Troubleshooting and Debugging

Omar Ismail

Solutions Developer, Google Cloud

CD

Agenda

Course Intro

Monitoring

Logging and Error Reporting

Troubleshooting and Debugging

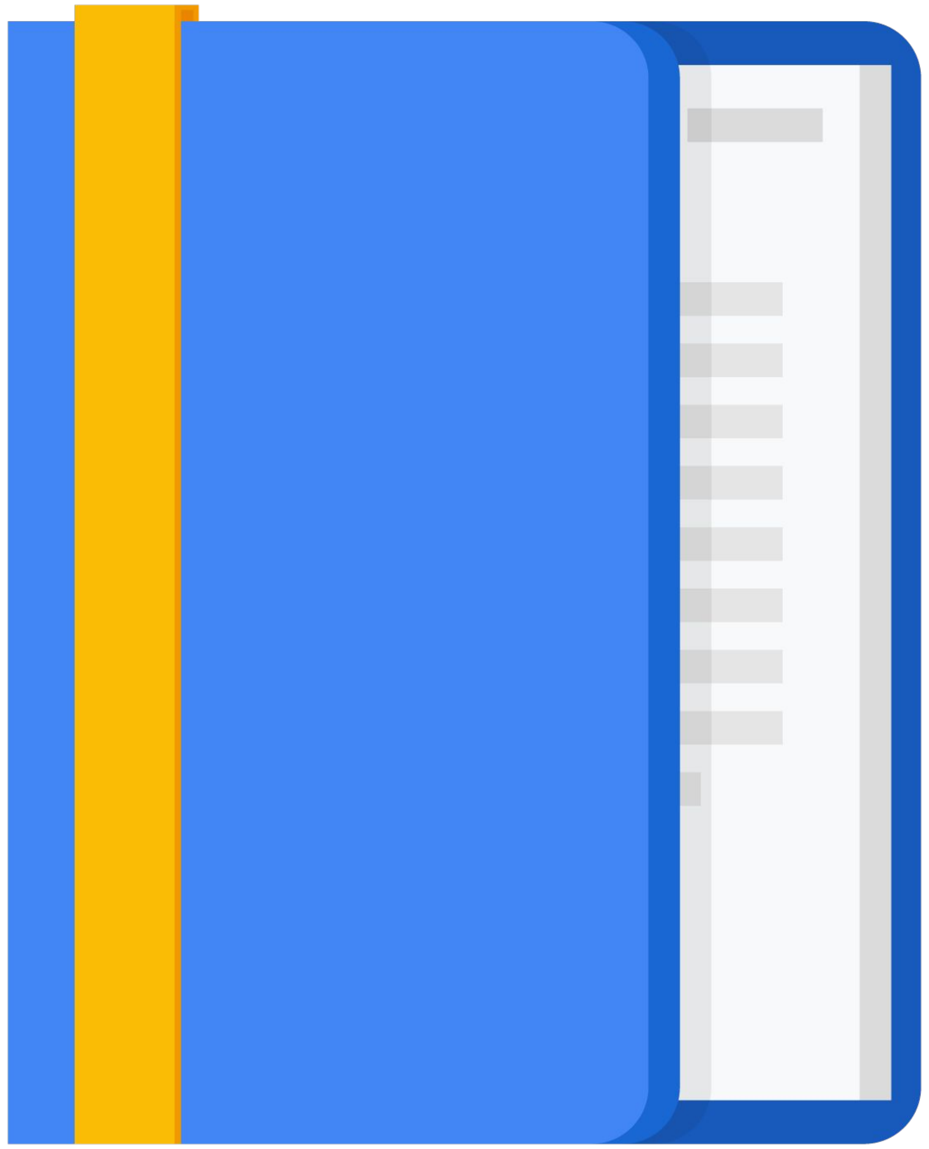
Performance

Testing and CI/CD

Reliability

Flex Templates

Course Summary



Troubleshooting and Debugging

Agenda

Troubleshooting workflow

Types of troubles

Google Cloud

Troubleshooting workflow

Troubleshooting workflow

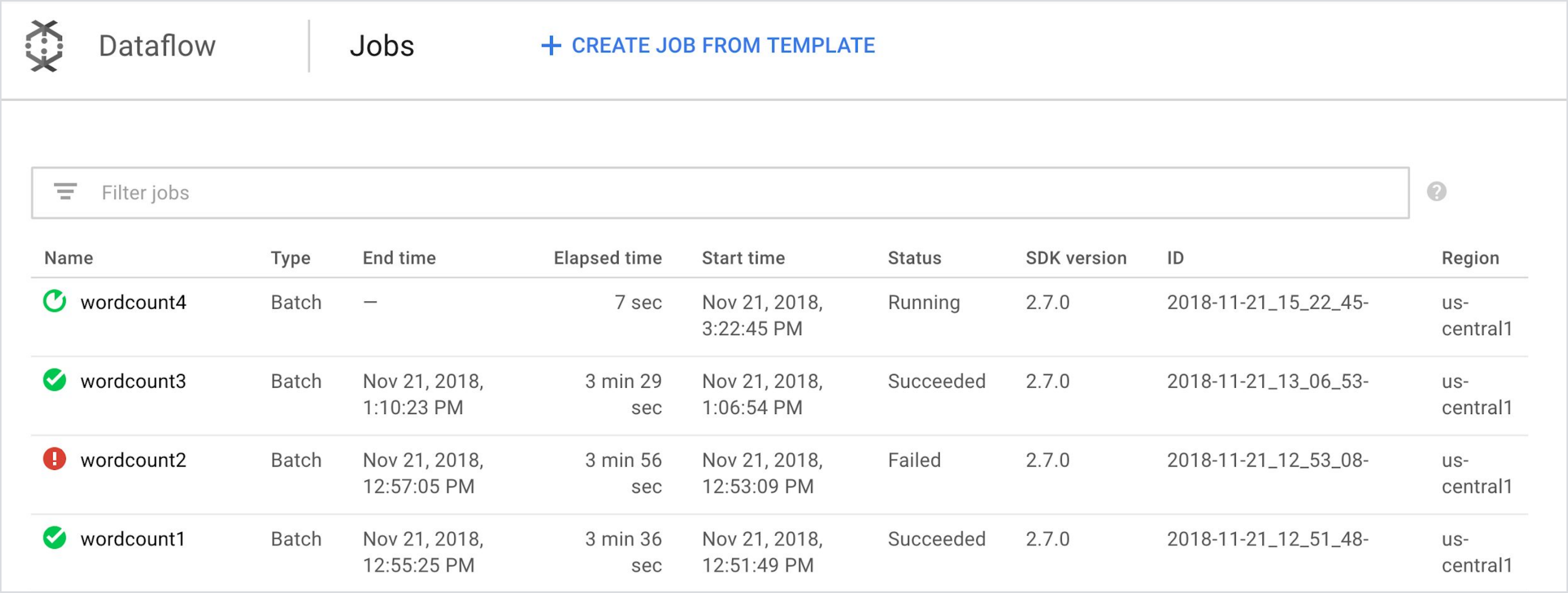
- Checking for errors
- Looking for anomalies in the Job Metrics tab

Troubleshooting workflow

Troubleshooting workflow

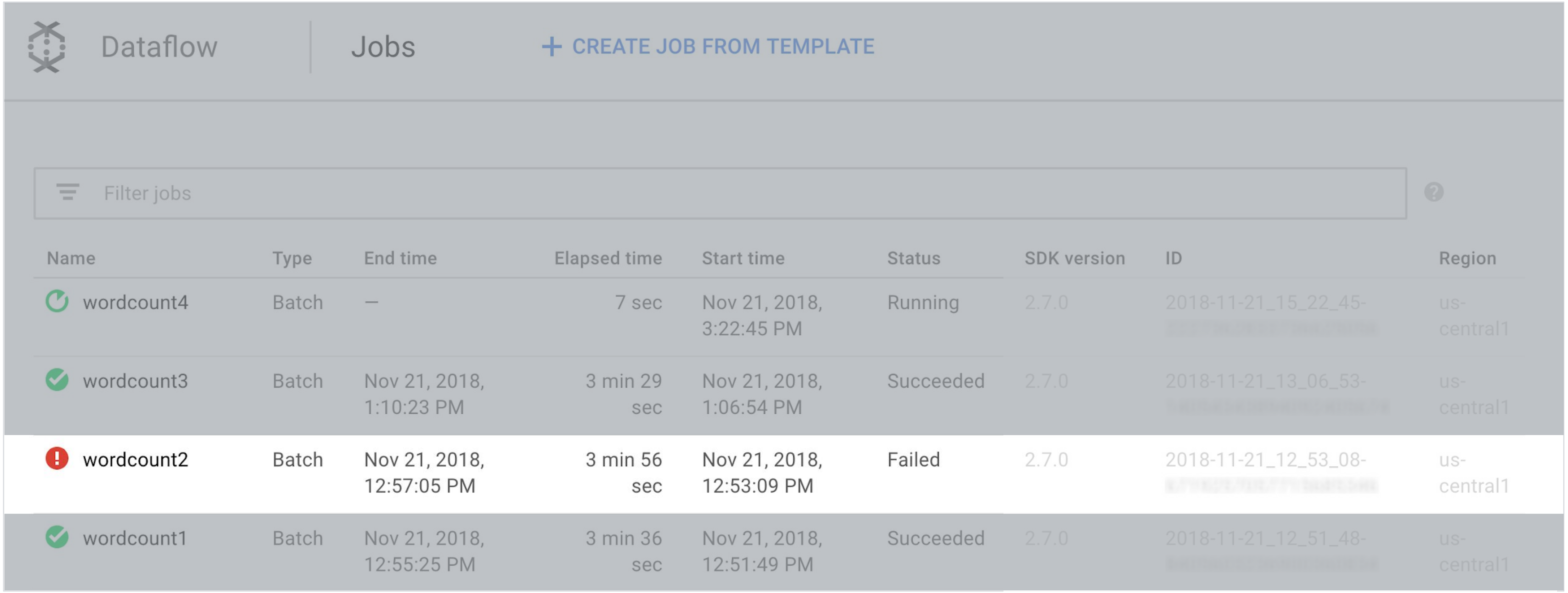
- Checking for errors
- Looking for anomalies in the Job Metrics tab

