Create Curated Data with Data Flows in Oracle Analytics

Before you Begin

This lab shows you how to modify columns and data in a data flow to create curated datasets in Oracle Analytics Cloud or Oracle Analytics Desktop. This tutorial uses a spreadsheet as the data source, however, you can use any supported data source.

Background

You might need to implement changes to you data before using that data in analyses. In a data flow, you can add, remove, change or merge columns, add calculations, modify the data contained in the columns, and create multiple datasets from one data flow. If you schedule the data flow to run periodically, you can capture updates in the data source, and enable persisting the transformations in your curated datasets.

After running the data flow, you can use the dataset to analyze the data by creating visualizations.

What Do You Need?

- Access to Oracle Analytics Cloud or Oracle Analytics Desktop
- Download the samp_revenue_denorm.xlsx to your computer

Create a Dataset and Data Flow

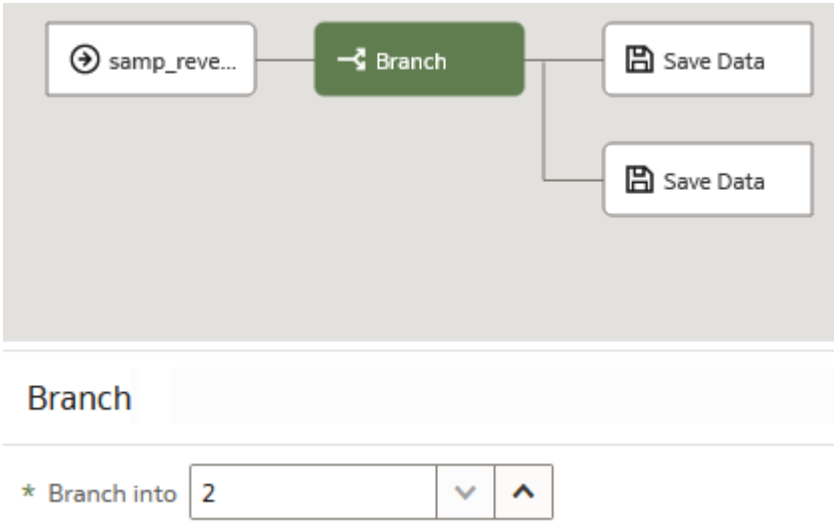
1. Sign in to Oracle Analytics Cloud.
2. On the Home page, click **Create**, and then click **Dataset**. In Create Dataset, click **Drop data file here or click to browse**, select the **samp_revenue_denorm.xlsx** file, and then click **Open**.
3. In Create Dataset Table from samp_revenue_denorm, click **OK**.
4. In the Join Diagram, click the **OFFICE_NUMBER** column, click **Measure** , and then click **Attribute**.
5. Click the **PROD_NUMBER** column, click **Measure** , and then click **Attribute**.
6. Click the **ORDER_NUMBER** column, click **Measure** , and then click **Attribute**.
7. Click **Save** . In Save Dataset As, enter samp_revenue_denorm, and then click **OK**.
8. Click the **samp_revenue_denorm** tab. In the dataset, select the **No data** column, select **Options** , and then select **Delete**.
9. Click **Go back** . In Save Changes, click **Save** .

Create Two Datasets

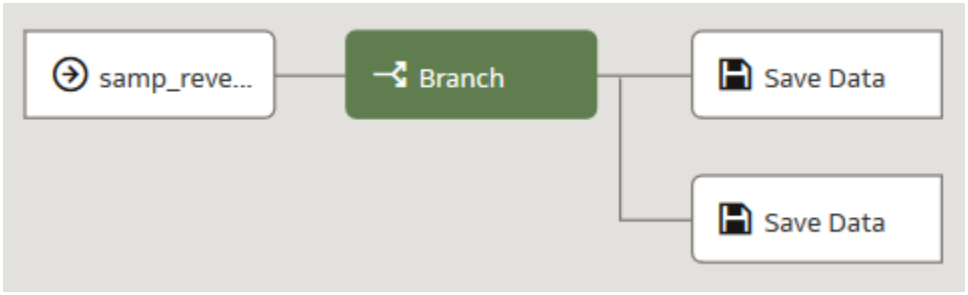
In this section, you use the sample_revenue dataset to create a PRODUCTS dataset.

