

# Dataflow Execution Settings

When you specify a Dataflow® job, you may pass to the running environment a set of property values to apply to the execution of the job. Overrides are defined in the Run Job page and are applied to the configured job.

- You can specify overrides for ad-hoc jobs through the Run Job page.
- You can specify overrides when you configure a scheduled job execution.

These property values override any settings applied to the project.

- Properties whose values are not specified in the dataflow execution overrides use the values that you set in the Execution Settings page.
- See *Execution Settings Page*.

### Run Job on Dataflow

#### Dataflow Execution Settings

**Region**

**Zone**

**Machine Type**

[Advanced Settings ^](#)

**VPC network mode**

**Autoscaling algorithms**

**Initial number of workers**

**Maximum number of workers**

**Service account**

**Labels**

[Add](#)

## Figure: Dataflow Execution Properties

### Default execution settings:

By default, Dataprep by Trifacta runs your job in the `us-central1` region on an `n1-standard-1` machine. As needed, you can change the geo location and the machine where your job is executed.

**Tip:** You can change the default values for the following in your Execution Settings. See *Execution Settings Page*.

**Making changes to these settings can affect performance times for executing your job.**

**Tip:** For more information on how the following settings affect your jobs, see *Run Job on Cloud Dataflow*.

Setting	Description
Regional Endpoint	A regional endpoint handles execution details for your Dataflow job, its location determines where the Dataflow job is executed.
Zone	A sub-section of region, a zone contains specific resources for a given region.  Select <code>Auto Zone</code> to allow the platform to choose the zone for you.
Machine Type	Choose the type of machine on which to run your job. The default is <code>n1-standard-1</code> .  Note: not all machine types supported directly through Dataprep by Trifacta.

For more information on these regional endpoints, see <https://cloud.google.com/dataflow/docs/concepts/regional-endpoints>.


For more information on machine types, <https://cloud.google.com/compute/docs/machine-types>.

### Advanced settings:

Setting	Description
VPC Network mode	<p>If the network mode is set to <code>Auto</code> (default), the job is executed over publicly available IP addresses. Do not set values for Network, Subnetwork, and Worker IP address configuration.</p> <p>As needed, you can override the default settings configured for your project for this job. Set this value to <code>Custom</code>.</p> <div style="border: 1px solid #ccc; padding: 10px; margin: 10px 0;"><p><b>NOTE:</b> Avoid applying overrides unless necessary. These network settings apply to job execution. Preview and sampling use the <code>default</code> network settings.</p></div> <ol style="list-style-type: none"><li>Specify the name of the VPC network in your region.</li><li>Specify the short or full URL of the Subnetwork. If both Network and Subnetwork are specified, Subnetwork is used. See <a href="https://cloud.google.com/dataflow/docs/guides/specifying-networks">https://cloud.google.com/dataflow/docs/guides/specifying-networks</a>.</li><li>Review and specify the Worker IP address configuration setting. See below.</li></ol> <p>For more information:</p> <ul style="list-style-type: none"><li><a href="https://cloud.google.com/vpc/docs/vpc">https://cloud.google.com/vpc/docs/vpc</a></li><li><a href="https://cloud.google.com/dataflow/docs/guides/specifying-networks">https://cloud.google.com/dataflow/docs/guides/specifying-networks</a></li></ul>

Network	To use a different VPC network, enter the name of the VPC network to use as an override for this job. Click <b>Save</b> to apply the override.
Subnetwork	<p>To specify a different subnetwork, enter the URL of the subnetwork. The URL should be in the following format:</p> <pre>regions/&lt;REGION&gt;/subnetworks/&lt;SUBNETWORK&gt;</pre> <p>where:</p> <ul style="list-style-type: none"> <li>• &lt;REGION&gt; is the region identifier specified under Regional Endpoint. These values must match.</li> <li>• &lt;SUBNETWORK&gt; is the subnetwork identifier.</li> </ul> <p>If you have access to another project within your organization, you can execute your Dataflow job through it by specifying a full URL in the following form:</p> <pre>https://www.googleapis.com/compute/v1/projects/&lt;HOST_PROJECT_ID&gt;/regions/&lt;REGION&gt;/subnetworks/&lt;SUBNETWORK&gt;</pre> <p>where:</p> <ul style="list-style-type: none"> <li>• &lt;HOST_PROJECT_ID&gt; corresponds to the project identifier. This value must be between 6 and 30 characters. The value can contain only lowercase letters, digits, or hyphens. It must start with a letter. Trailing hyphens are prohibited.</li> </ul> <p>Click <b>Save</b> to apply the override.</p>

For more information on these settings, see [Execution Settings Page](#).

 **Feature Availability:** This feature may not be available in all product editions.

Setting	Description
Worker IP address configuration	<p>If the VPC Network mode is set to <code>custom</code>, then choose one of the following:</p> <ul style="list-style-type: none"> <li>• <code>Allow public IP addresses</code> - Use Dataflow workers that are available through public IP addresses. No further configuration is required.</li> <li>• <code>Use internal IP addresses only</code> - Dataflow workers use private IP addresses for all communication. <ul style="list-style-type: none"> <li>• If a Subnetwork is specified, then the Network value is ignored.</li> <li>• The specified Network or Subnetwork must have Private Google Access enabled.</li> </ul> </li> </ul>
Autoscaling Algorithms	<p>The type of algorithm to use to scale the number of Google Compute Engine instances to accommodate the size of your job. Possible values:</p> <ul style="list-style-type: none"> <li>• <code>Throughput based</code> - Scaling is determined by the volume of data expected to be passed through Dataflow.</li> <li>• <code>None</code> - None algorithm is applied. <ul style="list-style-type: none"> <li>• If none is selected, use <code>initial number of workers</code> to specify a fixed number of Google Compute Engine instances.</li> </ul> </li> </ul>
Initial number of workers	Number of Google Compute Engine instances with which to launch the job. This number may be adjusted as part of job execution. This number must be an integer between 1 and 1000, inclusive.
Maximum number of workers	Maximum number of Google Compute Engine instances to use during execution. This value must be greater than the initial number of workers and must be an integer between 1 and 1000, inclusive.

Service account	<p>Every Dataprep by Trifacta job executed in Dataflow requires that the job be submitted through a service account. By default, Dataprep by Trifacta uses a default Compute Engine service account under which to run jobs.</p> <p>Optionally, you can specify a different service account under which to run your job.</p> <div data-bbox="313 258 1455 380" style="border: 1px solid #ccc; padding: 10px; margin: 10px 0;"> <p><b>NOTE:</b> When using a named service account to access data and run jobs in other projects, the user running the job must be granted the <code>roles/iam.serviceAccountUser</code> role on the service account to use it.</p> </div> <p>For more information on service account usage and requirements, see <i>Google Service Account Management</i></p>
Labels	<p>Create or assign labels to apply to the billing for the Dataprep by Trifacta jobs run in your project. You may reference up to 64 labels.</p> <div data-bbox="313 520 1455 604" style="border: 1px solid #ccc; padding: 10px; margin: 10px 0;"> <p><b>NOTE:</b> Each label must have a unique key name.</p> </div> <p>For more information, see <a href="https://cloud.google.com/resource-manager/docs/creating-managing-labels">https://cloud.google.com/resource-manager/docs/creating-managing-labels</a>.</p>

**Notes on behavior:**

- Values specified here are applied to the current job or to all jobs executed using the output object.
- Properties not specified here are not submitted, and the default values for Dataflow are used.