Apache Beam &
Google Cloud Dataflow

Slides by Frances Perry, April 2016

Beam Model: Asking the Right Questions

What results are calculated?

Where in event time are results calculated?

When in processing time are results materialized?

How do refinements of results relate?

What is Apache Beam?

1. The Beam Model: **What** / **Where** / **When** / **How**

2. SDKs for writing Beam pipelines -- starting with Java

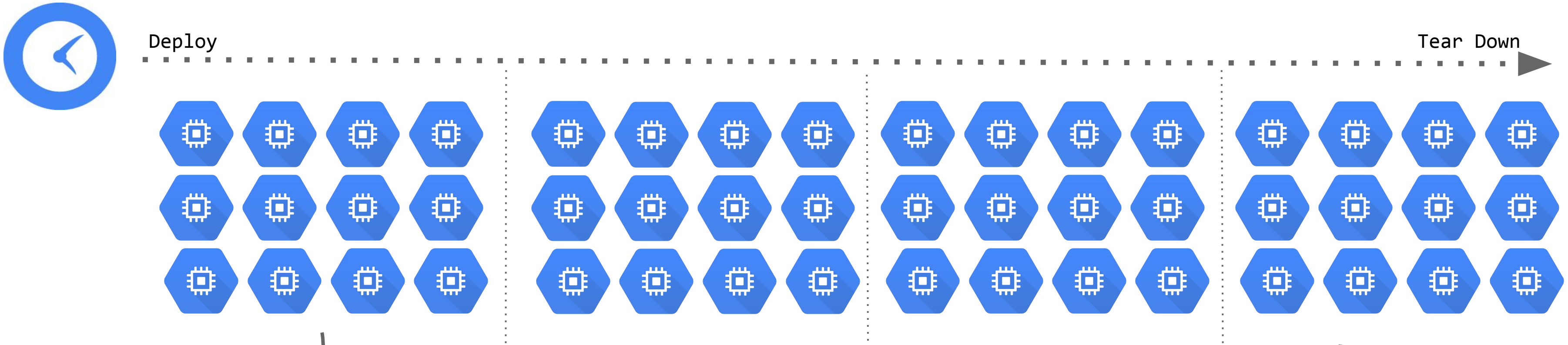
3. Runners for Existing Distributed Processing Backends
   a. Apache Flink (thanks to dataArtisans)
   b. Apache Spark (thanks to Cloudera)
   c. Google Cloud Dataflow (fully managed service)
   d. Local (in-process) runner for testing

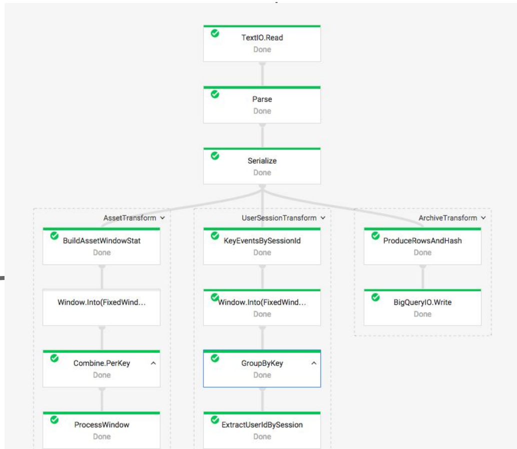
The Cloud Dataflow Service

A great place for executing Beam pipelines which provides:

- Fully managed, no-ops execution environment
- Integration with Google Cloud Platform
- Java support in GA. Python in Alpha.

Fully Managed: Worker Lifecycle Management





TextIO.Read
Done

Parse
Done

Serialize
Done

AssetTransform

BuildAssetWindowStat
Done

Window.Into(FixedWind...

