Using Stackdriver Monitoring for Cloud Dataflow pipelines

Beta

This product or feature is in a pre-release state and might change or have limited support. For more information, see the product launch stages (https://cloud.google.com/products/#product-launch-stages).

Note: Stackdriver Monitoring in the Cloud Console is now Generally Available and the default experience. For a limited period of time, you also have the option to use the classic Stackdriver Monitoring console. For more information, see Stackdriver Monitoring in the Cloud Console (https://cloud.google.com/monitoring/docs/monitoring_in_console).

Stackdriver provides powerful monitoring, logging, and diagnostics. Dataflow integration with Stackdriver Monitoring lets you access Dataflow job metrics such as Job Status, Element Counts, System Lag (for streaming jobs), and User Counters from the Stackdriver dashboards. You can also employ Stackdriver alerting capabilities to notify you of various conditions, such as long streaming system lag or failed jobs.

Before you begin

Follow one of the quickstarts (https://cloud.google.com/dataflow/docs/quickstarts) to get your Dataflow project set up and construct and run your pipeline (https://cloud.google.com/dataflow/pipelines/constructing-your-pipeline).

Custom metrics

Any metric you define in your Apache Beam pipeline is reported by Dataflow to Stackdriver as a custom metric. There are three types of Apache Beam pipeline metrics (https://beam.apache.org/documentation/programming-guide/#metrics): `Counter`, `Distribution`, and

`Gauge`. Dataflow currently only reports `Counter` and `Distribution` to Stackdriver. `Distribution` is reported as four sub-metrics suffixed with `_MAX`, `_MIN`, `_MEAN`, and `_COUNT`. Dataflow does not support creating a histogram from `Distribution` metrics.

Dataflow reports incremental updates to Stackdriver approximately every 30 seconds. All user metrics are exported as a `double` data type to avoid conflicts. Custom metrics in Dataflow appear in Stackdriver as `custom.googleapis.com/dataflow/metric-name` and are limited to 500 metrics per project.

Custom metrics reported to Stackdriver incurs charges based on the Stackdriver Monitoring pricing (https://cloud.google.com/stackdriver/pricing#monitoring-costs).

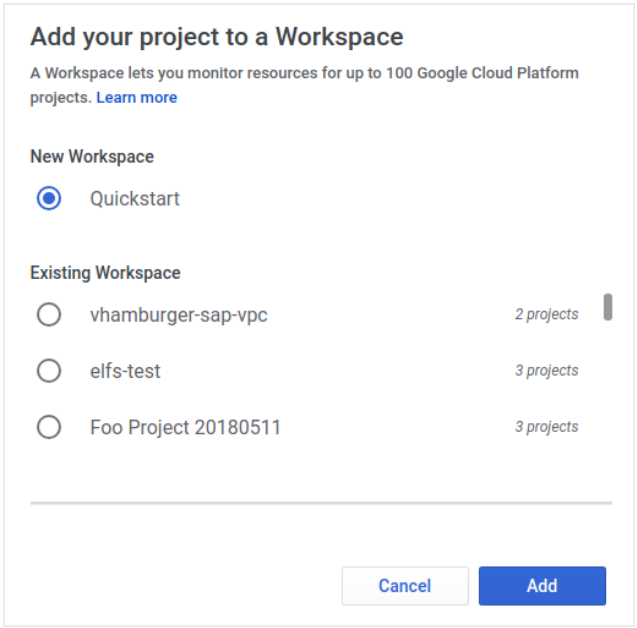
Explore metrics

You can explore Dataflow metrics using Stackdriver. Follow the steps in this section to observe the several standard metrics provided for each of your Apache Beam pipelines.