1. On the Home page, click **Create**, and then click **Data Flow**. In Add Dataset, select **sample_revenue_denorm**, and then click **Add**.
2. From the Data Flow Steps panel, drag **Branch** to the **Add a step** node.

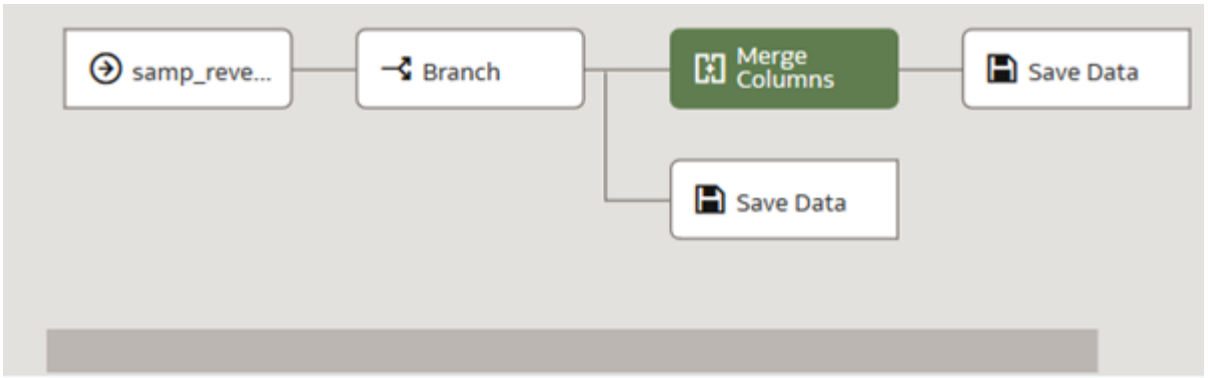
   Branch uses 2 as the default number of datasets created from the source dataset. You can increase the number of datasets created when the data flow is run.

