How to design a data flow (BETA)

A data flow consists of a data source, a chain of one or more operations that refine and shape that data source, and a target that the data moves to. As you design data flows, your activity is automatically saved.

Starting the Design Flow task

To start the Design Flow task, click **Design Flow** in the left navigation menu.

The Design Flow task comprises a toolbar along the top, the (initially blank) canvas where your nodes and data flows appear, and a right sidebar. See “Design Flow task basics” at the end of this document for detailed information about the actions you can take from each of these elements.

Adding data sources to data flows

To add a data source to a data flow:

1. On the Design Flow task canvas, click the empty canvas or the toolbar ⊕ **Add source** icon.
2. In the Select a Source window, select a connection from the **Connections** list.
   Tip: If you did not yet add a connection, click **Connections** in the left navigation menu, click **Add Connection**, and complete the Create Connection flow. Then begin again from “Starting the Design Flow task”.
3. Select a schema from the **Schemas** list.
4. Select one or more tables from the **Tables** list that you want to work with.
   By default, all columns in a selected table are included in the flow.
5. If you want to exclude any columns from the data flow, select each table in turn and use the **Columns** list to manage the columns .
6. Click **ADD**. The selected tables appear on the canvas as nodes.

Adding operations to data flows

Much of building a data flow involves shaping and refining the data to get it to your desired result. You start with a data source and then chain operations to shape the data. Operations can include joining two data sets, filtering rows, performing calculations, and so on.

To add operations to a data flow:

1. Select the node you want to chain an operation to.
2. From the sidebar, click **Operations**.
3. Select the operation that you want to apply and complete that operation flow.
   When completed, the operation is shown as a new chained node on the canvas.
4. Repeat steps 1 to 3 to chain more operations to the data flow.

Operations overview

The list of available operations continues to expand, but here are a few of the more common ones:

- **Prepare data set**: Displays an Excel-like grid that you can use to work interactively with the rows and columns of sample data. You can choose from among over 80 operations at the table or column level, such as cleansing data, type conversions, math functions, and more.
- **Join data sets**: Joins two data flows together to produce a joined data set.
- **Change schema**: Renames, reorders, and removes columns from the data source. In addition, you can perform basic type conversion.
- **Sort rows**: Sorts a column in ascending or descending order.
- **Remove duplicates**: Removes rows in which a column contains duplicate values.
- **Filter rows**: Selects specific column values that you want to keep.

Completing data flows

When you're done refining and shaping the data and you're satisfied with the results, you can then set a target and complete that data flow.

To complete a data flow:

1. On the canvas, select the end-point in the data flow that you want to complete.
2. On the node's vertical ellipse or on the toolbar, select **Set Target**.
3. In the Select a Target window, use the **Connections** list to select a target connection to write the data to.
4. Select a schema from the **Schemas** list.
5. Specify whether you want the data to be appended to the table, to re-create the entire table, to replace the table contents, or to merge with the table. Optionally, you can perform advanced mapping using the Advanced Mapping link.
6. Click **Save**.
   The data flow is now complete with a new target connection node.

Managing data flow activities

Use the Activities task to manage your existing activities, including running an activity, scheduling an activity, editing an activity, and more.

To start the Activities task, click **Activities** in the left navigation menu.

Design Flow task basics

The Design Flow task comprises a toolbar along the top, the canvas where your nodes and data flows appear, and a right sidebar. This section provides detailed information about the actions you can take from each of these elements.

Toolbar basics

At the top of the Design Data Flow task is a toolbar. Using this toolbar, you can perform the following actions, which are automatically saved as you work:

- Specify a name for the data flow activity. To change the name, click the pencil icon, specify a new name, and then click the check mark icon.
- Add more data sources by clicking the **Add source** icon.
- Complete a data flow by specifying where to write the data flow results by clicking the **Set Target** icon.
- Use the **Undo operation** to undo a recent change as you build a flow and the **Redo operation** to reapply the change.
- Close the data flow activity by either clicking the icon or clicking a different task in the left navigation menu.

Canvas basics

You can interact with the canvas. On the canvas you can perform the following actions:

- Select a node by clicking it. When you do, the sidebar appears and provides more actions for the selected node. When you point to the node, a vertical ellipse appears which, when clicked, opens a menu that contains actions specific to the selected node, such as **Preview**, which allows you to view sample data, or **Set Target**.
- Click and hold the mouse to drag the entire canvas.
- Zoom in/zoom out using either your mouse wheel or the icons near the mini-map view that is located in the lower right corner. To reset your view, click the RESET icon.
- Highlight nodes that contain a certain string by clicking the **Search** icon and specifying a search string. Nodes that contain the string are highlighted while nodes that do not contain the string are less prominent.
- Use the Settings icon to change the node layout and lines.

Sidebar basics

When you select a node on the canvas, the sidebar appears. The sidebar provides information such as the data source name and the quality score. Use the sidebar buttons to display the following items:

- **Operations**, such as Prepare data set, Sort rows, and more, that can be added to the end of the data flow.
- **History**, which shows information about the operation(s) that the node represents.
- **Details**, which displays information for the data source that the node represents, such as connection information, data quality, and more.
- **Discussion**, which you can use to comment on the data flow activity. You can add and delete your own comments as well as view prior discussion comments from yourself and others. This discussion is on the data flow activity itself.