SAMOAProposal

SAMOA Proposal

Abstract

SAMOA is an an open-source platform for mining big data streams.

Proposal

SAMOA provides a collection of distributed streaming algorithms for the most common data mining and machine learning tasks such as classification, clustering, and regression, as well as programming abstractions to develop new algorithms that run on top of distributed stream processing engines (DSPEs). It features a pluggable architecture that allows it to run on several DSPEs such as Apache Storm, Apache S4, and Apache Samza.

Background

Hadoop and its ecosystem have changed the way data are processed by allowing to push algorithms to unprecedented scale. As an example, Mahout allows to run data mining and machine learning algorithms on very large datasets. However, Hadoop and Mahout are not suited to handle streaming data. Simply put, the goal of SAMOA is to provide a streaming counterpart to Mahout.

Rationale

SAMOA aims to fill the current gap in tools for mining large scale streams. Simplifying, think of SAMOA as "Mahout for streaming". Many organizations can benefit from a scalable stream mining platform such as SAMOA.

SAMOA is a natural fit for the Apache Software Foundation. It is licensed under the ASL v2.0. It already interoperates with several existing Apache projects such as Storm, S4, and Samza. Furthermore, it is complementary to existing Apache projects such as Mahout. The initial committers are familiar with the Apache process and subscribes to the Apache mission. Indeed, the team includes multiple Apache committers. Finally, joining Apache will help coordinate the development effort of the growing number of organizations which contribute to SAMOA.

Initial Goals

- Move the existing codebase to Apache
- Integrate with the Apache development process
- Incremental development and releases per Apache guidelines

Current Status

SAMOA started as a research project at Yahoo Labs in 2013 and was open-sourced in October the same year. It has been under development on Yahoo's public GitHub repository since being open-sourced. It has undergone two releases (0.1, 0.2).

Meritocracy

The SAMOA project already operates on meritocratic principles. Today, SAMOA has several developers and has accepted multiple patches from outside of Yahoo Labs. However, our intent with this incubator proposal is to start building a more diverse developer community around SAMOA that follows the Apache meritocracy model. We will identify all committers and PPMC members for the project operating under the ASF meritocratic principles. We plan to continue support for new contributors and work with those who contribute significantly to the project to make them committers.

Community

SAMOA is currently being used internally at Yahoo. Acceptance into the Apache foundation would bolster the existing user and developer community around SAMOA. That community includes contributors from several institutions, active mostly on GitHub's pages. SAMOA has been starred more than 300 times and forked more than 50 times on GitHub as of November 2014.

Core Developers

The core developers are a diverse group, many of which already very experienced with open source. There are two existing Apache committers, along with people from various companies and universities.

Alignment

The ASF is the natural choice to host SAMOA. First, its goal of encouraging community-driven open-source projects fits with our vision for SAMOA. Additionally, many other projects that SAMOA is based on, such as Apache Storm, S4, Samza, and HDFS, are hosted by the ASF. Close proximity of SAMOA to these projects within the ASF will provide mutual benefit.

Known Risks

Orphaned Products

Given the current level of investment in SAMOA the risk of the project being abandoned is minimal. There are several constituents who are highly incentivized to continue development, and Yahoo Labs relies on SAMOA as a platform for a large number of long-term research projects. However, the small number of initial committers might be a concern. We plan to address this issue during incubation by growing the community and the number of committers.

Inexperience with Open Source

SAMOA has existed as a healthy open source project for one year. During this time, we have curated an open-source community successfully, attracting developers from a diverse group of universities and companies including Huawei, Yahoo, University of Porto, and Universitat Politecnica de Catalunya.

Gianmarco is a committer for Apache Pig, Matthieu for Apache S4. Albert is one of the lead developers of MOA, an open-source tool for streaming machine learning.

Homogenous Developers

The initial list of committers includes developers from several institutions, both academic and industrial. The committers are geographically distributed across Europe, America, and Asia.