1. In the Google Cloud Console, select **Stackdriver Monitoring**:

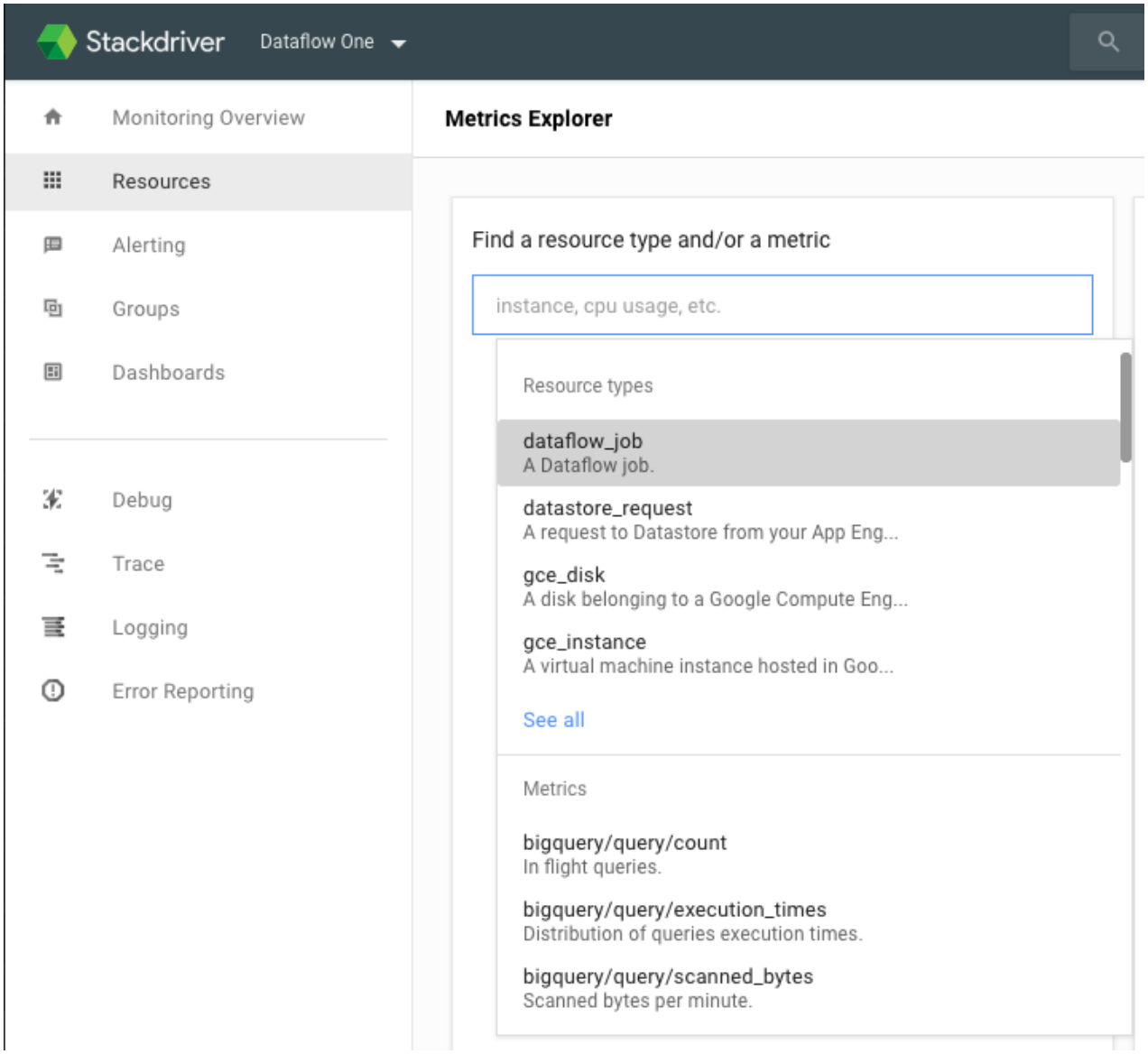
   GO TO MONITORING (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/MONITORING)

2. If the **Add your project to a Workspace** dialog is displayed, create a new Workspace by selecting your Google Cloud project under **New Workspace** and then clicking **Add**. In the following image, the Google Cloud project name is `Quickstart`:



The **Add your project to a Workspace** dialog is displayed only when you have at least one existing Workspace available to you. The Workspaces listed under **Existing Workspace** are Workspaces you've created or Workspaces for Google Cloud projects where you have editorial permission. You can choose between creating a new Workspace and adding your project to an existing Workspace by using this dialog.

3. In the **Resource** menu, select **Metrics Explorer**.
4. In the **Find a resource type and/or a metric** pane, select the `dataflow_job` resource type.



