

BΞΔM
S U M M I T

Troubleshooting Slow Running Beam Pipelines

By Mehak Gupta
Google Cloud, Canada

About Me



Hello!

I'm Mehak

Technical Solutions Specialist at Google Cloud



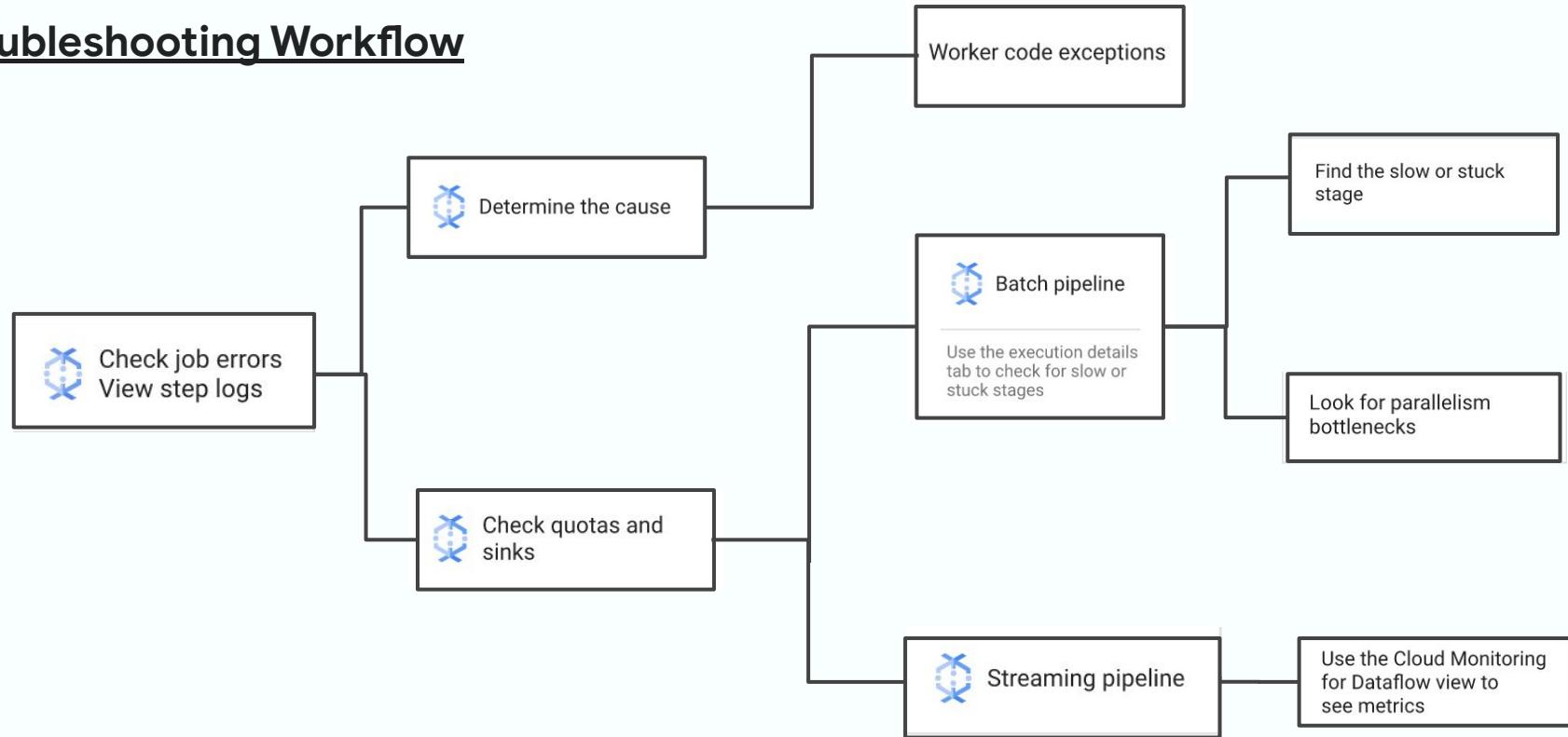
- Apache Beam pipeline troubleshooting techniques that would empower professionals to research and resolve Beam issues by themselves.
- Self service skills would reduce MTTR (Mean Time To Recover) significantly
- Share some tricks and samples of troubleshooting slow running beam pipelines using Dataflow as an example

How to identify if the beam pipeline is slow/stuck

- Pipeline is running from a long time without reporting results
- Increased data watermark or system latency
- Pipeline is not consuming input

Troubleshooting Slow Running Beam Pipelines

Troubleshooting Workflow



Troubleshoot slow/stuck dataflow jobs

The screenshot shows the Google Cloud Dataflow console interface. On the left, there's a sidebar with navigation links: Dataflow (selected), Overview, Jobs (selected), Pipelines, and Workbench. The main area is titled "Jobs" and contains two rows of job information. A red box highlights the first row for "wordcount6".

Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region	Insights
wordcount6	Streaming		48 days 19 hr	Apr 11, 2023, 3:08:03 PM	Running	2.43.0	2023-04-11_12_08_02-301837046300790666	us-east1	
wordcount5	Batch	May 28, 2023, 12:31:04 PM	21 hr 30 min	May 27, 2023, 3:00:29 PM	Succeeded	2.46.0	2023-05-27_12_00_28-10040782164894481412	us-east1	

Troubleshooting using Logs Explorer View

Troubleshoot slow/stuck dataflow jobs

Check logs here

JOB GRAPH EXECUTION DETAILS JOB METRICS RECOMMENDATIONS AUTOSCALING

Overview Jobs Pipelines Workbench Snapshots SQL Workspace

Metrics Data freshness Overview metrics Data freshness System latency Throughput Errors Streaming metrics Backlog Processing Parallelism