Checking for errors



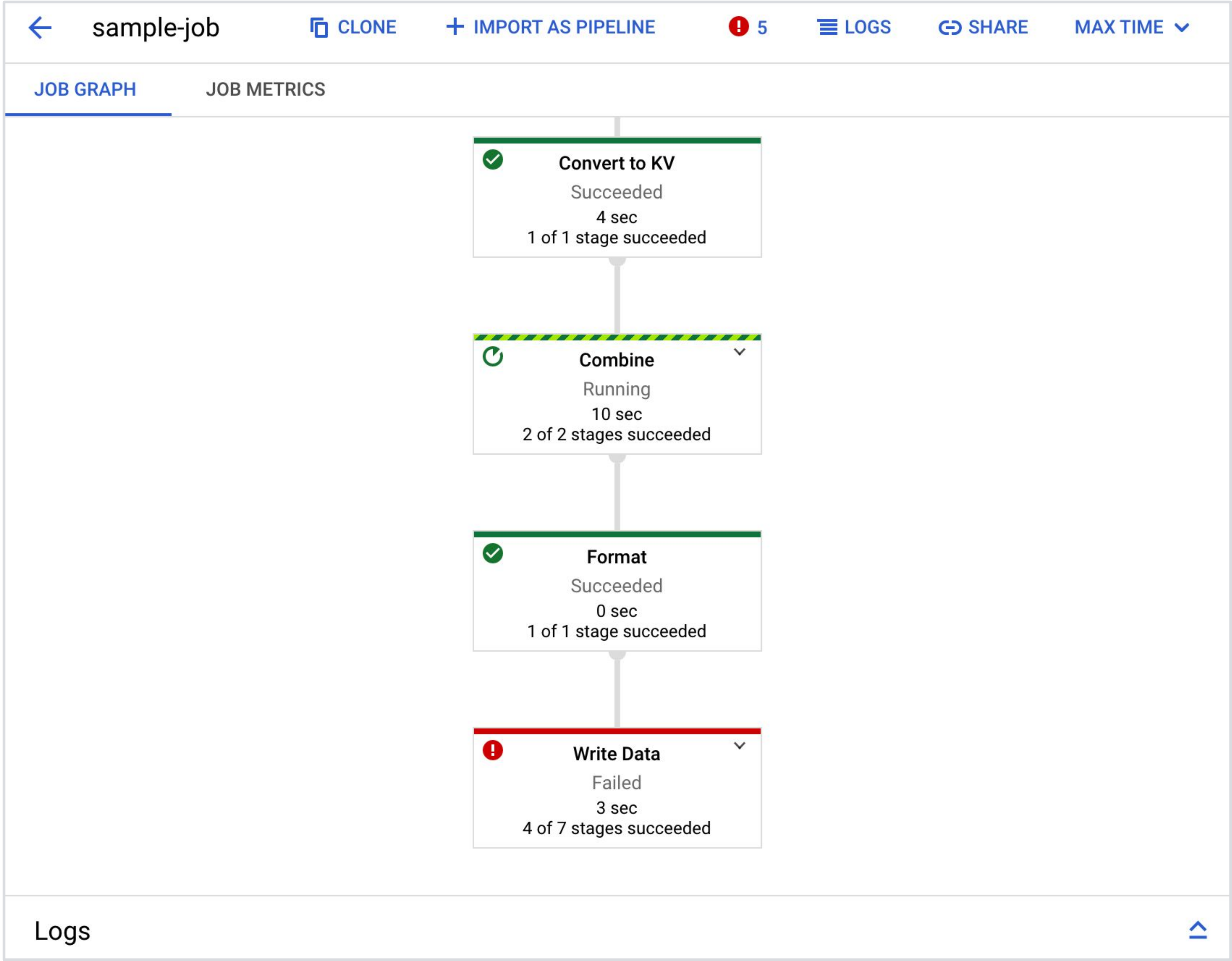
Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region
wordcount4	Batch	—	7 sec	Nov 21, 2018, 3:22:45 PM	Running	2.7.0	2018-11-21_15_22_45-	us-central1
wordcount3	Batch	Nov 21, 2018, 1:10:23 PM	3 min 29 sec	Nov 21, 2018, 1:06:54 PM	Succeeded	2.7.0	2018-11-21_13_06_53-	us-central1
wordcount2	Batch	Nov 21, 2018, 12:57:05 PM	3 min 56 sec	Nov 21, 2018, 12:53:09 PM	Failed	2.7.0	2018-11-21_12_53_08-	us-central1
wordcount1	Batch	Nov 21, 2018, 12:55:25 PM	3 min 36 sec	Nov 21, 2018, 12:51:49 PM	Succeeded	2.7.0	2018-11-21_12_51_48-	us-central1

Checking for errors



Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region
↻ wordcount4	Batch	—	7 sec	Nov 21, 2018, 3:22:45 PM	Running	2.7.0	2018-11-21_15_22_45-[redacted]	us-central1
✓ wordcount3	Batch	Nov 21, 2018, 1:10:23 PM	3 min 29 sec	Nov 21, 2018, 1:06:54 PM	Succeeded	2.7.0	2018-11-21_13_06_53-[redacted]	us-central1
! wordcount2	Batch	Nov 21, 2018, 12:57:05 PM	3 min 56 sec	Nov 21, 2018, 12:53:09 PM	Failed	2.7.0	2018-11-21_12_53_08-[redacted]	us-central1
✓ wordcount1	Batch	Nov 21, 2018, 12:55:25 PM	3 min 36 sec	Nov 21, 2018, 12:51:49 PM	Succeeded	2.7.0	2018-11-21_12_51_48-[redacted]	us-central1

Checking for errors



Convert to KV
Succeeded
4 sec
1 of 1 stage succeeded

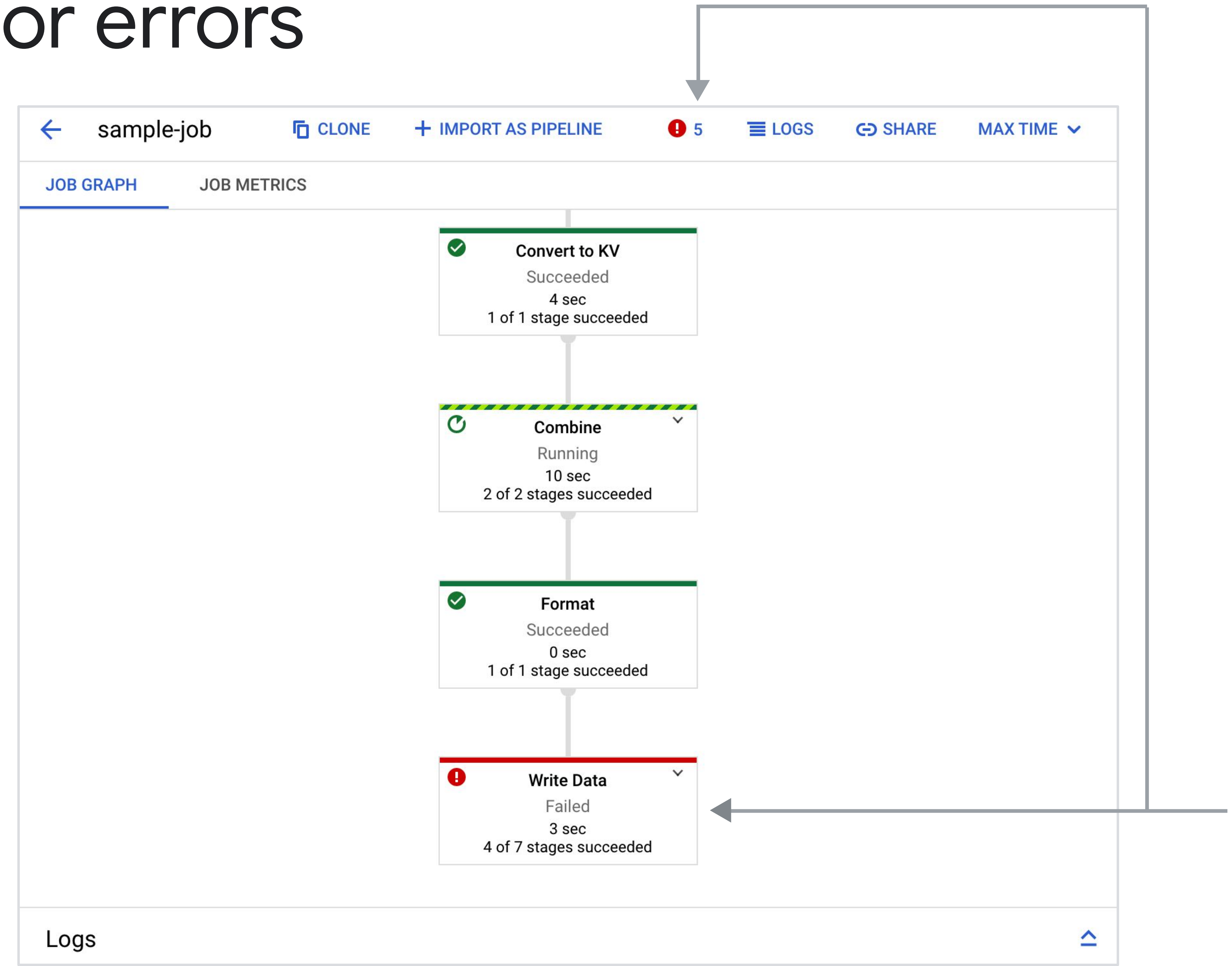
Combine
Running
10 sec
2 of 2 stages succeeded