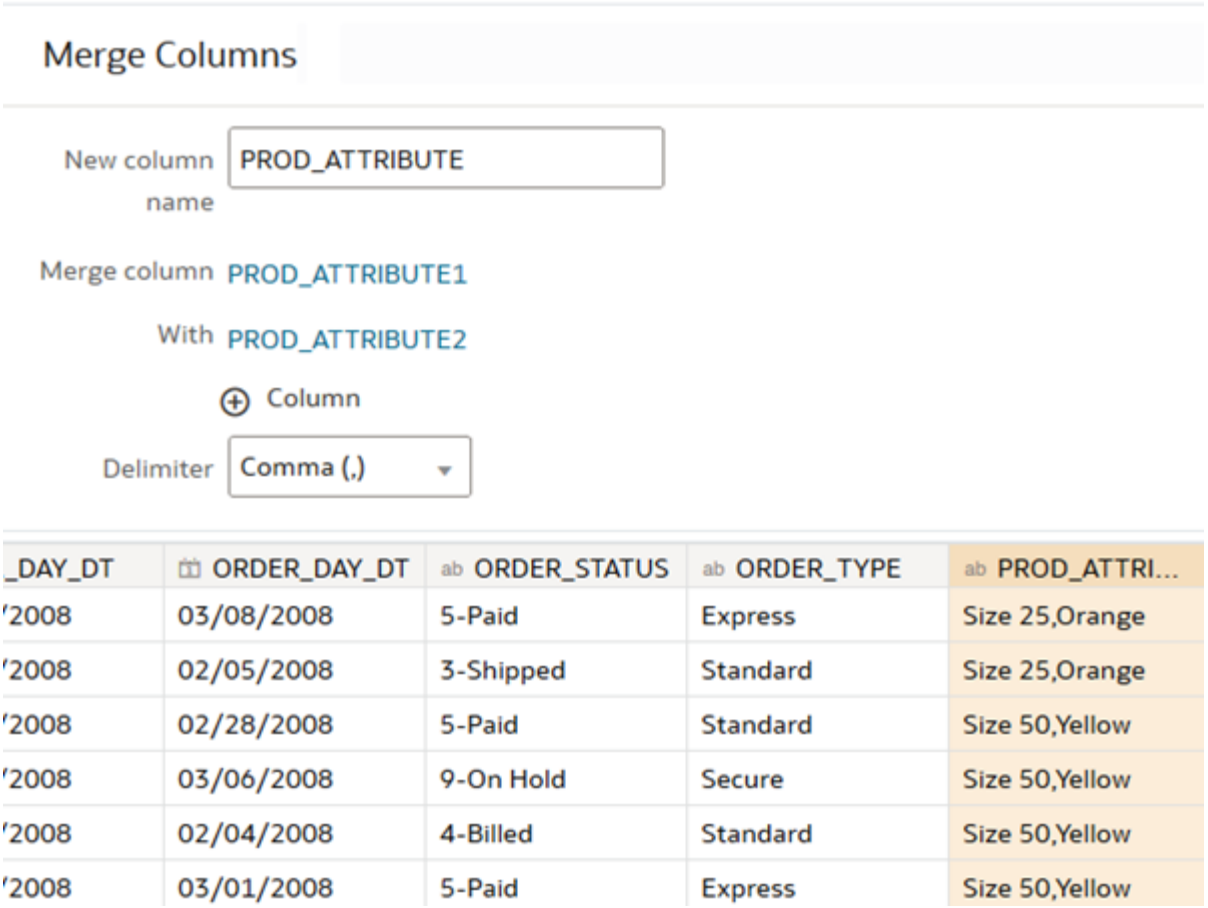


3. Click **Add a step** node on the top branch, and then select **Merge Columns**.



4. In Merge Columns, enter Prod_Attribute in **New column name**. Next to **Merge column**, click the hyperlink and then select **PROD_ATTRIBUTE1**. Next to **With**, click the hyperlink, and then select **PROD_ATTRIBUTE2**. From the **Delimiter** list, select **Comma (,)**.





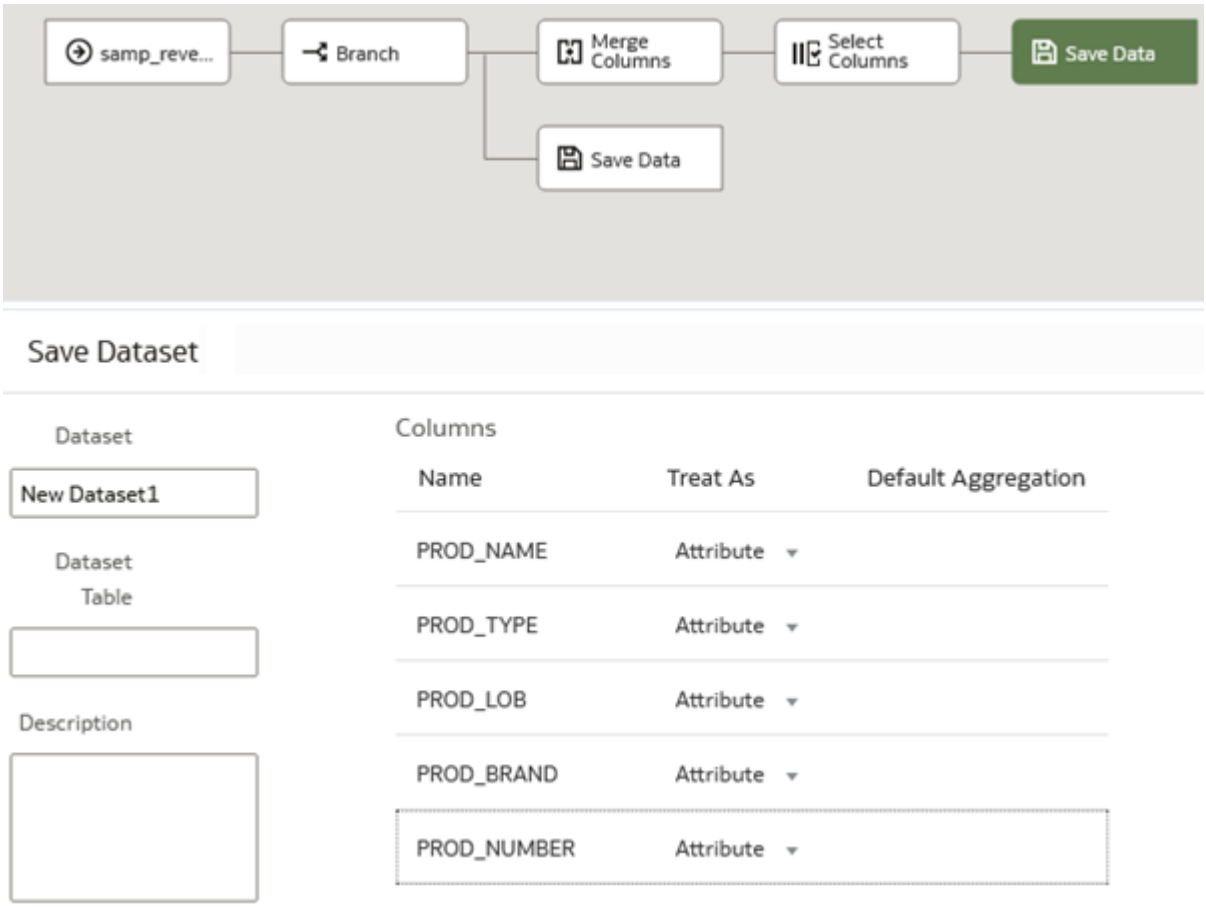
_DAY_DT	ORDER_DAY_DT	ORDER_STATUS	ORDER_TYPE	PROD_ATTRI...
'2008	03/08/2008	5-Paid	Express	Size 25,Orange
'2008	02/05/2008	3-Shipped	Standard	Size 25,Orange
'2008	02/28/2008	5-Paid	Standard	Size 50,Yellow
'2008	03/06/2008	9-On Hold	Secure	Size 50,Yellow
'2008	02/04/2008	4-Billed	Standard	Size 50,Yellow
'2008	03/01/2008	5-Paid	Express	Size 50,Yellow

