COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

AWS Glue Data Catalog

Tecnologías de procesamiento Big Data

Grado en Ingeniería Matemática e Inteligencia Artificial

comillas.edu

ÍNDICE

ÍNDICE

01 | ¿QUÉ ES AWS GLUE?

Introducción

- AWS Glue es un servicio de integración de datos 'serverless' que facilita la detección, preparación y la combinación de los datos para analítica, machine learning y desarrollo de aplicaciones.

- AWS Glue proporciona todas las funcionalidades necesarias para la integración de datos, pudiendo comenzar a analizarlos en minutos en vez de en meses.

- Es un servicio para ETL (Extract, Transform, Load) totalmente administrado por AWS.

- Facilita la preparación y carga de datos para el análisis.

- Servicio Serverless: no requiere administrar ni aprovisionar infraestructura.

- Automatiza el proceso de descubrimiento de datos.

ÍNDICE

02 | COMPONENTES PRINCIPALES DE AWS GLUE

Introducción

AWS Glue Data Catalog: Repositorio centralizado que almacena y organiza los metadatos de todas las fuentes de datos y tablas dentro del ecosistema de AWS.

ETL engine: Motor de procesamiento que automatiza la generación y ejecución de código para extraer, transformar y cargar datos entre diferentes fuentes.

Glue Crawler: Herramienta automatizada que escanea, descubre y cataloga metadatos de diversas fuentes de datos actualizando el Data Catalog.

Glue jobs: Scripts de procesamiento ETL que ejecutan las transformaciones necesarias para convertir los datos de origen en su formato de destino deseado.

ÍNDICE

Catálogo de metadatos persistente: Un sistema que mantiene y almacena de forma permanente la información estructural de tus datos, garantizando que no se pierda entre sesiones.

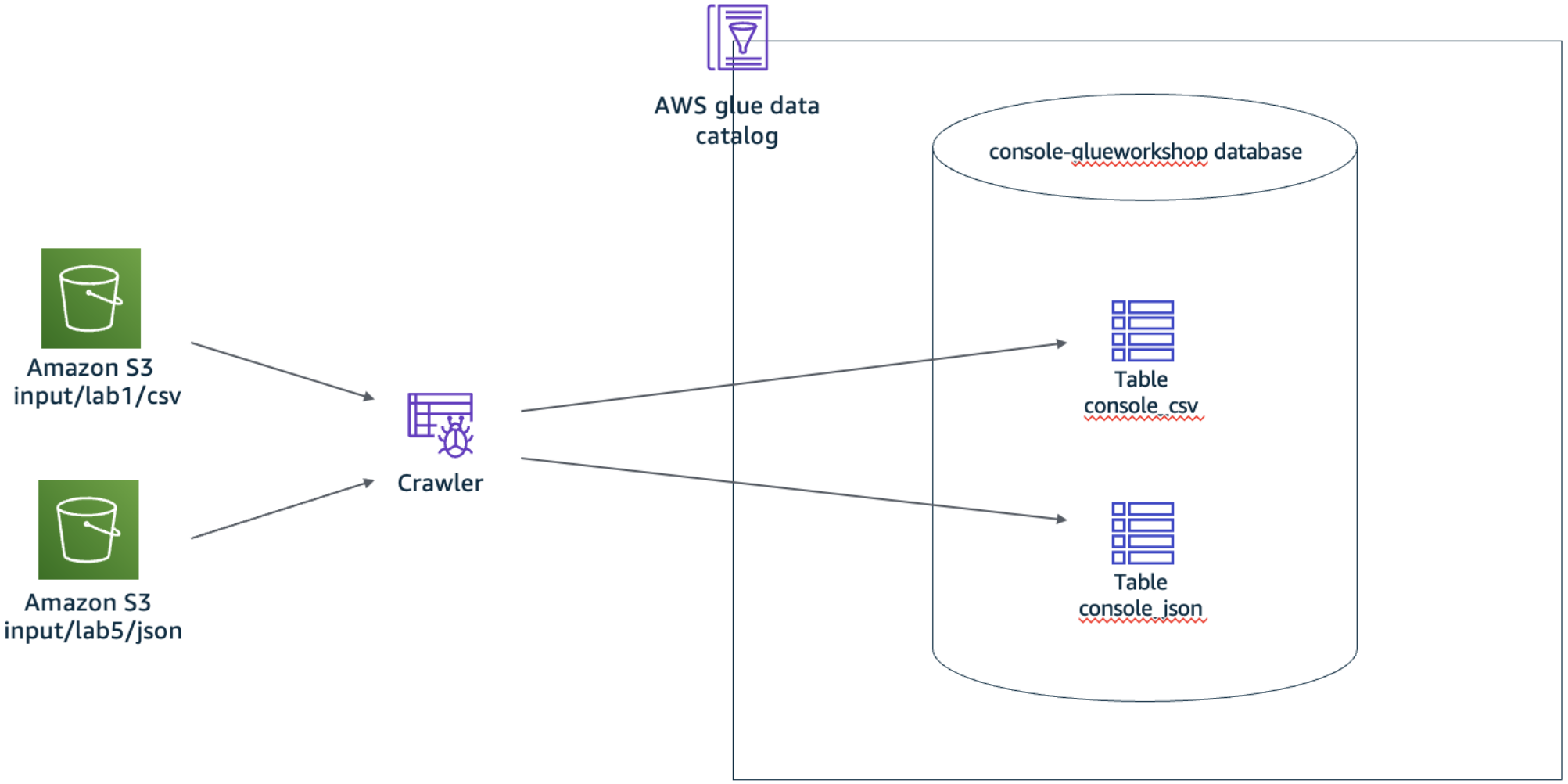
Almacén central de metadatos: Un repositorio unificado que actúa como punto único de referencia para almacenar y gestionar toda la información sobre tus datos en AWS.

