Configuraciones

En el problema planteado, hay varias configuraciones importantes que se deben tomar en cuenta para garantizar el correcto funcionamiento del flujo de datos entre los servicios de Google Cloud, MySQL y BigQuery.

A continuación, se describen las configuraciones a considerar:

1. **Configuración del Bucket en Google Cloud Storage (GCS)**

    - Descripcion: Asegurarse de que el bucket este correctamente configurado para almacenar los archivos CSV.

    - Pasos:

        i. Crea el bucket en Google Cloud Storage.
        ii. Organizar los archivos CSV dentro de las carpetas correspondientes (central, sucursal1, sucursal2).
        iii. Verificar permisos de acceso al bucket para permitir que Dataflow lea los archivos (configuración de IAM y permisos del bucket).
        iv. Si el proceso de configuración causa conflictos, se puede consultar la documentación de Google para la configuración de IAM en buckets (https://cloud.google.com/storage/docs/access-control/using-iam-permissions?hl=es-419).

2. **Configuración de Dataflow para Procesamiento de archivos CSV**

    - Descripción: Crear y configurar un pipeline de Apache Beam que se ejecutara en Dataflow para leer y procesar los archivos CSV desde GCS.

    - Pasos:

        i. Habilitar la API de Dataflow en el proyecto de Google Cloud.
        ii. Definir el pipeline de Dataflow con las siguientes etapas:

            - Source: Leer archivos CSV desde el bucket creado.
            - Transform: Parsear los archivos CSV a un formato estructurado (JSON, tablas).
            - Sink: Inssertar los datos parseados en las tablas correspondientes en MySQL.

        iii. Verificar que la configuración de la red y las credenciales de acceso permitan la conectividad entre Dataflow y MySQL.
        iv. Si el proceso de configuración causa conflictos, se puede consultar la documentación de Google para configurar las opciones de canalización de Dataflow (https://cloud.google.com/dataflow/docs/guides/setting-pipeline-options?hl=es-419).

3. **Configuración de MySQL**

- Descripción: Configurar el servidor MySQL para recibir los datos procesados por Dataflow.
- Pasos:
    i. Configurar MySQL en la instancia deseada, asegurarse de tener creadas las tablas que recibíran los datos.
    ii. Activar el registro binario (binlog) es MySQL, para capturar los eventos de CDC:
        - binlog_format=ROW
        - log_slae_updates=true
        - expire_logs_days=7
    iii. Asegurarse de que la instancia de MySQL este accesible desde la red para permitir conexiones entrantes desde Dataflow y Datastream.
    iv. Si el proceso de configuración causa conflictos, se puede consultar la documentación de Google para conectarse con un cliente MySQL (https://cloud.google.com/sql/docs/mysql/connect-admin-ip?hl=es-419).

4. **Configuración de Datastream para replicación de datos**

- Descripción: Configurar Google Cloud para replicar los datos desde MySQL hacia BigQuery utilizando CDC (Captura de Datos de Cambio).
- Pasos:
    i. Habilitar la API de Datastream en el proyecto de Google Cloud.
    ii. Crear un perfil de conexión de BigQuery, proporcionando la dirección IP, el puerto, nombre de usuario y contraseña de la instancia de MySQL.
    iii. Crear un perfil de conexión a BigQuery, proporcionando los detalles del poyecto, dataset y tabla de destino.
    iv. Configurar el flujo de replicación de Datastream:
        - Elegir el tipo de replicación: CDC o replicación completa + CDC.
        - Definir los objetos de MySQL que se replicarán (tablas especificas o bases de datos enteras)
    v. Configurar el destino en BigQuery especificando el dataset y la estructura de las tablas.
    vi. Si el proceso de configuración causa conflictos, se puede consultar la documentación de Google para consultar las opciones de conectividad de red (https://cloud.google.com/datastream/docs/network-connectivity-options?hl=es-419).

5. **Configuración de BigQuery como destino de replicación**

    - Descripción: Configurar el dataset en BigQuery que recibirá los datos replicados desde MySQL mediante Datastream.
    - Pasos:
        i. Crear el dataset en Bigquery.
        ii. Asegurarse de que las tablas en BigQuery coincidan en estructura con las tablas replicadas desde MySQL.
        iii. Habilitar permisos de IAM para Datastream y asegurarse de que Datastream tenga los permisos correctos para escribir en el dataset de BigQuery.
        iv. Monitorear las tareas de replicación en Datastream y verificar que los datos se están replicando correctamente en las tablas de BigQuery.

6. **Configuración de Seguridad y Permisos**

    - Descripción: Asegurarse de que todos los componentes puedan comunicarse entre sí de manera segura.
    - Pasos:
        i. Configurar correctamente las cuentas de servicio en Google Cloud para que Dataflow, Datastream y otros servicios puedan interactuar con GCS, MySQL y BigQuery.
        ii. Definir las reglas de firewall para permitir el acceso de Dataflow y Datastream a la instancia de MySQL.
        iii. Verificar los permisos de IAM para el acceso de los servicios a BigQuery, GCS y MySQL.

7. **Monitoreo y mantenimiento**

    - **Descripción:** Implementar las herramientas de monitoreo para supervisar el pipeline de Dataflow y la replicación de Datastream.
    - **Pasos:**
        i. Habilitar Stackdriver Monitoring o Cloud Monitoring para seguir el desempeño del pipeline de Dataflow y las réplicas de Datastream.
        ii. Configurar alertas para posibles fallos en la replicación o procesamiento de datos.
        iii. Revisar los logs de Dataflow y Datastream para asegurarse de que los flujos estén funcionando correctamente.

Estas son las configuracinoes clave que se deben de tener en cuenta para el flujo competo de procesamiento de datos desde GCS hasta BigQuery.