Practical Lesson: Getting Started with Dataform in BigQuery

Dataform is a tool that allows data teams to transform data directly in their BigQuery data warehouse. In this lesson, we will walk through the steps to set up Dataform, create a repository, and run a SQL query.

Step 1: Access Dataform and Enable the API

1. Navigate to BigQuery Dataform.
2. If prompted, enable the Dataform API. This process may take 2-3 minutes.
3. Open the Google Cloud Shell located at the bottom of the Google Cloud Console.

Step 2: Configure IAM Permissions

Run the following commands in Cloud Shell to grant the necessary permissions:

```
PROJECT_NUMBER=$(gcloud projects describe $GOOGLE_CLOUD_PROJECT
 ↪  --format="value(projectNumber)")

gcloud projects add-iam-policy-binding $GOOGLE_CLOUD_PROJECT \
    --member=serviceAccount:service-$PROJECT_NUMBER@gcp-sa-
     ↪  dataform.iam.gserviceaccount.com
     ↪  \
    --role=roles/bigquery.admin
```

Step 3: Create a Repository

1. In Dataform, click on “Create repository”.
2. Choose a repository ID in the format `zhaw-cas-de-2023-dataform-[SHORTNAME]`.
3. Select the region `europe-west6` (Zürich).
4. Click “Create”.

Step 4: Initialize a Development Workspace

1. Go to “Repositories” and select the repository you just created.
2. Click “Create development workspace”.
3. Enter a workspace ID in the format `zhaw-cas-de-2023-workspace-[SHORT NAME]`.
4. Select your workspace and click “Initialize Workspace”.

Step 5: Create and Execute a SQL File

1. In the workspace, delete the example files `first_view` and `second_view`.
2. Open `dataform.json` and update the following:
   - `"defaultSchema": "zhaw_adventureworks"`
   - `"defaultLocation": "europe-west6"`
3. Click on the three dots next to “Definitions” and choose “Create file”.
4. Name the file `sh-staging-view.sqlx`.
5. Copy and paste the SQL code **replace [YOUR PROJECT].[YOUR DATASET].[SALESORDERHEADER TABLE] with yours**:
6.
7.

```
config {
type: "view", // Creates a view in BigQuery. Try changing to "table"
 ↪  instead.
}

-- This is an example SQLX file to help you learn the basics of
 ↪  Dataform.
-- Visit https://cloud.google.com/dataform/docs/sql-workflows for more
 ↪  information on how to configure your SQL workflow.
-- You can delete this file, then commit and push your changes to your
 ↪  repository when you are ready.
-- Config blocks allow you to configure, document, and test your data
 ↪  assets.
-- The rest of a SQLX file contains your SELECT statement used to
 ↪  create the table.
-- Selecting address and region information

SELECT
salesorderid AS order_id,
revisionnumber AS revision_number,
orderdate AS order_date,
duedate AS due_date,
shipdate AS ship_date,
status AS order_status,
customerid AS customer_id,
salespersonid AS salesperson_id,
territoryid AS territory_id,
```

```
billtoaddressid AS billing_address_id,
shiptoaddressid AS shipping_address_id,
shipmethodid AS shipping_method_id,
creditcardid AS credit_card_id,
creditcardapprovalcode AS credit_card_approval_code,
currencyrateid AS currency_rate_id,
SAFE_CAST(subtotal AS FLOAT64) AS subtotal_amount,
SAFE_CAST(taxamt AS FLOAT64) AS tax_amount,
SAFE_CAST(freight AS FLOAT64) AS freight_amount,
SAFE_CAST(subtotal AS FLOAT64) + SAFE_CAST(taxamt AS FLOAT64) +
↪  SAFE_CAST(freight AS FLOAT64) AS total_amount,
comment AS order_comment,
TIMESTAMP_DIFF(shipdate, orderdate, DAY) AS days_to_ship
FROM
`[YOUR PROJECT].[YOUR DATASET].[SALESORDERHEADER TABLE]`
WHERE
SAFE_CAST(subtotal AS FLOAT64) IS NOT NULL
AND SAFE_CAST(taxamt AS FLOAT64) IS NOT NULL
AND SAFE_CAST(freight AS FLOAT64) IS NOT NULL
AND orderdate IS NOT NULL
AND shipdate IS NOT NULL
```

8. Commit the changes.
9. Click “Start execution” to run the query.
10. Go to “All executions” and click “Execute”.