Compatible con Apache Hive: Se integra perfectamente con Apache Hive, permitiendo usar sus funcionalidades de procesamiento de datos y consultas sin necesidad de migraciones.

Registro detallado de esquemas y propiedades de datos: Mantiene un inventario completo y detallado de las estructuras de datos, incluyendo tipos de columnas, formatos y características específicas de cada conjunto de datos.

03 | AWS GLUE DATA CATALOG

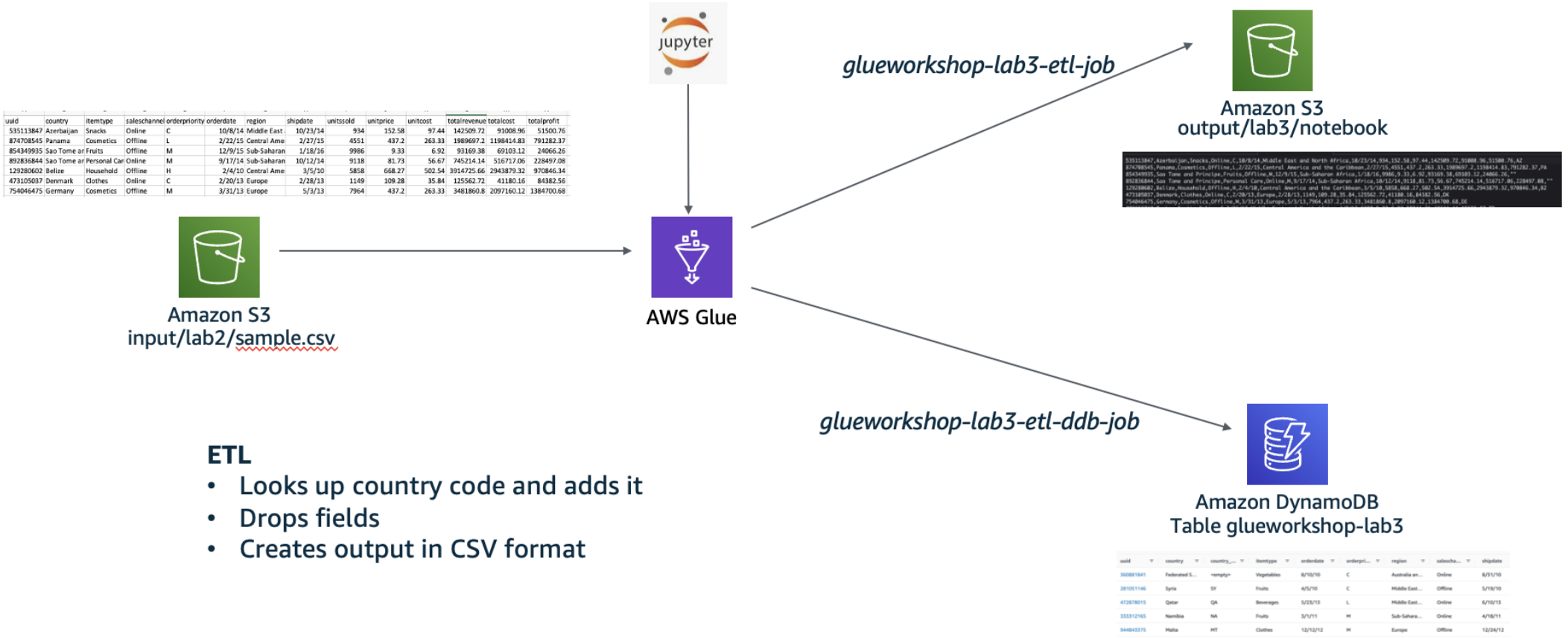
Introducción



03 | AWS GLUE DATA CATALOG

Introducción

First test using notebook
Then deploy code using Glue Job

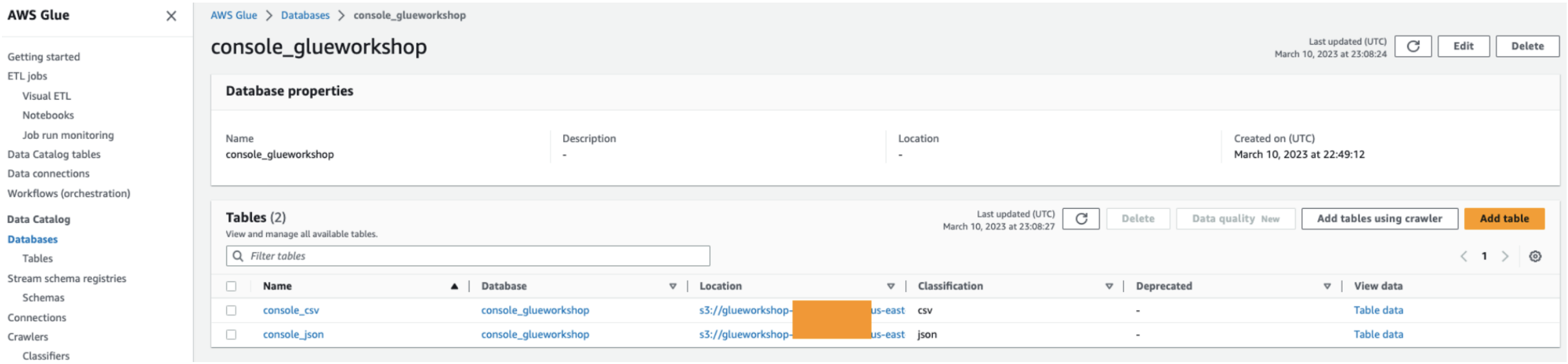


ETL

- Looks up country code and adds it
- Drops fields
- Creates output in CSV format

03 | AWS GLUE DATA CATALOG

Introducción



03 | AWS GLUE DATA CATALOG

Glue Crawler

Un Glue Crawler es una herramienta automatizada que escanea las fuentes de datos para descubrir, catalogar y organizar metadatos en el AWS Glue Data Catalog.

- Descubrir automáticamente el esquema de los datos
- Entender los formatos y tipos de datos
- Organizan los metadatos en tablas
- Actualizan el Data Catalog cuándo se necesite
- Programables para ejecución periódica
- Detectan cambios en los esquemas de los datos