Format
Succeeded
0 sec
1 of 1 stage succeeded

Write Data
Failed
3 sec
4 of 7 stages succeeded

Checking for errors

Checking for errors



Convert to KV
Succeeded
4 sec
1 of 1 stage succeeded

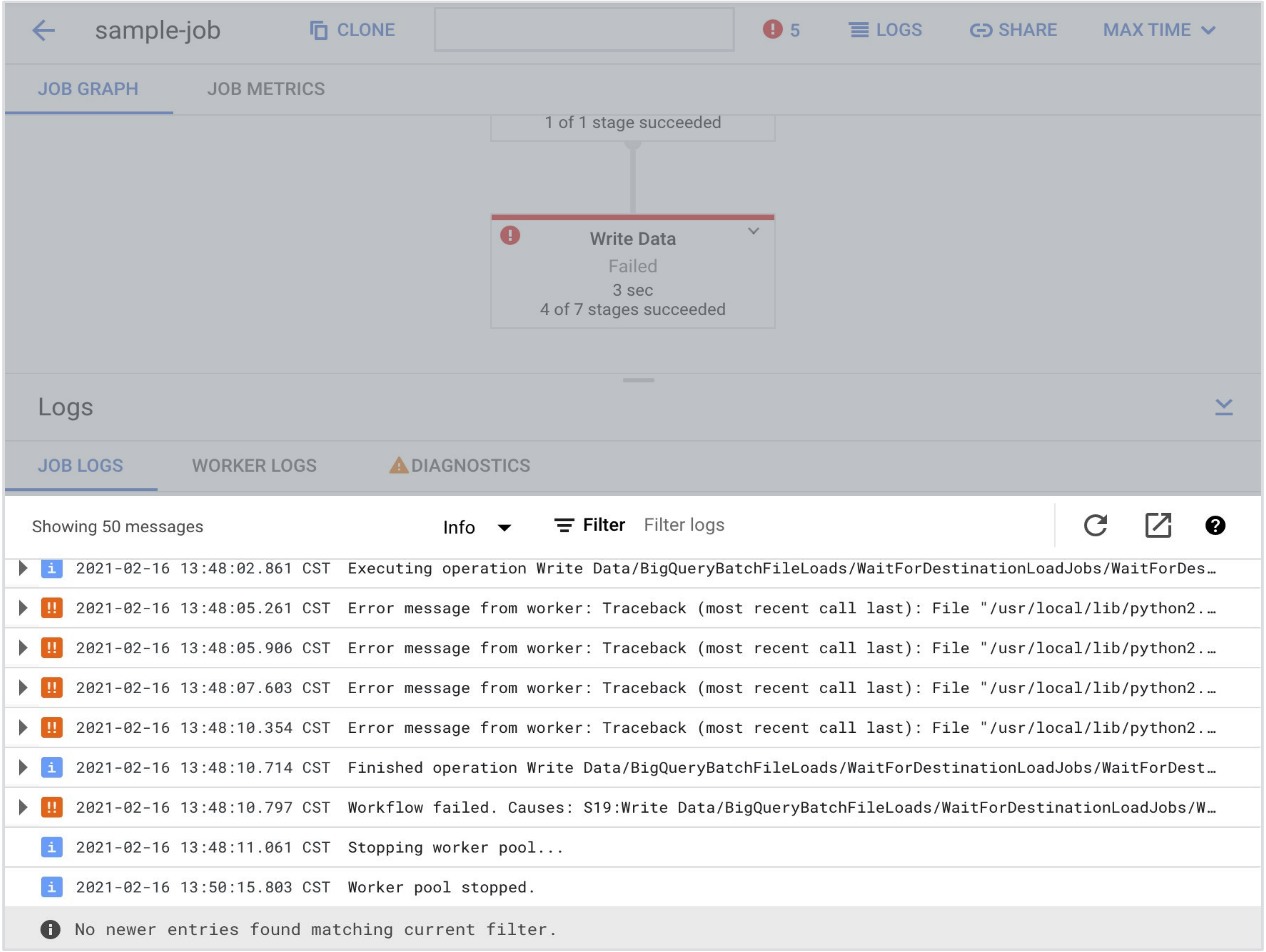
Combine
Running
10 sec
2 of 2 stages succeeded

Format
Succeeded
0 sec
1 of 1 stage succeeded

Write Data
Failed
3 sec
4 of 7 stages succeeded

Checking for errors

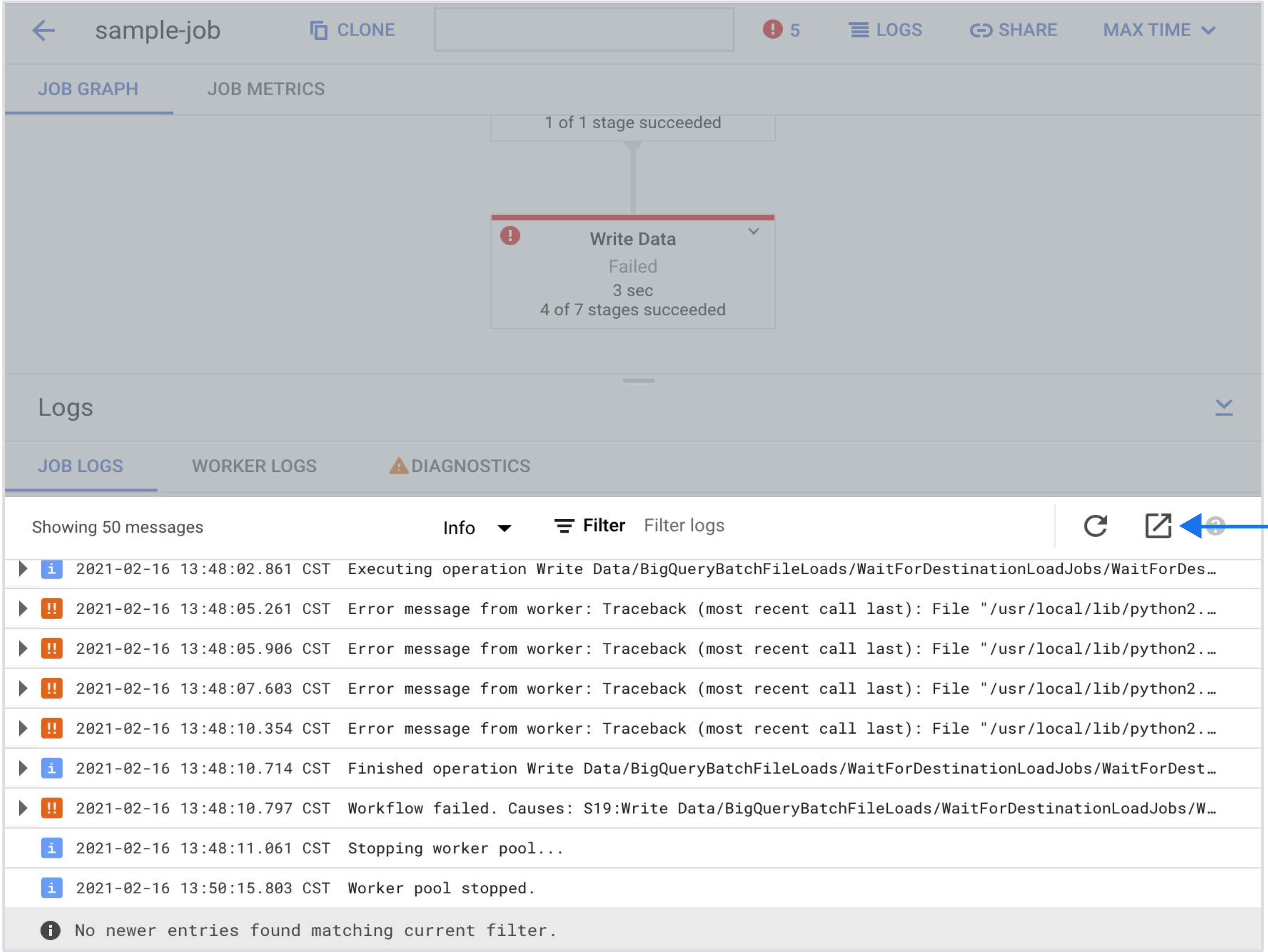
Checking for errors



Showing 50 messages

```
i   2021-02-16 13:48:02.861 CST  Executing operation Write Data/BigQueryBatchFileLoads/WaitForDestinationLoadJobs/WaitForDes…
!!  2021-02-16 13:48:05.261 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
!!  2021-02-16 13:48:05.906 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
!!  2021-02-16 13:48:07.603 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
!!  2021-02-16 13:48:10.354 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
i   2021-02-16 13:48:10.714 CST  Finished operation Write Data/BigQueryBatchFileLoads/WaitForDestinationLoadJobs/WaitForDest…
!!  2021-02-16 13:48:10.797 CST  Workflow failed. Causes: S19:Write Data/BigQueryBatchFileLoads/WaitForDestinationLoadJobs/W…
i   2021-02-16 13:48:11.061 CST  Stopping worker pool...
i   2021-02-16 13:50:15.803 CST  Worker pool stopped.
    No newer entries found matching current filter.
```

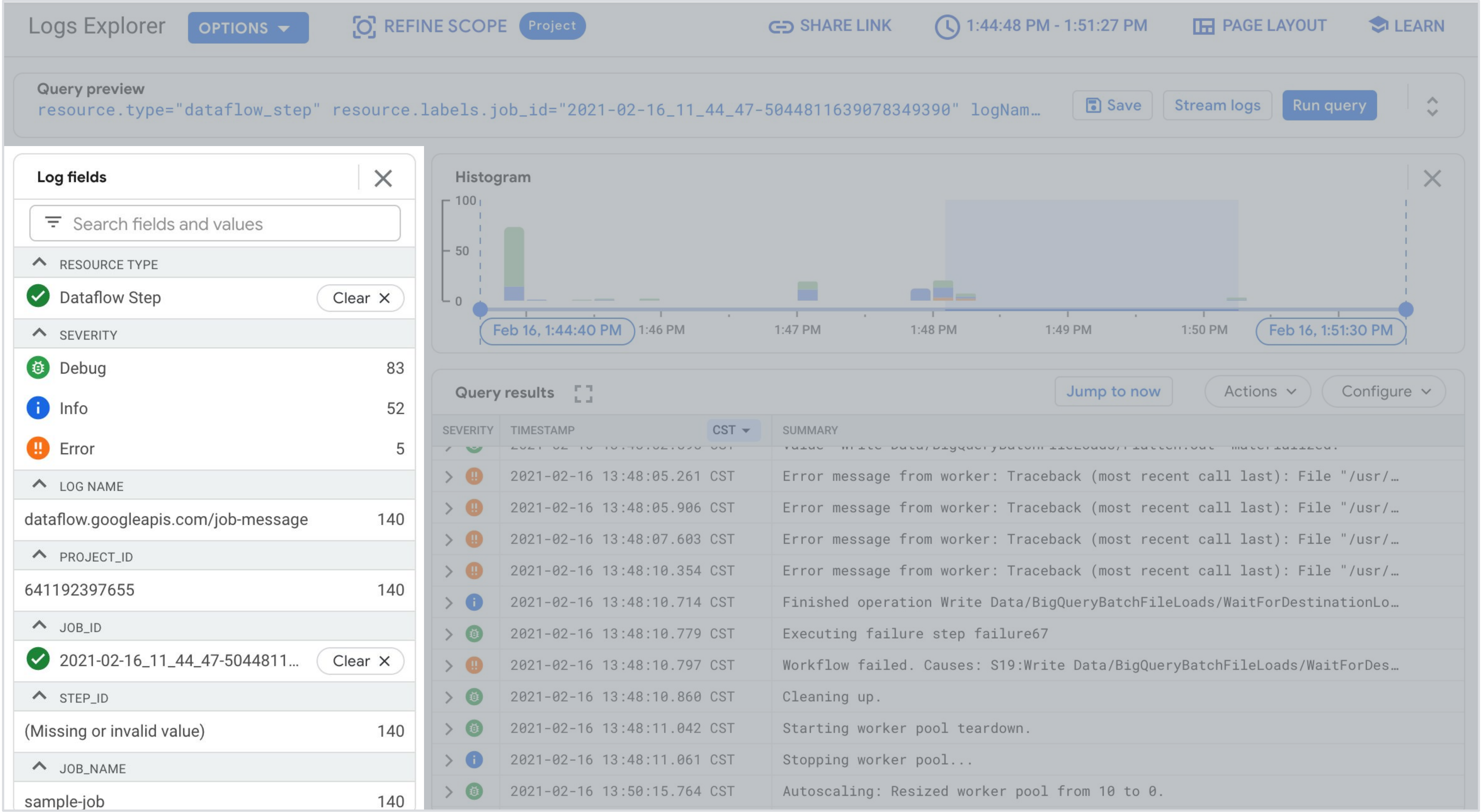
Checking for errors



```
2021-02-16 13:48:02.861 CST  Executing operation Write Data/BigQueryBatchFileLoads/WaitForDestinationLoadJobs/WaitForDes…
2021-02-16 13:48:05.261 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
2021-02-16 13:48:05.906 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
2021-02-16 13:48:07.603 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
2021-02-16 13:48:10.354 CST  Error message from worker: Traceback (most recent call last): File "/usr/local/lib/python2.…
2021-02-16 13:48:10.714 CST  Finished operation Write Data/BigQueryBatchFileLoads/WaitForDestinationLoadJobs/WaitForDest…
2021-02-16 13:48:10.797 CST  Workflow failed. Causes: S19:Write Data/BigQueryBatchFileLoads/WaitForDestinationLoadJobs/W…
2021-02-16 13:48:11.061 CST  Stopping worker pool...
2021-02-16 13:50:15.803 CST  Worker pool stopped.
No newer entries found matching current filter.
```

