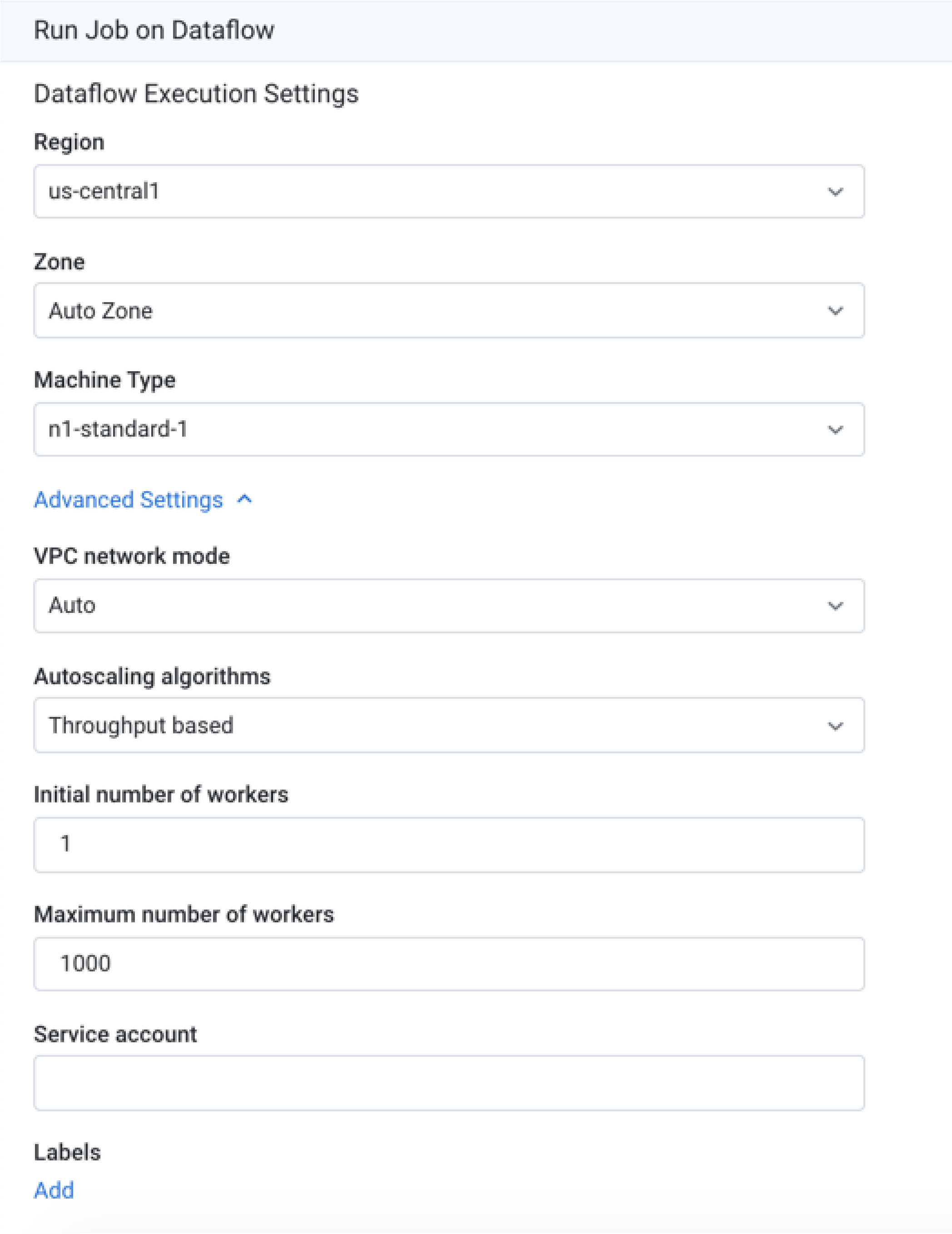
# Dataflow Execution Settings

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: us-central1

Zone: Auto Zone

Machine Type: n1-standard-1

Advanced Settings

VPC network mode: Auto

Autoscaling algorithms: Throughput based

Initial number of workers: 1

Maximum number of workers: 1000

Service account:

Labels

Add

Copyright © 2022 Trifacta Inc.

***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.<br>Select `Auto Zone` 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 `n1-standard-1`.<br>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 | If the network mode is set to `Auto` (default), the job is executed over publicly available IP addresses. Do not set values for Network, Subnetwork, and Worker IP address configuration.<br>As needed, you can override the default settings configured for your project for this job. Set this value to `Custom`.<br>**NOTE:** Avoid applying overrides unless necessary. These network settings apply to job execution. Preview and sampling use the `default` network settings.<br>1. Specify the name of the VPC network in your region.<br>2. Specify the short or full URL of the Subnetwork. If both Network and Subnetwork are specified, Subnetwork is used. See *https://cloud.google.com/dataflow/docs/guides/specifying-networks*.<br>3. Review and specify the Worker IP address configuration setting. See below.<br>For more information:<br>- *https://cloud.google.com/vpc/docs/vpc*<br>- *https://cloud.google.com/dataflow/docs/guides/specifying-networks* |

Copyright © 2022 Trifacta Inc.

| | |
|---|---|
| Network | To use a different VPC network, enter the name of the VPC network to use as an override for this job. Click **Save** to apply the override. |
| Subnetwork | To specify a different subnetwork, enter the URL of the subnetwork. The URL should be in the following format:<br>`regions/<REGION>/subnetworks/<SUBNETWORK>`<br>where:<br>- `<REGION>` is the region identifier specified under Regional Endpoint. These values must match.<br>- `<SUBNETWORK>` is the subnetwork identifier.<br>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:<br>`https://www.googleapis.com/compute/v1/projects/<HOST_PROJECT_ID>/regions/<REGION>/subnetworks/<SUBNETWORK>`<br>where:<br>- `<HOST_PROJECT_ID>` 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.<br>Click **Save** to apply the override. |

For more information on these settings, see *Execution Settings Page*.

**Feature Availability:** This feature is available in the following editions:

- Dataprep by Trifacta® Enterprise Edition
- Dataprep by Trifacta Professional Edition
- Dataprep by Trifacta Premium

| Setting | Description |
|---|---|
| Worker IP address configuration | If the VPC Network mode is set to `custom`, then choose one of the following:<br>- `Allow public IP addresses` - Use Dataflow workers that are available through public IP addresses. No further configuration is required.<br>- `Use internal IP addresses only` - Dataflow workers use private IP addresses for all communication.<br>  - If a Subnetwork is specified, then the Network value is ignored.<br>  - The specified Network or Subnetwork must have Private Google Access enabled. |
| Autoscaling Algorithms | The type of algorithm to use to scale the number of Google Compute Engine instances to accommodate the size of your job. Possible values:<br>- `Throughput based` - Scaling is determined by the volume of data expected to be passed through Dataflow.<br>- `None` - None algorithm is applied.<br>  - If none is selected, use `initial number of workers` to specify a fixed number of Google Compute Engine instances. |
| 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 | 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.<br>Optionally, you can specify a different service account under which to run your job.<br>**NOTE:** When using a named service account to access data and run jobs in other projects, the user running the job must be granted the `roles/iam.serviceAccountUser` role on the service account to use it.<br>For more information on service account usage and requirements, see *Google Service Account Management* |
| Labels | 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.<br>**NOTE:** Each label must have a unique key name.<br>For more information, see *https://cloud.google.com/resource-manager/docs/creating-managing-labels*. |

### 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.

Copyright © 2022 Trifacta Inc.