StormProposal

Storm Proposal

Abstract

Storm is a distributed, fault-tolerant, and high-performance realtime computation system that provides strong guarantees on the processing of data.

Proposal

Storm is a distributed real-time computation system. Similar to how Hadoop provides a set of general primitives for doing batch processing, Storm provides a set of general primitives for doing real-time computation. Its use cases span stream processing, distributed RPC, continuous computation, and more. Storm has become a preferred technology for near-realtime big-data processing by many organizations worldwide (see a partial list at https://github.com/nathanmarz/storm/wiki/Powered-By). As an open source project, Storm's developer community has grown rapidly to 46 members.

Background

The past decade has seen a revolution in data processing. MapReduce, Hadoop, and related technologies have made it possible to store and process data at scales previously unthinkable. Unfortunately, these data processing technologies are not realtime systems, nor are they meant to be. The lack of a "Hadoop of realtime" has become the biggest hole in the data processing ecosystem. Storm fills that hole.

Storm was initially developed and deployed at BackType in 2011. After 7 months of development BackType was acquired by Twitter in July 2011. Storm was open sourced in September 2011.

Storm has been under continuous development on its Github repository since being open-sourced. It has undergone four major releases (0.5, 0.6, 0.7, 0.8) and many minor ones.

Rationale

Storm is a general platform for low-latency big-data processing. It is complementary to the existing Apache projects, such as Hadoop. Many applications are actually exploring using both Hadoop and Storm for big-data processing. Bringing Storm into Apache is very beneficial to both Apache community and Storm community.

The rapid growth of Storm community is empowered by open source. We believe the Apache foundation is a great fit as the long-term home for Storm, as it provides an established process for community-driven development and decision making by consensus. This is exactly the model we want for future Storm development.

Initial Goals

- Move the existing codebase to Apache
- Integrate with the Apache development process
- Ensure all dependencies are compliant with Apache License version 2.0
- Incremental development and releases per Apache guidelines

Current Status

Storm has undergone four major releases (0.5, 0.6, 0.7, 0.8) and many minor ones. Storm 0.9 is about to be released. Storm is being used in production by over 50 organizations. Storm codebase is currently hosted at github.com, which will seed the Apache git repository.

Meritocracy

We plan to invest in supporting a meritocracy. We will discuss the requirements in an open forum. Several companies have already expressed interest in this project, and we intend to invite additional developers to participate. We will encourage and monitor community participation so that privileges can be extended to those that contribute.

Community

The need for a low-latency big-data processing platform in the open source is tremendous. Storm is currently being used by at least 50 organizations worldwide (see https://github.com/nathanmarz/storm/wiki/Powered-By), and is the most starred Java project on Github. By bringing Storm into Apache, we believe that the community will grow even bigger.

Core Developers

Storm was started by Nathan Marz at BackType, and now has developers from Yahoo!, Microsoft, Alibaba, Infochimps, and many other companies.

Alignment

In the big-data processing ecosystem, Storm is a very popular low-latency platform, while Hadoop is the primary platform for batch processing. We believe that it will help the further growth of big-data community by having Hadoop and Storm aligned within Apache foundation. The alignment is also beneficial to other Apache communities (such as Zookeeper, Thrift, Mesos). We could include additional sub-projects, Storm-on-YARN and Storm-on-Mesos, in the near future.

Known Risks

Orphaned Products

The risk of the Storm project being abandoned is minimal. There are at least 50 organizations (Twitter, Yahoo!, Microsoft, Groupon, Baidu, Alibaba, Alipay, Taobao, PARC, RocketFuel etc) are highly incentivized to continue development. Many of these organizations have built critical business applications upon Storm, and have devoted significant internal infrastructure investment in Storm.

Inexperience with Open Source

Storm has existed as a healthy open source project for several years. During that time, we have curated an open-source community successfully, attracting over 40 developers from a diverse group of companies including Twitter, Yahoo!, and Alibaba.

Homogenous Developers

The initial committers are employed by large companies (including Twitter, Yahoo!, Alibaba, Microsoft) and well-funded startups. Storm has an active community of developers, and we are committed to recruiting additional committers based on their contributions to the project.

