

BEAM  
SUMMIT

# Troubleshooting Slow Running Beam Pipelines

## About Me



**Hello!**

**I'm Mehak**

**Technical Solutions Specialist at Google Cloud**

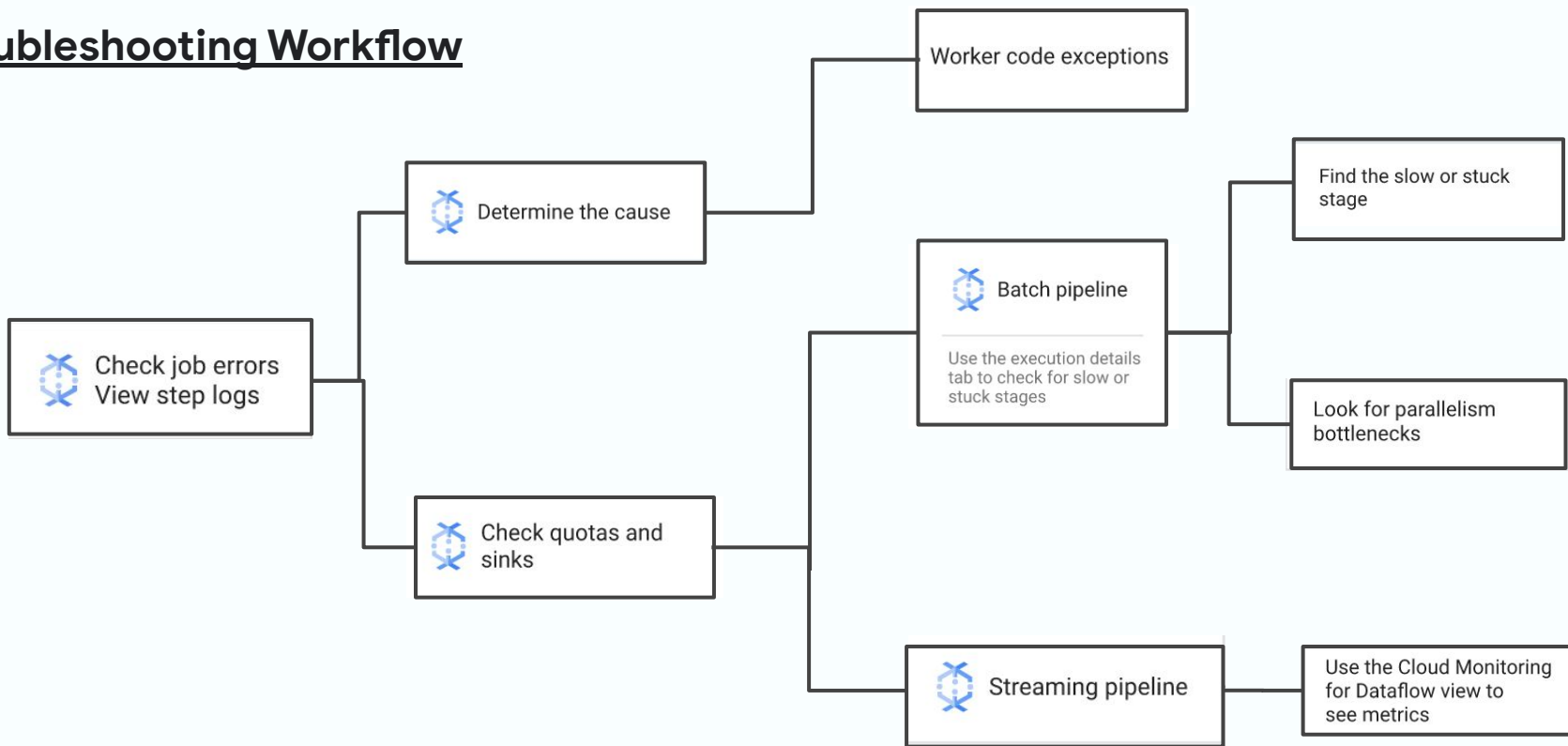


- Apache Beam pipeline troubleshooting techniques that would empower professionals to research and resolve Beam issues by themselves.
- Self service skills would reduce MTTR (Mean Time To Recover) significantly
- Share some tricks and samples of troubleshooting slow running beam pipelines using Dataflow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



