MRUnitProposal

MRUnit, a library to support unit testing of Hadoop MapReduce jobs

Abstract

MRUnit is a java library that provides mocks and infrastructure for writing unit tests for Hadoop MapReduce jobs and related components.

Proposal

MRUnit is a java library to facilitate unit testing of Hadoop MapReduce jobs by providing drivers and mock objects to simulate the Hadoop runtime environment of a map reduce job. This code base already exists as a subproject of the Apache Hadoop MapReduce project and lives in the "contrib" directory of the source tree.

Background

Writing unit tests of MapReduce jobs can be a tedious process. User code can quickly become entangled with Hadoop APIs making testing difficult and error prone. In many cases, users will simply forgo testing given the complexity of the environment. MRUnit was created as a simple library users can use in conjunction with test suites like JUnit to provide a harness for injecting appropriate mock objects.

Rationale

MRUnit has existed as a contrib component of Apache Hadoop. This has served to introduce users to the library and to provide necessary functionality to developers in the form of development support. That said, MRUnit is not necessarily an intrinsic component of Hadoop proper and could benefit from being a standalone project in that:

- A separate project would support an independent development and release schedule allowing for faster iteration and response to user requests.
- Separating adjunct projects from the core Hadoop codebase simplifies Hadoop's build and release.
- MRUnit users can get a simpler artifact in a way most appropriate to development time (i.e. Maven or Ivy repositories).
- MRUnit can build out independent support for different versions of Hadoop without requiring circular dependencies or testing issues.

Having greater development and tooling support for Hadoop makes the project accessible to a wider audience by reducing the chance of bugs.

Initial Goals

- Provide a new home for the existing codebase.
- Make artifacts available via Maven and / or Ivy.
- Expand test support for other Hadoop components (e.g. Partitioners)
- Establish a lightweight, independent release cycle.

Current Status

Meritocracy

MRUnit was originally created by Aaron Kimball, and has had some contributions from members of the Hadoop community. By becoming its own project, significant contributors to MRUnit would become committers, and allow the project to grow.

Community

The MRUnit community is predominantly composed of engineers who author MapReduce jobs running against Apache Hadoop. Given that this library appeals to a specific subset of the overall Apache Hadoop community, it makes sense to decouple its release cycle from that of Hadoop as a whole, to allow more rapid iteration in this space.

Core developers

Aaron Kimball wrote most of the original code and is familiar with open source and Apache-style development, being a Hadoop committer. A number of other contributors have provided patches to this codebase over time. Eric Sammer has worked as a committer on Flume, a github-based open source project.

Alignment

MRUnit aligns with Hadoop as it aims to be a testing harness and framework for the Hadoop MapReduce framework.

Known Risks

Orphaned products

All members of the team are committed to making MRUnit a success.

Inexperience with Open Source

The initial code comes from Hadoop where it was developed in an open-source, collaborative way. All the initial committers are committers on other Apache projects (with the exception of Eric who is experienced with open source development at Github and other communities), and are experienced in working with new contributors.

Homogenous Developers

The initial set of committers is from a diverse set of organizations, and geographic locations. They are all experienced with developing in a distributed environment.

Reliance on Salaried Developers

It is expected that MRUnit will be developed on a combination of volunteer and salaried time.

Relationships with Other Apache Products

MRUnit will depend on many other Apache Projects as already mentioned above (e.g. Hadoop).

A Excessive Fascination with the Apache Brand

We think that MRUnit will benefit from The Apache Incubator. There was discussion about moving this project entirely out of Apache Hadoop and into e.g., Github (as a fork), but after Chris Mattmann prompted some discussions on the Hadoop general list to stick around in the Incubator, many of the MRUnit enthusiasts (Eric, Patrick, Nige, etc.) thought it would be a great idea to stick around in Apache. We have all had experience working within the Apache community and benefiting from Apache-released software, and believe the Apache community remains the right home for this project.

Documentation

Information on the library can be found at

- http://svn.apache.org/viewvc/hadoop/mapreduce/trunk/src/contrib/mrunit/doc/overview.html?view=markup

Initial Source

- http://svn.apache.org/viewvc/hadoop/mapreduce/trunk/src/contrib/mrunit/
- https://github.com/esammer/mrunit

Source and Intellectual Property Submission Plan

The initial source is already in an Apache project's SVN repository (Hadoop), so there should be no action required here.

External Dependencies

The existing external dependencies all have Apache compatible licenses.

Cryptography

MRUnit code likely does not fall into this area.

Required Resources

Mailing lists

- mrunit-private (with moderated subscriptions)
- mrunit-dev
- mrunit-commits
- mrunit-user

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/mrunit

Issue Tracking

- JIRA MRUnit (MRUNIT)

Other Resources

The existing code already has unit and integration tests so we would like a Hudson instance to run them whenever a new patch is submitted. This can be added after project creation.

Initial Committers

- First Last (user at apache dot org)
- Eric Sammer (esammer at cloudera dot com)
- Aaron Kimball (kimballa at apache dot org)
- Konstantin Boudnik (cos at apache dot org)
- Garrett Wu (wugarrett at gmail dot com)

Affiliations

- Eric Sammer, Cloudera
- Nigel Daley, Proximal Labs
- Patrick Hunt, Cloudera
- Chris A. Mattmann, NASA Jet Propulsion Laboratory/University of Southern California
- Aaron Kimball, Odiago
- Garrett Wu, Odiago

Sponsors

Champion

- Patrick Hunt

Nominated Mentors

- Nigel Daley
- Patrick Hunt
- Chris Mattmann

Sponsoring Entity

- Incubator PMC