Select Columns to Create a Dataset

In this section, you select the columns used to create a PRODUCTS dataset.

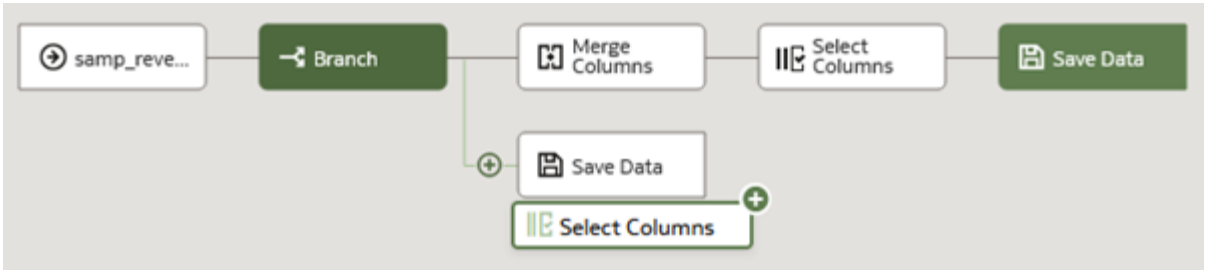
1. From the Data Flow Steps panel, drag **Select Columns** to **Add a step** between Merge Columns and the Save Data node.
2. In Select Columns, click **Remove all**. Hold down the **Ctrl** key and select the following columns:
   - **PROD_NAME**
   - **PROD_TYPE**
   - **PROD_LOB**
   - **PROD_BRAND**
   - **PROD_NUMBER**
   - **Prod_Attribute**

3. Click **Add selected**.
4. Click the top **Save Data** node. In Save Dataset, enter Products in the **Dataset** field. In the PROD_NUMBER row, click **Measure** in the Treat As column, and then select **Attribute**.