Combine.PerKey
Done

ProcessWindow
Done

UserSessionTransform

KeyEventsBySessionId
Done

Window.Into(FixedWind...
Done

GroupByKey
Done

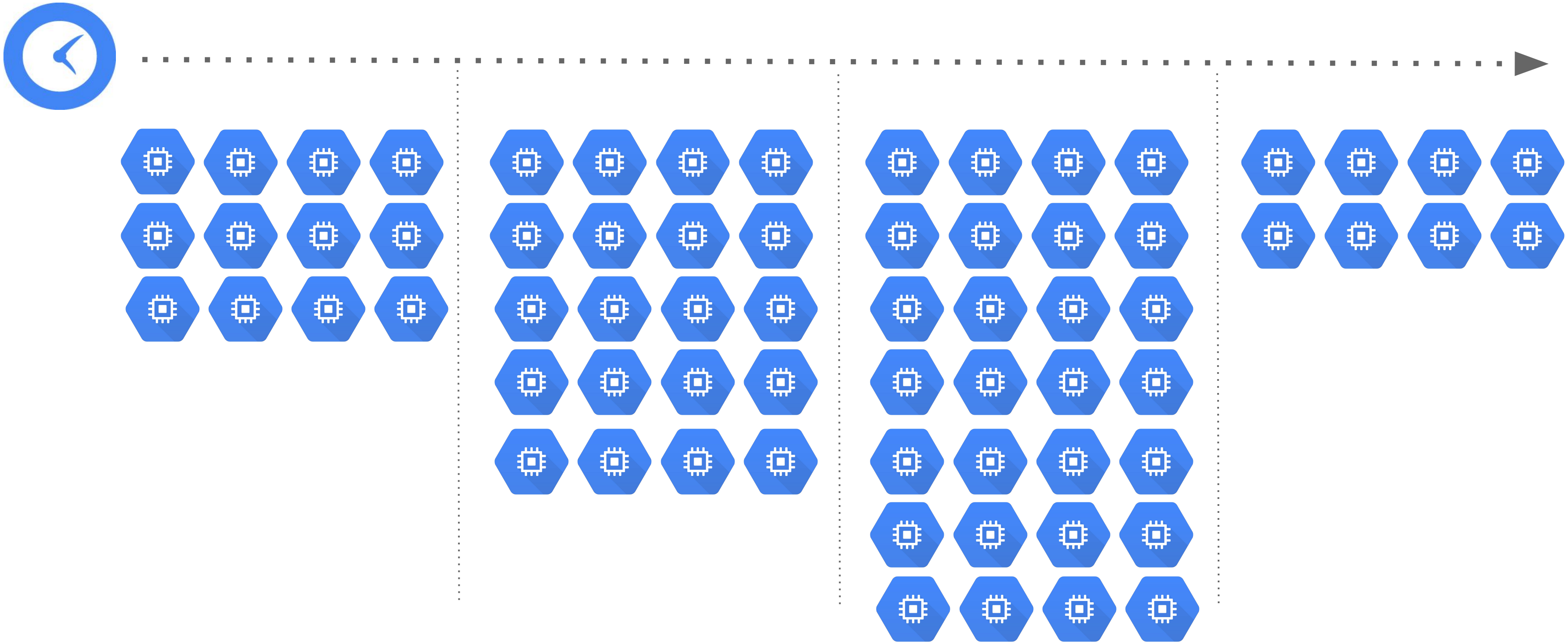
ExtractUserIdBySession
Done

ArchiveTransform