View
full logs

Checking for errors



Query preview

```
resource.type="dataflow_step" resource.labels.job_id="2021-02-16_11_44_47-5044811639078349390" logNam…
```

Log fields

RESOURCE TYPE
- Dataflow Step

SEVERITY

Debug	83
Info	52
Error	5

LOG NAME

dataflow.googleapis.com/job-message	140

PROJECT_ID

641192397655	140

JOB_ID
- 2021-02-16_11_44_47-5044811…

STEP_ID

(Missing or invalid value)	140

JOB_NAME

sample-job	140

Query results

```
SEVERITY  TIMESTAMP                    CST   SUMMARY
          2021-02-16 13:48:05.261 CST        Error message from worker: Traceback (most recent call last): File "/usr/…
          2021-02-16 13:48:05.906 CST        Error message from worker: Traceback (most recent call last): File "/usr/…
          2021-02-16 13:48:07.603 CST        Error message from worker: Traceback (most recent call last): File "/usr/…
          2021-02-16 13:48:10.354 CST        Error message from worker: Traceback (most recent call last): File "/usr/…
          2021-02-16 13:48:10.714 CST        Finished operation Write Data/BigQueryBatchFileLoads/WaitForDestinationLo…
          2021-02-16 13:48:10.779 CST        Executing failure step failure67
          2021-02-16 13:48:10.797 CST        Workflow failed. Causes: S19:Write Data/BigQueryBatchFileLoads/WaitForDes…
          2021-02-16 13:48:10.860 CST        Cleaning up.
          2021-02-16 13:48:11.042 CST        Starting worker pool teardown.
          2021-02-16 13:48:11.061 CST        Stopping worker pool...
          2021-02-16 13:50:15.764 CST        Autoscaling: Resized worker pool from 10 to 0.
```

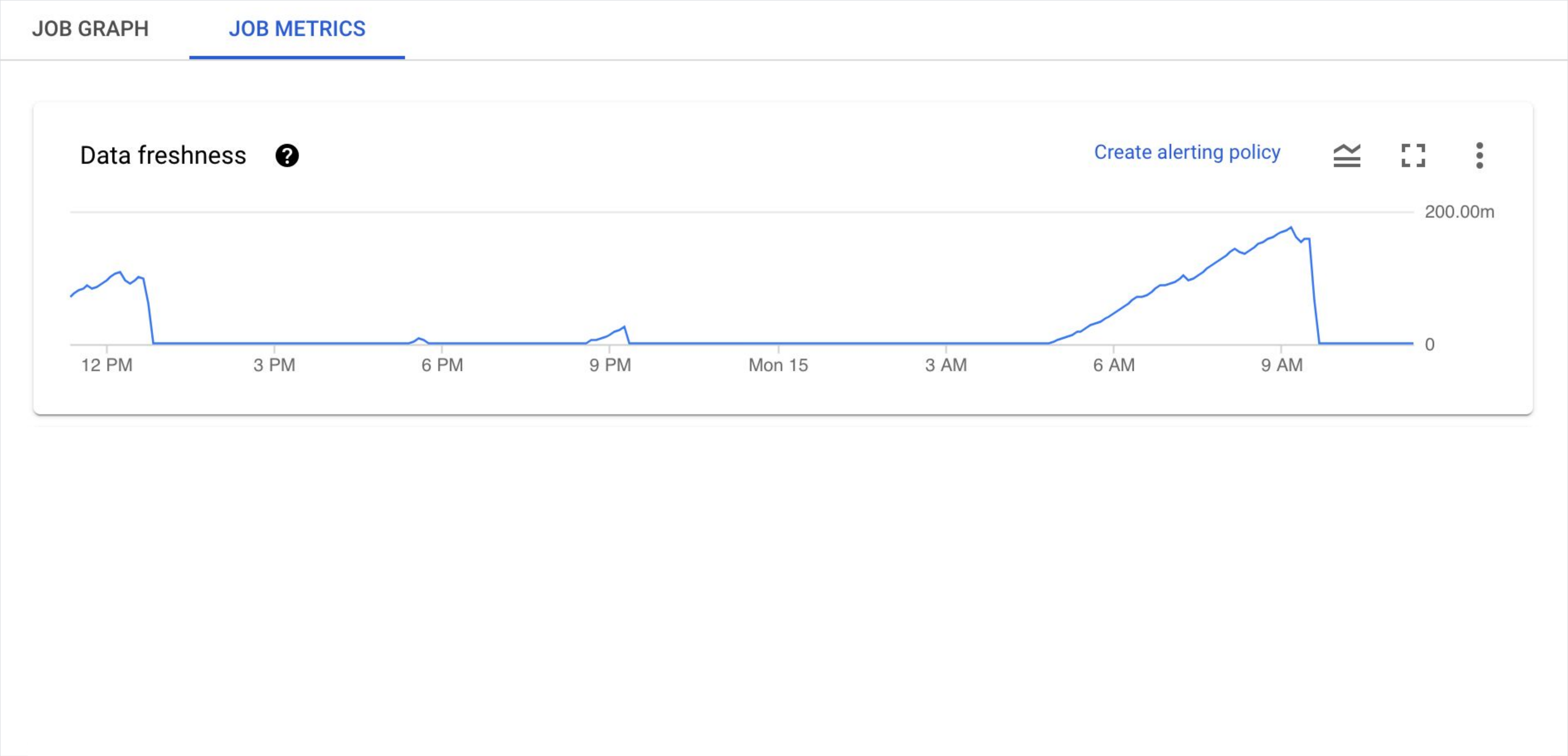
Troubleshooting workflow

Troubleshooting workflow

- Checking for errors
- Looking for anomalies in the Job Metrics tab

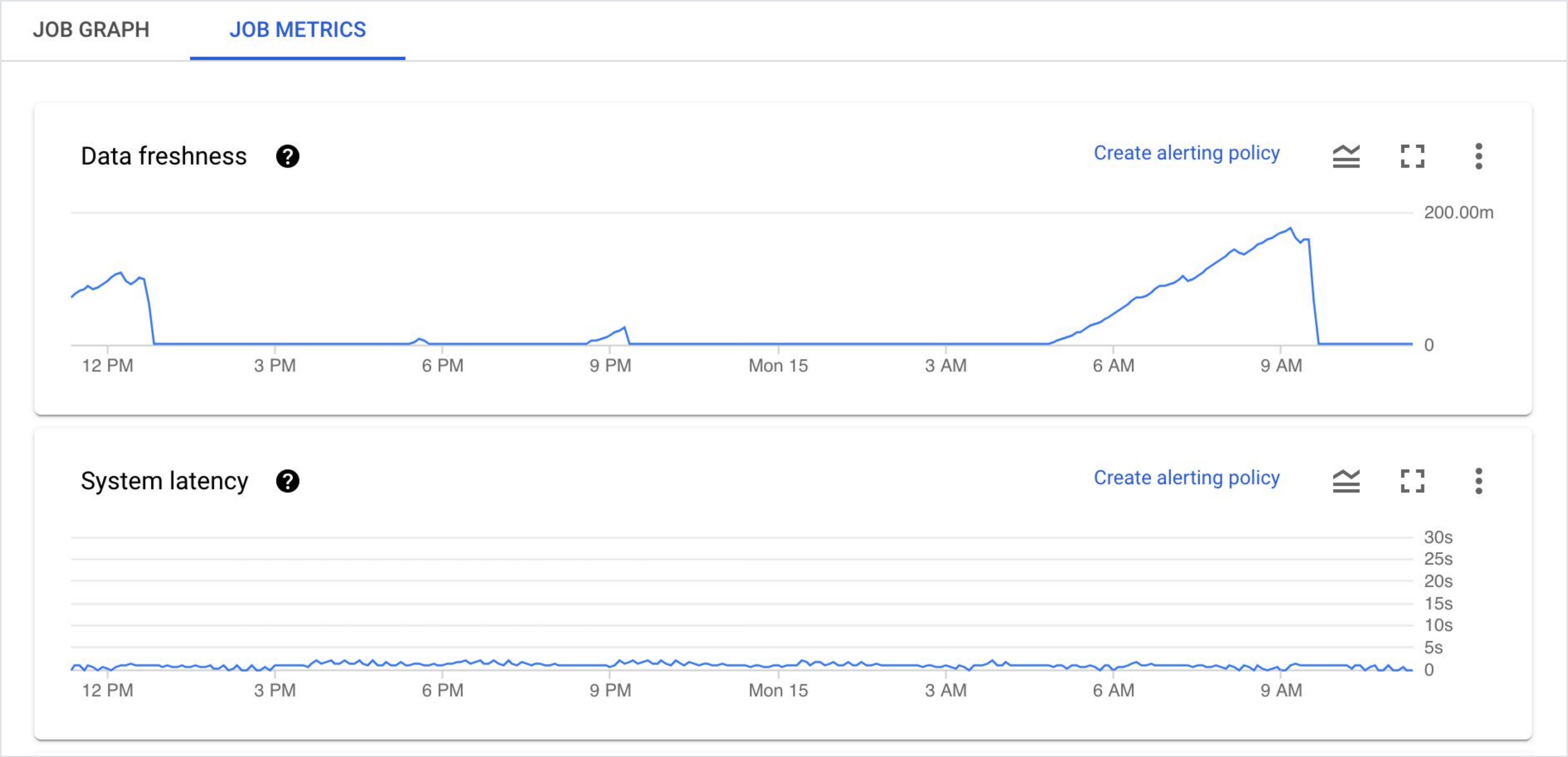
Looking for anomalies in the Job Metrics tab