# Troubleshoot slow/stuck dataflow jobs

Dataflow

Jobs [+ CREATE JOB FROM TEMPLATE](#) [+ CREATE MANAGED DATA PIPELINE](#) [ENABLE SORTING](#) [REFRESH](#) [LEARN](#)

Running [Filter](#) Filter jobs ? ||

Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region	Insights <span>?</span>
<b>wordcount6</b>	Streaming		48 days 19 hr	Apr 11, 2023, 3:08:03 PM	Running	<span>i</span> 2.43.0	2023-04-11 12 08 02-301837046300790666	us-east1	
<b>wordcount5</b>	Batch	May 28, 2023, 12:31:04 PM	21 hr 30 min	May 27, 2023, 3:00:29 PM	Succeeded	<span>i</span> 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

The screenshot shows the Google Cloud Dataflow console interface. The top navigation bar includes 'Overview', 'JOB GRAPH', 'EXECUTION DETAILS', 'JOB METRICS' (selected), 'RECOMMENDATIONS', and 'AUTOSCALING'. The left sidebar contains 'Jobs', 'Pipelines', 'Workbench', 'Snapshots', and 'SQL Workspace'. The main content area is titled 'Data freshness' and includes a 'Metrics' sidebar with options like 'System latency', 'Throughput', 'Errors', 'Backlog', 'Processing', and 'Parallelism'. A line chart shows 'Data freshness by stages' over time, with a sharp increase starting at 10:30 PM. Below the chart, the 'Logs' section is highlighted with a red box, showing 'JOB LOGS' and 'WORKER LOGS'. A red arrow points from the text 'Check logs here' to this section. The log table below shows several entries with timestamps and summaries. The last log entry is expanded, showing a detailed stack trace. A red box highlights the 'Open in Logs Explorer' button at the bottom right of the log entry, with a red arrow pointing to it.

Severity: Info Filter Search all fields and values

SEVERITY	TIMESTAMP	SUMMARY
Info	2023-02-28 23:20:18.161 EST	Operation ongoing in step ArchiveAndDeserializeToCdpMessage/Archiving/BigQueryIO.Write/StorageApiLoads/StorageApiWriteInconsistent/Write Records for at least 50m0s without outpu...
Info	2023-02-28 23:20:18.163 EST	Operation ongoing in step ArchiveAndDeserializeToCdpMessage/Archiving/BigQueryIO.Write/StorageApiLoads/StorageApiWriteInconsistent/Write Records for at least 50m0s without outpu...
Info	2023-02-28 23:20:18.165 EST	Operation ongoing in step ArchiveAndDeserializeToCdpMessage/Archiving/BigQueryIO.Write/StorageApiLoads/StorageApiWriteInconsistent/Write Records for at least 50m0s without outputting or completing in state finish at java.base@11.0.9/jdk.internal.misc.Unsafe.park(Native Method) at java.base@11.0.9/java.util.concurrent.locks.LockSupport.park(LockSupport.java:194) at java.base@11.0.9/java.util.concurrent.locks.AbstractQueuedSynchronizer.parkAndCheckInterrupt(AbstractQueuedSynchronizer.java:885) at java.base@11.0.9/java.util.concurrent.locks.AbstractQueuedSynchronizer.doAcquireSharedInterruptibly(AbstractQueuedSynchronizer.java:1039) at java.base@11.0.9/java.util.concurrent.locks.AbstractQueuedSynchronizer.acquireSharedInterruptibly(AbstractQueuedSynchronizer.java:1345) at 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) at app/org.apache.beam.sdk.io.gcp.bigquery.RetryManager\$Operation.await(RetryManager.java:139) at app/org.apache.beam.sdk.io.gcp.bigquery.RetryManager.await(RetryManager.java:258) at app/org.apache.beam.sdk.io.gcp.bigquery.StorageApiWriteUnshardedRecords\$WriteRecordsDoFn.flushAll(StorageApiWriteUnshardedRecords.java:664) at app/org.apache.beam.sdk.io.gcp.bigquery.StorageApiWriteUnshardedRecords\$WriteRecordsDoFn.finishBundle(StorageApiWriteUnshardedRecords.java:744) at app/org.apache.beam.sdk.io.gcp.bigquery.StorageApiWriteUnshardedRecords\$WriteRecordsDoFnDoFnInvoker.invokeFinishBundle(Unknown Source)