Data freshness by stages Create alerting policy

Logs HIDE 1

JOB LOGS WORKER LOGS DIAGNOSTICS

Severity Info Filter Search all fields and values

6 HOURS

Severity	Timestamp	Summary
Info	2023-02-28 23:28:10.161 EST	Operation ongoing in step ArchiveAndDeserializeToCdpMessage/Archiving/BigQueryIO.Write/StorageApiLoads/StorageApiWriteInconsistent/Write Records for at least 50m00s without outputting or completing in state finish at java.base@11.0.9/jdk.internal.misc.Unsafe.park(Native Method) at
Info	2023-02-28 23:28:10.163 EST	java.base@11.0.9/java.util.concurrent.locks.LockSupport.park(LockSupport.java:194) at
Info	2023-02-28 23:28:10.165 EST	java.base@11.0.9/java.util.concurrent.locks.AbstractQueuedSynchronizer.parkAndCheckInterrupt(AbstractQueuedSynchronizer.java:885) at
Info		java.base@11.0.9/java.util.concurrent.locks.AbstractQueuedSynchronizer.doAcquireSharedInterruptibly(AbstractQueuedSynchronizer.java:1039) at
Info		java.base@11.0.9/java.util.concurrent.locks.AbstractQueuedSynchronizer.acquireSharedInterruptibly(AbstractQueuedSynchronizer.java:1345) at
Info		java.base@11.0.9/java.util.concurrent.CountDownLatch.await(CountDownLatch.java:232) at app/org.apache.beam.sdk.io.gcp.bigquery.RetryManager\$Callback.await(RetryManager.java:156)
Info		at app/org.apache.beam.sdk.io.gcp.bigquery.RetryManager\$Operation.await(RetryManager.java:139) at
Info		app/org.apache.beam.sdk.io.gcp.bigquery.RetryManager.await(RetryManager.java:258) at
Info		app/org.apache.beam.sdk.io.gcp.bigquery.StorageApiWriteUnshardedRecords\$WriteRecordsDoFn.flushAll(StorageApiWriteUnshardedRecords.java:664) at
Info		app/org.apache.beam.sdk.io.gcp.bigquery.StorageApiWriteUnshardedRecords\$WriteRecordsDoFn.finishBundle(StorageApiWriteUnshardedRecords.java:747) at
Info		app/org.apache.beam.sdk.io.gcp.bigquery.StorageApiWriteUnshardedRecords\$WriteRecordsDoFn\$DoFnInvoker.invokeFinishBundle(Unknown Source)

Release Notes

Open in Logs Explorer

Troubleshoot slow/stuck dataflow jobs

Last 30 days Search all fields

1 resource.type="dataflow_step"
2 resource.labels.job_id="2023-04-11_08_02_27-301837046300790666"
3

Log fields Histogram

[Create metric](#) [Create alert](#) [Jump to now](#) More actions

Log fields Search fields and values

RESOURCE TYPE Dataflow Step [Clear X](#)

SEVERITY Info 16 Warning 1

LOG NAME [dataflow.googleapis.com/job-message](#) 17

PROJECT ID [60595549251](#) 17

JOB ID [2023-04-11_12_08_02_301837046300](#)

STEP ID Value not present 17

JOB NAME

Histogram

Query results 17 log entries

SEVERITY TIMESTAMP ↑ EDT SUMMARY EDIT

This query has been updated. Run it to view matching entries. [Run query](#)

Severity	Timestamp	Message
Info	2023-05-10 00:59:40.173 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Warning	2023-05-10 11:05:50.046 EDT	Internal Issue (a115679b95e60023): 63963027:24112
Info	2023-05-13 06:38:12.417 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-13 06:38:25.223 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
Info	2023-05-22 23:18:24.798 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-22 23:18:42.714 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
Info	2023-05-23 03:47:00.560 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-23 03:47:24.437 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
Info	2023-05-26 17:25:06.517 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-26 17:25:21.435 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
Info	2023-05-26 22:03:39.756 EDT	Worker configuration: n1-standard-4 in us-central1-c.

Troubleshoot slow/stuck dataflow jobs

Last 30 days Relative time (Ex: 15m, 1h, 1d, 1w)

08_02_27-301837046300790666*

Last 15 seconds 15s
Last 30 seconds 30s
Last 1 minute 1m
Last 5 minutes 5m
Last 10 minutes 10m
Last 15 minutes 15m
Last 30 minutes 30m
Last 45 minutes 45m
Last 1 hour 1h
Last 3 hours 3h
Last 6 hours 6h
Last 12 hours 12h
Last 1 day 1d
Last 2 days 2d
Last 7 days 7d
Last 14 days 14d
Last 30 days 30d

Start and end times >
 Around a time >
 Time zone: EDT (UTC-4) >

Dataflow Step Log name Severity +1 filter Show query

Create metric Create alert Jump to now More actions

Histogram

May 2, 2:00AM May 5 May 7 May 9 May 11 May 13 May 15 May 17 May 19 May 21 May 23 May 25 May 27 May 29 Jun 1, 8:00AM

Query results 17 log entries

SEVERITY TIMESTAMP ↑ EDIT SUMMARY EDIT

This query has been updated. Run it to view matching entries.

Severity	Timestamp	Message
> i	2023-05-10 00:59:40.173 EDT	Worker configuration: n1-standard-4 in us-central1-c.
> !	2023-05-10 11:05:50.046 EDT	Internal Issue (a115679b95e60023): 63963027:24112
> i	2023-05-13 06:30:12.417 EDT	Worker configuration: n1-standard-4 in us-central1-c.
> i	2023-05-13 06:30:25.223 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
> i	2023-05-22 23:18:24.798 EDT	Worker configuration: n1-standard-4 in us-central1-c.
> i	2023-05-22 23:18:42.714 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
> i	2023-05-23 03:47:00.560 EDT	Worker configuration: n1-standard-4 in us-central1-c.
> i	2023-05-23 03:47:24.437 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
> i	2023-05-26 17:25:06.517 EDT	Worker configuration: n1-standard-4 in us-central1-c.
> i	2023-05-26 17:25:21.435 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri-
> i	2023-05-26 22:03:39.756 EDT	Worker configuration: n1-standard-4 in us-central1-c.

Find in results Correlate by Download

Troubleshoot slow/stuck dataflow jobs

Last 30 days Search all fields Dataflow Step Log name Severity +1 filter Show query

resource.type="dataflow_step"
resource.labels.job_id="2023-04-11_08_02_27-301837046300790666"
...