Streaming pipelines



Looking for anomalies in the Job Metrics tab

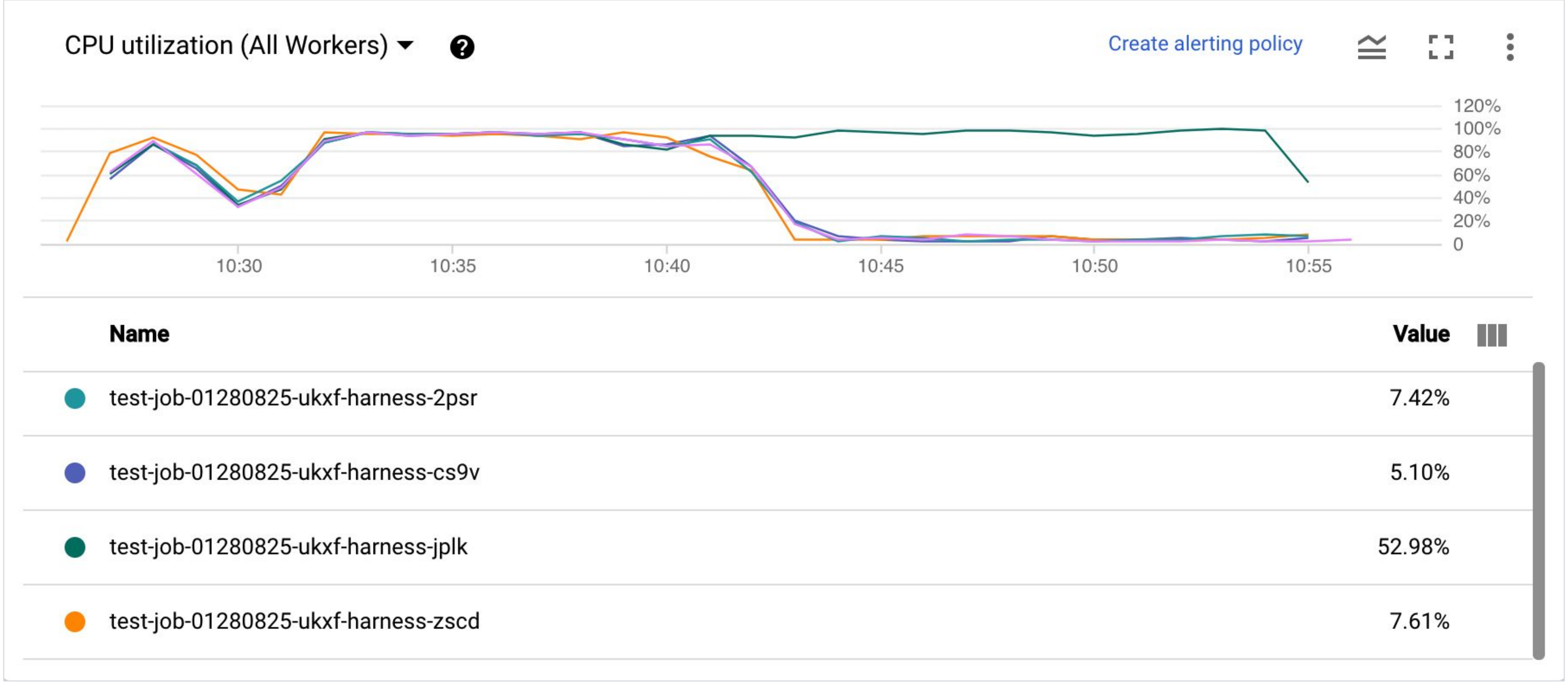
Streaming pipelines



Google Cloud

Looking for anomalies in the Job Metrics tab

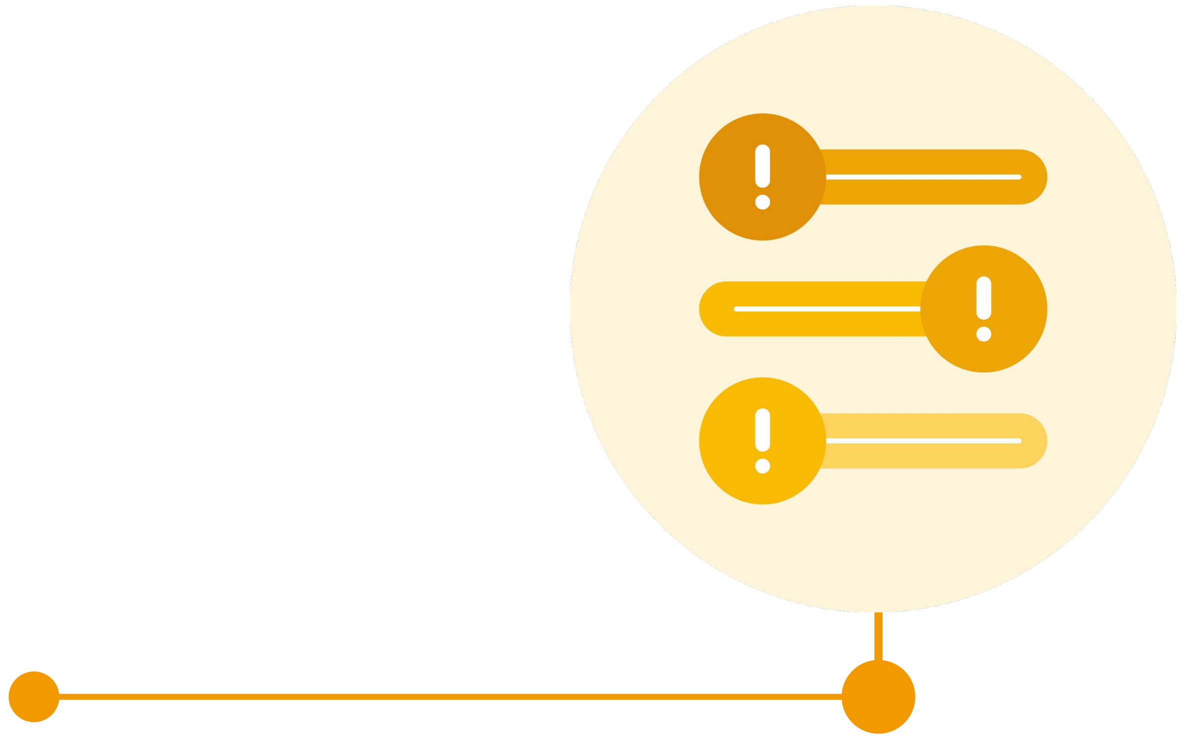
All Dataflow jobs



	Name	Value
●	test-job-01280825-ukxf-harness-2psr	7.42%
●	test-job-01280825-ukxf-harness-cs9v	5.10%
●	test-job-01280825-ukxf-harness-jplk	52.98%
●	test-job-01280825-ukxf-harness-zscd	7.61%

Troubleshooting and Debugging

Agenda



Troubleshooting workflow

Types of troubles

Types of troubles

Types of troubles

- Failure building the pipeline
- Failure starting the pipeline on Dataflow
- Failure during pipeline execution
- Performance problems

Types of troubles

Types of troubles

- Failure building the pipeline
- Failure starting the pipeline on Dataflow
- Failure during pipeline execution
- Performance problems

Failure building the pipeline

These errors occur when Apache Beam is building the pipeline:

- Validating "Beam Model" aspects of the pipeline
- Validating input/output specifications
- You can reproduce it with the direct runner, in a unit test

Failure building the pipeline

```
java.lang.IllegalStateException: GroupByKey cannot be applied to non-bounded PCollection
in the GlobalWindow without a trigger. Use a Window.into or Window.triggering transform
prior to GroupByKey.
    at org.apache.beam.sdk.transforms.GroupByKey.applicableTo (GroupByKey.java:153)
    at org.apache.beam.sdk.transforms.GroupByKey.expand (GroupByKey.java:185)
    at org.apache.beam.sdk.transforms.GroupByKey.expand (GroupByKey.java:107)
    at org.apache.beam.sdk.Pipeline.applyInternal (Pipeline.java:537)
    at org.apache.beam.sdk.Pipeline.applyTransform (Pipeline.java:471)
    at org.apache.beam.sdk.values.PCollection.apply (PCollection.java:357)
    at ...
```

Types of troubles

Types of troubles

- Failure building the pipeline
- Failure starting the pipeline on Dataflow
- Failure during pipeline execution
- Performance problems

Failure starting the pipeline on Dataflow

Once the Dataflow service has received your pipeline's graph, the service will attempt to validate your job:

- Making sure the service can access your job's associated Cloud Storage buckets for file staging and temporary output.

Failure starting the pipeline on Dataflow

Once the Dataflow service has received your pipeline's graph, the service will attempt to validate your job:

- Making sure the service can access your job's associated Cloud Storage buckets for file staging and temporary output.
- Checking for the required permissions in your Google Cloud project.

Failure starting the pipeline on Dataflow

Once the Dataflow service has received your pipeline's graph, the service will attempt to validate your job:

- Making sure the service can access your job's associated Cloud Storage buckets for file staging and temporary output.
- Checking for the required permissions in your Google Cloud project.
- Making sure the service can access input and output sources, such as files.