Open in Logs Explorer

# Troubleshoot slow/stuck dataflow jobs

🕒 Last 30 days 🔍 Search all fields

Dataflow Step Log name Severity +1 filter  Show query

```
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

605955549251 17

JOB ID

2023-04-11\_12\_08\_02-3018370463007 17

STEP ID

Value not present 17

JOB NAME

Histogram

Query results 17 log entries

Find in results Correlate by Download

SEVERITY	TIMESTAMP	SUMMARY
This query has been updated. Run it to view matching entries. <a href="#">Run query</a>		
Info	2023-05-10 00:59:40.173 EDT	Worker configuration: n1-standard-4 in us-central1-c.
Warning	2023-05-10 11:05:58.046 EDT	Internal Issue (a115679b95e60023): 63963027:24112
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.

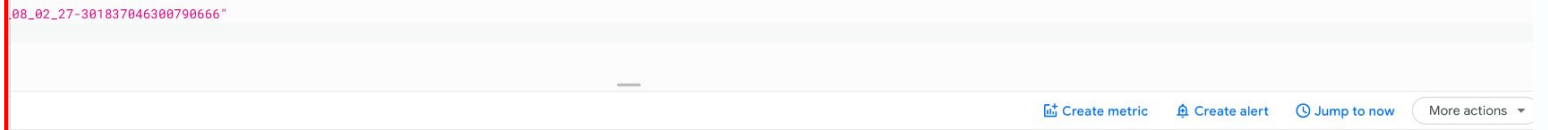
# Troubleshoot slow/stuck dataflow jobs

⌂ Last 30 days 🔍 Search all fields

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

- 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



Query results 17 log entries Find in results Correlate by Download

SEVERITY	TIMESTAMP ↑	EDT ▼	SUMMARY EDIT
🕒	This query has been updated. Run it to view matching entries. Run query		
>	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
>	2023-05-13 06:30:12.417 EDT		Worker configuration: n1-standard-4 in us-central1-c.
>	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-
>	2023-05-22 23:18:24.798 EDT		Worker configuration: n1-standard-4 in us-central1-c.
>	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-
>	2023-05-23 03:47:00.560 EDT		Worker configuration: n1-standard-4 in us-central1-c.
>	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-
>	2023-05-26 17:25:06.517 EDT		Worker configuration: n1-standard-4 in us-central1-c.
>	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-
>	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

```
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

- 605955549251 17

JOB ID

- 2023-04-11\_12\_08\_02\_3018370463007

STEP ID

Value not present 17

JOB NAME

2

Query results 17 log entries

Find in results Correlate by Download

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

SEVERITY	TIMESTAMP	SUMMARY
> 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.

# Troubleshoot slow/stuck dataflow jobs

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

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

Log fields Histogram

Log fields: Search fields and values

- RESOURCE TYPE
  - Dataflow Step
- 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

Query results 17 log entries

Find in results Correlate by Download

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

SEVERITY	TIMESTAMP	SUMMARY
> 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.

# Troubleshoot slow/stuck dataflow jobs

Logs Explorer REFINE SCOPE Project SHARE LINK

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

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

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

Log fields

RESOURCE TYPE Clear x

- Dataflow Step

SEVERITY

- Debug 52
- Info 10
- Error 2

Histogram 60 0 May 27, 3:00 PM May 28

Query results 65 log entries

SEVERITY	TIMESTAMP	SUMMARY	EDIT
Info		This query has been updated. Run it to view matching entries.	Run
Info		To view older entries: <span>Extend time by: 1 day</span> <span>Edit time</span>	

2023-05-27 15:23:47.727 EDT Autoscaling is enabled for

Select which logs you want to view from here:

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

Select log names Clear X

- 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
- ViolationAutoResolveEvent1 monitoring.googleapis.com/ViolationAutoR...
- ViolationOpenEvent1 monitoring.googleapis.com/ViolationOpenE...

