Adding Datasets

The first step in designing an AtScale model is to import the physical tables and views that it will be based on.

About Reusing Datasets

Reusing datasets between the dimension Canvas and the model Canvas - or between multiple dimensions - can cause the engine to take unexpected join-paths at deployment time.

If you want to reference the same table from multiple points in the same model, it is recommended that you define separate Query Datasets (QDSs) that select from the desired table. The separate datasets should be used to back each dimension or fact model component. Doing so ensures the generation of runtime join-paths that are readily identified in Design Center.

Note that each QDS must be created from scratch; converting an existing dataset to a QDS does not satisfy the uniqueness requirement.

About This Task

If you are not sure what datasets you need for your model, you can drag them on to the main canvas to look at them first. Consider the following:

- Adding datasets to the repository

  You can build up a dataset in your repository by dragging all the datasets from the **Data Sources** panel to the main Canvas. The new datasets are automatically added to the `datasets/` folder in the repository. This allows you to preview columns, add calculated columns, and plan how you will model relationships between the datasets. You can then remove all the datasets from the Canvas and begin modeling by adding them one by one from the **Repo Browser**.

- Using a table partitioned in BigQuery

  If you are using Google BigQuery as a data warehouse, if you import as a fact dataset a table that is partitioned in BigQuery, the `\_PARTITIONTIME` pseudo-column is listed among the columns of the dataset. When you are modeling data, you can build a date dimension from this pseudo-column. When users working in client BI applications query your published model, they can include `WHERE` clauses on `\_PARTITIONTIME`; BigQuery will perform partition pruning when serving the results.

  For more information about the `\_PARTITIONTIME` pseudo-column, see the Google BigQuery documentation here.

Adding A Table-Based Dataset To The Repository

You can add table-based datasets to the repository by either dragging them from the **Data Sources** panel directly onto the Canvas, or by creating them via the **Edit Dataset** panel. Creating them via the panel gives you the opportunity to edit them before they are added.

To add a dataset via the **Edit Dataset** panel:

1. In the **Data Sources** panel, click a data source name to expand it.
2. Click a database or schema name to expand it.
3. Locate the database table you want to create a dataset from, click its context menu, and select **Create Dataset**. The **Edit Dataset** panel opens.
4. Complete the following fields:

    - **Unique name**: A unique name for the dataset. This must be unique across all repositories and subrepositories.
    - **Label**: A label for the dataset.
    - **Description**: A description of the dataset.
    - **Source Type**: Determines whether the dataset is based on a table or a query. Select **Table**.
    - **Schema**: Select the schema to base the dataset on.
    - **Table**: Select the table to base the dataset on.

5. Click **Apply**.

The dataset is added to the repository and appears in the `datasets/` folder in the **Repo Browser**. From here, you can edit it, or add it to models and dimensions as needed.

Adding A Query Based Dataset To The Repository

A query-based dataset allows you to select data from one or more tables in your data warehouse and save the query in AtScale as a new dataset.

This enables you to manipulate the tables you have in your data warehouse and select just the rows you are interested in for your model. For example, you can select only the columns you need, join rows from multiple tables, and do other data manipulation functions. Query-based datasets do not alter your data warehouse.

> **Note:** You cannot add calculated columns to a query-based dataset after it is created. Do all of the required data manipulation in the `SELECT` list of the query.

To add a query-based dataset to the repository:

1. In Design Center, open the **Data Sources** panel.

2. Locate the table you want to create a dataset from, click its context menu, and select **Create Dataset**. The **Edit Dataset** panel opens.
3. Complete the following fields:

   - **Unique name**: A unique name for the dataset. This must be unique across all repositories and subrepositories.
   - **Label**: A label for the dataset.
   - **Description**: A description of the dataset.
   - **Source Type**: Determines whether the dataset is based on the physical table or a query. Select **Query**.

4. In the **Selection SQL** field, enter the SQL query you want to use to create the dataset.
5. Click **Preview SQL**.

   > Note: This button is disabled if your data warehouse is an instance of Google BigQuery. The reason is that Google BigQuery does not respect the LIMIT keyword in SQL statements that would limit the size of the result set in the preview. Rather, BigQuery performs a full table scan, which is costly.

   If results appear, the SQL statement is valid. If you see an error or do not see data, double-check your SQL syntax and try running the statement directly against your data warehouse.

6. Click **Apply**.

The dataset is added to the repository and appears in the `datasets/` folder in the **Repo Browser**. From here, you can edit it, or add it to models and dimensions as needed.

Adding A Dataset To A Model

1. Open the model you want to add datasets to and switch to the **Canvas** tab.
2. In the **Repo Browser**, locate the dataset you want to add and do one of the following:

   - Drag the dataset onto the model Canvas.
   - Click the context menu for the dataset and select **Add to model**.

The dataset appears on the model Canvas.

Adding A Dataset To A Dimension

1. Open the dimension you want to add datasets to and switch to the **Canvas** tab.
2. In the **Repo Browser**, locate the dataset you want to add and do one of the following:

   - Drag the dataset onto the dimension Canvas.
   - Click the context menu for the dataset and select **Add to dimension**.

The dataset appears on the dimension Canvas.