

CS2032 - Cloud Computing (Ciclo 2024-2)

Data Science

Semana 5 - Taller 2: Data Analytics en S3 - 2 Data Sources

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

Objetivo del taller 2:

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: Data Catalog en **Glue**
4. Ejercicio 3: Consultas con **Athena**
5. Ejercicio 4: Ejercicio propuesto
6. Cierre

Ejercicio 1:

Datos en S3

- **Paso 1:** Subir el archivo indicado por el docente

Amazon S3 > Buckets > gcr-univ-data




gcr-univ-data Información

Objetos | Propiedades | Permisos

Objetos (3) Información

Los objetos son las entidades fundamentales que se almacenan en S3. Más información [↗](#)

🔍 Buscar objetos por prefijo

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

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


notas/

Objetos | Propiedades

Objetos (1) Información

Los objetos son las entidades fundamentales que se almacenan en S3. Más información [↗](#)

🔍 Buscar objetos por prefijo

<input type="checkbox"/>	Nombre ▲	Tipo
<input type="checkbox"/>	 notas.csv	csv

Contenido

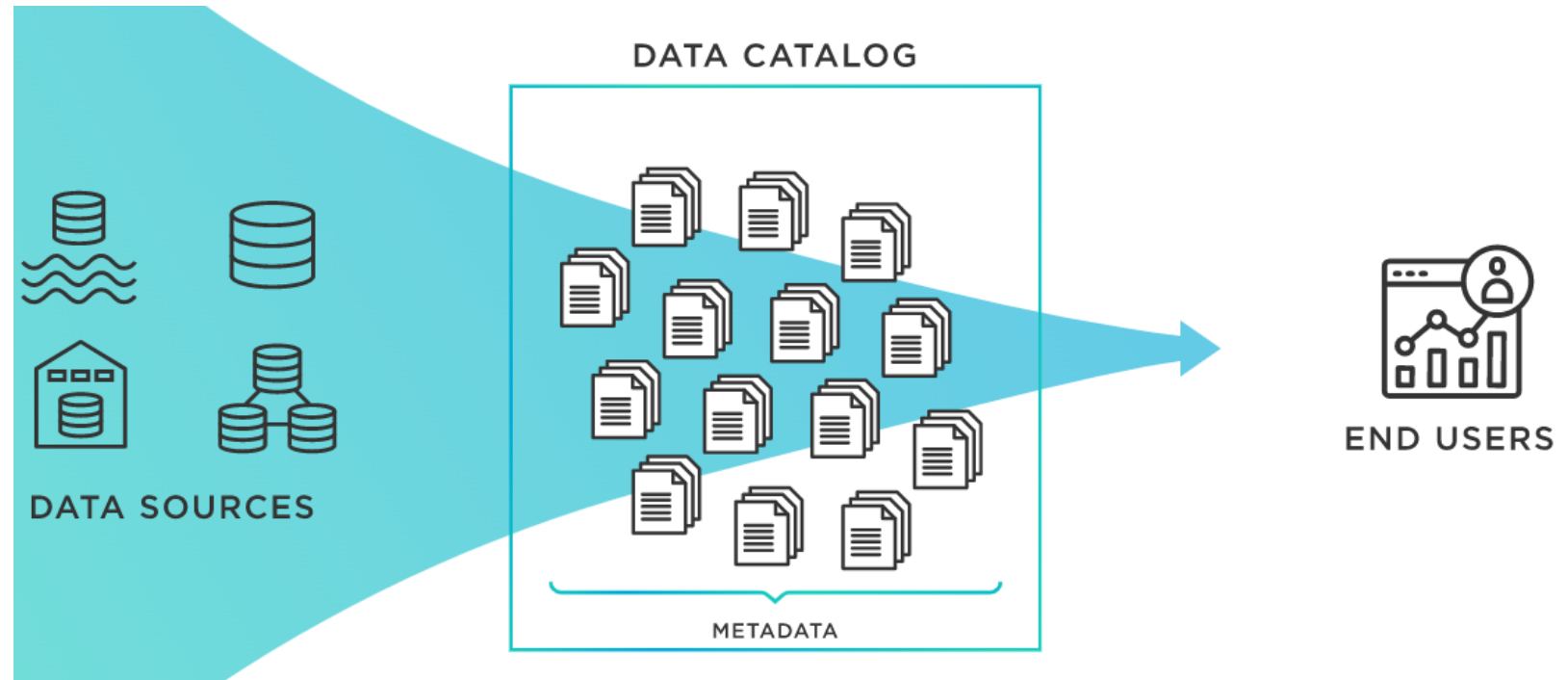
Data Analytics en S3

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

Ejercicio 2:

¿ 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 2:

Data Catalog en Glue

- **Paso 1:** Cree una tabla “notas” en el database “universidad” utilizando el schema entregado por el docente

Tables (2)
View and manage all available tables.

Filter tables

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

Schema (5)
View and manage the table schema.

