How to: Visualise spark output in the spark history server — AWS Glue

Craig Godden-Payne
Feb 7 · 3 min read ★

A guide on how to setup a standalone spark history server, for help when diagnosing issues with AWS Glue.

I recently had to try and spot why a job running in AWS Glue was not performing as expected, and there didn't seem to be any inbuilt tools in AWS to help with this.

I found a post which referred to a cloud formation stack, which would setup spark so that you could import the logs into that, but it seemed a bit overkill when all I wanted to do was view the history logs in a graphical way, rather than try and piece together bits from a huge json file.

AWS Glue outputs a huge amount of logging information, most of which I found to be irrelevant to what I was trying to diagnose, when my job was running slow and not as expected.

Getting the spark logs

When you run a glue job in AWS, there is an option for you to output the spark logs, to an S3 bucket, which can then be visualised using the spark history ui server. I would always select this option anyway, as the cost of S3 storage is so low, it seems like a no brainer, especially if you want to occasionally try and optimise the jobs.

Spark UI

I found that when trying to setup SparkUI locally, it was not the easiest application to setup, but I did manage to find a docker image running exactly what I needed, which was the SparkUI server with the History server built in.

I then I copied the spark logs into an events directory, then built and ran the container.

Dockerfile:

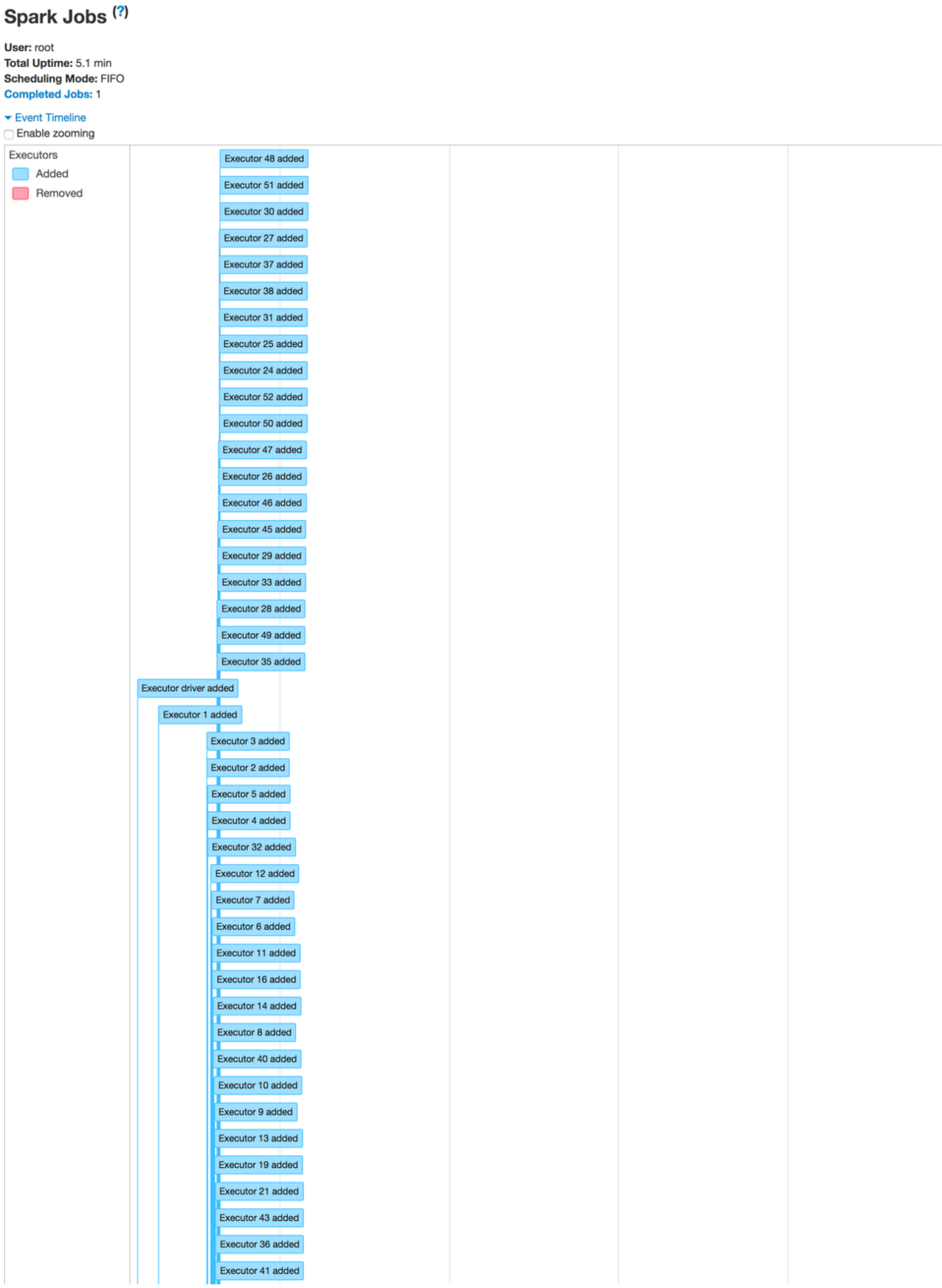
```
ARG SPARK_IMAGE=gcr.io/spark-operator/spark:v2.4.4
FROM ${SPARK_IMAGE}

RUN apk --update add coreutils

RUN mkdir /tmp/spark-events

ENV SPARK_NO_DAEMONIZE TRUE
ENTRYPOINT ["/opt/spark/sbin/start-history-server.sh"]
```

I built and ran the image

```
docker build . -t spark-history-server
docker run -it -v ${PWD}/events:/tmp/spark-events -p 18080:18080
spark-history-server
```

I could then view the results at http://localhost:18080

Here are some example logs which I have outputted, I find that it is certainly much easier to figure out what the hell is going on with performance of my glue job, over trying to figure out what is going on from the the Glue output!!