Failure starting the pipeline on Dataflow

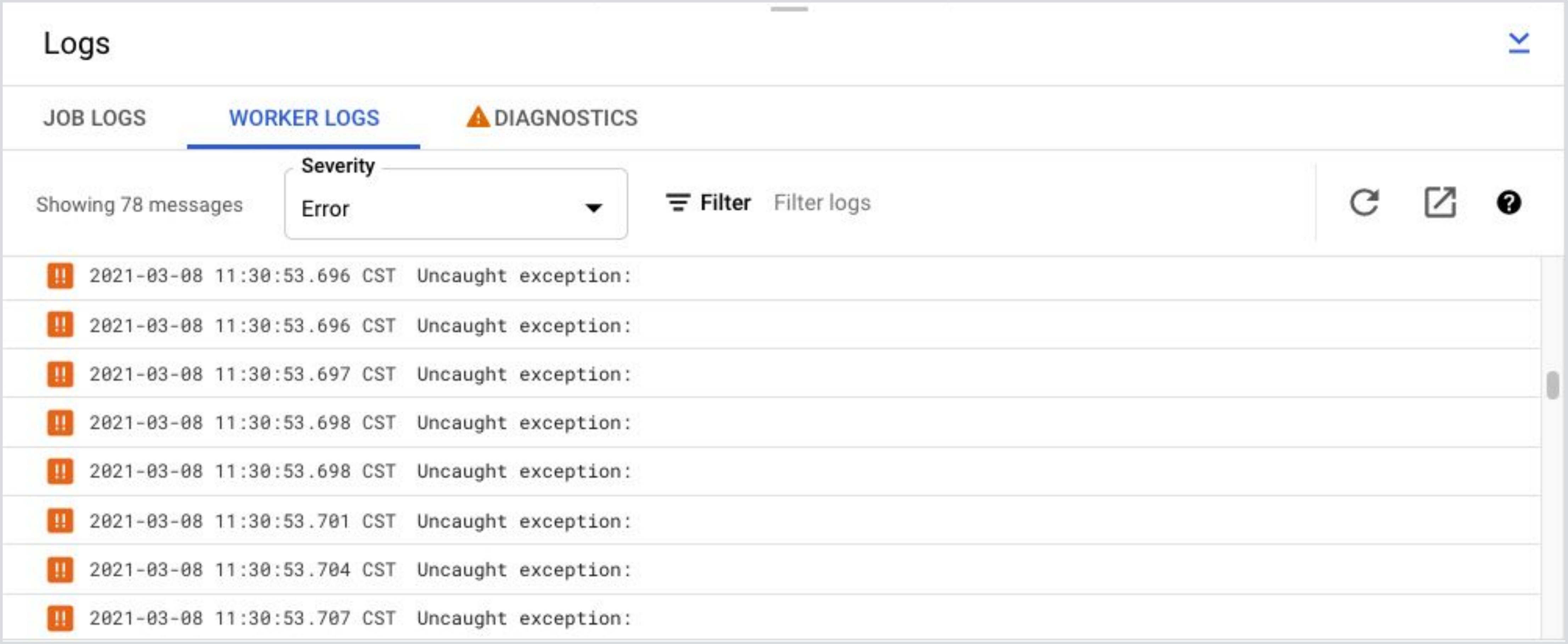
```
apitools.base.py.exceptions.HttpForbiddenError: HttpError accessing
<https://dataflow.googleapis.com/v1b3/projects/xyz/locations/us-central1/jobs?alt=json>:
response: <{'status': '403', 'content-length': '288', 'x-xss-protection': '0',
'x-content-type-options': 'nosniff', 'transfer-encoding': 'chunked', 'vary': 'Origin,
X-Origin, Referer', 'server': 'ESF', '-content-encoding': 'gzip', 'cache-control': 'private',
'date': 'Tue, 10 Dec 2019 17:56:00 GMT', 'x-frame-options': 'SAMEORIGIN', 'alt-svc':
'quic=":443"; ma=2592000; v="46,43",h3-Q050=":443"; ma=2592000,h3-Q049=":443";
ma=2592000,h3-Q048=":443"; ma=2592000,h3-Q046=":443"; ma=2592000,h3-Q043=":443"; ma=2592000',
'content-type': 'application/json; charset=UTF-8'}>, content <{
        "error": {
          "code": 403,
          "message": "(43abf11c2446750): Could not create workflow; user does not have write
      access to project: xyz Causes: (43abf11c244683b): Permission 'dataflow.jobs.create'
      denied on project: 'xyz'",
          "status": "PERMISSION_DENIED"
        }
      }
```

Types of troubles

Types of troubles

- Failure building the pipeline
- Failure starting the pipeline on Dataflow
- Failure during pipeline execution
- Performance problems

Failure during pipeline execution

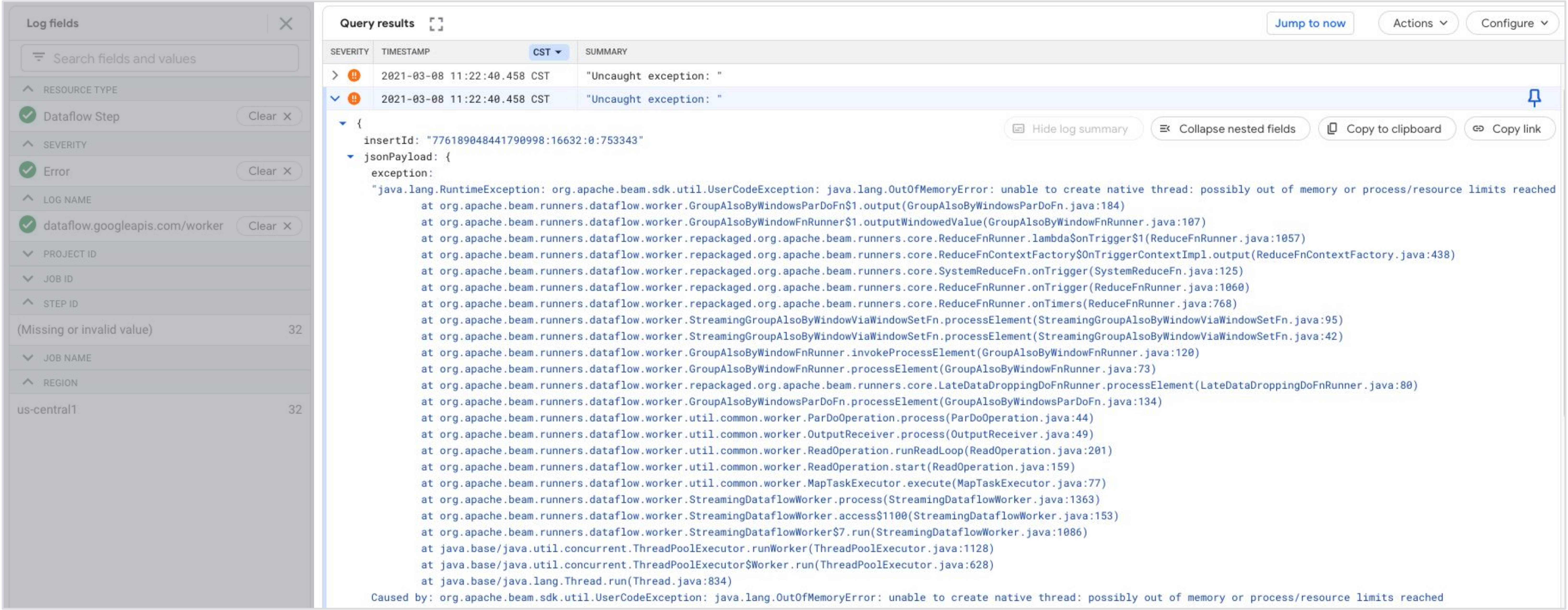


Showing 78 messages

Severity: Error

```
2021-03-08 11:30:53.696 CST  Uncaught exception:
2021-03-08 11:30:53.696 CST  Uncaught exception:
2021-03-08 11:30:53.697 CST  Uncaught exception:
2021-03-08 11:30:53.698 CST  Uncaught exception:
2021-03-08 11:30:53.698 CST  Uncaught exception:
2021-03-08 11:30:53.701 CST  Uncaught exception:
2021-03-08 11:30:53.704 CST  Uncaught exception:
2021-03-08 11:30:53.707 CST  Uncaught exception:
```

Failure during pipeline execution



RESOURCE TYPE
Dataflow Step — Clear ✕

SEVERITY
Error — Clear ✕

LOG NAME
dataflow.googleapis.com/worker — Clear ✕

PROJECT ID

JOB ID

STEP ID
(Missing or invalid value) 32

JOB NAME

REGION
us-central1 32

SEVERITY	TIMESTAMP CST	SUMMARY
!	2021-03-08 11:22:40.458 CST	"Uncaught exception: "
!	2021-03-08 11:22:40.458 CST	"Uncaught exception: "

```
{
  insertId: "776189048441790998:16632:0:753343"
  jsonPayload: {
   exception:
   "java.lang.RuntimeException: org.apache.beam.sdk.util.UserCodeException: java.lang.OutOfMemoryError: unable to create native thread: possibly out of memory or process/resource limits reached
          at org.apache.beam.runners.dataflow.worker.GroupAlsoByWindowsParDoFn$1.output(GroupAlsoByWindowsParDoFn.java:184)
          at org.apache.beam.runners.dataflow.worker.GroupAlsoByWindowFnRunner$1.outputWindowedValue(GroupAlsoByWindowFnRunner.java:107)
          at org.apache.beam.runners.dataflow.worker.repackaged.org.apache.beam.runners.core.ReduceFnRunner.lambda$onTrigger$1(ReduceFnRunner.java:1057)
          at org.apache.beam.runners.dataflow.worker.repackaged.org.apache.beam.runners.core.ReduceFnContextFactory$OnTriggerContextImpl.output(ReduceFnContextFactory.java:438)
          at org.apache.beam.runners.dataflow.worker.repackaged.org.apache.beam.runners.core.SystemReduceFn.onTrigger(SystemReduceFn.java:125)
          at org.apache.beam.runners.dataflow.worker.repackaged.org.apache.beam.runners.core.ReduceFnRunner.onTrigger(ReduceFnRunner.java:1060)
          at org.apache.beam.runners.dataflow.worker.repackaged.org.apache.beam.runners.core.ReduceFnRunner.onTimers(ReduceFnRunner.java:768)
          at org.apache.beam.runners.dataflow.worker.StreamingGroupAlsoByWindowViaWindowSetFn.processElement(StreamingGroupAlsoByWindowViaWindowSetFn.java:95)
          at org.apache.beam.runners.dataflow.worker.StreamingGroupAlsoByWindowViaWindowSetFn.processElement(StreamingGroupAlsoByWindowViaWindowSetFn.java:42)
          at org.apache.beam.runners.dataflow.worker.GroupAlsoByWindowFnRunner.invokeProcessElement(GroupAlsoByWindowFnRunner.java:120)
          at org.apache.beam.runners.dataflow.worker.GroupAlsoByWindowFnRunner.processElement(GroupAlsoByWindowFnRunner.java:73)
          at org.apache.beam.runners.dataflow.worker.repackaged.org.apache.beam.runners.core.LateDataDroppingDoFnRunner.processElement(LateDataDroppingDoFnRunner.java:80)
          at org.apache.beam.runners.dataflow.worker.GroupAlsoByWindowsParDoFn.processElement(GroupAlsoByWindowsParDoFn.java:134)
          at org.apache.beam.runners.dataflow.worker.util.common.worker.ParDoOperation.process(ParDoOperation.java:44)
          at org.apache.beam.runners.dataflow.worker.util.common.worker.OutputReceiver.process(OutputReceiver.java:49)
          at org.apache.beam.runners.dataflow.worker.util.common.worker.ReadOperation.runReadLoop(ReadOperation.java:201)
          at org.apache.beam.runners.dataflow.worker.util.common.worker.ReadOperation.start(ReadOperation.java:159)
          at org.apache.beam.runners.dataflow.worker.util.common.worker.MapTaskExecutor.execute(MapTaskExecutor.java:77)
          at org.apache.beam.runners.dataflow.worker.StreamingDataflowWorker.process(StreamingDataflowWorker.java:1363)
          at org.apache.beam.runners.dataflow.worker.StreamingDataflowWorker.access$1100(StreamingDataflowWorker.java:153)
          at org.apache.beam.runners.dataflow.worker.StreamingDataflowWorker$7.run(StreamingDataflowWorker.java:1086)
          at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128)
          at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:628)
          at java.base/java.lang.Thread.run(Thread.java:834)
   Caused by: org.apache.beam.sdk.util.UserCodeException: java.lang.OutOfMemoryError: unable to create native thread: possibly out of memory or process/resource limits reached
```

Failure during pipeline execution

```python
class FilterMessagesFn(beam.DoFn):
    BAD_MESSAGE_TAG = 'bad_message'
    GOOD_MESSAGE_TAG = 'good_message'

    def process(self, element, window=beam.DoFn.WindowParam):
        try:
            data = element.decode()
            # tag the elements accordingly
            if 'bad' in data:
                yield pvalue.TaggedOutput(self.BAD_MESSAGE_TAG, element)
            else:
                yield pvalue.TaggedOutput(self.GOOD_MESSAGE_TAG, element)

        # handle any exceptions in the processing
        except Exception as exp:
            logging.getLogger.warning(exp)
            yield pvalue.TaggedOutput(self.BAD_MESSAGE_TAG, element)
```

Failure during pipeline execution

To track failing elements:

- Log the failing elements and check the output using Cloud Logging.