Create a Second Dataset

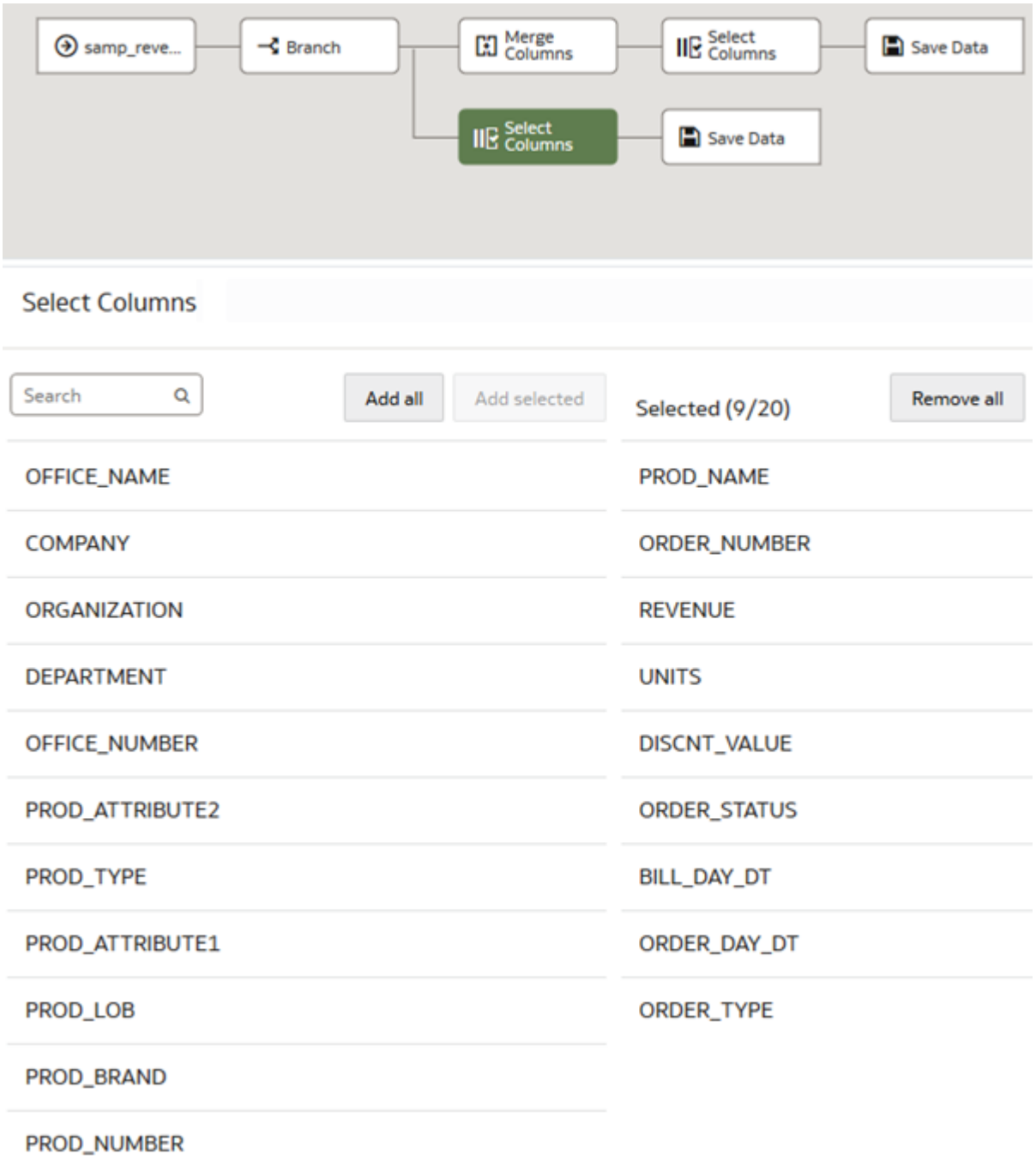
1. From the Data Flow Steps panel, drag **Select Columns** to **Add a step** ⊕ between the Branch and the second Save Data nodes.