Log fields Histogram Create metric Create alert Jump to now More actions

Log fields Search fields and values RESOURCE TYPE Dataflow Step Clear X SEVERITY Info 16 Warning 1 LOG NAME dataflow.googleapis.com/job-message 17 PROJECT ID 605955549251 17 JOB ID 2023-04-11_12_08_02_301837046300 STEP ID Value not present 17 JOB NAME

Histogram May 2, 2:00AM Jun 1, 8:00AM

Query results 17 log entries

SEVERITY	TIMESTAMP	SUMMARY
Info	2023-05-10 00:59:40.173 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Warning	2023-05-10 11:05:50.046 EDT	Internal Issue (a115679b95e60023): 63963827:24112
Info	2023-05-13 06:30:12.417 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-13 06:30:25.223 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
Info	2023-05-22 23:18:24.798 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-22 23:18:42.714 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
Info	2023-05-23 03:47:00.560 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-23 03:47:24.437 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
Info	2023-05-26 17:25:06.517 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Info	2023-05-26 17:25:21.435 EDT	Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
Info	2023-05-26 22:03:39.756 EDT	Worker configuration: n1-standard-4 in us-central1-c.

Troubleshoot slow/stuck dataflow jobs

Last 30 days Search all fields Dataflow Step Log name Severity +1 filter Show query

resource.type="dataflow_step"
resource.labels.job_id="2023-04-11_08_02_27-301837046300790666"
3

Log fields Histogram

CREATE metric CREATE alert JUMP to now More actions

Log fields Search fields and values RESOURCE TYPE Dataflow Step Clear X SEVERITY

Info 16 Warning 1 LOG NAME dataflow.googleapis.com/job-message 17

PROJECT ID 605955549251 17 JOB ID 2023-04-11_12_08_02_3018370463007 STEP ID Value not present 17 JOB NAME

Histogram

May 2, 2:00 AM May 5 May 7 May 9 May 11 May 13 May 15 May 17 May 19 May 21 May 23 May 25 May 27 May 29 Jun 1, 8:00 AM

Query results 17 log entries

SEVERITY TIMESTAMP ↑ EDIT SUMMARY EDIT

This query has been updated. Run it to view matching entries. Run query

> i 2023-05-10 00:59:40.173 EDT Worker configuration: n1-standard-4 in us-central1-c.
> ! 2023-05-10 11:05:58.046 EDT Internal Issue (a115679b95e68023): 63963027:24112
> i 2023-05-13 06:38:12.417 EDT Worker configuration: n1-standard-4 in us-central1-c.
> i 2023-05-13 06:38:25.223 EDT Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
> i 2023-05-22 23:18:24.798 EDT Worker configuration: n1-standard-4 in us-central1-c.
> i 2023-05-22 23:18:42.714 EDT Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
> i 2023-05-23 03:47:00.560 EDT Worker configuration: n1-standard-4 in us-central1-c.
> i 2023-05-23 03:47:24.437 EDT Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
> i 2023-05-26 17:25:06.517 EDT Worker configuration: n1-standard-4 in us-central1-c.
> i 2023-05-26 17:25:21.435 EDT Your project already contains 100 Dataflow-created metric descriptors, so new user metrics of the form custom.googleapis.com/* will not be created. However, all user metri...
> i 2023-05-26 22:03:39.756 EDT Worker configuration: n1-standard-4 in us-central1-c.

Find in results Correlate by Download

Troubleshoot slow/stuck dataflow jobs

You are missing one or more permissions required to use the query library. [Learn more](#).

Query Saved (0) Suggested (3) Library

5/27/23, 3:23 PM – 5/30/23, 11:46 AM Search all fields

Dataflow Step Log name Severity +1 filter

resource.type="dataflow_step"
resource.labels.job_id="2023-04-11_08_02_27-30183704300790666"

Log fields Histogram

RESOURCE TYPE: Dataflow Step

SEVERITY: Debug (52), Info (10), Error (2)

Histogram: May 27, 3:00 PM – May 28

Query results: 65 log entries

SEVERITY: EDT SUMMARY EDIT

This query has been updated. Run it to view matching entries.

To view older entries: Extend time by: 1 day, Edit time

Select log names

Search log names

- system
- vm-health
- vm-monitor
- worker
- worker-startup
- CLOUD MONITORING API
- ViolationAutoResolveEventV1
- ViolationOpenEventV1

Cancel Apply

Select which logs you want to view from here:

- worker-startup
- worker
- docker & kubelet
- shuffler

Troubleshoot slow/stuck dataflow jobs

Logs Explorer REFINE SCOPE Project SHARE LINK

Query Saved (0) Suggested (3) Library Clear query Stream logs

5/27/23, 3:23 PM – 5/30/23, 11:46 AM Search all fields Dataflow Step Log name Severity +1 filter ✓

1 resource.type="dataflow_step"
2 resource.labels.job_id="2023-04-11_08_02_27-30183704300790666"
3

Log fields Histogram

Search fields and values

RESOURCE TYPE

Dataflow Step Clear X

SEVERITY

Severity	Count
Debug	52
Info	10
Error	2

Histogram

May 27, 3:00 PM May 28 May 29

Query results 65 log entries Find in results

SEVERITY EDT SUMMARY EDIT

This query has been updated. Run it to view matching entries. Run query

To view older entries: Extend time by: 1 day Edit time

2023-05-27 15:23:47.727 FDT Autoscaling is enabled for job 2023-05-27 12:23:45-449341954761765565. The number of workers will be between 1 and 60.