Failure during pipeline execution

To track failing elements:

- Log the failing elements and check the output using Cloud Logging.
- Check the Dataflow worker and worker startup logs for warnings or errors related to work item failures.

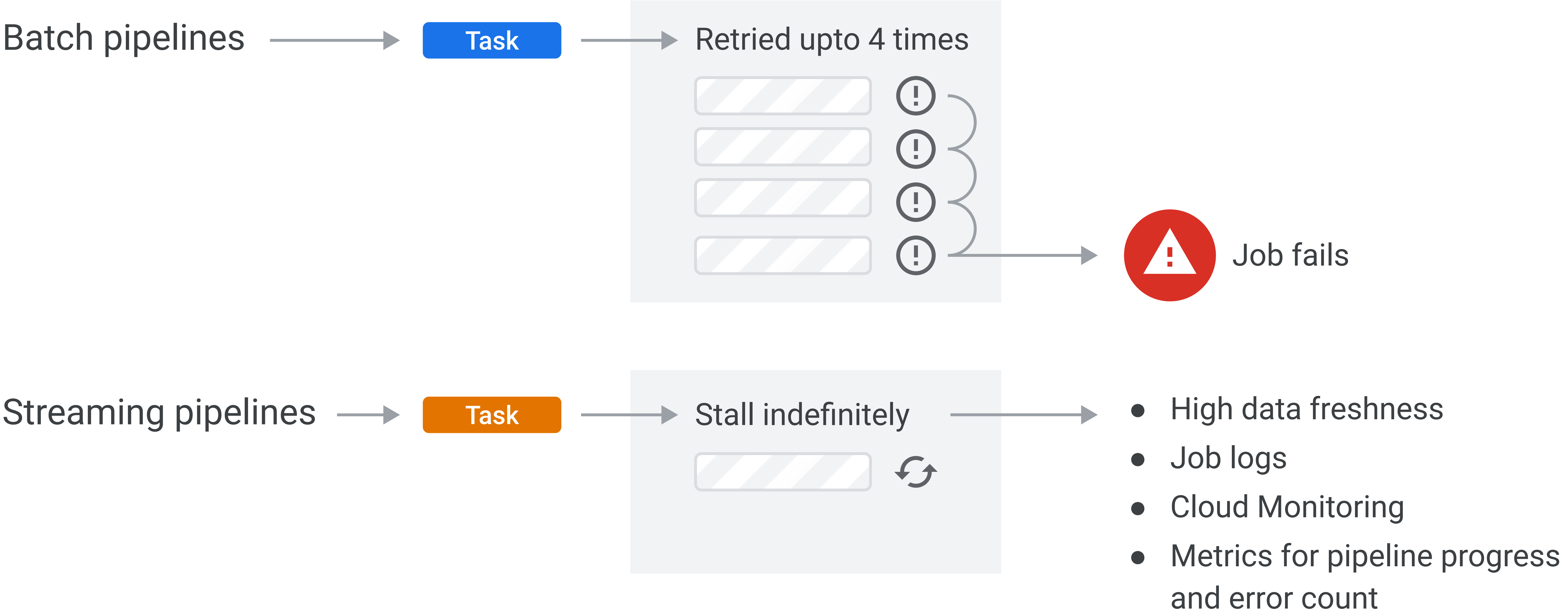
Failure during pipeline execution

To track failing elements:

- Log the failing elements and check the output using Cloud Logging.
- Check the Dataflow worker and worker startup logs for warnings or errors related to work item failures.
- Set your ParDo to write the failing elements to an additional output for later inspection.

Failure during pipeline execution

Batch vs streaming



Batch pipelines → Task → Retried upto 4 times → Job fails

Streaming pipelines → Task → Stall indefinitely →

- High data freshness
- Job logs
- Cloud Monitoring
- Metrics for pipeline progress and error count

Types of troubles

Types of troubles

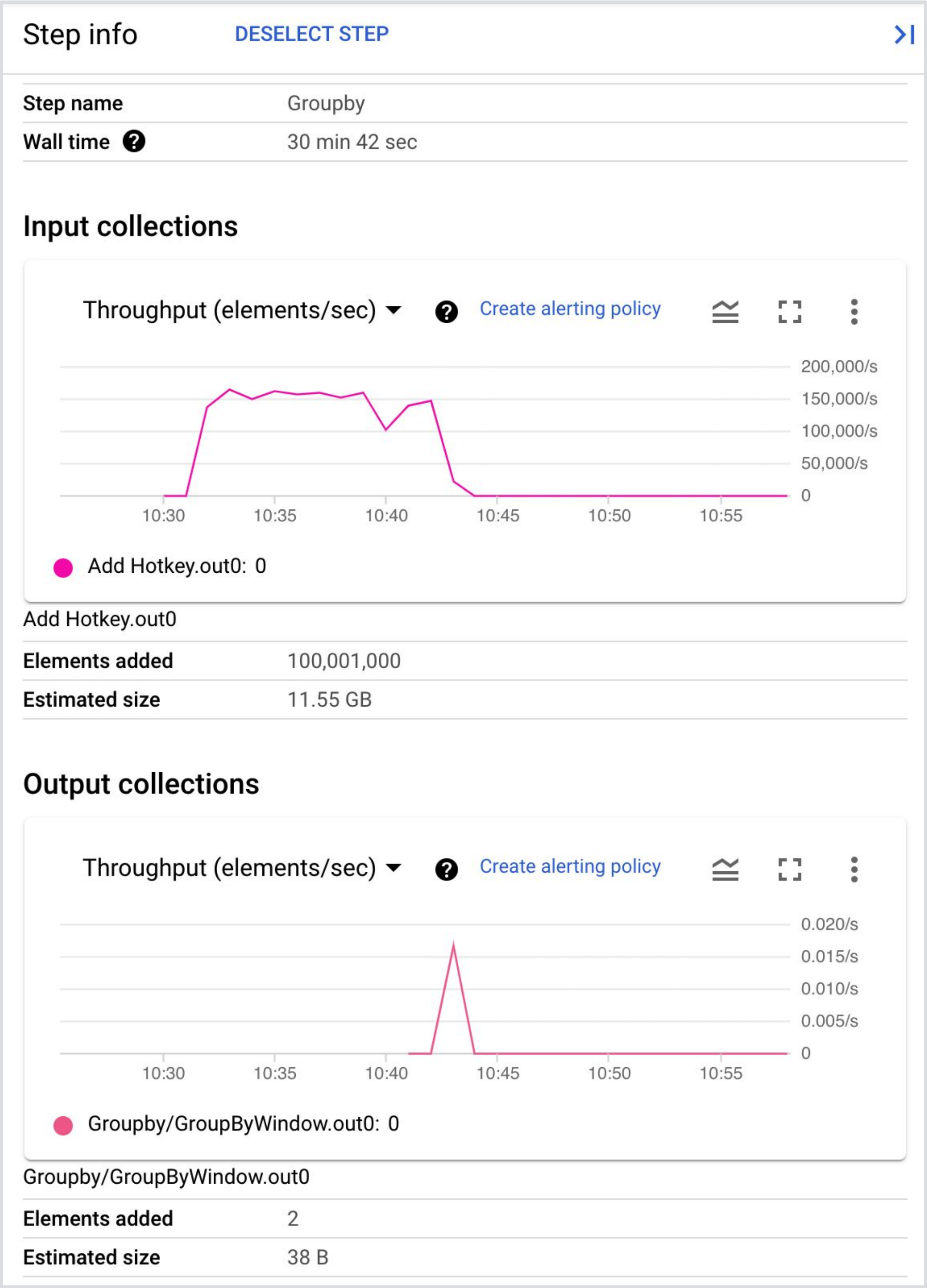
- Failure building the pipeline
- Failure starting the pipeline on Dataflow
- Failure during pipeline execution
- Performance problems

Performance problems

Factors that can influence the performance of a pipeline:

- Pipeline design
- Data shape
- Interactions with sources, sinks and external systems

