Documentación:

1. Realiza una descripción detallada del pipeline de datos construido.

**Pipeline de Datos para Lake Formation**

**1. Ingesta de Datos**

**Fuentes de Datos:**

* + Archivos CSV

(clientes – proveedores transacciones)

**Procesos de Ingesta:**

* + **AWS Glue Jobs:** Extraen datos desde la fuente y los cargan en S3 en formato **Parquet**
  + **ETL simplificado:** En Glue, se configuran transformaciones básicas para limpiar y formatear datos antes de almacenarlos.

**Almacenamiento en Amazon S3**

S3 actúa como la capa de almacenamiento del Data Lake.

* **Estructura del Bucket S3:**
  + Datos organizados jerárquicamente por fuente, tipo, y partición:

bash

s3://data-lake-electra/

├── bronce/ # Datos sin procesar

├── plata/ # Datos procesados y limpios

└── oro/ # Datos optimizados para análisis

**Características Clave:**

* + **Cifrado con AWS KMS** para proteger los datos.
  + **Versionado** activado para rastrear cambios en los datos.
  + **Políticas de Acceso:** Bucket Policies restringen accesos no autorizados.

**Gestión de Permisos con Lake Formation**

Lake Formation centraliza el control de acceso a los datos del Data Lake.

**Registro de Ubicaciones:**

* + Los buckets de S3 se registran en Lake Formation, lo que permite administrar permisos desde un único lugar.

**Control de Acceso Granular:**

* + Permisos otorgados por base de datos, tabla, columna o fila.

**Column-Level y Row-Level Permissions:**

* + Lake Formation aplica filtros automáticamente para garantizar que solo los datos autorizados se entreguen al consumidor.

**Transformación de Datos con AWS Glue**

AWS Glue realiza las tareas de transformación, limpieza y enriquecimiento de los datos.

**Glue Jobs:**

* + Transforman los datos sin procesar (raw) en conjuntos de datos listos para análisis (curated).
  + Ejemplo: Cambiar formatos de fecha, normalizar nombres de columnas, calcular métricas agregadas.

**Glue Crawlers:**

* + Automatizan la catalogación de datos almacenados en S3.
  + Generan o actualizan esquemas en el **Glue Data Catalog**.

**Integración con Lake Formation:**

* + Glue respeta las políticas de Lake Formation para garantizar acceso seguro.

**Orquestación del Pipeline**

AWS Glue también actúa como el orquestador principal para coordinar tareas y dependencias.

**Glue Workflows:**

* + Organizan los Glue Jobs, Crawlers y otras tareas en un pipeline estructurado.
  + Ejemplo de flujo:
    - **Job 1:** Extrae datos de Local y los almacena en S3.
    - **Job 2:** Limpia y transforma datos en S3.
    - **Crawler:** Actualiza el Glue Data Catalog con los datos transformados.

**Triggers y Programación:**

* + Workflows y Jobs se activan con triggers:
    - Basados en tiempo ( cada noche a las 10:00 PM).
    - Basados en eventos (carga de nuevos datos en S3 ).
    - Garantiza la correcta ejecución e los job

**Consulta y Consumo de Datos**

Los datos procesados y catalogados están disponibles para análisis y reporting.

**Athena para Consultas SQL:**

* + Los usuarios pueden ejecutar consultas SQL directamente sobre el Data Lake.

**Diagrama del Pipeline**

El pipeline puede visualizarse como un flujo:

**Ingesta:** Fuentes de datos → Glue ETL Jobs → S3 (raw).

**Transformación:** S3 (raw) → Glue ETL Jobs → S3 (curated).

**Catalogación:** Glue Crawlers → Glue Data Catalog → Lake Formation.

**Consumo:** Lake Formation → Athena/BI Tools.

1. Indica que proceso es necesario seguir para configurar permisos y políticas para los diferentes servicios de AWS utilizados.

Creo que antes de configurar los permisos es necesario

* **Identificar los recursos** involucrados (S3 buckets, Glue Jobs, RDS,likeformation).
* **Determinar los roles o usuarios** que accederán a estos recursos.
* **Definir las acciones permitidas** (lectura, escritura, administración).

Para ello es necesario en

* S3: Configurar los Bucket Policies, el Bloqueo de Acceso Público, la Encriptación y los IAM Policies
* Glue: configurar IAM Roles con acceso a S3, Glue Catalog, bases de datos, athena
* Likeformation

**Asignar un Administrador de Lake Formation:** Por defecto, la cuenta raíz tiene control total, pero es mejor asignar un usuario o rol específico como administrador.

**Registrar Ubicaciones de Almacenamiento (Buckets de S3):**

* + Registra los buckets de S3 donde estarán los datos del Data Lake.
  + En la consola de Lake Formation: Selecciona el bucket S3 y asocia un IAM Role con los permisos necesarios:

**Configurar AWS Glue Data Catalog:**

* + Habilita el control de acceso centralizado para los datos registrados en el Glue Data Catalog desde Lake Formation.
  + En la consola de Lake Formation, ve a **"Configuración"** y activa **"Usar Lake Formation para permisos de acceso"**.

Ve a la pestaña **"Permisos"** de Lake Formation.

Selecciona **"Otorgar"** y defina:

* La base de datos o tabla del Glue Data Catalog.
* Los permisos específicos (e.g., SELECT, INSERT, ALTER).
* El usuario, grupo, o IAM Role que recibirá los permisos.
* Opcional: Agrega condiciones como **particiones** o **columnas específicas**.

**Configurar Integración con Otros Servicios**

**Amazon S3**

* Confirma que los permisos de bucket S3 (Bucket Policies o IAM) no contradicen los permisos definidos en Lake Formation.
* Evita otorgar acceso completo a S3 para que los usuarios deban pasar por las políticas de Lake Formation.

**Athena**

* Habilita los permisos para que los usuarios puedan ejecutar consultas sobre el Data Lake.
* En Lake Formation:
  + Otorga permisos SELECT sobre las tablas necesarias.
* En IAM, asegura que el usuario o rol tenga: permisos de ejecución de querys y consulta de resultados

**Redshift Spectrum**

* Configura Lake Formation para gestionar las tablas externas en Redshift Spectrum.
* Asegúrate de que el IAM Role usado por Redshift tenga permisos para leer las tablas registradas.

**Automatización con IaC (Infrastructure as Code)**

Usa herramientas como **AWS CloudFormation**, **AWS CDK** o **Terraform** para definir y desplegar permisos de Lake Formation de forma consistente.

**Ejemplo con AWS CloudFormation:**

**Buenas Prácticas**

* **Principio de Menor Privilegio:** Otorga acceso granular solo a los datos necesarios.
* **Control de Acceso Centralizado:** Usa Lake Formation como la única fuente de permisos, evitando conflictos con Bucket Policies o permisos directos de Glue.