Select log names Clear X

Search log names

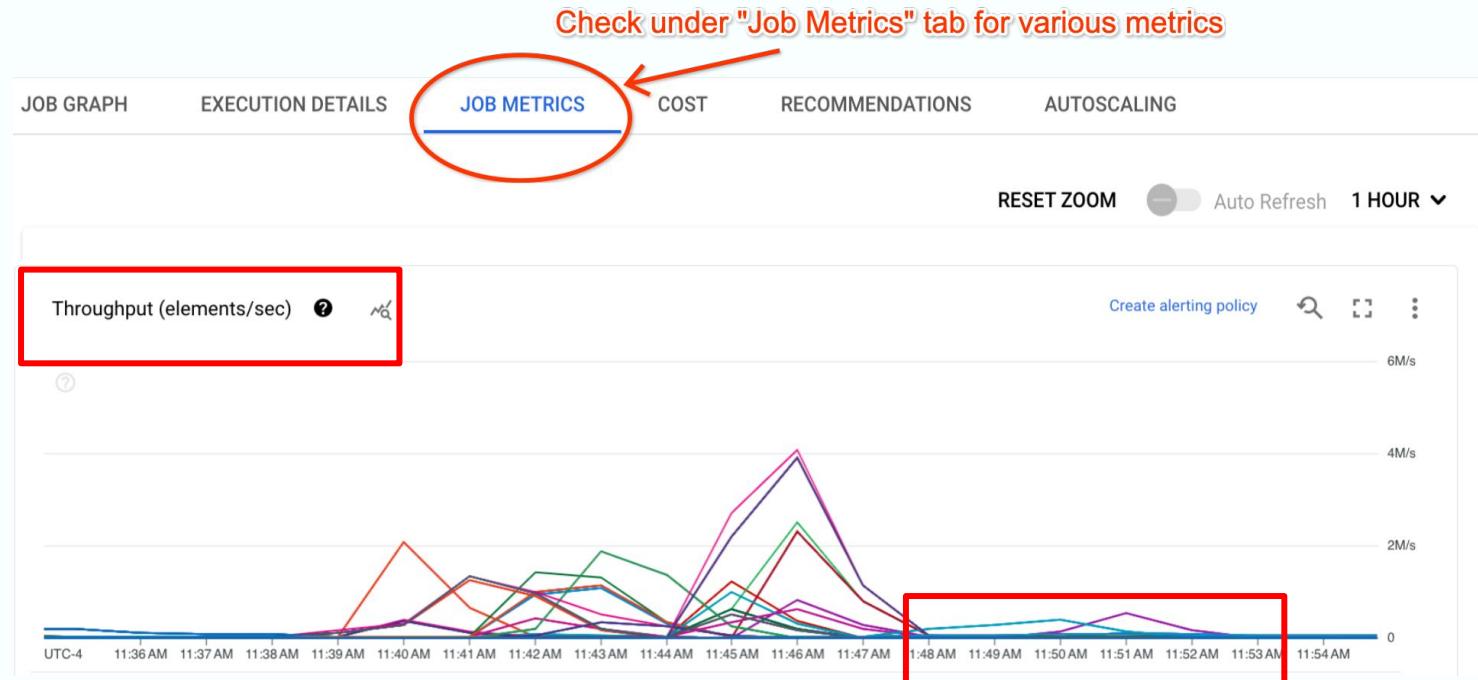
- system dataflow.googleapis.com/system
- vm-health dataflow.googleapis.com/vm-health
- vm-monitor dataflow.googleapis.com/vm-monitor
- worker dataflow.googleapis.com/worker
- worker-startup dataflow.googleapis.com/worker-startup
- CLOUD MONITORING API
- ViolationAutoResolveEventv1 monitoring.googleapis.com/ViolationAutoR...
- ViolationOpenEventv1 monitoring.googleapis.com/ViolationOpenE...