Reliance on Salaried Developers

It is expected that Storm development will occur on both salaried time and on volunteer time, after hours. The majority of initial committers are paid by their employer to contribute to this project. However, they are all passionate about the project, and we are confident that the project will continue even if no salaried developers contribute to the project. We are committed to recruiting additional committers including non-salaried developers.

Relationships with Other Apache Products

As mentioned in the Alignment section, Storm is closely integrated with Hadoop, Zookeeper, Thrift, YARN and Mesos in a numerous ways. We look forward to collaborating with those communities, as well as other Apache communities (including Apache S4 which focuses on stateful low-latency processing).

An Excessive Fascination with the Apache Brand

Storm is already a healthy and 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.

Documentation

The reader will find these websites highly relevant:

- Storm website: http://storm-project.net
- Storm documentation: https://github.com/nathanmarz/storm/wiki
- Codebase: https://github.com/nathanmarz/storm
- User group: https://groups.google.com/group/storm-user

Source and Intellectual Property Submission Plan

The Storm codebase is currently hosted on Github: https://github.com/nathanmarz/storm. This is the exact codebase that we would migrate to the Apache foundation.

The Storm source code is currently licensed under Eclipse Public License Version 1.0. Some source code was contributed under a contributor agreement based on the Sun contributor agreement (v1.5). More recent code has been contributed under an Apache style agreement (see https://dl.dropboxusercontent.com/u/133901206/storm-apache-style-cla.txt).

Upon entering Apache, Storm will migrate to an Apache License 2.0 with all contributions licensed 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.

storm-kafka, which lets one use Kafka as a source for Storm, will also be submitted under the contrib folder for the Apache Storm project.

Yahoo! is also willing to move Storm-on-YARN code from github to be a subproject of Apache Storm project. Storm-on-YARN is currently licensed under Apache License 2.0 and receive contribution under Apache style CLA. Upon entering Apache, Yahoo! will sign over copyright to Apache foundation.

External Dependencies

To the best of our knowledge, all of Storm dependencies (except 0MQ/JMQ) 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).

Storm has used 0MQ and JMQ as the default mechanism for internal messaging layer, and 0MQ/JMQ is licensed under GNU Lesser General Public License. Recently, we have made Storm messaging layer pluggable, and plan to use Netty (which is licensed under Apache License v2) as our default messaging plugin (while keep 0MQ as an optional plugin).

Cryptography

We do not expect Storm to be a controlled export item due to the use of encryption. Storm enable encryptions via 2 plugins:

- SASL authentication plugins … Currently, we have provide “no-op” authentication and digest authentication. In near future, we will introduce Kerberos authentication.
- Tuple payload serialization plugins … Storm provides plugins for plain-object serialization and blowfish encryption.

Required Resources

Mailing lists

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

Subversion Directory

Git is the preferred source control system: git://git.apache.org/storm

Issue Tracking

JIRA Storm (STORM)

Initial Committers

- Nathan Marz <nathan at nathanmarz dot com>
- James Xu <xumingmingv at gmail dot com>
- Jason Jackson <jason at cvk dot ca>
- Andy Feng <afeng at yahoo-inc dot com>
- Flip Kromer <flip at infochimps dot com>
- David Lao <davidlao at microsoft dot com>
- P. Taylor Goetz <ptgoetz at gmail dot com>

Affiliations

- Nathan Marz - Nathan’s Startup
- James Xu - Alibaba
- Jason Jackson - Twitter
- Andy Feng - Yahoo!
- Flip Kromer - Infochimps
- David Lao - Microsoft
- P. Taylor Goetz - Health Market Science

Sponsors

Champion

- Doug Cutting <cutting at apache dot org>

Nominated Mentors

- Ted Dunning <tdunning at maprtech dot com>
- Arvind Prabhakar <arvind at apache dot org>
- Devaraj Das <ddas at hortonworks dot com>
- Matt Franklin <m.ben.franklin at gmail dot com>
- Benjamin Hindman <benjamin.hindman at gmail dot com>

Sponsoring Entity

The Apache Incubator