",
```

Cancel

✓ Save

Ejercicio 3:

Data Catalog en Glue

- **Paso 2:** Cree una tabla “alumnos”

[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 for data processing jobs.

Tables (1)

View and manage all available tables.

<input type="checkbox"/>	Name ▲	Database ▼	Location ▼	Classification
<input type="checkbox"/>	alumnos	universidad	s3://gcr-univ-data/athena/	CSV

Contenido

Data Analytics en S3

1. Objetivo del taller 1
2. Ejercicio 1: Datos en S3
3. Ejercicio 2: Consultas con **S3** Select
4. Ejercicio 3: Data Catalog en **Glue**
5. **Ejercicio 4: Consultas con Athena**
6. Ejercicio 5: Ejercicio propuesto
7. Cierre

Tomar en cuenta: Athena permite hacer consultas SQL sobre varios archivos

Ejercicio 4:

Consultas en Athena

- **Paso 1:** Ejecute esta consulta en Athena

`SELECT * FROM "universidad"."alumnos" order by alumno_id;`

Amazon Athena > Editor de consultas

Editor Consultas recientes Consultas guardadas Configuración Grupo de trabajo primary

Datos

Origen de datos
AwsDataCatalog

Base de datos
universidad

Tablas y vistas

▼ Tablas (1) < 1 >

✚ alumnos

► Vistas (0) < 1 >

Consulta 8

1 `SELECT * FROM "universidad"."alumnos" order by alumno_id;`

SQL Ln 1, Col 58

Ejecutar de nuevo Explicar Cancelar Borrar Crear

☐ Volver a utilizar los resultados de la consulta hasta hace 60 minutos

Resultados de la consulta Estado de la consulta

✓ Completado Tiempo en cola: 65 ms Tiempo de ejecución: 549 ms Datos analizados: 0.60 KB

Resultados (9) Copiar Descargar resultados

#	alumno_id	first_name	last_name	birthday	street_address	city	state	phone
1	1	Alejandro	Rosalez	2000/12/12	123 Main St.	Any Town	MD	301-555-0158
2	2	Jane	Doe	2004/10/05	456 State St.	Anywhere	WA	360-555-0163
3	3	John	Stiles	2006/09/20	1980 8th St.	Nowhere	NY	914-555-0122
4	4	Li	Juan	2001/06/29	1323 22nd Ave.	Anytown	NY	914-555-0149

Ejercicio 4:

Consultas en Athena

- **Paso 2:** Ejecute estas consultas en Athena
 - `SELECT * FROM "universidad"."alumnos" where state = 'NY';`
 - `SELECT * FROM "universidad"."alumnos" where birthday >= '2004';`
 - `SELECT substr(birthday,1,4) as year, count(*) as alumnos FROM "universidad"."alumnos" group by 1 order by 1 desc;`

Contenido










Data Analytics en S3

1. Objetivo del taller 1
2. Ejercicio 1: Datos en S3
3. Ejercicio 2: Consultas con S3 Select
4. Ejercicio 3: Data Catalog en Glue
5. Ejercicio 4: Consultas con Athena
6. **Ejercicio 5: Ejercicio propuesto**
7. Cierre

Ejercicio 5:

Ejercicio propuesto - Casa

- Modifique el formato y contenido de los archivos de csv a **json** y los sube a nueva carpeta en bucket S3
- Cree una nueva tabla “alumnos_json” en Base de Datos “universidad” en Data Catalog de Glue
- Realice consultas en Athena
- Suba la evidencia al padlet

Nombre ▲	Tipo
 alumno001.csv	CSV
 alumno002.csv	CSV
 alumno003.csv	CSV
 alumno004.csv	CSV
 alumno005.csv	CSV
 alumno006.csv	CSV
 alumno007.csv	CSV
 alumno008.csv	CSV
 alumno009.csv	CSV

Data format

Classification

Choose the format of the data in your table.

- ☐ Avro
- ☐ CSV
- ☒ JSON
- ☐ XML
- ☐ Parquet
- ☐ ORC

Contenido

Data Analytics en S3

1. Objetivo del taller 1
2. Ejercicio 1: Datos en S3
3. Ejercicio 2: Consultas con **S3** Select
4. Ejercicio 3: Data Catalog en **Glue**
5. Ejercicio 4: Consultas con **Athena**
6. Ejercicio 5: Ejercicio propuesto
7. Cierre

Cierre:

Data Analytics en S3 - Qué aprendimos?

- Analizar datos de archivos CSV en S3 con SQL
- Entender y usar Catálogo de Datos

Gracias

Elaborado por docente: Geraldo Colchado