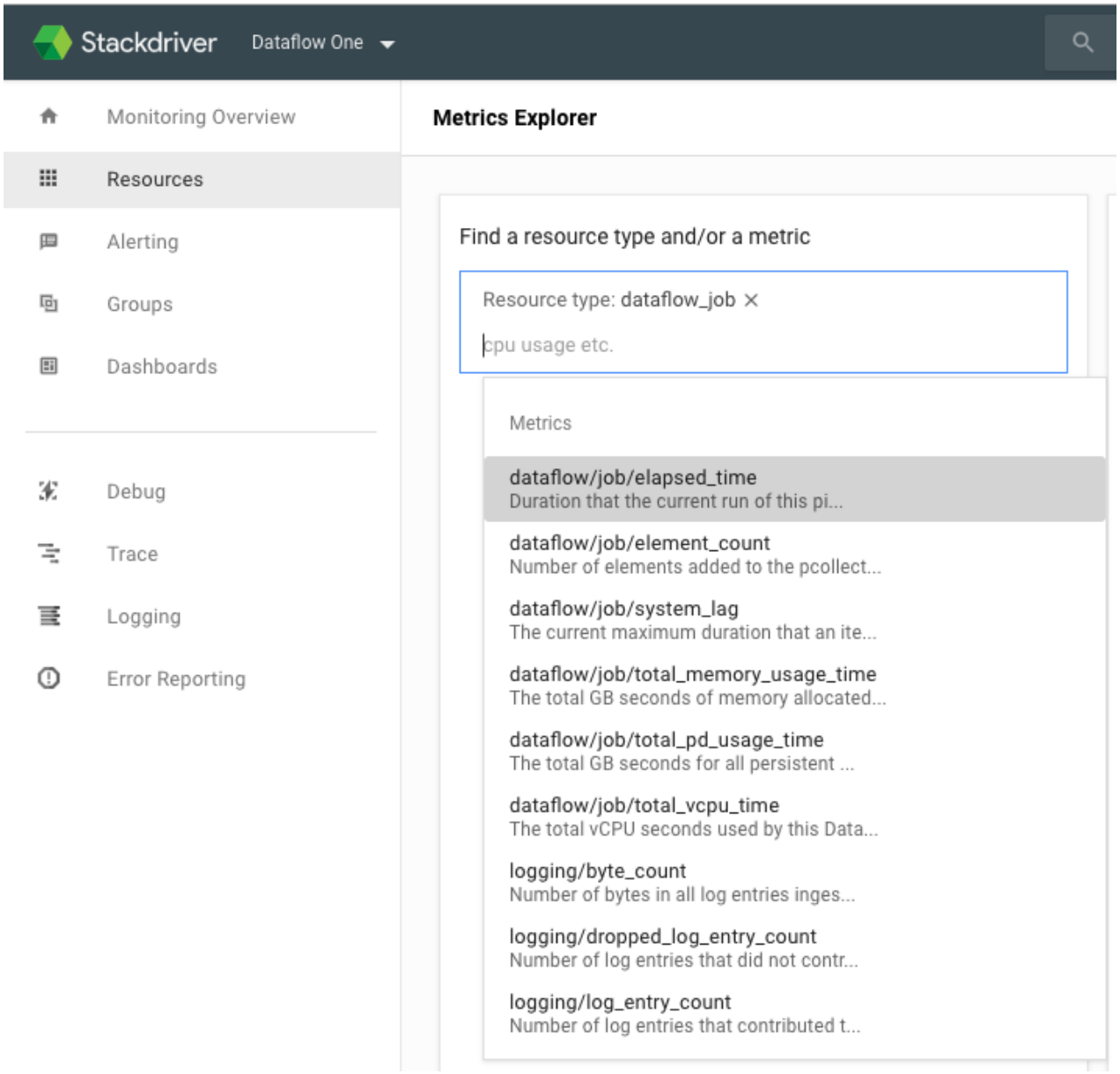
5. From the list that appears, select a metric you'd like to observe for one of your jobs.

Click to read about the **Cloud Dataflow-related metrics** available.

- **Job status:** Provide job status (Failed, Successful) as an enum every 30 seconds and on update. Note: `enum`s may not be charted or used for alerts, but this value may