Reliance on Salaried Developers

Like most open source projects, SAMOA receives a substantial support from salaried developers. In addition, those working from within corporations often devote “after hours” or spare time in the project - and these come from several organizations. We will work to ensure the ability for the project to continuously be stewarded and to proceed forward independently of salaried developers.

Relationship with Other Apache Products

SAMOA interoperates with several existing Apache projects, mainly by using them as stream processing engines: Apache Storm, Apache S4, and Apache Samza. It is a counterpart of Apache Mahout for streaming. It also uses several other Apache components, including Apache Maven and several Apache Commons libraries.

A Excessive Fascination with the Apache Brand

SAMOA is already a healthy and relatively well known open source project. This proposal is not for the purpose of generating publicity. Rather, the primary benefits to joining Apache are those outlined in the Rationale section. We are more interested in establishing a strong community that can drive the project independently of Yahoo.

Documentation

The reader will find these websites relevant:

- SAMOA website: http://samoa-project.net/
- SAMOA documentation: https://github.com/yahoo/samoa/wiki/
- Issue tracking: https://github.com/yahoo/samoa/issues
- Codebase: https://github.com/yahoo/samoa
- User group: http://groups.google.com/group/samoa-user

Initial Source

The SAMOA codebase is currently hosted on GitHub: https://github.com/yahoo/samoa. This is the exact codebase that we would migrate to the Apache foundation.

Source and Intellectual Property Submission Plan

Currently, the SAMOA codebase is distributed under an Apache license v2.0. The vast majority of code has copyright held by Yahoo. Upon entering the Incubator, Yahoo will grant a license to the Apache foundation. In certain cases where individuals or organizations hold copyright, we will ensure they grant a license to the Apache foundation. Going forward, all commits will be licensed directly to the Apache foundation through our signed Individual Contributor License Agreements for all committers on the project.

Cryptography

We do not expect SAMOA to be a controlled export item due to the use of encryption.

External Dependencies

To the best of our knowledge, all dependencies of SAMOA are distributed under Apache compatible licenses. Upon acceptance to the incubator, we would begin a thorough analysis of all transitive dependencies to verify this fact and introduce license checking into the build and release process (for instance integrating Apache Rat).

Required Resources

Mailing Lists

We will migrate the existing SAMOA mailing lists as follows:

- samoa-users@googlegroups --> users@samoa.incubator.apache.org
- samoa-developers@googlegroups --> dev@samoa.incubator.apache.org

SAMOA commits are hosted on GitHub, so we would request the following mailing list:

- commits@samoa.incubator.apache.org

We would also request the following mailing list:

- private@samoa.incubator.apache.org (with moderated subscription)

Source control

The SAMOA team would like to use Git for source control, due to our current use of Git. We request a writeable Git repo for SAMOA, and mirroring to be set up to GitHub through INFRA.

https://git-wip-us.apache.org/repos/asf/incubator-samoa.git

Issue Tracking

SAMOA currently uses GitHub for issue tracking. We will migrate to the Apache JIRA instance. http://issues.apache.org/jira/browse/SAMOA

Initial Committers & Affiliations

- Albert Bifet, Huawei, <abifet at waikato dot ac dot nz>
- Gianmarco De Francisci Morales, Yahoo Labs, <gdfm at apache dot org>
- Nicolas Kourtellis, Yahoo Labs, <nkourtellis at gmail dot com>
- Matthieu Morel, Yahoo Labs, <mmorel at apache dot org>
- Arinto Murdopo, Living Analytics Research Centre, <arintom at smu dot edu dot sg>
- Olivier Van Laere, BlueShift Labs, <olivier at getblueshift dot com>

Sponsors

Champion

- Daniel Dai <daijy at apache dot org>

Nominated Mentors

- Alan Gates <gates at apache dot org>
- Ted Dunning <tdunning at apache dot org>
- Ashutosh Chauhan <hashutosh at apache dot org>
- Enis Soztutar <enis at apache dot org>

Sponsoring Entity

The Apache Incubator