Configuring internet access and firewall rules

This page explains how to provide routes and define your Google Cloud firewall rules for the network associated (/dataflow/docs/guides/specifying-networks) with your Dataflow jobs.

The **default** network has configurations that allow Dataflow jobs to run. However, other services might also twork. Make sure your changes to **default** are compatible with all of your services. Alternatively, create a sep rk for Dataflow.

Internet access for Dataflow

Dataflow worker virtual machines (VMs) need to be able to reach Google Cloud APIs and services. You can either configure worker VMs with an external IP address so that they meet the Internet access requirements (/vpc/docs/vpc#internet_access_reqs), or you can use Private Google Access (/vpc/docs/private-access-options#pga).

With Private Google Access, VMs that have only internal IP addresses can access select public IPs for Google Cloud and services. Read Configuring Private Google Access (/vpc/docs/configure-private-google-access) for information about the routing and firewall rules requirements (/vpc/docs/configure-private-google-access#requirements) and configuration steps (/vpc/docs/configure-private-google-access#configuring_access_to_google_services_from_internal_ips).

Jobs that access APIs and services outside of Google Cloud require internet access. For example, Python SDK jobs need access to the Python Package Index (PyPI). In this case, you must either configure worker VMs with external IP addresses or use a Network Address Translation solution, such as Cloud NAT (/nat/docs/overview). Read Managing Python Pipeline Dependencies (https://beam.apache.org/documentation/sdks/python-pipeline-dependencies/) on the Apache Beam website for more details.

Domain Name System (DNS) limitations

Custom BIND is not supported when using Dataflow. To customize DNS resolution when using Dataflow with VPC Service Controls, use Cloud DNS private zones (/dns/docs/overview#dns-private-zones) instead of using custom BIND servers. To use your own

on-premises DNS resolution, consider using a Google Cloud DNS forwarding method (/dns/docs/overview#dns-forwarding-methods).

Firewall rules

Firewall rules let you allow or deny traffic to and from your VMs. This page assumes familiarity with how Google Cloud firewall rules work as described on the Firewall Rules Overview (/vpc/docs/firewalls) and Using Firewall Rules (/vpc/docs/using-firewalls) pages, including the implied firewall rules (/vpc/docs/firewalls#default_firewall_rules).

Firewall rules required by Dataflow

Dataflow requires that worker VMs communicate with one another using specific TCP ports within the VPC network that you specify in your pipeline options (/dataflow/docs/guides/specifying-networks). You need to configure firewall rules in your VPC network to allow this type of communication.

Some VPC networks, like the automatically created `default` network, include a `default-allow-internal` rule that meets the firewall requirement for Dataflow.

Because all worker VMs have a network tag with the value `dataflow`, you can create a more specific firewall rule for Dataflow. A project owner, editor, or Security Admin (/compute/docs/access/iam#compute.securityAdmin) can use the following `gcloud` command to create an ingress allow rule that permits traffic on TCP ports 12345 and 12346 from VMs with the network tag `dataflow` to VMs with the same tag:

```
d compute firewall-rules create FIREWALL_RULE_NAME \
-network NETWORK \
-action allow \
-direction DIRECTION \
-target-tags dataflow \
-source-tags dataflow \
-priority 0 \
-rules tcp:12345-12346
```

In the preceding example, replace the following variables:

- ***FIREWALL_RULE_NAME*** with a name for the firewall rule
- ***NETWORK*** with the name of the network that your worker VMs use
- ***DIRECTION*** with the direction (/vpc/docs/firewalls#direction_of_the_rule) of the firewall rule

For further guidance about firewall rules, refer to Using Firewall Rules (/vpc/docs/using-firewalls). For specific TCP ports used by Dataflow, you can view the project container manifest (/deployment-manager/docs/deployments/viewing-manifest#view_a_manifest). The container manifest explicitly specifies the ports in order to map host ports into the container. You can also view network configuration and activity by opening a SSH session on one of your workers (https://cloud.google.com/compute/docs/tutorials/service-account-ssh#sa_ssh) and running `iproute2`. Read the `iproute2` page (https://wiki.linuxfoundation.org/networking/iproute2) for more information.

SSH access to worker VMs

Dataflow does not require SSH; however, SSH is useful for troubleshooting.

If your worker VM has an external IP address, you can connect to the VM (/compute/docs/instances/connecting-to-instance) through either the Cloud Console or by using `gcloud` command-line tool. To connect using SSH, you must have a firewall rule that allows incoming connections on TCP port 22 from at least the IP address of the system on which you're running `gcloud` or the system running the web browser you use to access the Cloud Console.

If you need to connect to a worker VM that only has an internal IP address, see Connecting to instances that do not have external IP addresses (/compute/docs/instances/connecting-advanced#sshbetweeninstances).

Except as otherwise noted, the content of this page is licensed under the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the Apache 2.0 License (https://www.apache.org/licenses/LICENSE-2.0). For details, see the Google Developers Site Policies (https://developers.google.com/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2020-06-26 UTC.