Cancel Apply

Troubleshooting using Job Metrics Tab

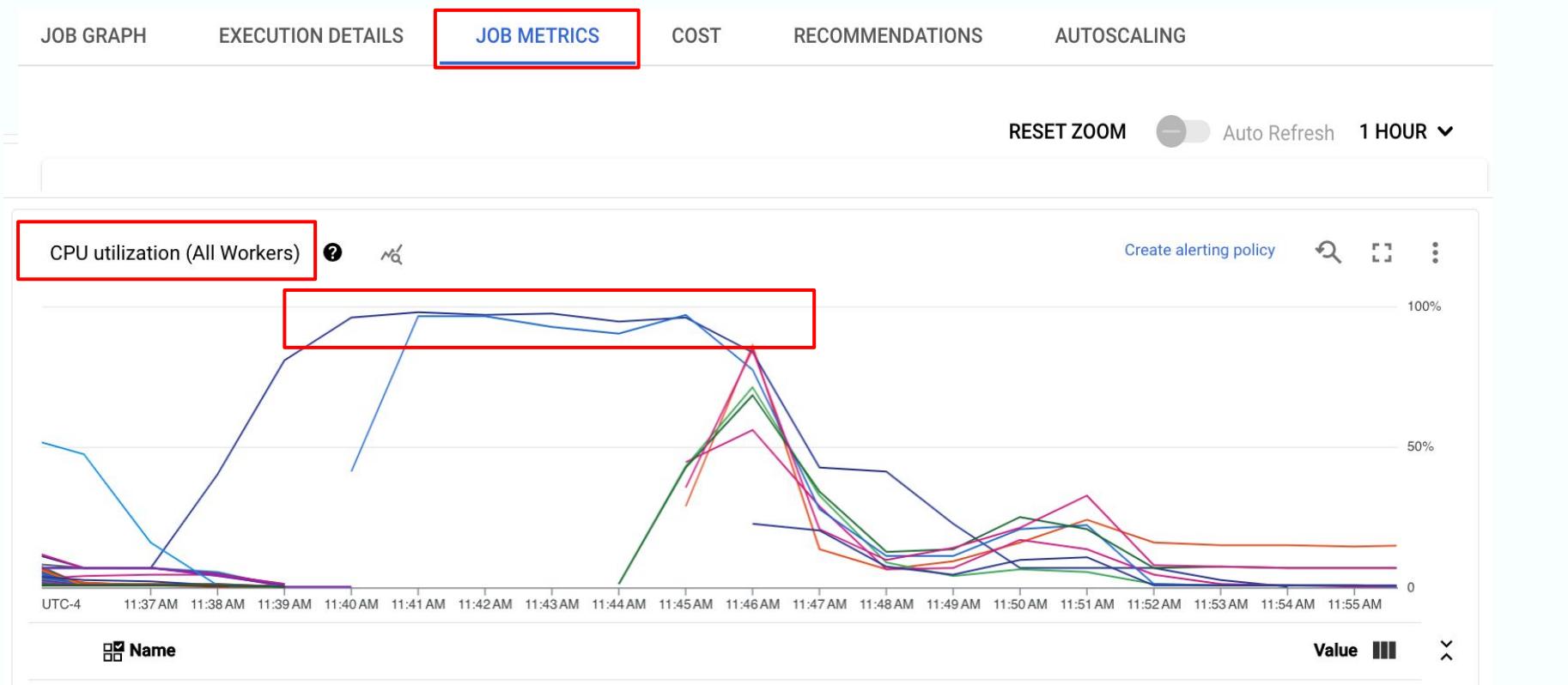
Troubleshoot slow/stuck dataflow jobs

Throughput dropping to zero



Troubleshoot slow/stuck dataflow jobs

High CPU Utilization



Troubleshoot slow/stuck dataflow jobs

High CPU Utilization



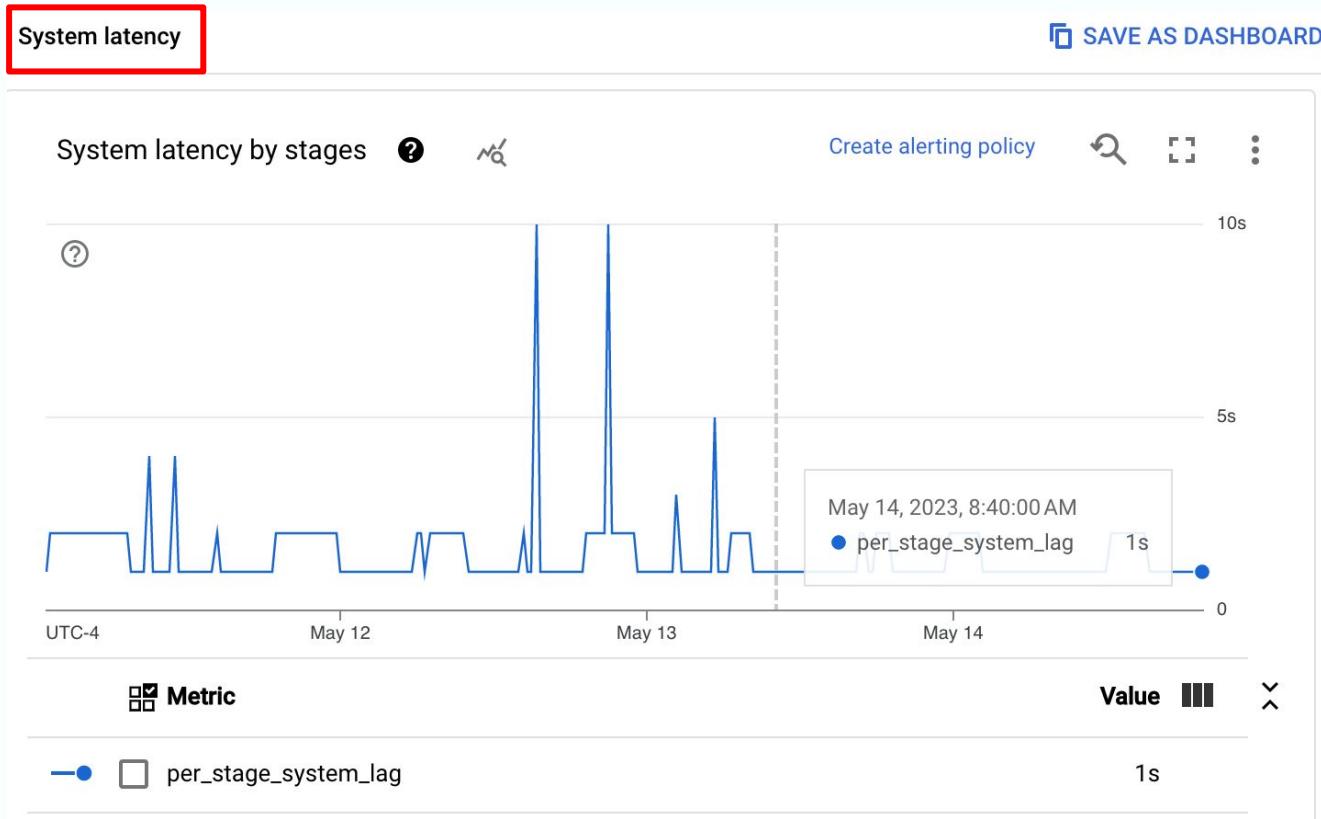
Troubleshoot slow/stuck dataflow jobs

Data Freshness



Troubleshoot slow/stuck dataflow jobs

System Latency



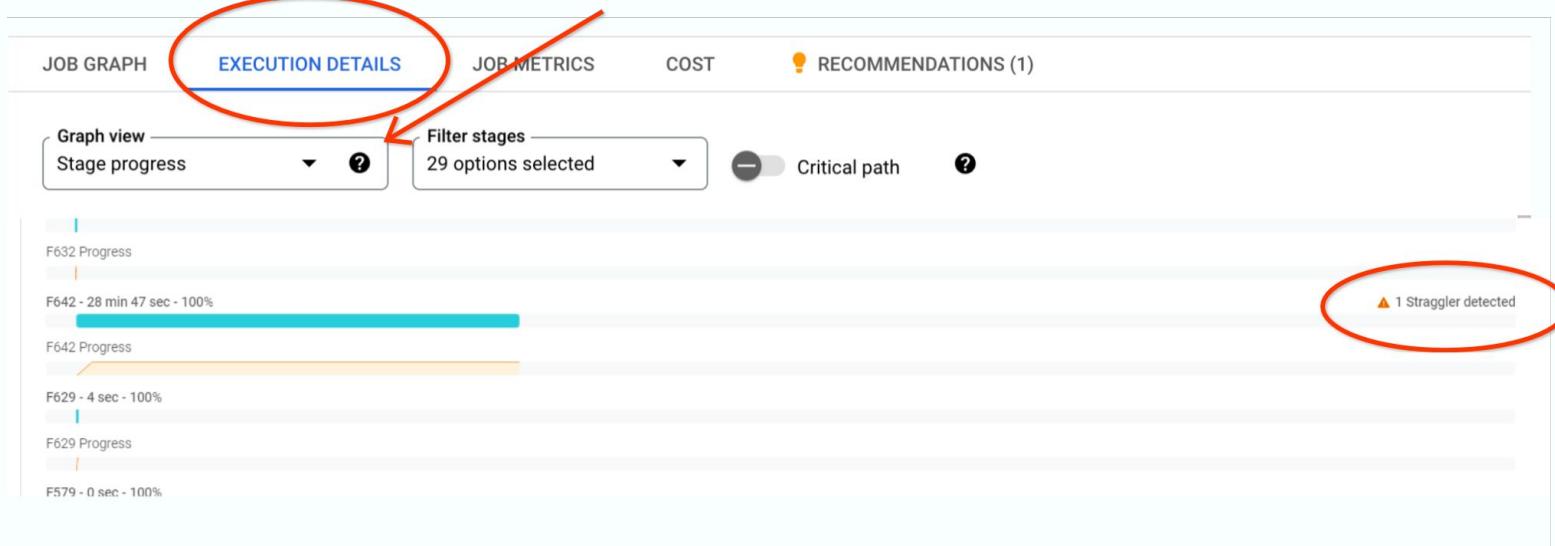
Stragglers in batch job

Troubleshoot slow/stuck dataflow jobs

When a batch job takes a long time to process data, it would be best to check on the [Straggler Workers](#)

How to check it?