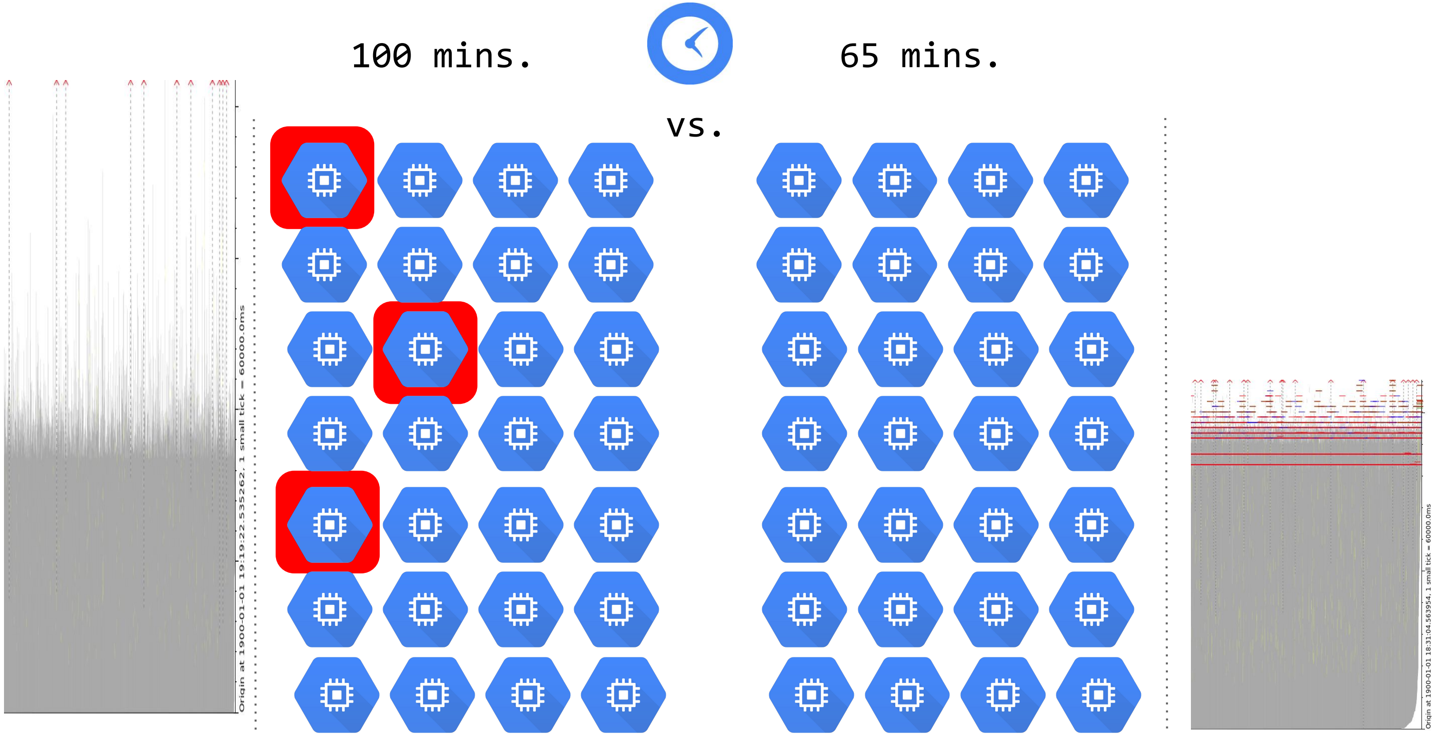
ProduceRowsAndHash
Done

BigQueryIO.Write
Done

Fully Managed: Dynamic Worker Scaling



Fully Managed: Dynamic Work Rebalancing



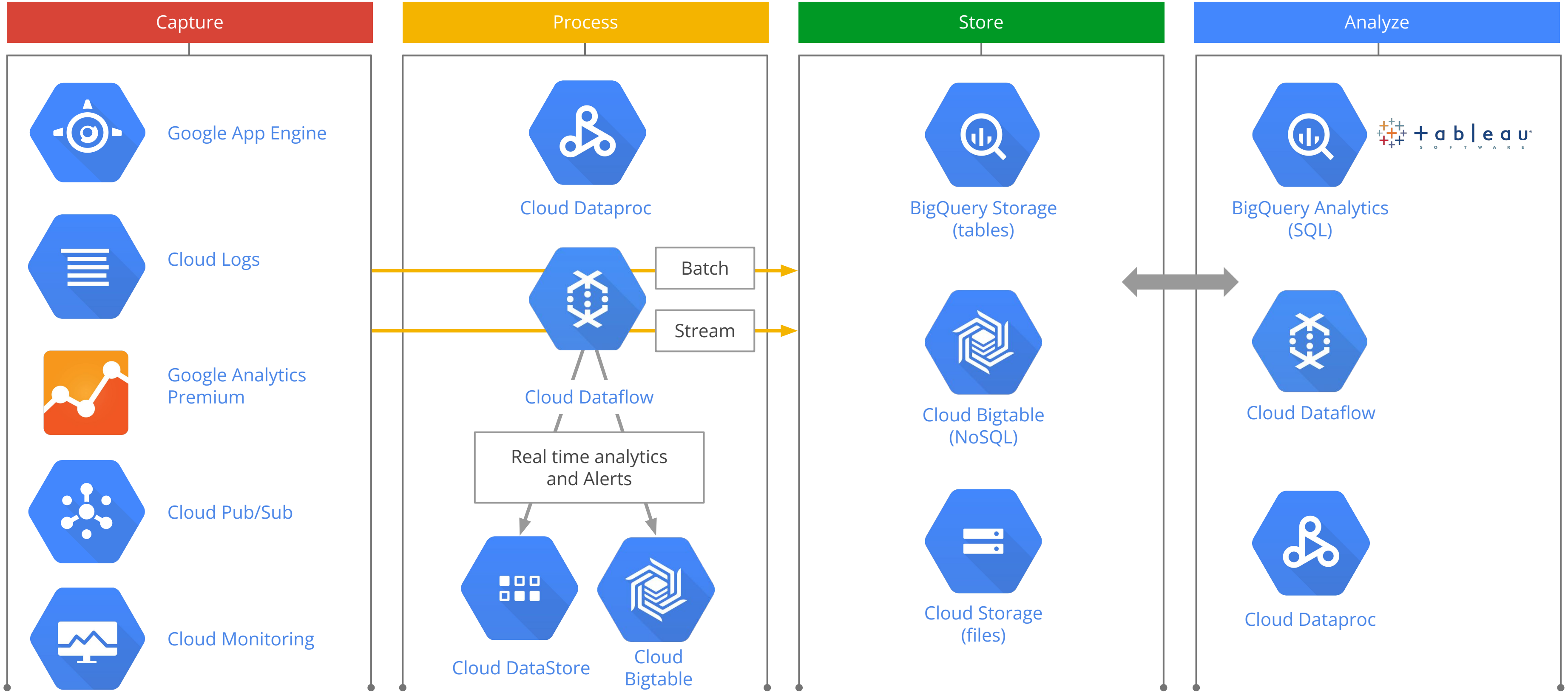
100 mins.