Cancel Apply

# Troubleshoot slow/stuck dataflow jobs

Logs Explorer

REFINE SCOPE Project

SHARE LINK

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

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

Log fields

Search fields and values

RESOURCE TYPE

Dataflow Step

SEVERITY

Debug	52
Info	10
Error	2

Histogram



Query results 65 log entries

SEVERITY TIMESTAMP 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 EDT 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

Search log names

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

Cancel Apply

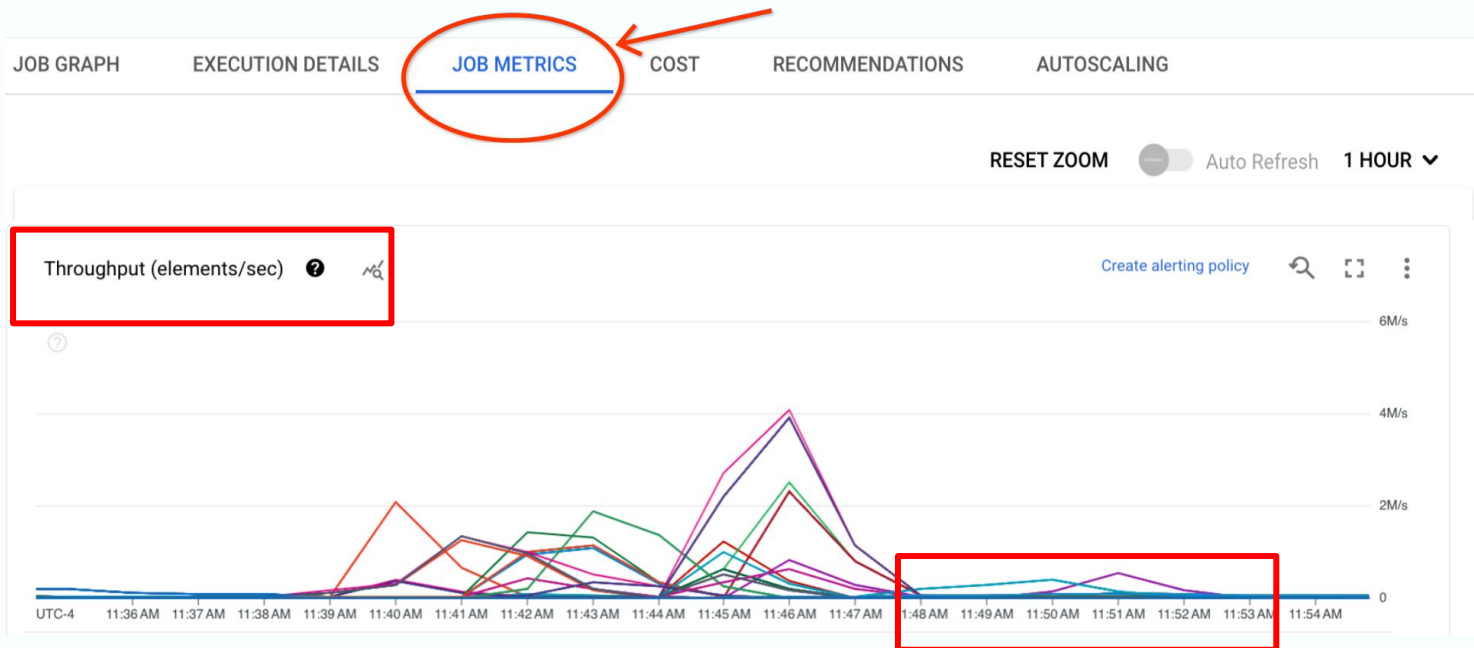
# Troubleshooting using Job Metrics Tab



# Troubleshoot slow/stuck dataflow jobs

## Throughput dropping to zero

Check under "Job Metrics" tab for various metrics



