Towards a Big Data Debugger in Apache Spark

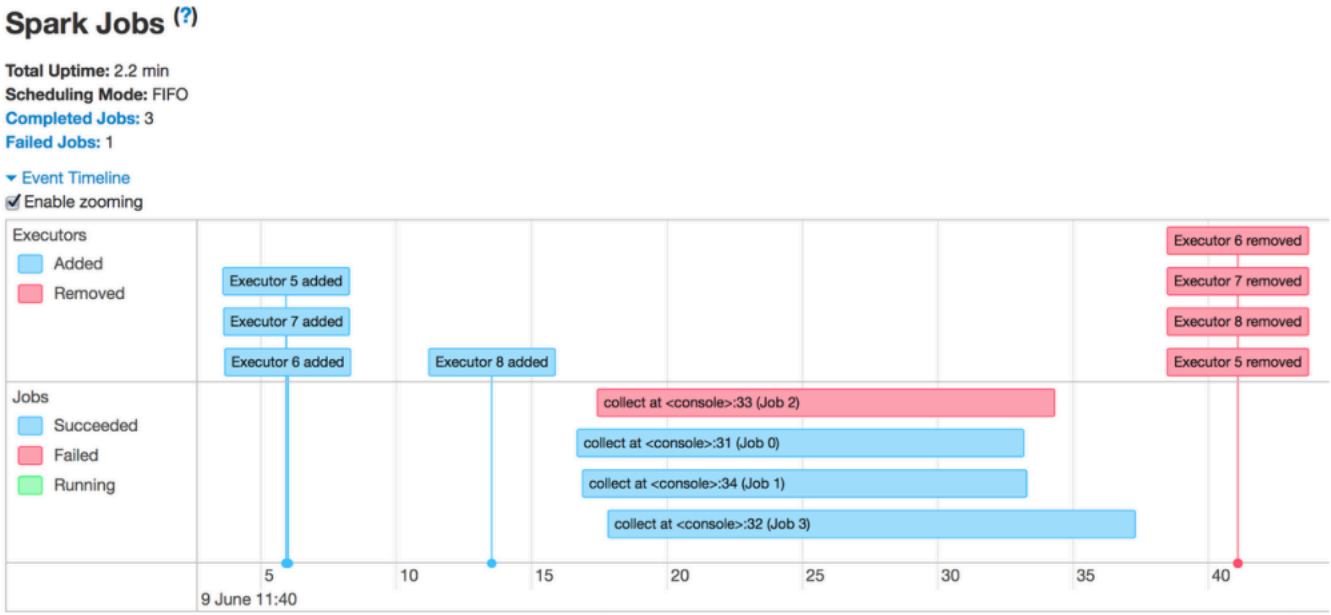
Tyson Condie, UCLA

Tuning Spark Applications

- Commonly through visualization tools

Tuning Spark Applications

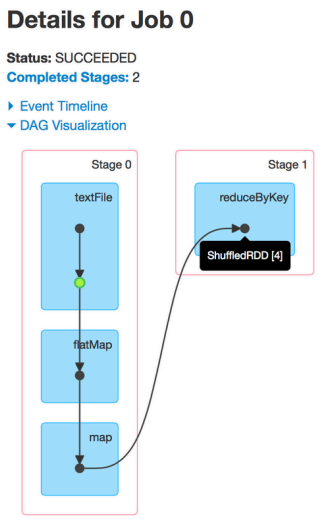
- Commonly through visualization tools
  - Timeline view of Spark events

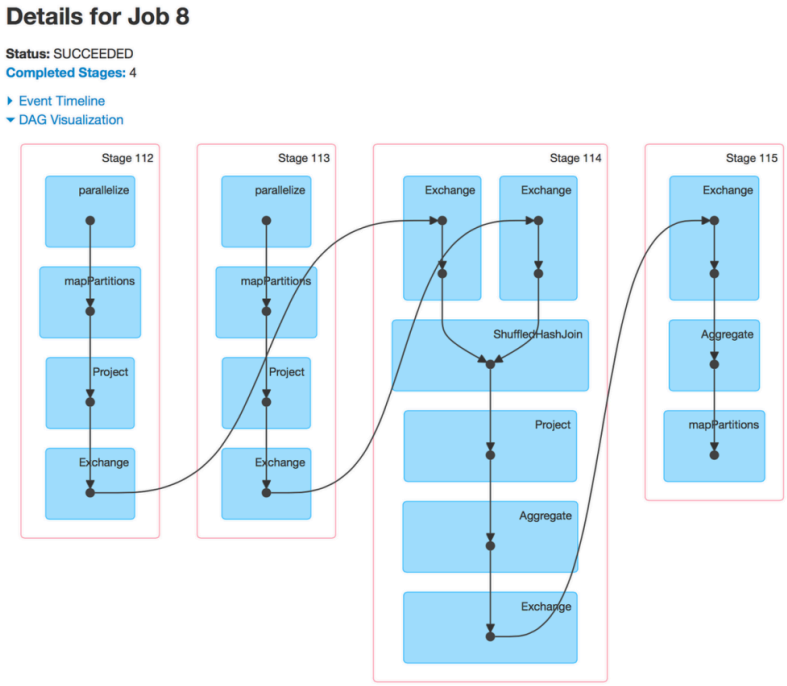


Taken from https://databricks.com/blog/2015/06/22/understanding-your-spark-application-through-visualization.html