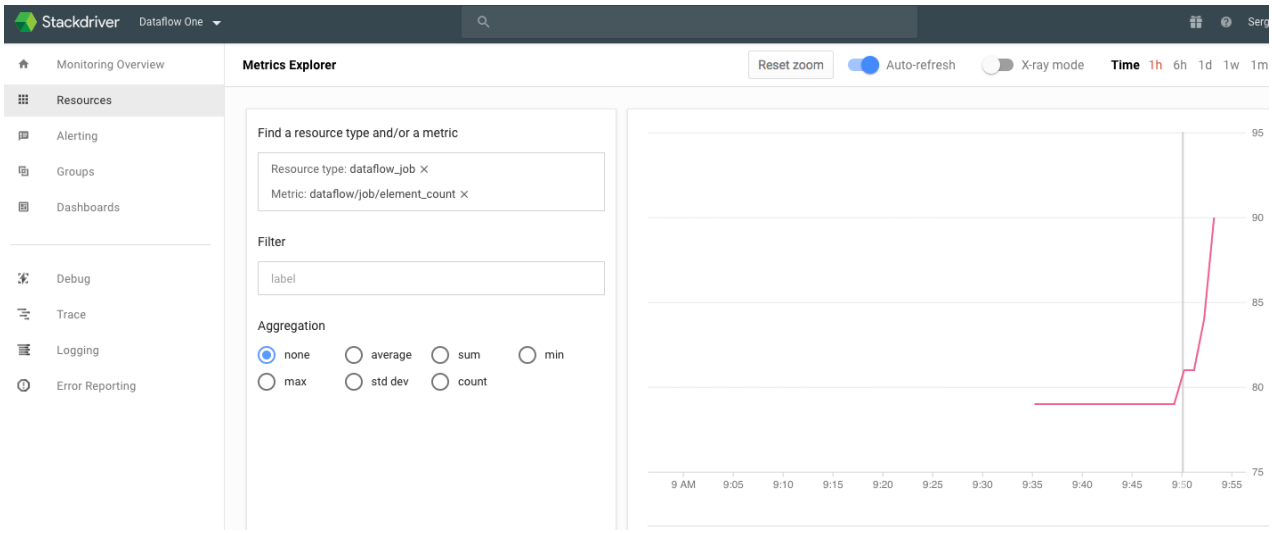
be retrieved through the Cloud Monitoring UI. For alerting use the "Failed" metric that sets to 1 if a job fails.

- **Failed:** Failed sets 1 if a job exits with a failure. Use this metric to alert on and chart the number of failed pipelines.
- **Elapsed time:** Job elapsed time (measured in seconds), reported every 30 seconds.
- **System lag:** Max lag across the entire pipeline, reported in seconds.
- **Current vCPU count:** Current # of virtual CPUs used by job and updated on value change.
- **Total vCPU usage:** Total # of virtual CPUs used by a job and updated on value change.
- **Total Persistent Disk usage:** Cumulative total of Persistent Disk used by job, measured in GB-seconds and updated on value change. Note: there are two different types of persistent disk (SSD and HDD). They are both reported using the same metric name and use a metric label to differentiate.
- **Total memory usage:** Cumulative total memory allocated to job, measured in GB-seconds and updated on value change.
- **Element count:** Number of elements per PCollection. Note: This is a per-PCollection metric, not a job-level metric, so it is not yet available for alerting.
- **Estimated byte count:** Number of bytes processed per PCollection. Note: This is a per-PCollection metric, not a job-level metric, so it is not yet available for alerting.



For example: This example shows a streaming pipeline that reads from a Cloud Pub/Sub topic and writes to BigQuery. It has 5 steps, one of which is `PubsubIO.Read`. The image below displays the `dataflow/job/element_count` for the `PubsubIO.Read` step of the

pipeline.



Create alerts and dashboards

Stackdriver does not only provide you with access to Dataflow-related metrics, but also lets you to create alerts and dashboards so you can chart time series of metrics and choose to be notified when these metrics reach specified values.