vs.

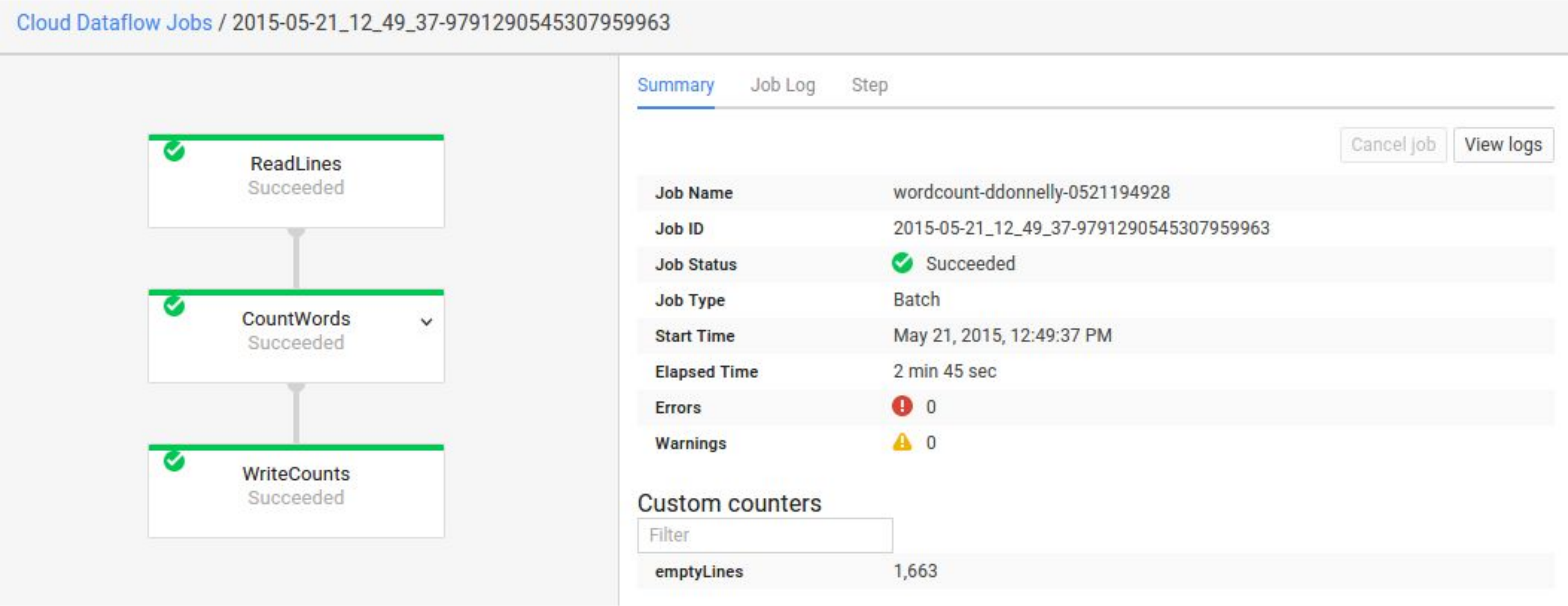
65 mins.