03 | AWS GLUE DATA CATALOG

Glue Crawler

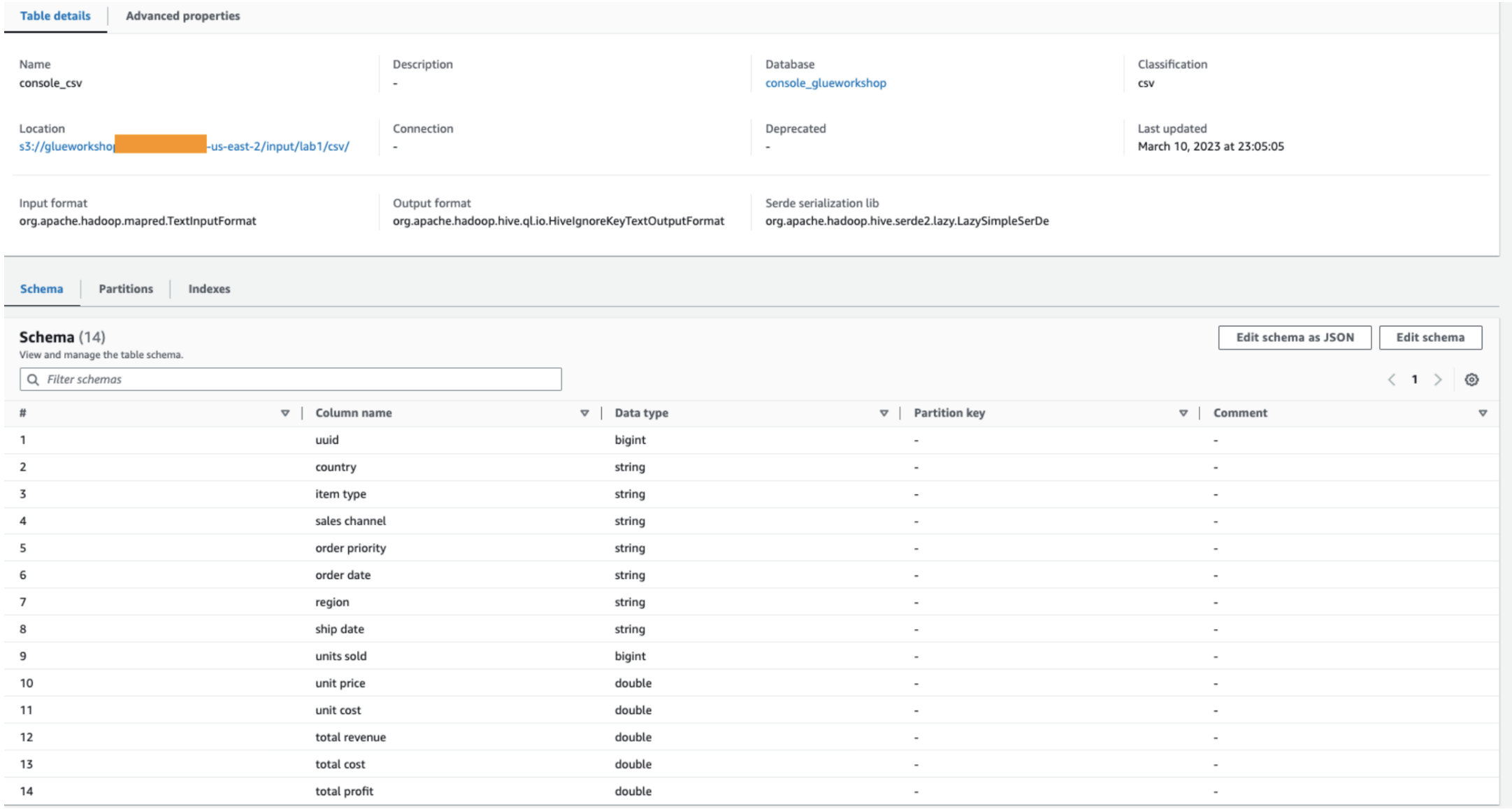


Table details | Advanced properties

Name console_csv	Description -	Database console_glueworkshop	Classification csv
Location s3://glueworkshop[redacted]-us-east-2/input/lab1/csv/	Connection -	Deprecated -	Last updated March 10, 2023 at 23:05:05
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema | Partitions | Indexes

Schema (14)

View and manage the table schema.

Edit schema as JSON | Edit schema

Filter schemas

1

#	Column name	Data type	Partition key	Comment
1	uuid	bigint	-	-
2	country	string	-	-
3	item type	string	-	-
4	sales channel	string	-	-
5	order priority	string	-	-
6	order date	string	-	-
7	region	string	-	-
8	ship date	string	-	-
9	units sold	bigint	-	-
10	unit price	double	-	-
11	unit cost	double	-	-
12	total revenue	double	-	-
13	total cost	double	-	-
14	total profit	double	-	-

ÍNDICE

04 | ORÍGENES DE DATOS COMPATIBLES CON AWS GLUE

AWS Glue puede integrarse con más de 80 fuentes de datos en AWS, en centros de datos locales y en otros proveedores de nube.

AWS Glue es compatible de forma nativa con los datos almacenados en Amazon Aurora, Amazon RDS para MySQL, Amazon RDS para Oracle, Amazon RDS para PostgreSQL, Amazon S3 y DynamoDB entre otros.

AWS Glue también es compatible con Amazon MSK, Amazon Kinesis Data Streams y Apache Kafka.

ÍNDICE

04 | ORÍGENES DE DATOS COMPATIBLES CON AWS GLUE
DEMO 1

- Crear base de datos y tablas en AWS Glue de forma manual
  - Comandos útiles a utilizar
    - MSCK REPAIR TABLE <tabla>;
    - ALTER TABLE <tabla>

      SET TBLPROPERTIES (

      'skip.header.line.count'='1'

      );
- Crear y ejecutar un AWS Glue Crawler
  - Ver AWS Glue workshop (Lab 01) para seguir una guía con todas las funcionalidades
  - Es necesario en la creación del Crawler indicar un role con permisos de lectura sobre el bucket donde están los datos