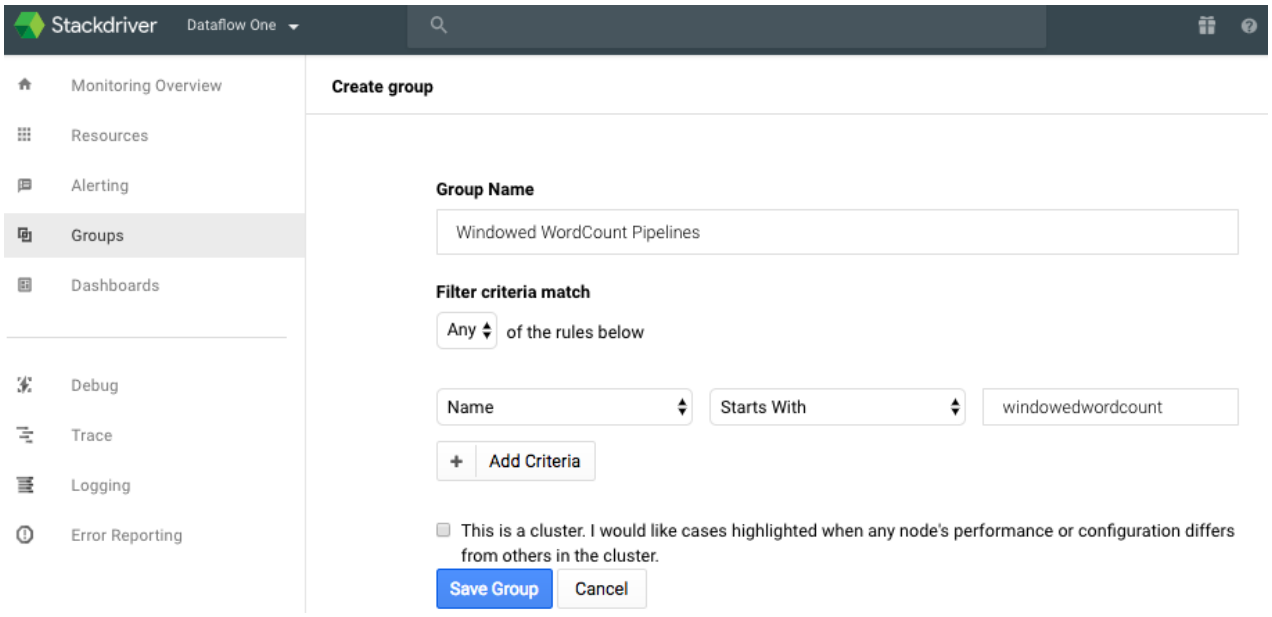
Create groups of resources

You can create resource groups that include multiple Apache Beam pipelines so that you can easily set alerts and build dashboards.

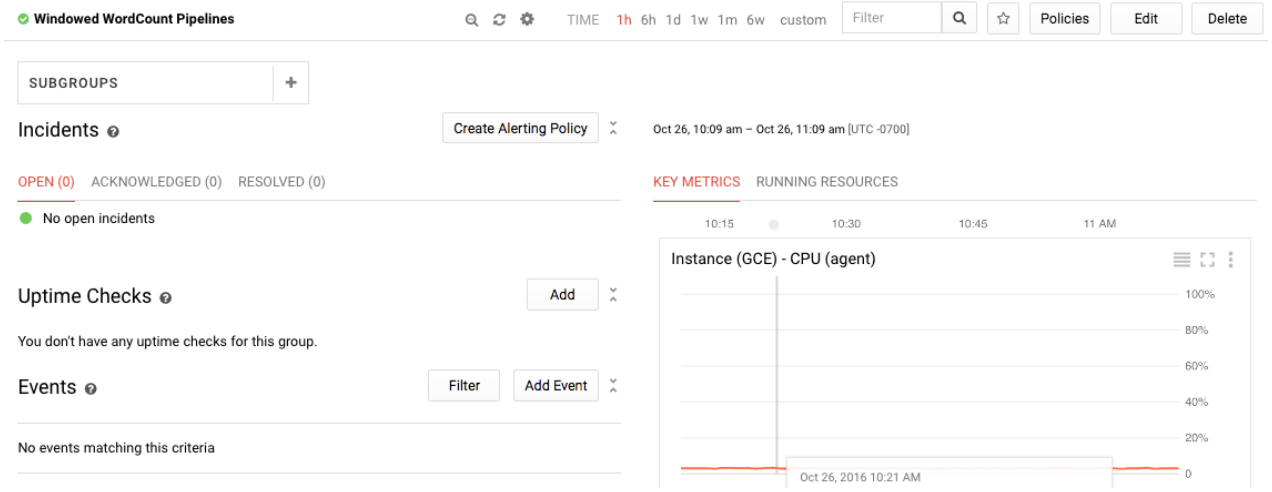
1. In the Google Cloud Console, select **Stackdriver Monitoring**:

   GO TO MONITORING (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/MONITORING)

2. In the **Groups** menu, select **Create Groups**.
3. Add filter criteria that define the Dataflow resources included in the group. For example, one of your filter criteria can be the name prefix of your pipelines.