# Troubleshoot slow/stuck dataflow jobs

## High CPU Utilization

JOB GRAPH

EXECUTION DETAILS

**JOB METRICS**

COST

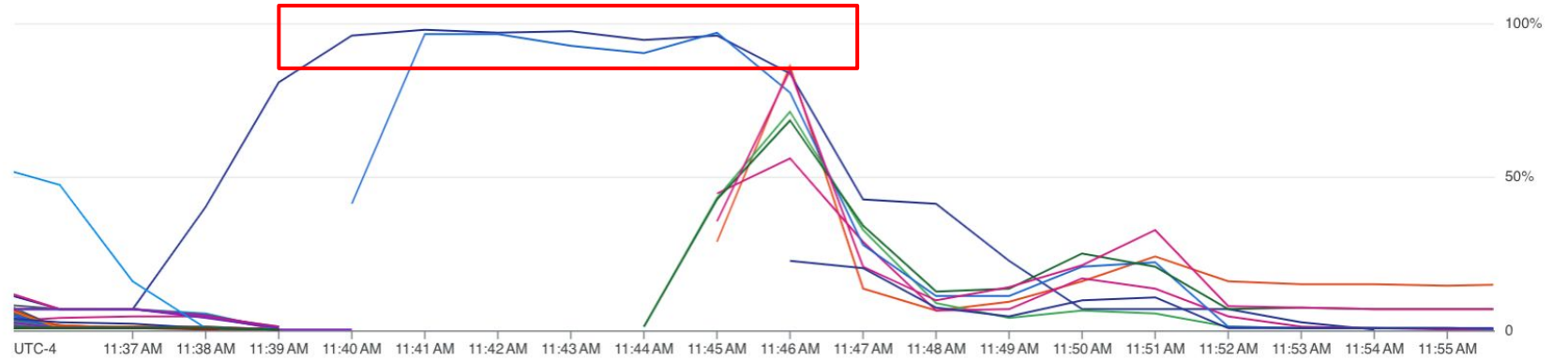
RECOMMENDATIONS

AUTOSCALING

RESET ZOOM  Auto Refresh 1 HOUR ▾

**CPU utilization (All Workers)** ?

Create alerting policy 🔍 🗲 ⋮



# Troubleshoot slow/stuck dataflow jobs

## High CPU Utilization

JOB GRAPH

EXECUTION DETAILS

**JOB METRICS**

COST

RECOMMENDATIONS

AUTOSCALING

RESET ZOOM



Auto Refresh

1 HOUR ▾

Metrics



CPU utilization

OVERVIEW METRICS

Throughput

Errors

RESOURCE METRICS

CPU utilization

Memory utilization

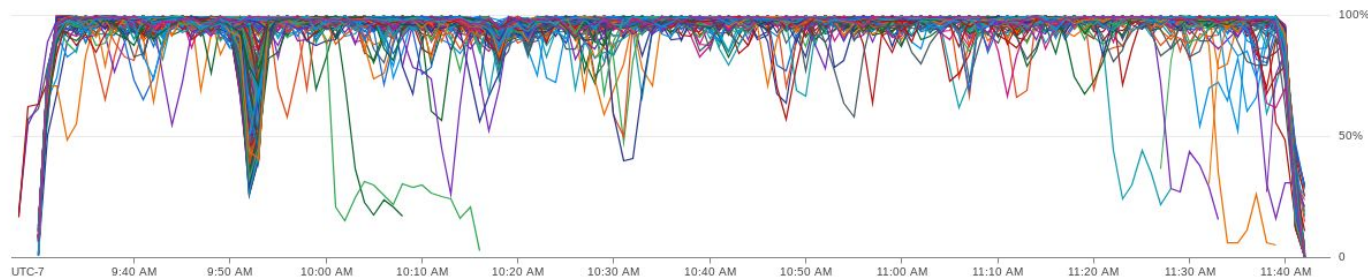
CPU utilization (All Workers) ⓘ



by instance name (mean)

1 min interval (mean)

