DATA ENGINEER CHALLANGE

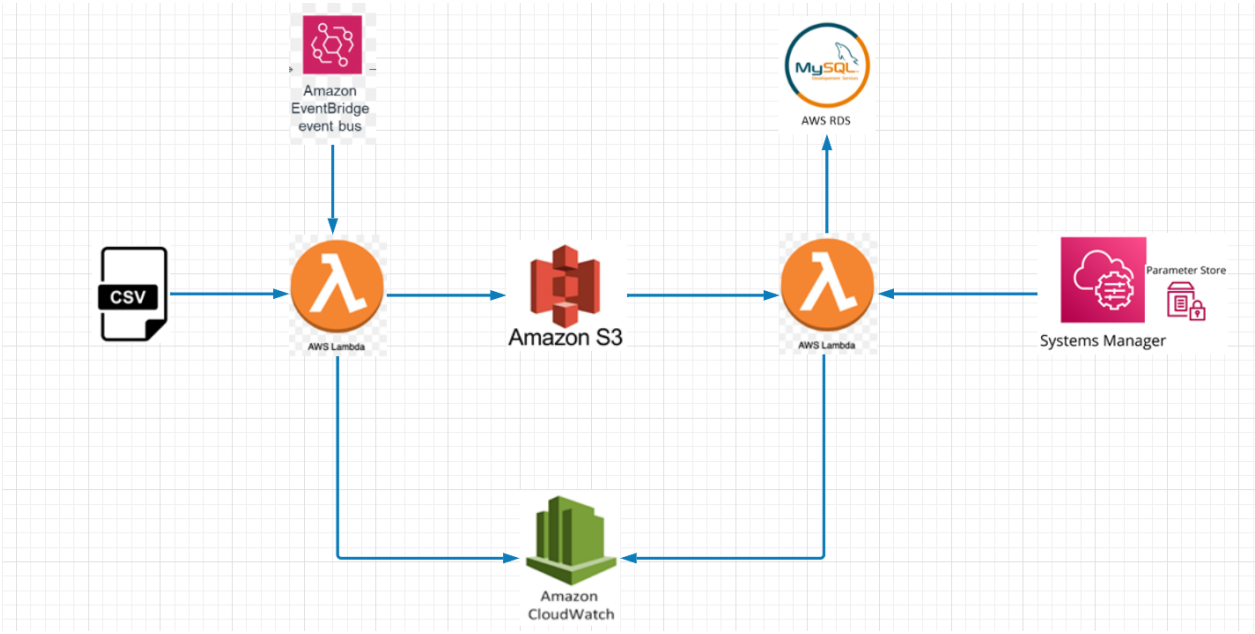
Federico De Grazia

Marzo 2023

Índice

Arquitectura

El siguiente diagrama presenta una propuesta de solución en la plataforma de aws:



Servicios involucrados

AWS Lambda:

Es el servicio encargado de ejecutar la lógica de procesamiento de los archivos csv suministrados. La lógica de la misma consiste en:

1. Establecer conexión con los servicios de aws a interactuar mediante la SDK de Python, para nuestro caso aws s3 y aws system manager.
2. Conexión y ejecución de consultas contra la base de datos.
3. Procesamiento de la información del csv, armado de los datasets a impactar en la base destino.

La primer función Lambda realiza el request a la url para descargar el archivo con la nueva información de la tabla “unificado”, guardando el archivo en un bucket de S3. La configuración de ejecución de esta función es un cron definido por los eventos de aws cloudwatch.

La segunda función toma el archivo generado en el paso anterior y realiza las modificaciones del enunciado:

1. Completar el campo “Fecha_Copia” con el datetime del momento.
2. Eliminar los duplicados según la lógica: “considerando que un registro será duplicado si los campos [ID], [MUESTRA] y [RESULTADO] son iguales en dos filas distintas.”

AWS S3:

Es el filesystem donde se almacenarán los archivos que involucran esta solución. La función Lambda luego de descargar el archivo csv, la acción de crear un nuevo objeto en este bucket es el disparador de la ejecución de la segunda función Lambda, encargada del procesamiento del mismo e impactar el resultado en la base destino.

AWS CloudWatch:

En él son almacenados los logs de ejecución de los servicios funciones Lambda, posibilitando el seguimiento, debugging y registro histórico de los mismos.

El archivo “example.png” es el log de una de las ejecuciones de prueba a modo de ejemplo.

AWS Parameter Store:

Funciona como almacenamiento, potencialmente compartido entre varios recursos, de información clave valor. Para nuestro caso, aws lambda consulta este servicio para conocer los datos de conexión a la base de datos destino, en específico estos son:

1. Host.
2. Nombre de la db.
3. Usuario.
4. contraseña.

De los cuales los últimos dos están encriptados por seguridad.

AWS RDS:

Es la base de datos utilizada en esta PoC como almacenamiento final de los datos procesados. Se trata de una base MySQL relacional donde fueron creadas las tablas. Entre la configuración de la misma, se estableció la generación automática de backups por medio de snapshots cada semana. La lógica de las ddls de la tabla “Unificado” se puede encontrar en el archivo “DDL.sql”.