4. After the group is created, you can see the basic metrics related to resources in that group.



Create alerts for Cloud Dataflow metrics

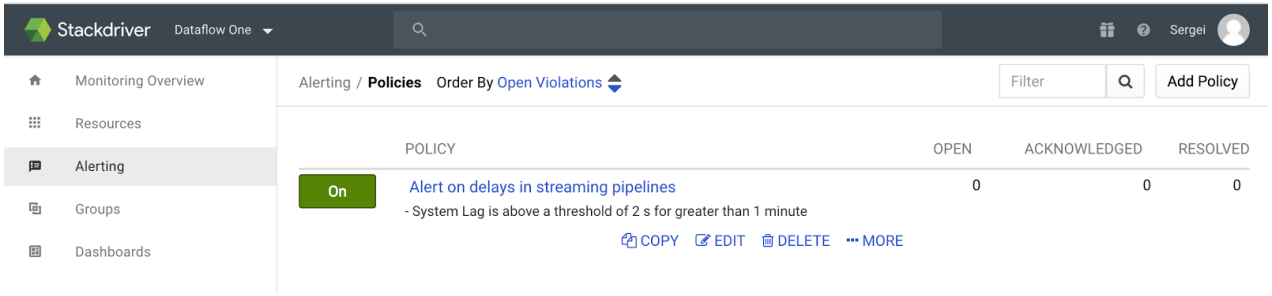
Stackdriver gives you the ability to create alerts and be notified when a certain metric crosses a specified threshold. For example, when **System Lag** of a streaming pipeline increases above a predefined value.

1. In the Google Cloud Console, select **Stackdriver Monitoring**:

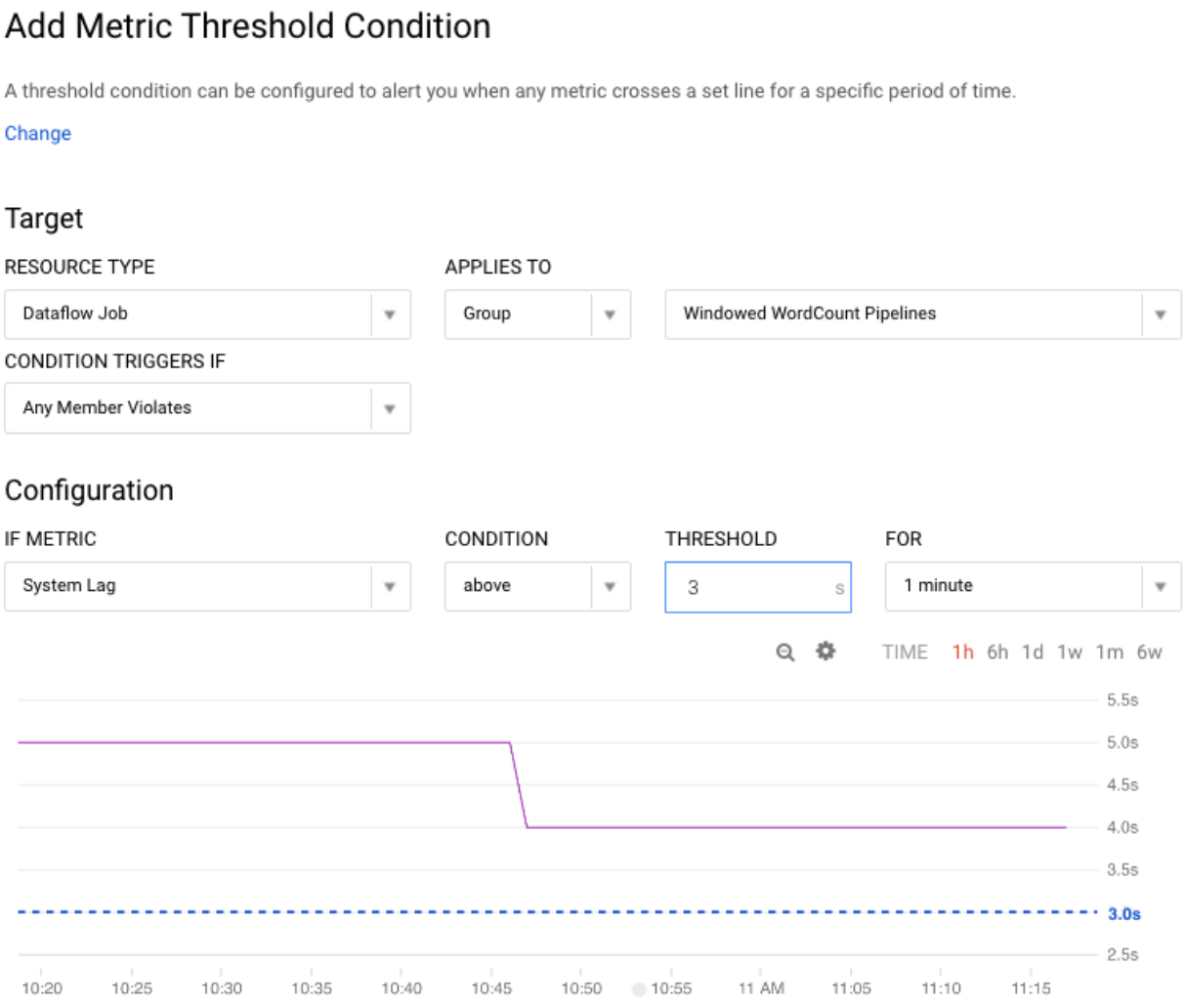
   GO TO MONITORING (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/MONITORING)

