Adding Google BigQuery Data Warehouses

A Google BigQuery data warehouse is an instance of Google BigQuery that contains the tables and views that you want to access as model facts and dimensions. It also contains aggregate table instances that either you or the AtScale engine creates in a BigQuery dataset that you specify.

Before You Begin

- Ensure that you are logged in to AtScale as an admin user.
- Aggregate Dataset: Ensure that you know the BigQuery dataset and ProjectID to use for aggregate tables to be built in the data warehouse. AtScale reads and writes aggregate data to this dataset. The AtScale service account user must have ALL privileges for this dataset.
- Assign your Google BigQuery service account these roles: `BigQuery Data Editor` and `BigQuery Job User`. The service account used will be the one used for setting up system query role. If aggregates are being stored in a project different from the project where fact data resides, then this service account needs `BigQuery Data Viewer` and `BigQuery Job User` for all the projects that contain fact data.
- The Google BigQuery service account used for setting up large/small query roles needs to have read access on all the projects with Fact data, plus on the project where aggregate dataset resides. At the minimum, the `BigQuery Data Viewer` and `BigQuery Job User` roles.
- Consider that Google BigQuery Views cannot be previewed in Design Center.

Authentication

You can choose to authenticate with a JSON credential file, or with Application Default Credentials (ADC)/Environment-provided service account. For additional details about Google authentication types, see Google Cloud Authentication. In both cases you need to perform some steps in advance.

When using the JSON credential file for your Google Cloud Platform service account:

- Ensure that the file can be accessed by AtScale.
- The file for the service accounts can be created from the Google Cloud Platform console from the Menu > IAM & admin > Service accounts. The file contains seven keys and their associated values. For example:

{ "name": "projects/my-project-123/serviceAccounts/my-sa-123@my-project-123.iam.gserviceaccount.com", "projectId": "my-project-123", "uniqueId": "113948692397867021414", "email": "my-sa-123@my-project-123.iam.gserviceaccount.com", "displayName": "my service account", "etag": "BwUp3rVlzes=", "oauth2ClientId": "117249000288840666939" }

The authentication that is based on Google Compute Node service account credentials can be used with the following setup:

- A user-managed service account is attached to the Google Compute Node that runs AtScale.
- The service account has access to project(s) or impersonates another service account that has access in Google BigQuery.
- If the service account has access to multiple Google BigQuery projects, then multiple AtScale data warehouses can be set up corresponding to the needed projects.

You can use the following commands to check which projects the compute node service account has access to:

- `gcloud config list`
- `gcloud projects list`

ADC-Based Authentication From On-Premise Node

The `GOOGLE_APPLICATION_CREDENTIALS` environment variable is used for providing the location of a credential JSON. You can use it in AtScale in the following way:

1. Update the ~/.bash_profile file of the `atscale` user, and set the environment variable `GOOGLE_APPLICATION_CREDENTIALS` to point to the location of the service account key or credential file:

```
cat ~/.bash_profile | grep GOOGLE
export GOOGLE_APPLICATION_CREDENTIALS=/home/atscale/SAcctCreds.json
```

2. Reload `bash_profile` using the `source` command, or start a new terminal.

```
source ~/.bash_profile
```

3. Restart all AtScale services; this needed to pick up the environment variable:

```
/opt/atscale/bin/atscale_stop
/opt/atscale/bin/atscale_start
```

4. Check if the environment variable is picked up:

```
cat /proc/`/opt/atscale/current/pkg/jdk/bin/jcmd -l | grep engine | awk '{print $1}'`/environ | tr '\0' '\n' | grep GOOGLE
```

You should get a result like this:

```
GOOGLE_APPLICATION_CREDENTIALS=/home/atscale/SAcctCreds.json
```

Procedure

To add a new Google BigQuery data warehouse:

1. In Design Center, open the main menu and select **Settings**. The **Settings** page opens.
2. In the **Settings** panel, click **Data Warehouses**.
3. In the **Add Data Warehouse** section, click the **BigQuery** icon. The **Add Data Warehouse** panel opens.
4. Enter a unique **Name** for the data warehouse.
5. (Optional) Enter the **External connection ID**. The value of this field defaults to the data warehouse name but can be overriden by enabling the **Override generated value?** option.
6. In the **Aggregate Project ID** field, enter name of the ProjectID in which the Aggregate Schema resides.

