Practical Lesson: Data Replication using Google Datastream

Hint for Naming Resources

When naming resources, use the format `de-2023-[RESOURCE NAME]-[YOUR NAME]` where `**[YOUR NAME]**` is your account name. Adjust the `**[RESOURCE NAME]**` accordingly based on the specific resource you're creating.

Exercise: Replicating Data from Cloud SQL to BigQuery using Datastream

In this exercise, you'll set up Google Datastream to replicate data from your Cloud SQL instance to BigQuery.

Prerequisites

- Completion of the previous exercise on setting up Cloud SQL and CDC.
- A Google Cloud Platform account.

Step 1: Set Up Datastream

1. **Navigate to Datastream**:
    - Open Google Datastream Console.
2. **Set up PostgreSQL Connection Profile**:
    - Click on `Connection Profiles`.
    - Choose `Create` followed by `PostgreSQL`.
    - Define the connection settings:
        1. Name & ID: `de-2023hs-psql-[SHORTNAME]`.
        2. Region: Europe-west6 (Zürich).
        3. IP: Use the public IP address from your Cloud SQL Instance.
        4. User/Password: Credentials from your Cloud SQL Instance.
        5. Database: `adventureworks`.
        6. Confirm all other fields.
        7. Click `Create`.
3. **Set up BigQuery Connection Profile**:

- In `Connection Profiles`, select `Create` followed by `BigQuery`.
- Define the connection settings:
    1. Name & ID: `de-2023hs-bq-[SHORTNAME]`.

4. **Create a Datastream**:
    - Click on `Streams`.
    - Choose `Create Stream`.
    - Define the stream settings:
        1. Stream name & ID: `de-2023hs-psql-bq-[SHORTNAME]`.
        2. Region: Europe-west6 (Zürich).
        3. Source type: PostgreSQL.
        4. Destination type: BigQuery.
        5. Click `Continue`.
        6. Select your Postgres Connection Profile as the source and run a test.
        7. Click `Continue`.
        8. Replication slot name: `psqlreplslot`.
        9. Publication name: `psqlrepl`.
        10. Select the following schemas/tables:
            - sales.customer
            - sales.salesorderheader
            - sales.salesorderdetail
            - person.person
            - person.address
            - person.stateprovince
            - person.countryregion *(Note: You can add more tables, but initial replication might take longer)*
        11. Click `Continue`.
        12. Select your BigQuery Connection Profile as the destination.
        13. Click `Continue`.
        14. Region: Europe-west6 (Zürich).
        15. Selection option: `single Dataset`.
        16. Create a dataset named `adventureworks` in the Zurich region.
        17. Staleness limit: 5 minutes.
        18. Click `Continue`, run validation, then `Create & Start`.

5. **Verify Replication**:
    - Navigate to BigQuery.
    - Check the destination tables. It might take up to 5 minutes for the data to appear.