Under Execution details, select Stage progress in graph view list



Troubleshoot slow/stuck dataflow jobs

There can be various causes of stragglers:

- **Hot Keys:** Hot keys can create stragglers because they limit ability of Dataflow to process elements in parallel.
 - a. Re-key your data. Apply a ParDo transform to output new key-value pairs.
- **Re-shuffle your data** to avoid a single worker having extra load

Scenario 1: Long active user operation

Troubleshoot slow/stuck dataflow jobs

Processing Stuck/ Operation ongoing



Error

Operation ongoing in step {step name} for at least {duration}

OR

Processing stuck in step {step name} for at least {duration}

Troubleshoot slow/stuck dataflow jobs



Processing Stuck/ Operation ongoing

From Logs Explorer

Query:

Query Saved (0) Suggested (2) Library

Last 14 days Search all fields

```
1 resource.type="dataflow_step"
2 resource.labels.job_id=$JOB_ID
3 logName:"/logs/dataflow.googleapis.com%2Fworker"
```

Results:

⌚ Operation ongoing in step Write to BQ/BatchLoads/SinglePartitionWriteTables/ParMultiDo(WriteTables) for at least 02h20m00s without outputting or completing in state finish.

```
at java.base@11.0.9/java.lang.Thread.sleep(Native Method)
at app//com.google.api.client.util.Sleeper$1.sleep(Sleeper.java:42)
at app//com.google.api.client.util.BackOffUtils.next(BackOffUtils.java:48)
at app//org.apache.beam.sdk.io.gcp.bigquery.BigQueryHelpers$PendingJobManager.nextBackOff(BigQueryHelpers.java:162)
at app//org.apache.beam.sdk.io.gcp.bigquery.BigQueryHelpers$PendingJobManager.waitForDone(BigQueryHelpers.java:148)
at app//org.apache.beam.sdk.io.gcp.bigquery.WriteTables$WriteTablesDoFn.finishBundle(WriteTables.java:380)
at app//org.apache.beam.sdk.io.gcp.bigquery.WriteTables$WriteTablesDoFn$DoFnInvoker.invokeFinishBundle(Unknown Source)
```

Troubleshoot slow/stuck dataflow jobs

Processing Stuck/ Operation ongoing



From Logs Explorer

```
⌚ Operation ongoing in step Write to BQ/BatchLoads/SinglePartitionWriteTables/ParMultiDo(WriteTables) for at least 02h20m00s without outputting or completing  
in state finish  
at java.base@11.0.9/java.lang.Thread.sleep(Native Method)  
at app//com.google.api.client.util.Sleeper$1.sleep(Sleeper.java:42)  
at app//com.google.api.client.util.BackOffUtils.next(BackOffUtils.java:48)  
at app//org.apache.beam.sdk.io.gcp.bigquery.BigQueryHelpers$PendingJobManager.nextBackOff(BigQueryHelpers.java:162)  
at app//org.apache.beam.sdk.io.gcp.bigquery.BigQueryHelpers$PendingJobManager.waitForDone(BigQueryHelpers.java:148)  
at app//org.apache.beam.sdk.io.gcp.bigquery.WriteTables$WriteTablesDoFn.finishBundle(WriteTables.java:380)  
at app//org.apache.beam.sdk.io.gcp.bigquery.WriteTables$WriteTablesDoFn$DoFnInvoker.invokeFinishBundle(Unknown Source)
```

<https://github.com/apache/beam/blob/master/sdks/java/io/google-cloud-platform/src/main/java/org/apache/beam/sdk/io/gcp/bigquery/BigQueryHelpers.java>