Origin at 1900-01-01 19:19:22.535262, 1 small tick = 60000.0ms
Origin at 1900-01-01 18:31:04.563954, 1 small tick = 60000.0ms

Integrated: Part of Google Cloud Platform



Integrated: Monitoring UI



Cloud Dataflow Jobs / 2015-05-21_12_49_37-9791290545307959963
Summary
Job Log
Step
Cancel job
View logs

ReadLines
Succeeded

CountWords
Succeeded

WriteCounts
Succeeded

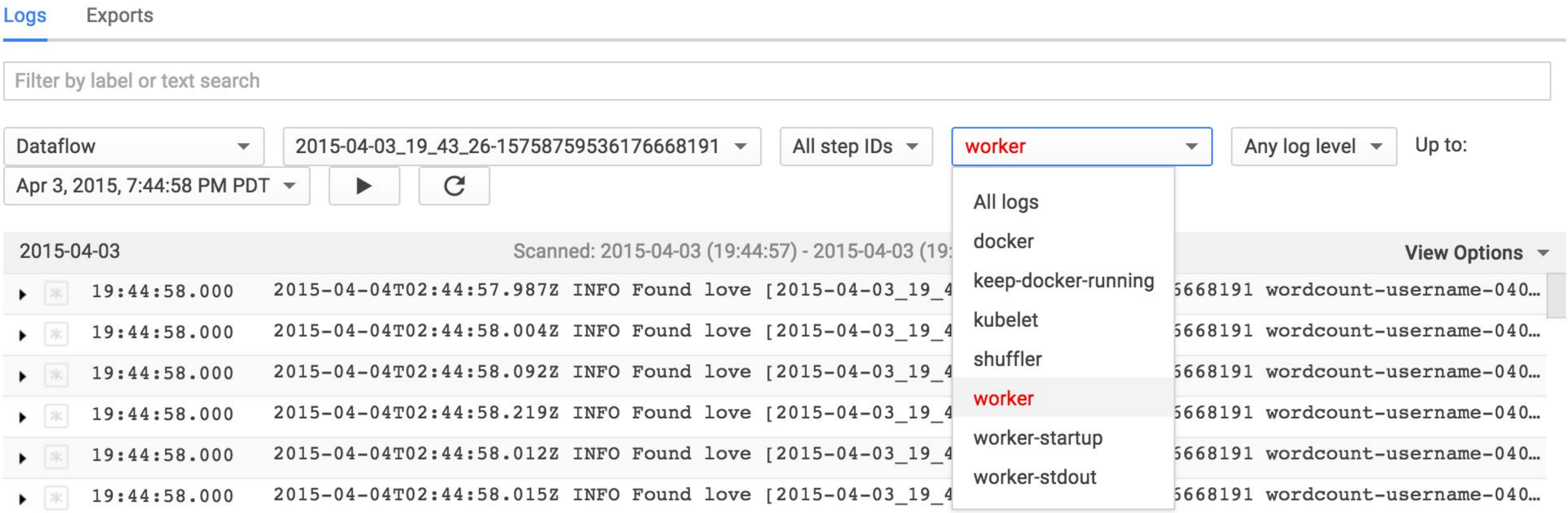
Job Name	wordcount-ddonnelly-0521194928
Job ID	2015-05-21_12_49_37-9791290545307959963
Job Status	✓ Succeeded
Job Type	Batch
Start Time	May 21, 2015, 12:49:37 PM
Elapsed Time	2 min 45 sec
Errors	0
Warnings	0

