Documentación - Pipeline Datalake comercializadora de energía

Objetivo

Diseñar e implementar una estrategia de *Data Lake* en AWS S3 que permita almacenar, transformar y consultar datos provenientes del sistema interno de una comercializadora de energía. Los datos incluyen información de **proveedores, clientes y transacciones**. El enfoque busca mantener trazabilidad, control de calidad, y disponibilidad para análisis futuros.

Arquitectura

La solución esta estructurada en tres capas:

1. Raw Layer

Es la primera capa, en la cual se almacenan los datos tal y como llegan desde las fuentes, en este caso, se almacenan tal cual los archivos *csv* de los proveedores, clientes y transacciones los cuales provienen del sistema que tiene la compañía. Esta etapa con el fin de tener una trazabilidad completa de la información, la utilidad de almacenar estos datos por temas de auditoria y tener un respaldo integro de los datos en caso de ser necesario.

Ubicación: s3://s3-energy-<stage>/raw/

2. Staging Layer

Es la capa intermedia donde los datos se limpian, normalizan y validan parcialmente. En este caso se aplican varias transformaciones importantes:

1. Todos los campos tipo *string* se llevan a minúsculas y se les elimina posibles espacios al inicio y al final
2. Se normaliza el nombramiento de las columnas, se reemplazan los espacios en blanco por '_' y todo en minúscula
3. Los campos precio y cantidad de las transacciones se evalúa que no sean nulos ni que su valor sea menor a =.

Ubicación: s3://s3-energy-<stage>/staging/

3. Trusted Layer

Es la capa donde los datos están validados y listos para ser consumidos por los usuarios. En este caso, se aplicaron unas transformaciones finales a los datos para garantizar la calidad e integración completa de estos.

1. Los valores del campo de *fecha transaccion* se estandarizan en su formato. Al igual que los valores de *precio* y *cantidad* se vuelven tipo double con el fin de considerar cifras decimales en los cálculos.
2. Se crea una columna nueva llamada '*monto_total*' la cual surge multiplicar el *precio* de la transacción por la *cantidad*, esta información es útil porque nos permite conocer rápidamente cual fue el total de la transacción.
3. Se eliminan posibles registros duplicados de los clientes y las transacciones

Ubicación: s3://s3-energy-<stage>/trusted/

Automatización y Orquestación

Carga automática y periódica

Se utiliza **AWS Lambda** en conjunto con **Amazon EventBridge** para orquestar las cargas diarias de la información al datalake. Cada día, los nuevos archivos *csv* de clientes, transacciones y proveedores son cargados automáticamente a la capa Raw.

Transformación con AWS Glue

Se desarrollan diferentes Jobs de Glue que ejecutan las transformaciones mencionadas anteriormente sobre los diferentes datasets.

Recursos Definidos en la Infraestructura(`iac/`)

- `functions/` → Contiene las funciones Lambda utilizadas.
  - `raw_loader.yml` → Lambda encargada de almacenar los archivos csv de clientes, transacciones y proveedores en la capa raw.
- `resources/` → Contiene los recursos requeridos por el pipeline.
  - `bucket.yml` → Bucket S3 que corresponde al datalake donde se almacenan los datos de la compañía de energía.
  - `crawler.yml` → Rastreador de AWS Glue para rellenar el Catálogo de Datos de AWS Glue con la base de datos y tablas correspondientes al proceso de energía.
  - `events.yml` → Se han definido dos eventos en EventBridge: uno para activar la carga automática y diaria de los archivos de la compañía, y otro para iniciar la ejecución del rastreador (crawler).
  - `glue_database.yml` → Base de datos de AWS Glue.
  - `jobs.yml` → Definición de los jobs de AWS Glue de la capa *stage* y *trusted*.
  - `roles.yml` → **Roles de IAM** necesarios para los diversos recursos de AWS dentro de la infraestructura. Cada rol está diseñado con el principio de **mínimos privilegios**, es decir, solo se otorgan los permisos necesarios para que los servicios y recursos puedan interactuar entre sí de manera segura y eficiente .

Configuración de permisos y políticas

Para realizar el proceso de configurar los permisos y políticas para los diferentes servicios de AWS utilizados se debe seguir una serie de pasos.

1. Identificar los servicios de AWS

Primero, es necesario identificar todos los servicios de AWS que forman parte del pipeline.

2. Crear roles de IAM

Los roles de Identity and Access Management (IAM) son necesarios para dar permisos a las instancias, funciones o servicios de AWS para interactuar con otros servicios. Cada Servicio debe tener un rol que defina las políticas necesarias, ya sean predefinidas o personalizadas para definir las acciones que el servicio puede realizar.

3. Configurar políticas de IAM

Las políticas predefinas o personalizadas definen qué acciones un servicio o usuario puede realizar en un recurso de AWS. Es por eso que, deben ser configuradas de manera adecuada y teniendo en cuenta el principio de Mínimo privilegio, es decir, solo otorgar los permisos estrictamente necesarios para cada servicio.

4. Configurar políticas de seguridad de Red

Es importante asegurarnos que los recursos en la red de AWS tengan configuraciones que permitan a los servicios comunicarse de manera segura entre sí, como por ejemplo definir la configuración VPC para controlar el trafico hacia y desde los servicios.

Siguiendo esto, para los recursos utilizados por el *Pipeline Datalake comercializadora de energía* se definen sus respectivos roles IAM en los cuales se configura los permisos y políticas necesario como por ejemplo, lectura sobre un S3, ejecución de un Crawler, entre otras.