Filter schemas

#	Column name ▼	Data type
1	codigo_alumno	int
2	ciclo	string
3	codigo_curso	string
4	tipo_nota	string
5	nota	int

Contenido

Data Analytics en S3

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

Ejercicio 3:

Consultas en Athena

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

```
SELECT * FROM "universidad"."notas";
```

Amazon Athena > Editor de consultas

Editor Consultas recientes Consultas guardadas Configuración Grupo de trabajo primari

Datos ↻ <

Origen de datos
AwsDataCatalog

Base de datos
universidad

Tablas y vistas Crear ⌵ ⚙

🔍 Filtrar tablas y vistas

▼ Tablas (2) < 1 >

- alumnos ⋮
- notas ⋮

► Vistas (0) < 1 >

Consulta 8 : ✕ Consulta 9 : ✕ Consulta 10 : ✕ Consulta 11 : ✕ Consulta 12 : ✕ Consulta 13 : ✕ Consulta 14 : ✕ **Consulta 15 : ✕**

1 **SELECT * FROM "universidad"."notas";**

SQL Ln 1, Col 1

Ejecutar de nuevo Explicar 🔗 Cancelar Borrar Crear ⌵

⏻ Volver a utilizar los resultados hasta h

Resultados de la consulta Estado de la consulta

✅ Completado Tiempo en cola: 65 ms Tiempo de ejecución: 460 ms Datos ana

Resultados (16) Copiar Descarga

🔍 Filas de búsqueda

#	codigo_alumno	ciclo	codigo_curso	tipo_nota	nota
1	1	202402	CLOUDCOMP	TI	15
2	1	202402	CLOUDCOMP	PA1	15
3	1	202402	CLOUDCOMP	TP	18
4	1	202402	CLOUDCOMP	EP	18
5	2	202402	CLOUDCOMP	TI	13
6	2	202402	CLOUDCOMP	PA1	7

Ejercicio 3:

Consultas en Athena

- **Paso 2:** Ejecute esta consulta en Athena para unir las tablas “alumnos” con “notas”

```
SELECT b.codigo_alumno, a.first_name, a.last_name, b.ciclo, b.codigo_curso, b.tipo_nota, b.nota
FROM "universidad"."alumnos" a, "universidad"."notas" b
WHERE a.alumno_id = b.codigo_alumno
ORDER BY b.codigo_alumno, b.tipo_nota;
```

1 SELECT b.codigo_alumno, a.first_name, a.last_name, b.ciclo, b.codigo_curso, b.tipo_nota, b.nota

2 FROM "universidad"."alumnos" a, "universidad"."notas" b

3 WHERE a.alumno_id = b.codigo_alumno

4 ORDER BY b.codigo_alumno, b.tipo_nota;

SQL Ln 4, Col 39

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: 86 ms

Tiempo de ejecución: 1.119 sec

Datos analizados: 1.01 KB

Resultados (16)

Copiar

Descargar resultados

🔍 Filas de búsqueda

< 1 >

⚙️

#	codigo_alumno	first_name	last_name	ciclo	codigo_curso	tipo_nota	nota
1	1	Alejandro	Rosalez	202402	CLOUDCOMP	EP	18
2	1	Alejandro	Rosalez	202402	CLOUDCOMP	PA1	15
3	1	Alejandro	Rosalez	202402	CLOUDCOMP	TI	15
4	1	Alejandro	Rosalez	202402	CLOUDCOMP	TP	18
5	2	Jane	Doe	202402	CLOUDCOMP	EP	14
6	2	Jane	Doe	202402	CLOUDCOMP	PA1	7

Ejercicio 3:

Consultas en Athena

- **Paso 3:** Crear vista con el query anterior para consultas futuras

```
CREATE OR REPLACE VIEW view_notas_alumnos AS
SELECT b.codigo_alumno, a.first_name, a.last_name, b.ciclo, b.codigo_curso, b.tipo_nota, b.nota
FROM "universidad"."alumnos" a, "universidad"."notas" b
WHERE a.alumno_id = b.codigo_alumno
ORDER BY b.codigo_alumno, b.tipo_nota;
```

Crear ▲ ⚙

Crear una tabla a partir de un origen de datos

Datos del bucket de S3

Rastreador de AWS Glue 🔗

Crear con SQL

CREATE TABLE

CREATE TABLE AS SELECT

CREATE TABLE AS SELECT(ICEBERG)

CREATE VIEW ✓

Datos ↻ <

Origen de datos

AwsDataCatalog ▼

Base de datos

universidad ▼

Tablas y vistas Crear ▼ ⚙

🔍 Filtrar tablas y vistas

▼ **Tablas** (2) < 1 >

+ alumnos ⋮

+ notas ⋮

▼ **Vistas** (1) < 1 >

+ **view_notas_alumnos** ⋮

Consulta 8 : ✕ | Consulta 9 : ✕ | Consulta 10 : ✕ | ✓ Consulta 11 : ✕ | ✓ Consulta 12 : ✕ | ↻

1 -- View Example

2 CREATE OR REPLACE VIEW **view_notas_alumnos** AS

3 SELECT b.codigo_alumno, a.first_name, a.last_name, b.ciclo, b.codigo_curso, b.tipo_nota, b.nota

4 FROM "universidad"."alumnos" a, "universidad"."notas" b

5 WHERE a.alumno_id = b.codigo_alumno

6 ORDER BY b.codigo_alumno, b.tipo_nota;

SQL Ln 6, Col 39

Ejecutar de nuevo Explicar 🔗 Cancelar Borrar Crear ▼

Resultados de la consulta Estado de la consulta

✓ Completado

La consulta se ha realizado correctamente.