Create alerting policy



Name

Value

Logs

HIDE



1



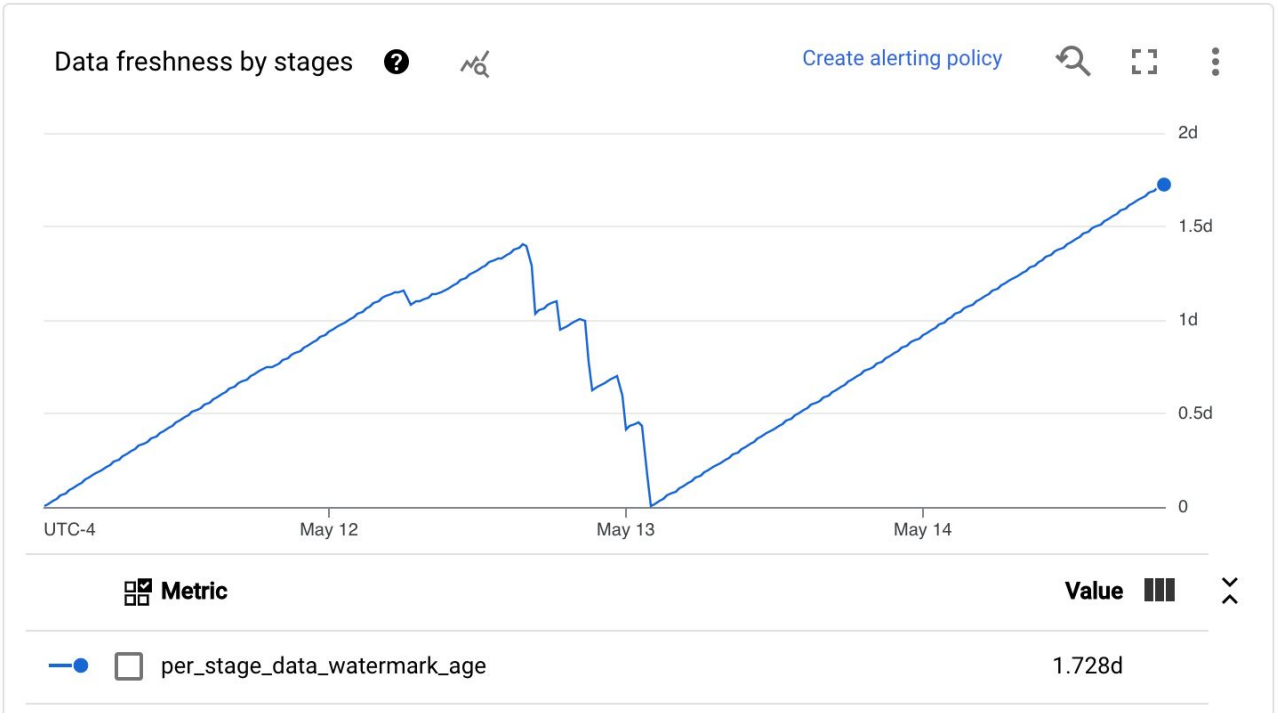
1

# Troubleshoot slow/stuck dataflow jobs

## Data Freshness

Data freshness

SAVE AS DASHBOARD

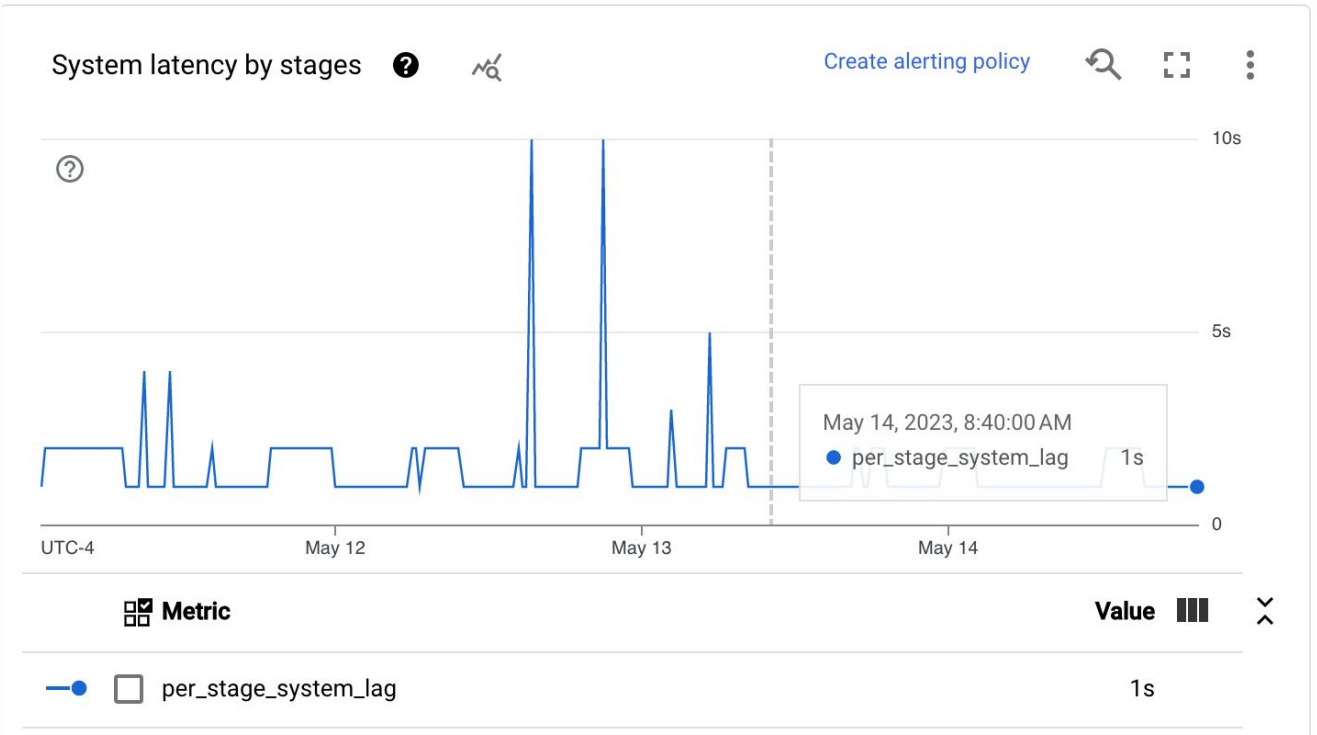


# Troubleshoot slow/stuck dataflow jobs

## System Latency

System latency

SAVE AS DASHBOARD



# 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

The screenshot shows the Databricks job execution details interface. The 'EXECUTION DETAILS' tab is selected and circled in red. An arrow points from this tab to the 'Graph view' dropdown menu, which is also circled in red and set to 'Stage progress'. Below the menu, a list of stages is shown with progress bars. The stage 'F642 - 28 min 47 sec - 100%' is highlighted in blue and has a red circle around the text '1 Straggler detected' on the right side of its progress bar.

Job ID	Stage Name	Duration	Progress	Status
F632	Progress			
F642	28 min 47 sec - 100%		100%	1 Straggler detected
F642	Progress			
F629	4 sec - 100%		100%	
F629	Progress			
F579	0 sec - 100%		100%	

## 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	SUMMARY	EDIT
🔍 49% of results are similar and can be hidden. <a href="#">Hide similar entries</a> <a href="#">Preview</a>			
> 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]_00001_00000	
