Business Intelligence - DataFusion

In this exercise, we will create a DataFusion instance and set up a pipeline to transfer data from Cloud-SQL to BigQuery. Follow the steps below:

Step 1: Create Instance

1. **Enter Name**: Provide a suitable name for your instance (e.g. 'bi-hs2023-df-[SHORTNAME]").
2. **Select Region**: Choose "Zürich" from the list.
3. **Edition**: Go with "Enterprise".
4. **Wait**: Once you've filled in the details, your instance will begin initializing. This might take up to 20 minutes.

Step 2a: Increase Your Quota

1. Navigate to IAM Quotas or from the GCP console, go to IAM → Quotas.
2. Use the filter option and search for "Compute Engine".
3. Increase the following quotas:
    - Persistent Disk Standard (GB): Set to 20000. (Firlter for "Compute Engine", "region:europe-west6", "disk" –> scroll down to "Persistent Disk Standard")
    - CPUs: Set to 40. (Firlter for "Compute Engine", "region:europe-west6", "CPUS" –> scroll to "CPUS")
    - In-use IP addresses: Set to 24. (Firlter for "Compute Engine", "region:europe-west6", "IP" –> scroll to "In-use IP addresses")
    - CPUs (all regions): Set to 60. (Firlter for "Compute Engine", "all regions" –> scroll to "In-use IP addresses")
4. Wait for an approval email confirming your quota increase.

Step 2b: set authorization

1. Go to IAM & Admin → IAM.
    1. Select "Include Google-provided role grants"
    2. Choose "Data Fusion Service Account" (xxx@gcp-sa-datafusion.iam.gserviceaccount.com) and then "Edit" (small pen on the right).
    3. Add Role "Cloud SQL Client" and then "Save".
    4. Choose "Compute Engine Default Account" (xxx@developer.gserviceaccount.com)

5. Make sure the user has following roles: “Cloud Data Fusion Runner”, “Cloud SQL Client”, “Dataproc Worker” and “Editor”

Step 3: View Instance

Once your instance is up, click on the “Instance” link to view it and ensure everything is set correctly. A separate window might open; if there are any pop-ups, confirm them to proceed.

Step 4: Hub

In the new window, navigate to the “Hub” section located on the top right.

Step 5: Search for PostgreSQL

In the search bar, input “PostgreSQL” to find relevant plugins and drivers.

Step 6: CloudSQL PostgreSQL JDBC Driver

1. Follow the provided instructions to **Download** the necessary file.
2. **Deploy** the driver: Simply drag and drop the downloaded file into the interface and press “Finish”.

Step 7: CloudSQL PostgreSQL Plugins

1. Press “Deploy” and then confirm by pressing “Finish”.

Step 8: Studio

1. Navigate to the “Hamburger Menu” and select “Studio”.

Step 9: Source Configuration

2. In the left pane, select “CloudSQL PostgreSQL” as your source.
3. Within the CloudSQL frame, select **Properties**.
    1. Set “Use connection” to “YES”.

2. Click on “Browse Connections” and then “Add Connection”.
3. Choose “CloudSQL PostgreSQL” and give it the name `cloudsql-postgresql-conn`.
4. For the JDBC driver, there should be only one option available. Select it.
5. Set the database to “adventureworks”.
6. Enter your connection details, including user and password.
7. Test your connection and then press “Create”.
8. Add a reference name (e.g. `psql-customer-bq`).
9. Input the following query: `select * from sales.customer;`
10. Press “Get Schema”.
11. Exclude the “rowguid” column by deselecting it. Select all other fields.
12. Validate the configuration and then close the window.

Step 10: Sink Configuration

1. From the left pane, select “BigQuery” as your sink.
2. Within the BigQuery frame, select **Properties**.
    1. Set “Use connection” to “YES”.
    2. Browse and select the “BigQuery Default” connection.
    3. Choose the “adventureworks” dataset.
    4. Set the table name to “customer”.
    5. Set “Truncate table” to “YES”.

Step 11: Connect Source to Sink

Draw a line connecting the “PostgreSQL” box to the “BigQuery” box.

Step 12: Pipeline Configuration

1. Provide a suitable name for your pipeline (e.g. `postgres-customer-bq`).
2. Save your pipeline configuration.

Step 13: Deployment

Deploy the pipeline to get it ready for execution.

Step 14: Execute Pipeline

Run the pipeline and wait for it to finish. This might take between 3-5 minutes.

Step 15: Verify Data Transfer

1. Navigate to BigQuery in your GCP console.
2. Check for the presence of the “customer” table in the “adventureworks” dataset.
3. Verify that the table contains the expected data.

Import Pipelines

1. Go to the “Hamburger Menu” (three horizontal bars) on the left top side.
2. Click on the big green “+” button on the top right.
3. From Pipeline, select “Import”.
4. Select the pipeline to import —> the pipeline opens in the designer.
5. Click “Deploy”.
6. Repeat for all pipelines.

> **⚠ Important Warning ⚠**
>
> **Don’t forget to stop your Cloud SQL instance when you are done.** Otherwise, it will use up all your Cloud Credits, and you can’t continue with the course!

> **⚠ Important Warning ⚠**
>
> **Don’t forget to delete your DataFusion when you are done.** Otherwise, it will use up all your Cloud Credits, and you can’t continue with the course!