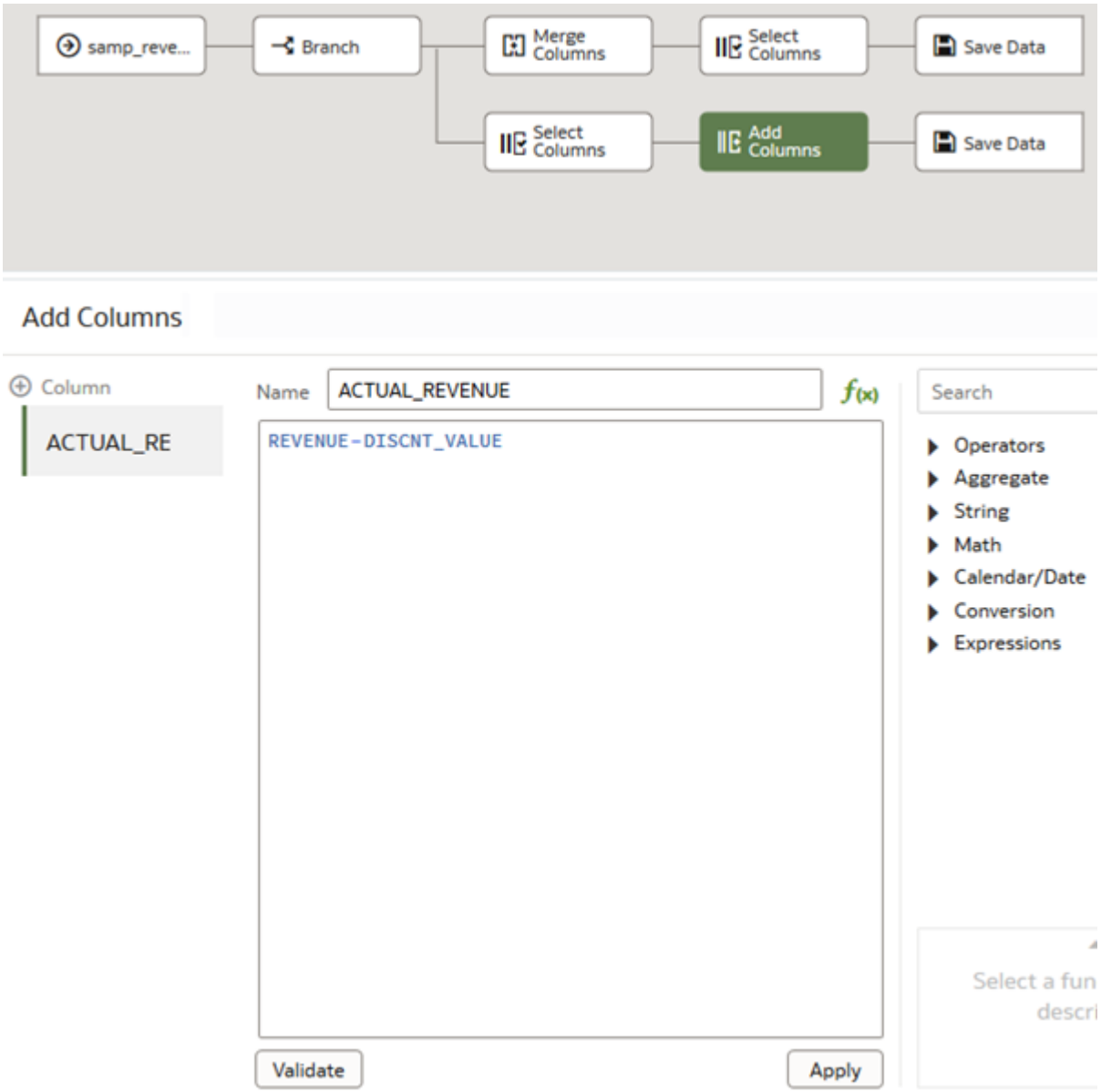
2. In Select Columns, click **Remove all**. Hold down the **Ctrl** key and select the following columns:
   - **PROD_NAME**
   - **ORDER_NUMBER**
   - **REVENUE**
   - **UNITS**
   - **DISCNT_VALUE**