> i	2023-05-23 19:04:08.886 EDT	job id beam_bq_job_LOAD_[REDACTED]_00001_00000-72 not found, so ...	
> !	2023-05-23 19:04:08.886 EDT	Load job beam_bq_job_LOAD_[REDACTED]_00001_00000-71 failed, will...	
> i	2023-05-23 19:04:08.886 EDT	Job beam_bq_job_LOAD_[REDACTED]_00001_00000-72 pending. retrying.	

# Troubleshoot slow/stuck dataflow jobs

## Apache Beam Issues/Feature Request



Product Solutions Open Source Pricing Search / Sign in Sign up

apache / beam Public Notifications Fork 4k Star 6.9k

Code Issues 4.1k Pull requests 208 Actions Projects Security Insights

is:issue is:open Labels 168 Milestones 2 **New issue**

4,052 Open 1,629 Closed Author Label Projects Milestones Assignee Sort

- 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** **java** **new feature** **P2**  
#27069 opened 6 hours ago by blakehice4 2 of 15 tasks
- [Bug]: Python KafkaIO 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**

Occurrences	Count	Error	First
	8	<b>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...</b> The worker was shut down after a long period of high memory pressure.	Dec 2022
	1	<b>StatusRuntimeException: UNAVAILABLE: keepalive watchdog timeout</b>	Jan 2023



# Troubleshoot slow/stuck dataflow jobs



## GC Thrashing/OOM

JOB GRAPH   EXECUTION DETAILS   **JOB METRICS**   💡 RECOMMENDATIONS (1)