Troubleshoot slow/stuck dataflow jobs

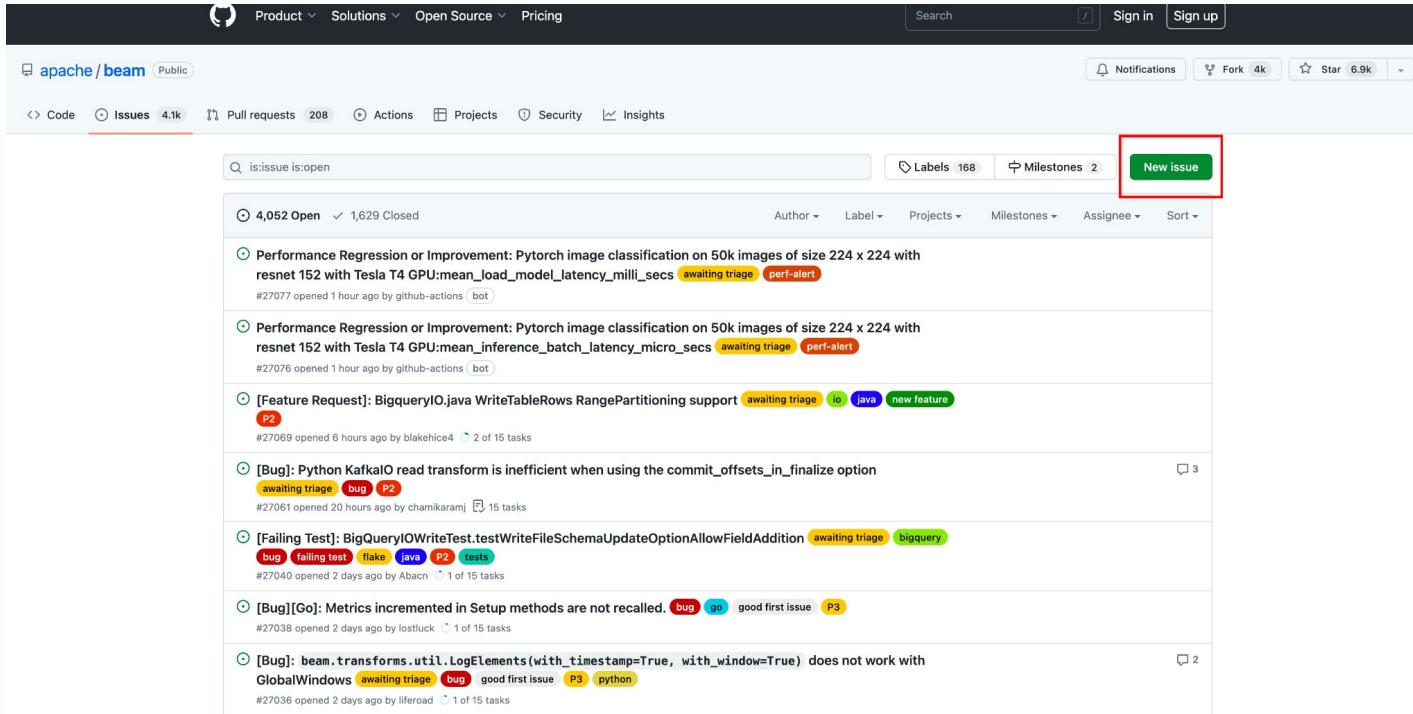
Processing Stuck/ Operation ongoing



SEVERITY	TIMESTAMP	EDT	SUMMARY	EDIT
?	49% of results are similar and can be hidden.			Hide similar entries Preview
> i	2023-05-23 19:04:08.886 EDT		Detected missing event columns in [REDACTED] BigQuery schema. Schema must be updated manually, if required. Dropping/Missing attributes from Event payload. Details	
> i	2023-05-23 19:04:08.886 EDT		Detected missing event columns in [REDACTED] BigQuery schema. Schema must be updated manually, if required. Dropping/Missing attributes from Event payload. Details	
> i	2023-05-23 19:04:08.886 EDT		No BigQuery job with job id beam_bq_job_LOAD_[REDACTED]_0001_0000	
> i	2023-05-23 19:04:08.886 EDT		job id beam_bq_job_LOAD_[REDACTED]_0001_0000-72 not found, so ...	
> !	2023-05-23 19:04:08.886 EDT		Load job beam_bq_job_LOAD_[REDACTED]_0001_0000-71 failed, will...	
> i	2023-05-23 19:04:08.886 EDT		Job beam_bq_job_LOAD_[REDACTED]_0001_0000-72 pending. retrying.	

Troubleshoot slow/stuck dataflow jobs

Apache Beam Issues/Feature Request



The screenshot shows the Apache Beam GitHub repository page. The 'Issues' tab is selected, displaying 4,052 open issues. A search bar at the top contains the query 'is:issue is:open'. To the right of the search bar is a green 'New issue' button, which is highlighted with a red box. Below the search bar, there are filters for Labels (168), Milestones (2), and a 'New issue' button. The main list of issues includes:

- Performance Regression or Improvement: Pytorch image classification on 50k images of size 224 x 224 with resnet 152 with Tesla T4 GPU:mean_load_model_latency_milli_secs awaiting triage perf-alert #27077 opened 1 hour ago by github-actions (bot)
- Performance Regression or Improvement: Pytorch image classification on 50k images of size 224 x 224 with resnet 152 with Tesla T4 GPU:mean_inference_batch_latency_micro_secs awaiting triage perf-alert #27076 opened 1 hour ago by github-actions (bot)
- [Feature Request]: BigqueryIO.java WriteTableRows RangePartitioning support awaiting triage io java new feature P2 #27069 opened 6 hours ago by blakehice 2 of 15 tasks
- [Bug]: Python Kafkalo read transform is inefficient when using the commit_offsets_in_finalize option awaiting triage bug P2 #27061 opened 20 hours ago by chamikaramj 15 tasks
- [Failing Test]: BigQueryIOWriteTest.testWriteFileSchemaUpdateOptionAllowFieldAddition awaiting triage bigquery bug failing test flake java P2 tests #27040 opened 2 days ago by Abacn 1 of 15 tasks
- [Bug][Go]: Metrics incremented in Setup methods are not recalled. bug go good first issue P3 #27038 opened 2 days ago by lostluck 1 of 15 tasks
- [Bug]: beam.transforms.util.LogElements(with_timestamp=True, with_window=True) does not work with GlobalWindows awaiting triage bug good first issue P3 python #27036 opened 2 days ago by liferoad 1 of 15 tasks