Performance problems



Step info DESELECT STEP

Step name	Groupby
Wall time	30 min 42 sec

Input collections

Add Hotkey.out0

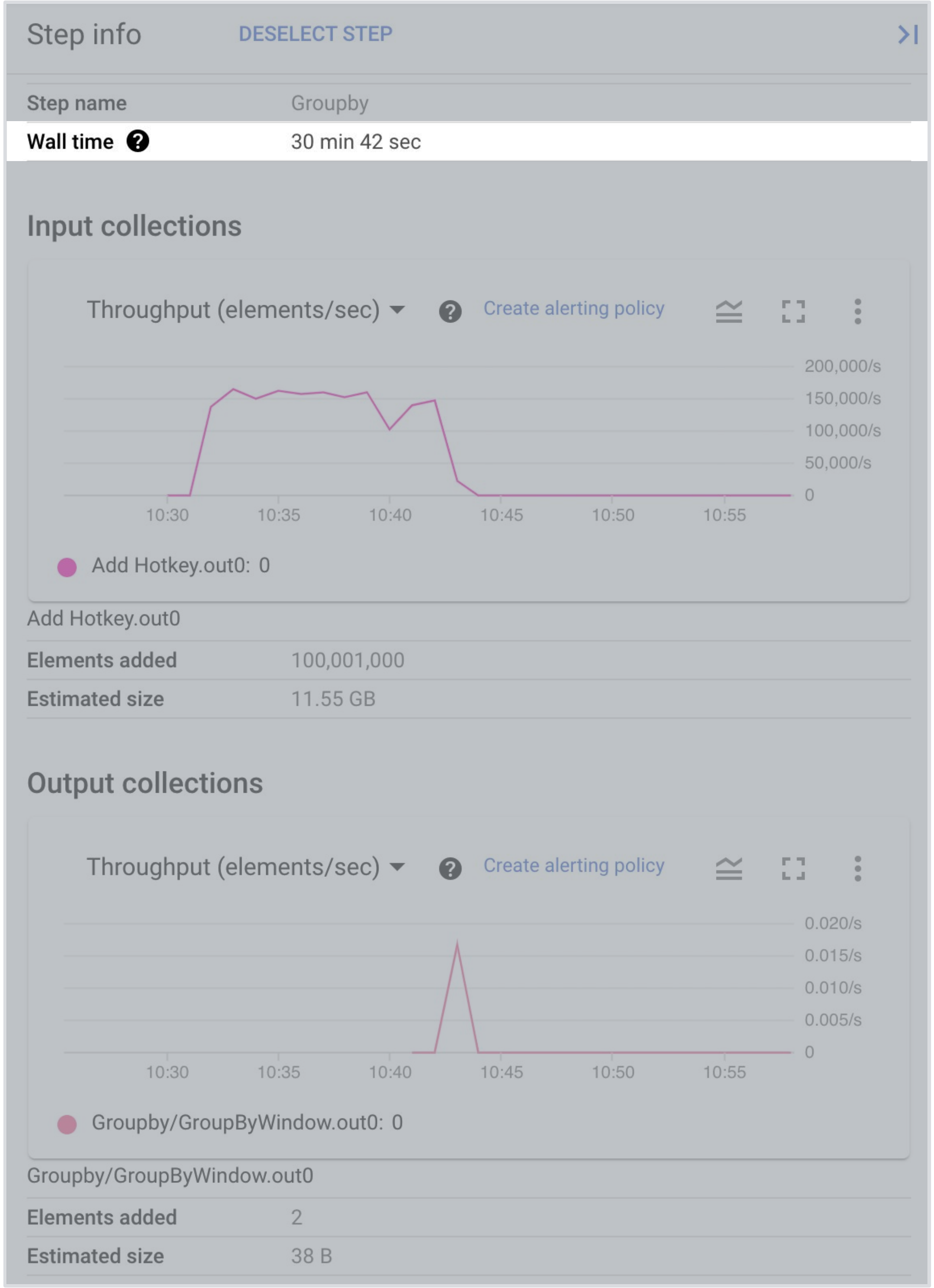
Elements added	100,001,000
Estimated size	11.55 GB

Output collections

Groupby/GroupByWindow.out0

Elements added	2
Estimated size	38 B

Performance problems



Step name	Groupby
Wall time ❓	30 min 42 sec

Input collections

Add Hotkey.out0

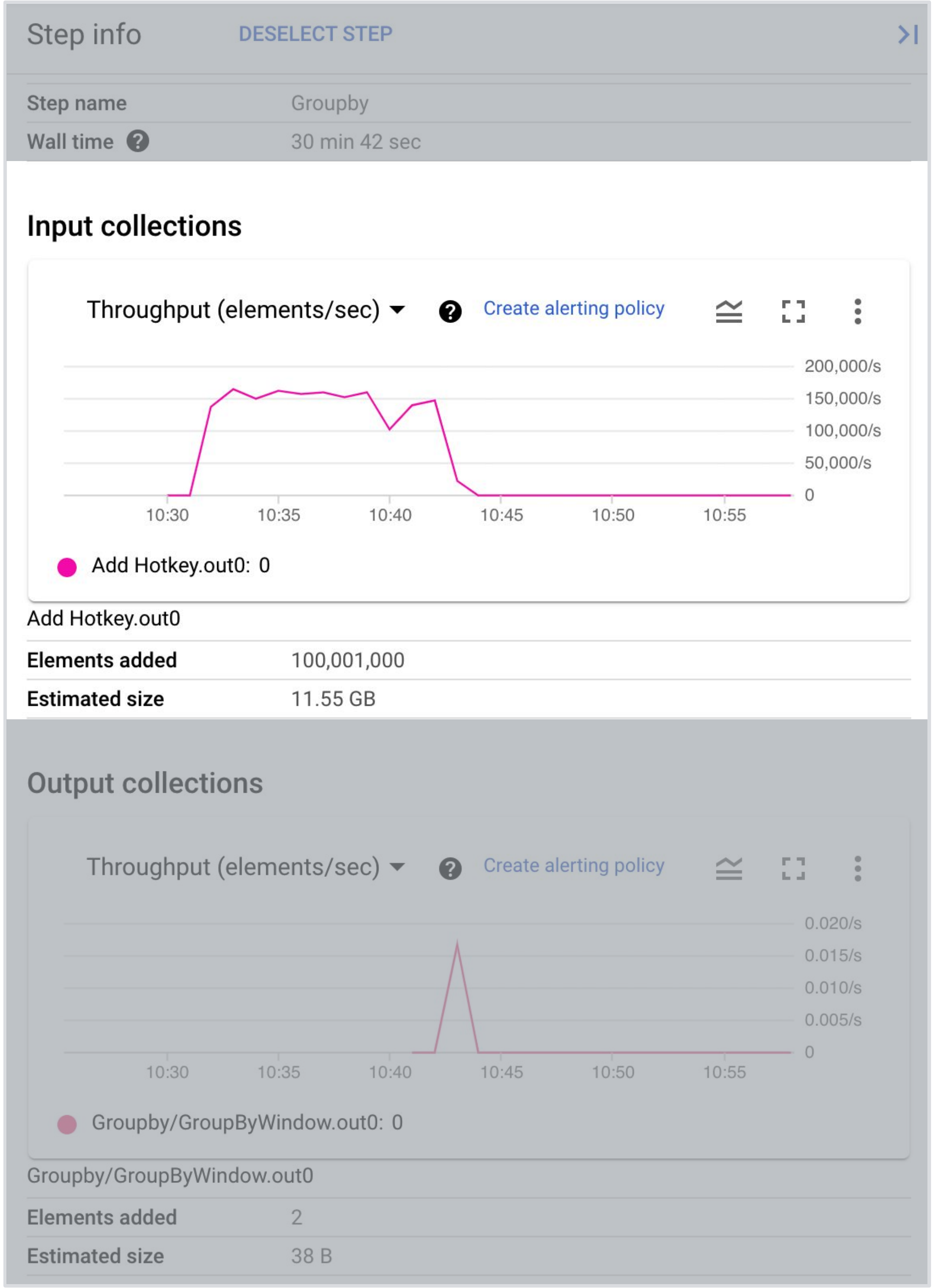
Elements added	100,001,000
Estimated size	11.55 GB

Output collections

Groupby/GroupByWindow.out0

Elements added	2
Estimated size	38 B

Performance problems



Step info DESELECT STEP

Step name	Groupby
Wall time	30 min 42 sec

Input collections

Add Hotkey.out0

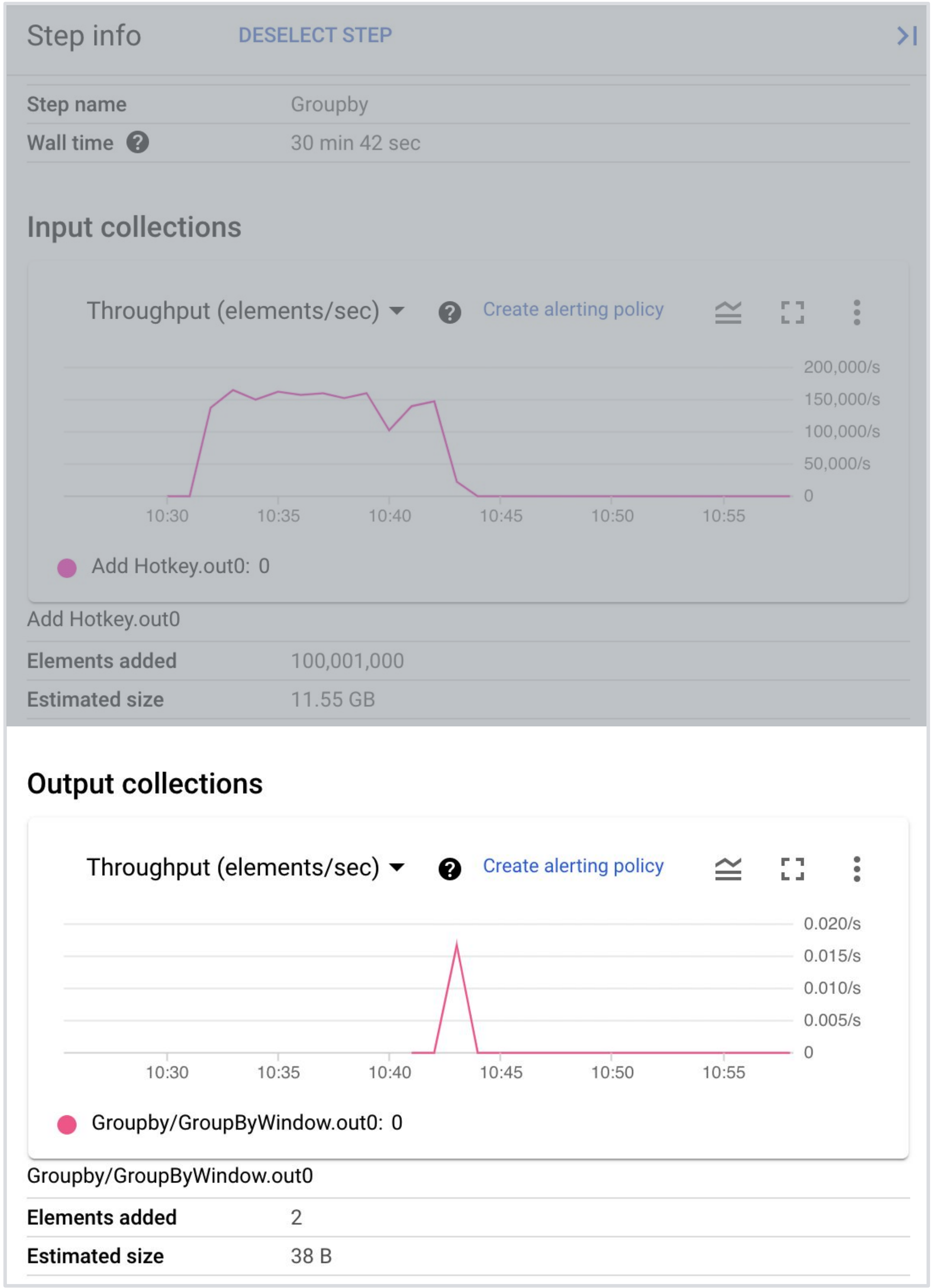
Elements added	100,001,000
Estimated size	11.55 GB

Output collections

Groupby/GroupByWindow.out0

Elements added	2
Estimated size	38 B

Performance problems



Step info DESELECT STEP

Step name	Groupby
Wall time	30 min 42 sec

Input collections

Add Hotkey.out0

Elements added	100,001,000
Estimated size	11.55 GB

Output collections

Groupby/GroupByWindow.out0

Elements added	2
Estimated size	38 B