Tuning Spark Applications

- Commonly through visualization tools
  - Execution DAG

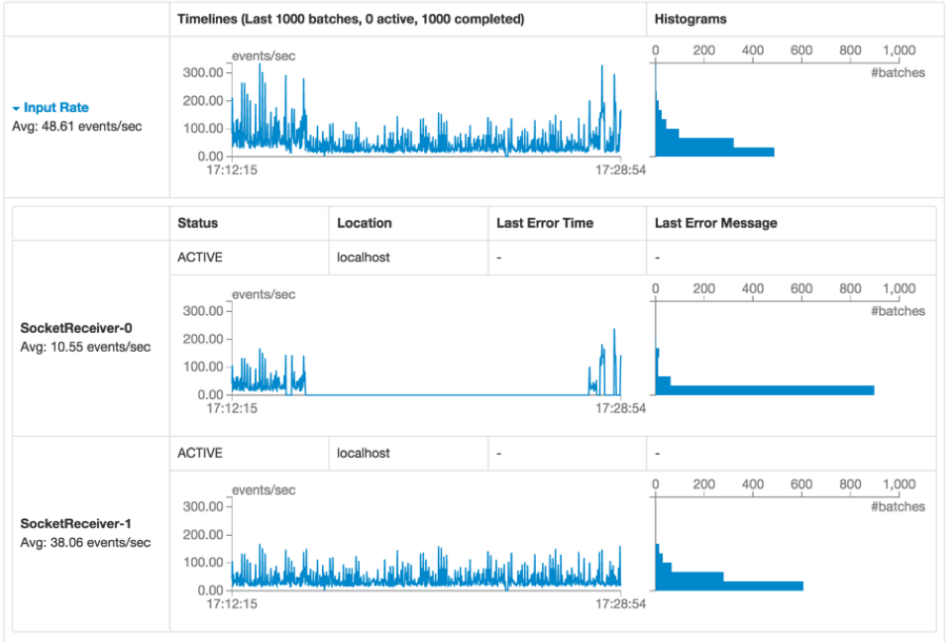




Taken from https://databricks.com/blog/2015/06/22/understanding-your-spark-application-through-visualization.html

Tuning Spark Applications

- Commonly through visualization tools
  - Visualization of Spark Streaming statistics



Taken from https://databricks.com/blog/2015/06/22/understanding-your-spark-application-through-visualization.html


stackoverflow
Questions
Tags
Users
Trying to debug a Spark Application on a cluster…
“I would like … through the Spark … code on the worker nodes when I submit my application … I am assuming I should setup Spark on Eclipse … to enable stepping through Spark source code on the worker nodes.”
./bin/spark-class org.apache.spark.deploy.worker.worker mas
command for submitting application
./sbin/spark-submit --class Application --master URL ~/a
Now, I would like to understand the flow of control through … source code on the worker nodes when I submit my application(I just want to use one … given examples that use reduce()). I am assuming I should setup Spark on Eclipse. The Eclipse setup link on the Apache Spark website seems to be broken. I would appreciate some guidance on setting up Spark and Eclipse to enable stepping through Spark source code on the worker nodes.
Thanks!
eclipse
debugging
apache-spark
share improve this question
asked Mar 17 at 3:19
AndroidDev93
698 3 15 38

After 5 months, still no good answers!

Add the relevant spark jars to the eclipse project. And then set the master in the code. And now

Have you tried passing remote debug parameters to worker JVM? I think its something like

You could run the Spark application in local mode if you just need to debug the logic of your transformations. This can be run in your IDE and you'll be able to debug like any other application:

When you run a spark application on yarn, there is an option like this:

0

```
YARN_OPTS="-agentlib:jdwp=transport=dt_socket,server=y,suspend=n,address=5455 $YARN_OPTS"
```

You can add it to `yarn-env.sh` and remote debugging will be available via `port 5455`.

If you use spark in standalone mode, I believe this can help:

```
export SPARK_JAVA_OPTS=-agentlib:jdwp=transport=dt_socket,server=y,suspend=n,address=5005
```

share improve this answer

edited Jul 20 at 21:23

Cleb

1,510 • 2 • 12 • 24

answered Jul 3 at 9:14

user3504158

1

Towards Interactive Debugging for Apache Spark

Goal: Develop "debugging toolkits" on Apache Spark where features operate at scale and impose minimal overheads on (normal) program execution

- Data provenance similar to databases but without going to offline
- Traditional debugging features: Breakpoints; Watchpoints; Stepping
- Record-level error/exception handling: Crash culprit; Outlier identification
- Execution replay: On input or intermediate data; Leading to a given result e.g., outlier, crash culprit

Titian Library: Provides capture and interactive analysis of data provenance

- Data Provenance supported by enabling record-level tracing in Spark's dataflow
  - Provenance recorded as data records are pipelined through transformations
  - Provenance exposed to the programmer as Resilient Distributed Datasets (RDDs) for analysis
- To appear in PVLDB Volume 9, Issue 3.
- Following on work: Breakpoints, Watchpoints and Stepping is under submission to ICSE 2016.

Supported through
- BD2K (Big Data to Knowledge) NIH Grant
- Industry grants from IBM Research and Intel

Collaborators: Matteo Interlandi (post-doc), Muhammad Ali Gulzar, Sai Deep Tetali, and UCLA Faculty---Miryung Kim, Todd Millstein