Scenario 2: GC Thrashing/OOM

Troubleshoot slow/stuck dataflow jobs



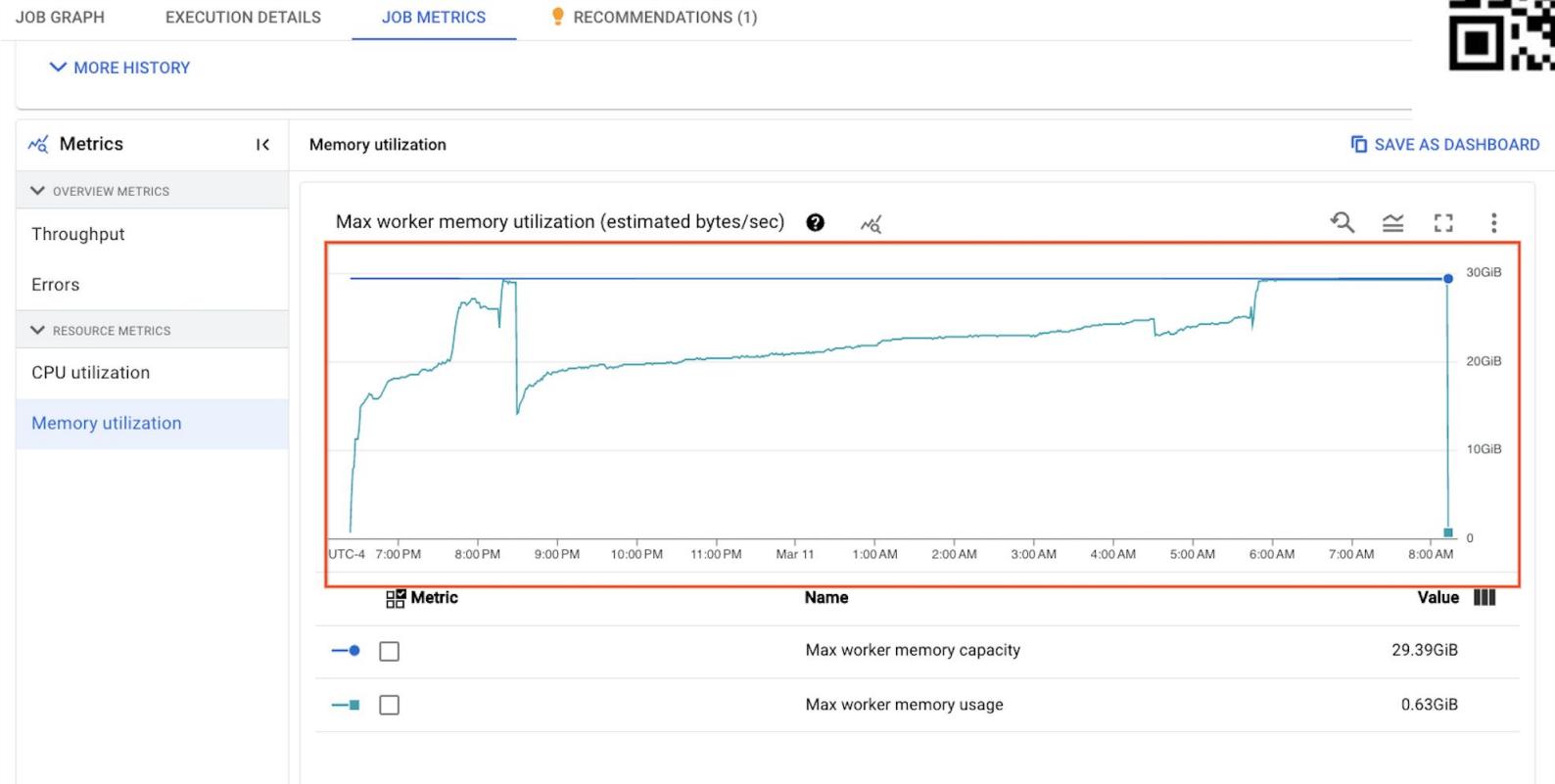
GC Thrashing/OOM: Diagnostics Tab

Logs			HIDE	! 30	▼
JOB LOGS		WORKER LOGS		⚠ DIAGNOSTICS	First
Occurrences	Count	Error	Dec 1 2022		
	8	Shutting down JVM after 8 consecutive periods of measured GC thrashing. Memory is used/total/max = 7904/20103/37513 MB, GC last/max = 90.03/95.7...			
		The worker was shut down after a long period of high memory pressure.			
	1	StatusRuntimeException: UNAVAILABLE: keepalive watchdog timeout			

Troubleshoot slow/stuck dataflow jobs



GC Thrashing/OOM





General Recommendations

- Use machine types with higher memory
 - [Link: goo.gle/45USWe3](https://goo.gle/45USWe3)
- Decrease the parallelism of processing by reducing the number of worker harness threads
 - [Link: goo.gle/45RM6WT](https://goo.gle/45RM6WT)
- Do vertical autoscaling (Enable Dataflow Prime)
 - [Link: goo.gle/3r3KZjv](https://goo.gle/3r3KZjv)

Performance Optimization using Dataflow profiling

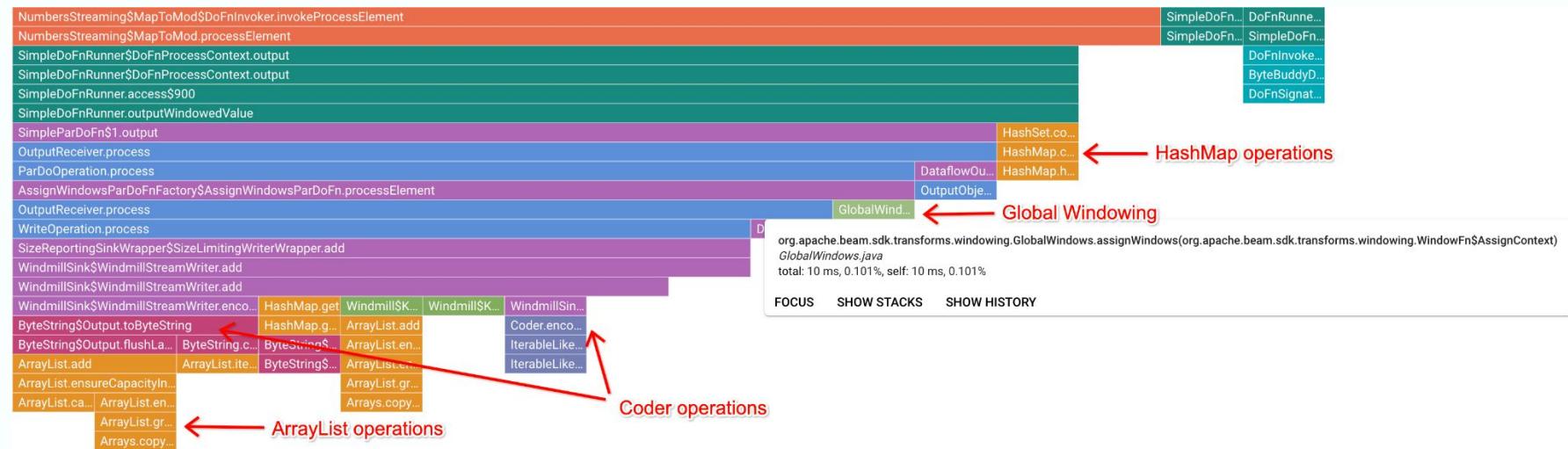
Cloud Profiler is available for Dataflow pipelines written in Apache Beam SDK for Java and Python, version 2.33.0 or later.



It can be enabled at pipeline start time.

E.g. For Java SDK, to enable CPU profiling, start the pipeline with the following option

```
--dataflowServiceOptions=enable_google_cloud_profiler
```



QUESTIONS?

mhkgupta@google.com

linkedin.com/in/mhkgupta