FluoProposal

Fluo Proposal

Abstract

Fluo is a distributed system for incrementally processing large data sets stored in Accumulo.

Proposal

Fluo is a distributed transaction and notification system that enables the incremental processing of large data sets. Its transaction system allows for concurrent, cross-node updates to data stored in Accumulo. Its notification system enables developers to write code to be executed when observed data changes. Fluo provides a core API to perform transactional updates using minimalistic get/set methods. Fluo also provides a higher order recipes API that builds on the core API to support more complex methods for transactional updates.

Background

Several frameworks exist for batch (i.e Spark, MapReduce) and stream (i.e Storm, Spark Streaming) processing of data. While batch and stream processing have strong use cases, they are not suited for joining incoming data in real-time to a large existing data set. To fill this need, Google developed an incremental processing system called Percolator and described it in the paper, *Large-scale Incremental Processing Using Distributed Transactions and Notifications* [1] .

Rationale

Fluo fills the need for cross-row (and cross-node) transactions in Accumulo by providing it with an open source implementation of Percolator. Fluo also satisfies a gap in Accumulo's ability to incrementally process data. Fluo also provides a novel recipes API which offers higher level abstractions for transactional updates.

Current Status

Fluo currently exists as an open source project on GitHub and has been in active development since 2013. The project has made an alpha release and two beta releases. The major features of Fluo outlined in this proposal have been implemented. Several example Fluo applications have been created and run successfully on clusters (up to 24 nodes).

Meritocracy

The Fluo project operates as a meritocracy and will continue to do so because we feel that a project comprised of a diverse set of committers will thrive. Therefore, we welcome new contributors and encourage them on their path to committership.

Community

Fluo is currently being used by a subset of the Accumulo community. The initial developers have been responsive to external contributions through pull requests and issues on GitHub. As Fluo releases a stable 1.0 version that is production-ready, we expect this community to grow. To encourage growth, we have created a project website with documentation, given talks at Meetups and the Accumulo Summit, and engaged with new users on GitHub and the Fluo mailing list.

Core Developers

The project was started by Keith Turner (an Apache Member and committer/PMC on Gora and Accumulo) in 2013, and the development has primarily consisted of his and Mike Walch's continued efforts. Additional developers have contributed over time, which has led to new committers.

Alignment

Fluo is closely linked to the Accumulo community, and fits well within the larger Hadoop ecosystem at Apache. Fluo utilizes several Apache projects, such as Accumulo, YARN, Twill, and ZooKeeper. Enabling closer collaboration between these communities through its coexistence within the ASF would help further drive the success of them all.

In addition to our technical ties to other ASF projects, our development philosophy aligns with Apache philosophies. Based on our experience with existing Apache projects, we are interested in establishing formal governance with a PMC and community bylaws, which we feel would best be done within Apache.

Known Risks

Orphaned Products

Fluo could be orphaned if the project fails to gain adoption and the core developers abandon their interest (this is not anticipated). This risk can be mitigated by attracting more committers and developing further documentation to ease adoption.

Inexperience with Open Source

Fluo has been an open source project on GitHub from the start of its development. Several Fluo developers are committers on other ASF projects as well as open source projects outside ASF, and understand open source development.

Homogeneous Developers

The initial committers work for different employers. We hope add more developers from other employers and industries.

Reliance on Salaried Developers

While most of the initial committers are paid to work on Fluo, there have been many contributions from developers working independently.

Relationships with Other Apache Products

Fluo uses Accumulo, Hadoop (HDFS & YARN), Twill, ZooKeeper, Curator, Thrift, and various Commons libraries. During development, contributions have been made to some of these Apache projects to better support Fluo use cases.

Apache Brand

While we recognize the impact of the Apache brand, we feel that Fluo would fit well in Apache because of its relationship to other Apache projects and because we share the ASF values of meritocracy and community over code.

Documentation

Information about Fluo can be found on the project website at http://fluo.io/. This includes:

- General documentation - http://fluo.io/docs/
- API documentation - http://fluo.io/apidocs/
- Release notes - http://fluo.io/release-notes/
- Blog posts - http://fluo.io/blog/

Initial Source

The initial source code is publicly available as an open source project on GitHub at https://github.com/fluo-io/fluo

Supplemental repositories also exist on GitHub at https://github.com/fluo-io and some of those will become part of the initial code base (perhaps in separate repositories).

Source and Intellectual Property Submission Plan

All of the Fluo's source code is available under the Apache License, Version 2.

The Fluo logo was designed and contributed to the Fluo project, for use by the project, and the contributors would like it to remain the logo of the project within the ASF, granting any necessary rights to the ASF, while continuing to use the logo on Fluo-related historical sites and project pages (such as Fluo's current GitHub site).

External Dependencies

Fluo has made it a point from its beginning to use dependencies which are compatible with the expectations of an ASF project. The following are its current dependencies, grouped by license.

Apache License, Version 2.0

- accumulo
- commons-{collections,configuration,io}
- curator
- dropwizard metrics
- easymock
- guava
- hadoop
- jcommander
- maven
- thrift
- twill
- zookeeper

BSD License (2-Clause)

- HdrHistogram

Eclipse Public License - v 1.0

- junit (not bundled)
- logback (binary bundling only)

MIT License (Expat)

- slf4j

Cryptography

none

Required Resources

Mailing Lists

- private at fluo.incubator.apache.org
- dev at fluo.incubator.apache.org
- notifications at fluo.incubator.apache.org

Git Repository

- https://git-wip-us.apache.org/repos/asf/incubator-fluo.git
  (The developers will use a git-based site for project documentation in the *asf-site* branch of the repo.)
- https://git-wip-us.apache.org/repos/asf/incubator-fluo-recipes.git

Issue Tracking

- https://issues.apache.org/jira/browse/FLUO
  (Currently, the developers rely on GitHub issues. If possible, GitHub integration for issue tracking would be preferred. If this is possible, the Fluo developers could work with INFRA to transfer the existing GitHub repositories to the Apache GitHub organization to bring the existing GitHub issues.)

Continuous Integration

- Travis CI on the GitHub mirror is fine (flag set to build only if *.travis.yml* file is present)

Initial Committers

- Keith Turner (kturner at apache dot org)
- Mike Walch (mike.walch at ptech-llc dot com)
- Corey Nolet (cjnolet at apache dot org)
- Christopher Tubbs (ctubbsii at apache dot org)
- Josh Elser (elserj at apache dot org)

Affiliations

- Keith Turner (Peterson Technologies, ASF Member, Accumulo PMC, Gora PMC)
- Mike Walch (Peterson Technologies)
- Corey Nolet (Tetra Concepts LLC, Accumulo PMC)
- Christopher Tubbs (U.S. Government, ASF Member, Accumulo PMC)
- Josh Elser (Hortonworks, ASF Member, Accumulo PMC, Calcite PMC, IPMC)

Sponsors

Champion

- Billie Rinaldi (billie at apache dot org) has volunteered to be our Champion

Nominated Mentors

- Drew Farris (drew at apache dot org)
- Josh Elser (elserj at apache dot org)

- Billie Rinaldi (billie at apache dot org)

Sponsoring Entity

- The Fluo team requests sponsorship from the Incubator PMC

1. USENIX (2010), http://research.google.com/pubs/pub36726.html