2. In the **Alerting** menu, select **Policies Overview**.

3. Click on **Add Policy**.

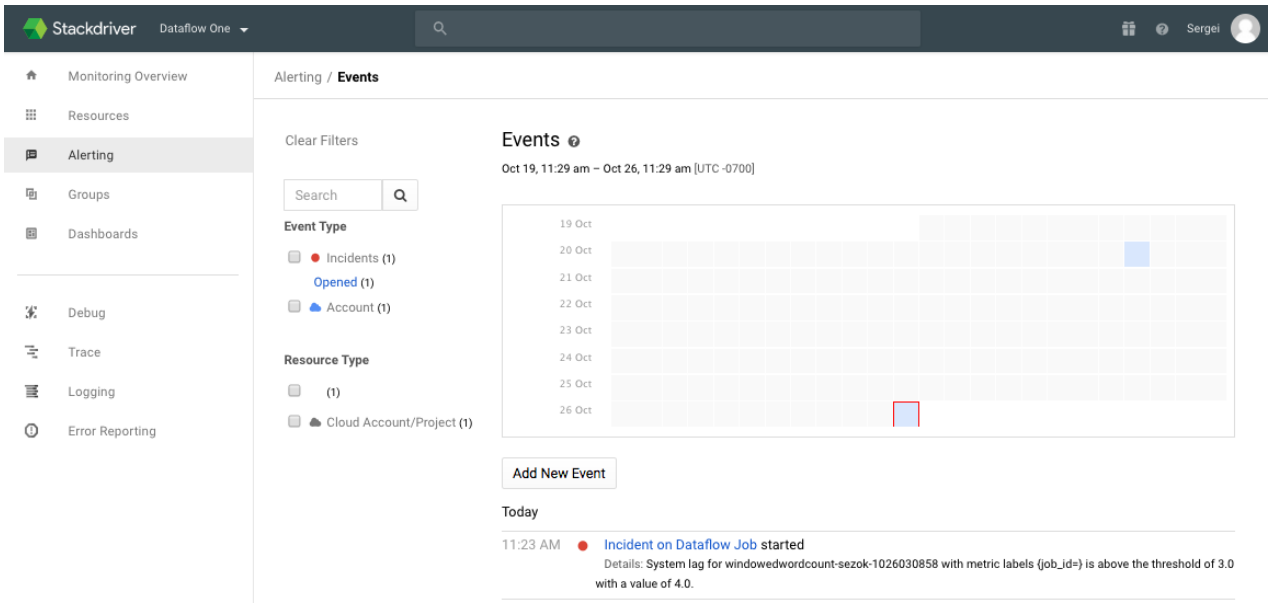


4. In the **Create new alerting policy** page, you can define the alerting conditions and the channels of communication for alerts.
   For example, to set an alert on the System Lag for the `WindowedWordCount` Apache Beam pipeline group, select 'Dataflow Job' in the **Resource Type** dropdown, 'Group' in the **Applies To** dropdown, and 'System Lag' in the **If Metric** dropdown.



5. After you've created an alert, you can review the events related to Dataflow by navigating to **Alerting** > **Events**. Every time an alert is triggered by a **Metric Threshold** condition, an

Incident and a corresponding **Event** are created in Stackdriver. If you specified a notification mechanism in the alert, such as email or SMS, you will also receive a notification.