Custom counters

Filter

emptyLines	1,663

Integrated: Distributed Logging

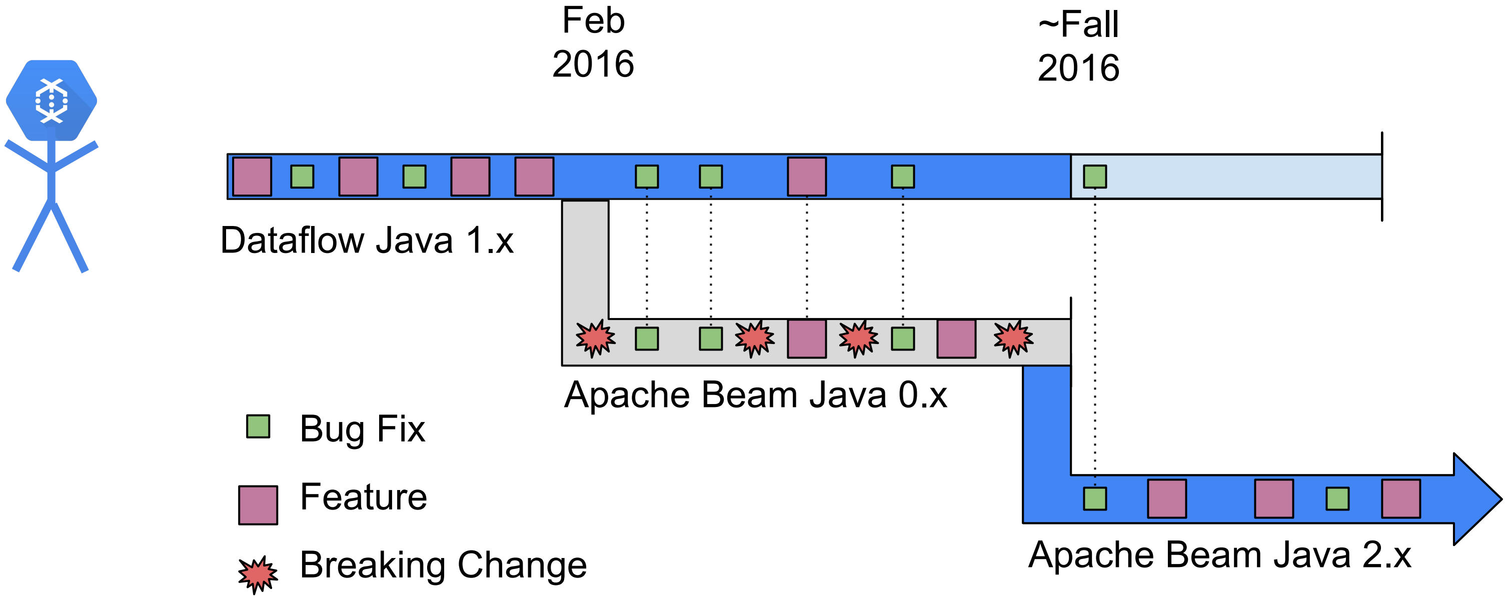


Dataflow | 2015-04-03_19_43_26-1575875953617666819 1 | All step IDs | worker | Any log level | Up to: Apr 3, 2015, 7:44:58 PM PDT

2015-04-03 Scanned: 2015-04-03 (19:44:57) - 2015-04-03 (19:

```
19:44:58.000   2015-04-04T02:44:57.987Z INFO Found love [2015-04-03_19_4   5668191 wordcount-username-040…
19:44:58.000   2015-04-04T02:44:58.004Z INFO Found love [2015-04-03_19_4   5668191 wordcount-username-040…
19:44:58.000   2015-04-04T02:44:58.092Z INFO Found love [2015-04-03_19_4   5668191 wordcount-username-040…
19:44:58.000   2015-04-04T02:44:58.219Z INFO Found love [2015-04-03_19_4   5668191 wordcount-username-040…
19:44:58.000   2015-04-04T02:44:58.012Z INFO Found love [2015-04-03_19_4   5668191 wordcount-username-040…
19:44:58.000   2015-04-04T02:44:58.015Z INFO Found love [2015-04-03_19_4   5668191 wordcount-username-040…
```

Transitioning from Dataflow 1.x to Beam



Learn More!

Cloud Dataflow

http://cloud.google.com/dataflow/

Cloud Dataflow on Stack Overflow

http://www.stackoverflow.com/questions/tagged/google-cloud-dataflow

Apache Beam

https://beam.apache.org

+

Thank you!