ChukwaProposal

Chukwa Proposal

Abstract

Chukwa is a log collection and analysis framework base on Hadoop Map/Reduce.

Proposal

Chukwa will develop a open source data collection system for monitoring large distributed systems. Chukwa is built on top of the Hadoop Distributed File System (HDFS) and Map/Reduce framework and inherits Hadoop’s scalability and robustness. Chukwa also includes a exible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.

Background

Apache Hadoop, lacks a good procedure to monitor and troubleshoot large distributed systems. Chukwa was initially developed at Yahoo Inc headed by Mac Yang, Sunnyvale in 2008. Chukwa was designed as a reference implementation for monitoring large distributed system on top of Hadoop. Since 2009 major parts of the development comes from Internet community contribution. Chukwa is current a Hadoop subproject.

Rationale

The maintainers and developers of Chukwa are interested in joining the Apache Software Foundation top level project for several reasons:

- Apache provide a great community for open source software development environment.
- It might open the door for sharing ideas or cooperation with other Apache projects, such as Avro and Hadoop.
- Chukwa would like to benefit from Apache's infrastructure.

Initial Goals

Though the bulk of Chukwa initial development is complete and the framework is running stable, there are still some large areas for future development. Some area we hope to focus on in Apache:

- Improve Chukwa Demux map/reduce Job
- Refine automated log analysis algorithms
- Remove dependency on relational database for reporting

Current Status

Meritocracy

The initial developers are very familiar with meritocratic open source development, both at Apache and elsewhere. Apache was chosen specifically because the initial developers want to encourage this style of development for the project.

Community

Chukwa is used in many organization which are interested in the advancement of the Chukwa development. Many of these have at least one developer that joined the Chukwa mailing list and so the mailing list is the most important communication platform. The Chukwa community encourages suggestions and contributions from any potential user and developer.

Core Developers

The initial set of Chukwa committers includes folks from the Hadoop communities.
We have varying degrees of experience with Apache-style open source development.

Alignment

Chukwa is a framework for Apache Hadoop. This is why Apache Hadoop is the most important dependency for Chukwa. And Chukwa is also a particularly good fit for Apache due to integration potential with other projects specifically Avro and Log4j.

Known Risks

Orphaned products

Most of the active developers would like to become Chukwa Committers or PMC Members and have long term interest to develop/maintain and **use** the code.

Inexperience with Open Source

Chukwa was started as an open source contribute project to Hadoop in 2008. Many of the committers have experience working on open source projects and there are also at least one developer which has experience as committer on other Apache projects.

Homogenous Developers

As mentioned above, the current list of committers includes developers from at least two different companies plus many independent volunteers.

Reliance on Salaried Developers

At this time, many of the code comes from different companies like RAD Lab. Because RAD Lab is a research facility, many of the work is done by students working on their diploma thesis.

Relationships with Other Apache Products

At this time, the only dependency to other Apache projects is Apache Hadoop. When dependency on relational database is removed, Avro will become the standard serialization framework for Chukwa.

A Excessive Fascination with the Apache Brand

The Chukwa project exist quite successful on their own and could continue on that path with no problems at all. We expect the Apache top level project brand could help to increase the visibility of the project and so maybe more developers could be interested in the project.

Documentation

- The existing project page could be found here: http://hadoop.apache.org/chukwa
- The Chukwa Architecture: http://hadoop.apache.org/chukwa/docs/current/design.html
- The Chukwa mailing list with archive: http://hadoop.apache.org/chukwa/mailing_lists.html

Initial Source

Source and Intellectual Property Submission Plan

The complete Chukwa code is under Apache Software License 2. The complete codebase is already hosted in ASF Repository.

External Dependencies

The dependencies all have Apache compatible licenses. These include BSD, CDDL, and MIT licensed dependencies.

Cryptography

None

Required Resources

Mailing lists

- dev AT chukwa DOT apache DOT org
- commits AT chukwa DOT apache DOT org
- user AT chukwa DOT apache DOT org
- private AT chukwa DOT apache DOT org

Subversion Directory

https://svn.apache.org/repos/asf/chukwa

Issue Tracking

JIRA CHUKWA

Initial Committers

- Jerome Boulon (jboulon AT apache DOT org)
- Chris Douglas (cdouglas AT apache DOT org)
- Bill Graham (billgraham AT gmail DOT com)
- Ari Rabkin (asrabkin AT apache DOT org)
- Jiaqi Tan (tanjiaqi AT gmail DOT com)
- Eric Yang (eyang AT apache DOT org)

Affiliations

- Jerome Boulon (Netflix)
- Chris Douglas (Yahoo Inc)
- Bill Graham (CBS Interactive)
- Owen O'Malley (Yahoo Inc)
- Ari Rabkin (RAD Lab)
- Jiaqi Tan (DSO National Laboratories)
- Eric Yang (Yahoo Inc)

Sponsors

Champion

Chris Douglas (and Mentor) for the project, (as defined in http://incubator.apache.org/incubation/Roles_and_Responsibilities.html)

Nominated Mentors

- Chris Douglas
- Owen O'Malley
- William A. Rowe Jr.
- Bernd Fondermann

Sponsoring Entity

- Incubator