Ejercicio 3:

Consultas en Athena

- **Paso 4:**
Consultar la
vista creada

```
SELECT * FROM "universidad"."view_notas_alumnos";
```

1 SELECT * FROM "universidad"."view_notas_alumnos";

SQL Ln 1, Col 50

Ejecutar de nuevo

Explicar

Cancelar

Borrar

Crear

☐ Volver a utilizar los resultados de la hasta hace 60 mi

Resultados de la consulta

Estado de la consulta

✓ Completado

Tiempo en cola: 78 ms

Tiempo de ejecución: 968 ms

Datos analizados:

Resultados (16)

Copiar

Descargar result

Q Filas de búsqueda

< 1 >

#	codigo_alumno	first_name	last_name	ciclo	codigo_curso	tipo_nota	nota
1	1	Alejandro	Rosalez	202402	CLOUDCOMP	EP	18
2	1	Alejandro	Rosalez	202402	CLOUDCOMP	PA1	15
3	1	Alejandro	Rosalez	202402	CLOUDCOMP	TI	15
4	1	Alejandro	Rosalez	202402	CLOUDCOMP	TP	18
5	2	Jane	Doe	202402	CLOUDCOMP	EP	14
6	2	Jane	Doe	202402	CLOUDCOMP	PA1	7

Contenido

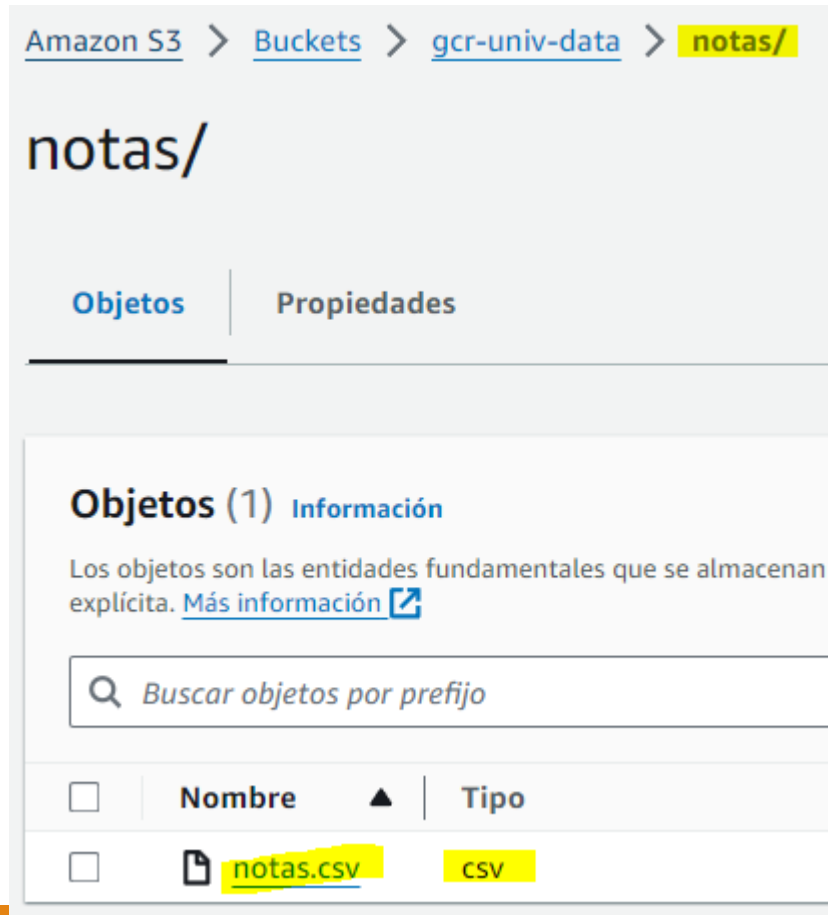
Data Analytics en S3

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

Ejercicio 4:

Ejercicio propuesto - Casa

- Modifique el formato y contenido del archivo “notas” de csv a **json** y lo sube a nueva carpeta en bucket S3
- Cree una nueva tabla “notas_json” en Base de Datos “universidad” en Data Catalog de Glue
- Realice consultas en Athena uniendo tabla “notas_json” con tabla “alumnos” o “alumnos_json”
- Suba la evidencia al padlet



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: Data Catalog en **Glue**
4. Ejercicio 3: Consultas con **Athena**
5. Ejercicio 4: Ejercicio propuesto
6. 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