Build your own custom monitoring dashboard

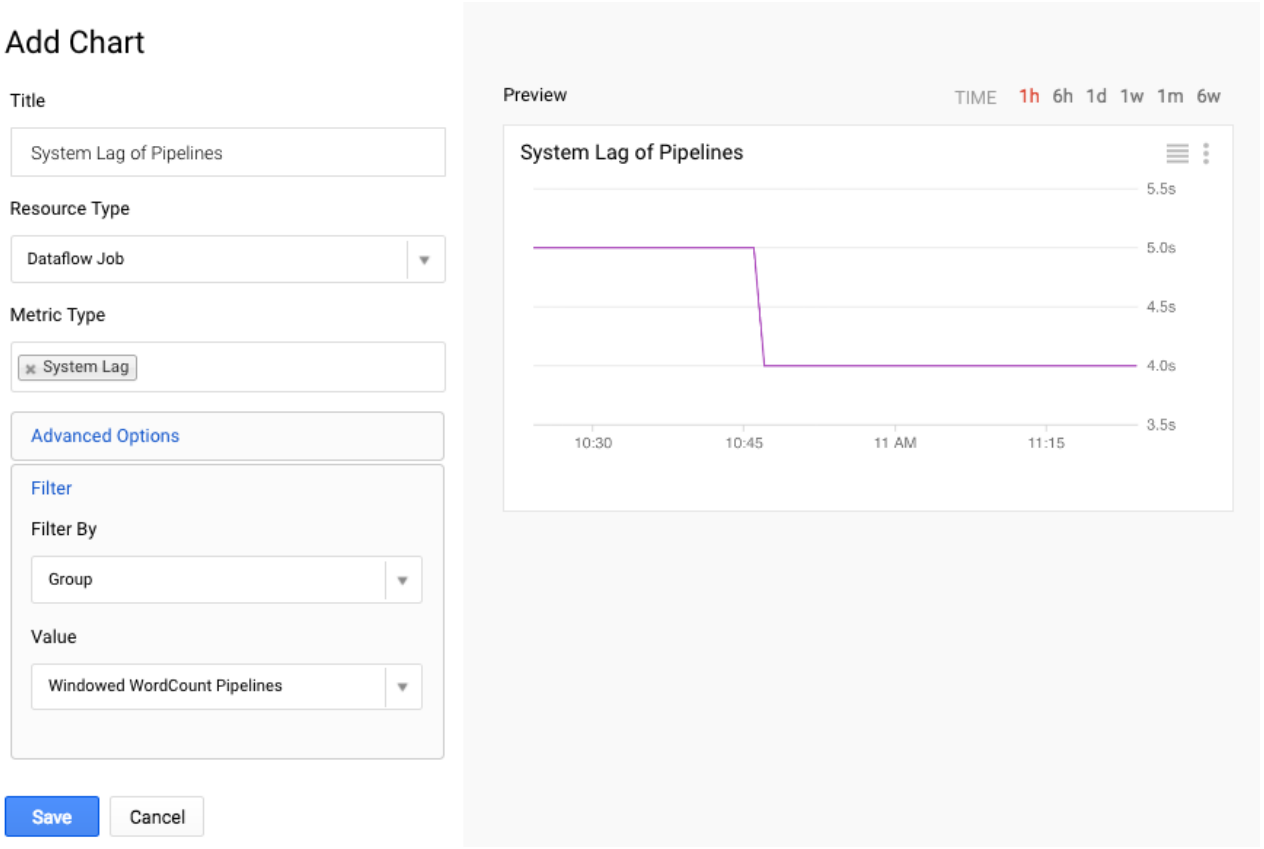
You can build Stackdriver monitoring dashboards with the most relevant Dataflow-related charts.

1. Go to the Google Cloud Console, and select **Stackdriver Monitoring**:

   GO TO MONITORING (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/MONITORING)

2. Select **Dashboards** > **Create Dashboard**.
3. Click on **Add Chart**.
4. In the **Add Chart** window, select "Dataflow Job" as the **Resource Type**, select a metric you want to chart in the **Metric Type** field, and select a group that contains Apache Beam

pipelines in the **Filter** panel.



You can add as many charts to the dashboard as you like.

Receive worker VM metrics from Stackdriver Monitoring agent

If you would like to monitor persistent disk, CPU, network, and process metrics from your Dataflow worker VM instances, you can enable the Stackdriver Monitoring Agent (https://cloud.google.com/monitoring/agent/) when you run your pipeline. See the list of available Monitoring agent metrics (https://cloud.google.com/monitoring/api/metrics_agent).

To enable the Monitoring agent, use the `--experiments=enable_stackdriver_agent_metrics` option when running your pipeline.

To disable the Monitoring agent without stopping your pipeline, update your pipeline by launching a replacement job (https://cloud.google.com/dataflow/pipelines/updating-a-pipeline#launching-your-replacement-job) and without specifying the `--experiments=enable_stackdriver_agent_metrics` parameter.

What's next

To learn more, consider exploring these other resources:

- Stackdriver documentation (https://cloud.google.com/stackdriver/docs/)
- Using the Cloud Dataflow Monitoring UI (https://cloud.google.com/dataflow/pipelines/dataflow-monitoring-intf)

Last updated January 9, 2020.