▼ MORE HISTORY

Metrics

IK

Memory utilization

SAVE AS DASHBOARD

OVERVIEW METRICS

Throughput

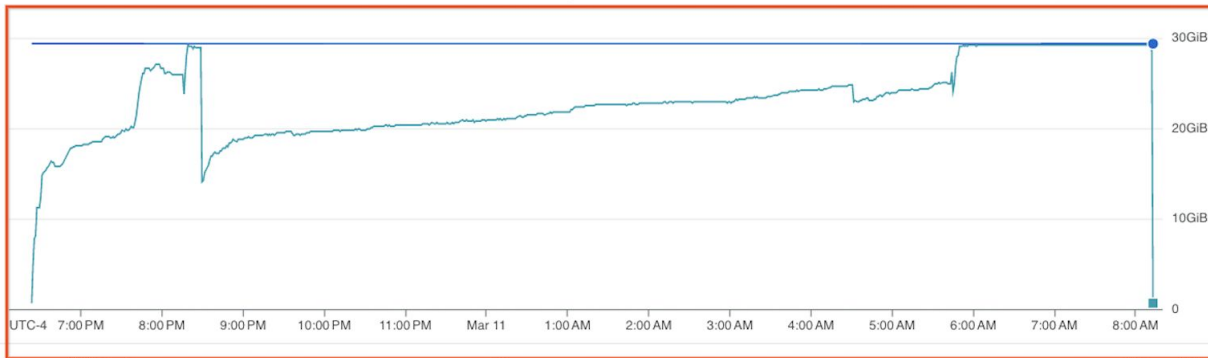
Errors

RESOURCE METRICS

CPU utilization

Memory utilization

Max worker memory utilization (estimated bytes/sec) ?



Metric

Name

Value



Max worker memory capacity

29.39GiB



Max worker memory usage

0.63GiB



### 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)**

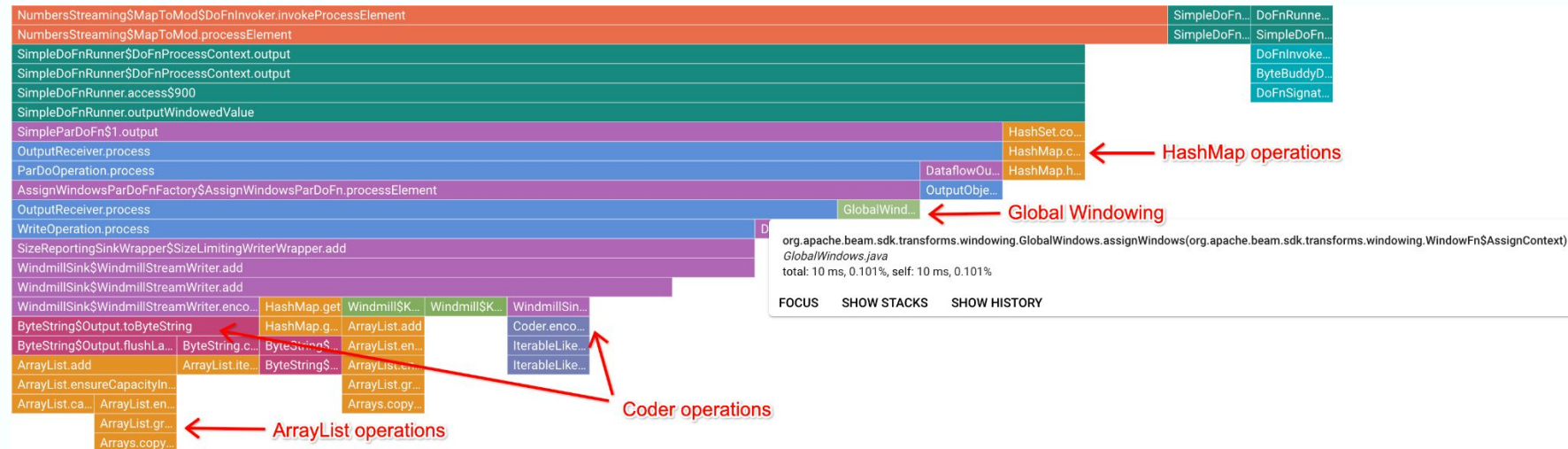
## Performance Optimization using Dataflow profiling

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](mailto:mhkgupta@google.com)

[linkedin.com/in/mhkgupta](https://www.linkedin.com/in/mhkgupta)