   The service account used for setting up system query role will be the one that creates aggregates and needs have All privileges on the project.
7. In the **Aggregate Schema** field, enter the name of the BigQuery dataset to use when creating aggregate tables. You must create a dataset; AtScale does not create one automatically.
8. (Optional) Specify whether the data warehouse is a **Read-only source**.

   - Select this option if AtScale will be configured to access this database with credentials that do not have permission to create new tables, or if you need to prevent AtScale from placing any aggregate tables in this data warehouse.
   - If **Read-only source** is enabled upon editing an existing data warehouse with aggregates present, all existing aggregates stored in this data warehouse will be deactivated.

9. (Optional) Enable impersonation for the data warehouse by selecting the **Enable to impersonate the client when connecting to data warehouses** checkbox. This allows the data warehouse administrator to apply more detailed security policies when securing data. When enabled, you can also enable the following options:

- **Always use Canary Queries**: Determines whether canary queries are required in the event of aggregate misses.
- **Allow Partial Aggregate Usage**: Enables mixed aggregate and raw data queries.

> AtScale cannot import data-loader bundles while impersonation is enabled. It is suggested that you import your desired data-loader bundles, commonly used for training, before enabling impersonation or use a separate AtScale installation for training purposes.

10. In the **Authentication** section, select one of the following sources for the Service Account:

    - **Use service account key JSON file**: When set, you need to upload the JSON credential file as described below.
    - **Use Application Default Credentials (ADC)**
    - **Use Application Default Credentials (ADC) impersonated with Service Account**

11. (Optional) Specify a connection to a Google Cloud Bucket:

    1. Enable the **Connect a Google Cloud Bucket** option.
    2. If you selected **Use service account key JSON file** for authentication, upload the JSON file in the **Service account key json file** field.
    3. If you have chosen **Use Application Default Credentials (ADC) impersonated with Service Account** for authentication, enter the **Service principal**.
    4. Enter the **Bucket name**. Ensure that the bucket you use is multi-regional.

12. Click **Test Bucket Connection** to test the data warehouse connection.
13. Click **Apply**.

Add A GBQ Connection

After setting up your data warehouse, you must add a connection to the data warehouse before you can run queries.

To add a Google BigQuery connection:

1. On the **Data Warehouses** page, click the expand icon for your data warehouse.
2. In the **Add Connection** field, click the plus icon. The **Add Big Query Connection** panel opens.
3. Enter a unique **Name** for the connection.
4. If authenticating with a JSON credential file:

    1. (Optional) Enable the **Project Id for query execution** option, then enter the BigQuery project in which queries are executed. This value is passed when creating a job. If not specified, AtScale uses the project associated with the service account key JSON.
    2. Upload the file associated with your Google BigQuery Connection.

5. (Optional) If authenticating with ADC, enable the **Project Id for query execution** option, then enter the BigQuery project in which queries are executed. This value is passed when creating a job. If you do not provide this value, AtScale uses either the projectID associated with the service account attached to the compute node, or the service account defined by the `GOOGLE_APPLICATION_CREDENTIALS` variable.
6. If authenticating with ADC impersonated with Service Account:

    1. (Optional) Enable the **Project Id for query execution** option, then enter the BigQuery project in which queries are executed. This value is passed when creating a job. If you do not provide this value, AtScale uses either the projectID associated with the service account attached to the compute node, or the service account defined by the `GOOGLE_APPLICATION_CREDENTIALS` variable.
    2. Enter the **Service Principal**.

7. Click **Test connection** to test the connection.
8. Click **Apply**.

Multiple connections can be set up, such that aggregate creation is handled by a different service account than the one that handles user query execution.