- **BILL_DAY_DT**
- **ORDER_DAY_DT**
- **ORDER_STATUS**
- **ORDER_TYPE**

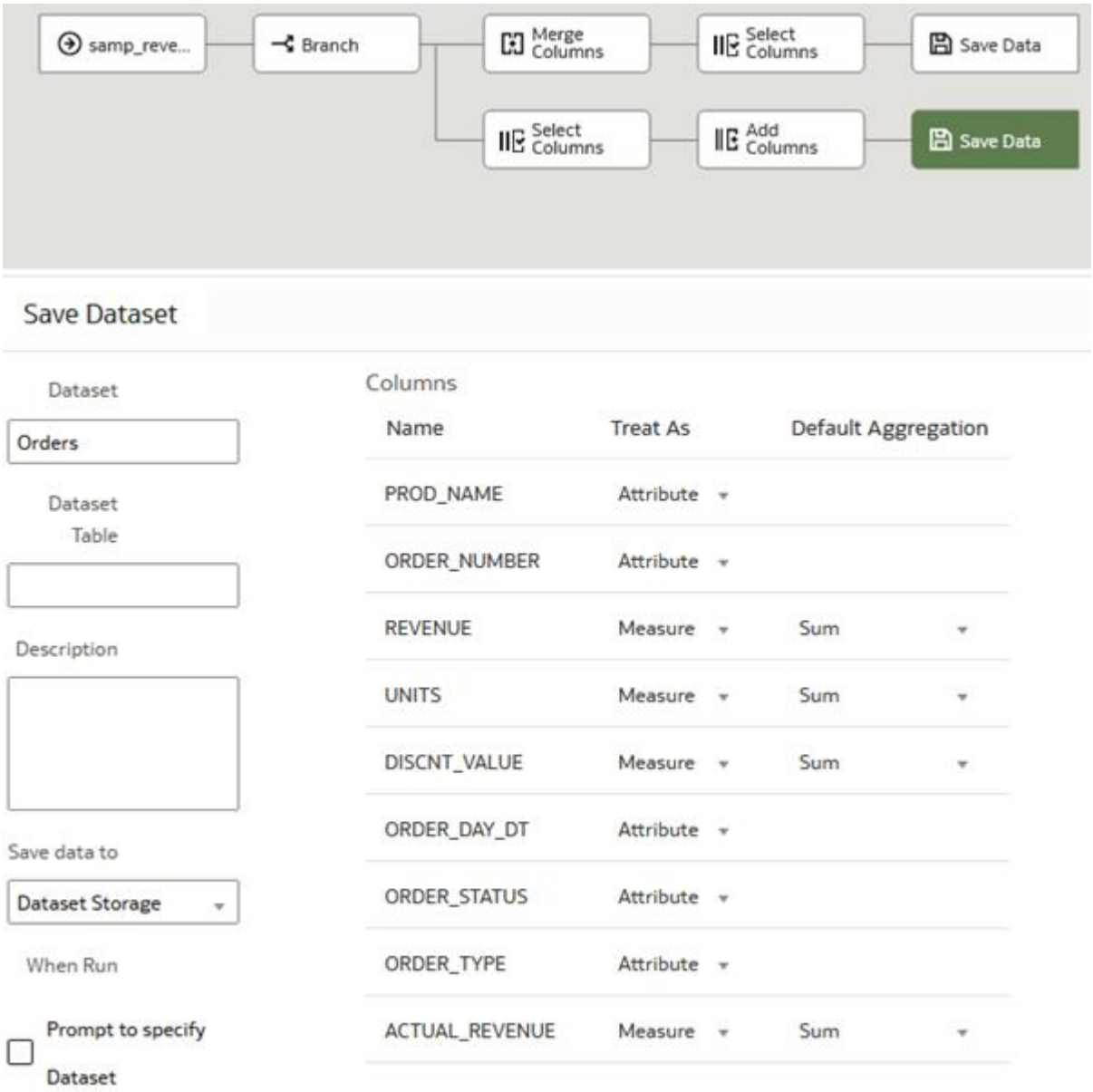
3. Click **Add selected**.



4. Drag **Add Columns** to the **Add a Step** node between Select Columns and Save Data. In Add Columns, enter ACTUAL_REVENUE in **Name**.
5. In the Expression field, start entering Revenue, and then select REVENUE from Available Data. Expand **Operators**, and double-click the **minus sign (-)**. After the minus sign, start entering DIS, and then select **DISCNT_VALUE** from Available Data.
6. Click **Validate**, and then click **Apply**.



7. Click the **Save Data** node on the branch with Add Columns. In Save Dataset, enter ORDERS in **Dataset**. In the ORDER_NUMBER row, click **Measure** in the Treat As column, and then select **Attribute**.



8. Click **Save**. In Save Data Flow As, enter Sample Revenue DF in Name, and then click **OK**.
9. Click **Run Data Flow**.
10. After the data flow run completes, click **Back**.
11. On the Home page, select **PRODUCTS**, click the **Action menu**, and then select **Inspect**. In the PRODUCTS dataset, click **Data Elements** to review the dataset. Click **Close**.
12. (Optional) On the Home page, select **ORDERS**, click the **Action menu**, and then select **Inspect**. In the ORDERS dataset, click **Data Elements** to review the dataset. Click **Close**.

Schedule the Data Flow Run

In this section, you schedule the data flow to run by defining the repetition, the duration, and interval. Your data might not change frequently, so you could define a schedule that meets your needs.

1. On the Home page, click the **Data** search tag, enter Sample_Revenue to located the data flow, and then click **Search**.
2. Select the **Sample_Revenue** data flow, click **Actions menu** ⋮, and select **New schedule**.
3. In Schedule, enter a **Name** for the schedule. In **Start**, click the calendar, and then select a month and day. In **Time**, click the clock to select the hour and minutes for the run's start time.
4. From **Repeat**, select **Hourly** as the frequency to use for running the data flow. In **End**, click the calendar to select the ending day for the data flow run schedule. From **Every**, select **1** to set the interval from running the data flow, and then click **OK**.