

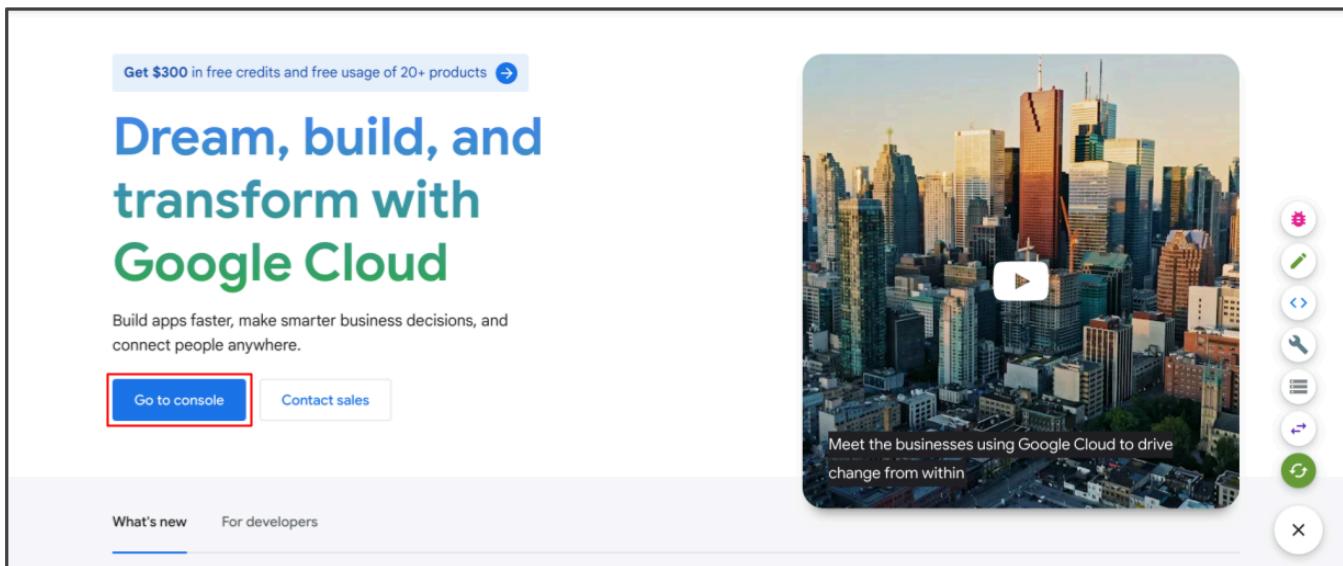


# Reading: Guide to Dataflow

As you have been learning, Dataflow is a serverless data-processing service that reads data from the source, transforms it, and writes it in the destination location. Dataflow creates pipelines with open source libraries, with which you can interact using different languages, including Python and SQL. This reading provides information about accessing Dataflow and its functionality.

## Navigate the homepage

If you completed the optional [Create a Google Cloud account](#) activity, you can follow along with the steps of this reading in your Dataflow console. Go to the [Dataflow Google Cloud](#) homepage and sign in to your account to access Dataflow. Then click the **Go to Console** button or the **Console** button. Here, you will be able to create new jobs and access Dataflow tools.



## Jobs

When you first open the console, you will find the **Jobs** page. The **Jobs** page is where your current jobs are in your project space. There are also options to CREATE JOB FROM TEMPLATE or CREATE MANAGED DATA PIPELINE from this page, so that you can get started on a new project in your Dataflow console. This is where you will go anytime you want to start something new.

Jobs

No jobs to display.  
Get started by creating a job from a template.

CREATE TAKE THE QUICKSTART

## Pipelines

Open the menu pane to navigate through the console and find the other pages in Dataflow. The **Pipelines** menu contains a list of all the pipelines you have created. If this is your first time using Dataflow, it will also display the processes you need to enable before you can start building pipelines. If you haven't already enabled the APIs, click **Fix All** to enable the API features and set your location.

Pipelines

| Name                | Status | Type | Managed | Region | Total Runs | Schedule | Start time | Elapsed time | End time |
|---------------------|--------|------|---------|--------|------------|----------|------------|--------------|----------|
| No rows to display. |        |      |         |        |            |          |            |              |          |

## Workbench

The **Workbench** section is where you can create and save shareable Jupyter notebooks with live code. This is helpful for first-time ETL tool users to check out examples and visualize the transformations.

The screenshot shows the Dataflow Workbench interface. On the left, a sidebar menu includes 'Overview', 'Jobs', 'Pipelines', 'Workbench' (which is selected and highlighted with a red box), 'Snapshots', and 'SQL Workspace'. At the top, there are buttons for 'NEW INSTANCE', 'REFRESH', 'START', 'STOP', 'RESET', 'UPGRADE', and 'DELETE'. A prominent message box at the top right encourages migrating to the Notebooks API. Below it, a note states that legacy instances don't receive updates. A table header for 'Instances' lists columns: 'Instance name' (sorted by ascending), 'Zone', 'Environment', 'Machine type', 'GPUs', 'Owner', and 'Last modified'. The main area displays a large globe icon with colored dots (yellow, red, green, blue) representing instances. A message at the bottom says 'No matched instances found.' and a 'Show debug panel' link.

## Snapshots

**Snapshots** save the current state of a pipeline to create new versions without losing the current state. This is useful when you are testing or updating current pipelines so that you aren't disrupting the system. This feature also allows you to back up and recover old project versions. You may need to enable APIs to view the **Snapshots** page; you will learn more about APIs in an upcoming activity.

The screenshot shows the Dataflow Snapshots page. The sidebar menu is identical to the Workbench page. The main content area has a heading 'Solutions' and a note explaining what Dataflow snapshots are used for. Below this is a table header for 'Solutions' with columns: 'Snapshot ID', 'Creation time' (sorted by descending), 'Expiration time', 'Source job ID', and 'Location'. The main area features a globe icon with colored dots. A message at the bottom says 'No snapshots are available' and provides instructions on how to create a snapshot. A 'Show debug panel' link is at the bottom right.

## SQL Workspace

Finally, the **SQL Workspace** is where you interact with your Dataflow jobs, connect to BigQuery functionality, and write necessary SQL queries for your pipelines.

The screenshot shows the Google Cloud Dataflow SQL Editor. On the left, there's a sidebar with navigation links: Overview, Jobs, Pipelines, Workbench, Snapshots, and SQL Workspace, with SQL Workspace highlighted by a red box. The main area is titled "Dataflow SQL Editor". It contains a search bar ("Type to search resources") and a collapsible tree view showing three tables from a dataset: "location\_latest..", "class\_activity\_q", and "covid19\_open\_c". Below the tree view, there are fields for "dfsqljob name" (with a note about uniqueness), "Regional endpoint" (with a note about deployment options), and "SQL query text". A section for "Required parameters" is present, followed by a link to "SHOW OPTIONAL PARAMETERS". Under "Destination", there's a dropdown for "Select an output" and a checkbox for "Additional output". At the bottom right is a blue "CREATE" button.

Dataflow also gives you the option to interact with your databases using other coding languages, but you will primarily be using SQL for these courses.

Dataflow is a valuable way to start building pipelines and exercise some of the skills you have been learning in this course. Coming up, you will have more opportunities to work with Dataflow, so now is a great time to get familiar with the interface!

---