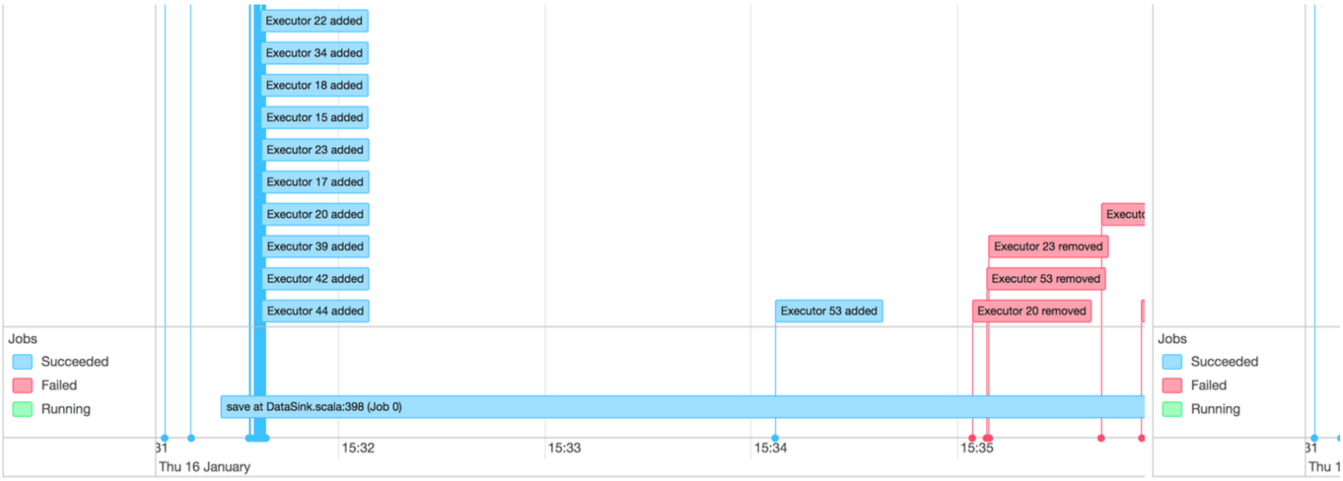


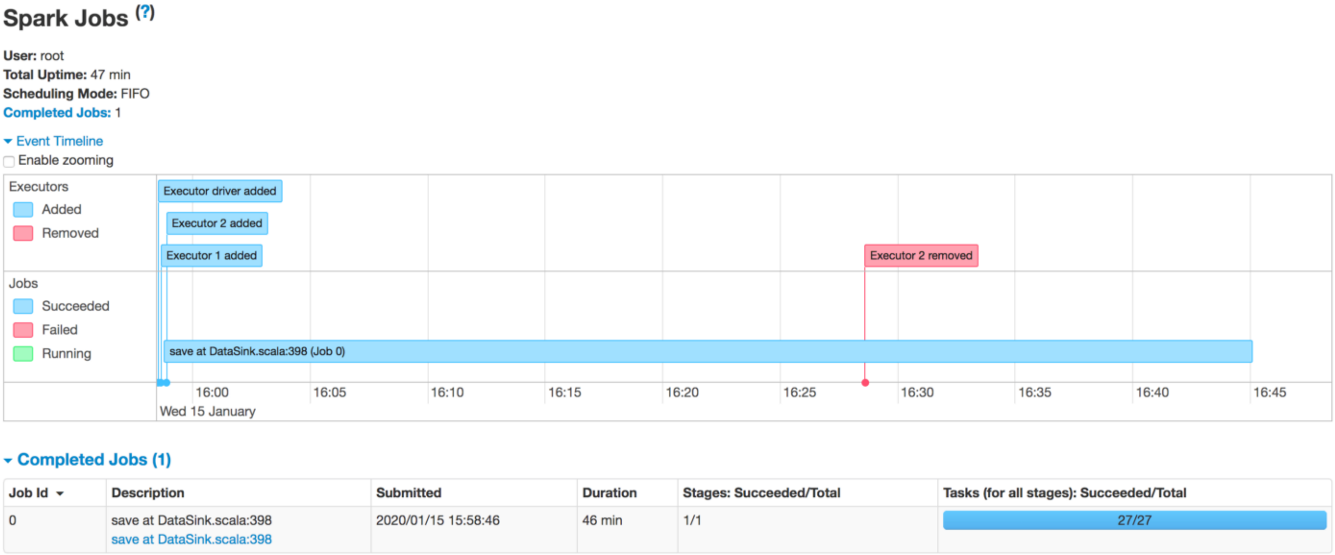
Spark Jobs (?)

User: root
Total Uptime: 5.1 min
Scheduling Mode: FIFO
Completed Jobs: 1

Spark Jobs (?)

User: root
Total Uptime: 5.1 min
Scheduling Mode: FIFO
Completed Jobs: 1





Spark Jobs (?)

User: root
Total Uptime: 47 min
Scheduling Mode: FIFO
Completed Jobs: 1

Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	save at DataSink.scala:398 save at DataSink.scala:398	2020/01/15 15:58:46	46 min	1/1	27/27

Written on January 16, 2020.

Originally published on: https://craig.goddenpayne.co.uk/visualising-glue-history-in-spark-history-server/

AWS Glue Apache Spark