DrElephantProposal

ABSTRACT

Dr. Elephant is a performance monitoring and tuning service for Apache Hadoop and Apache Spark jobs and workflows. While the system is primarily aimed at developers, we have discovered that it is also popular with cluster operators who use it to monitor the health of workloads running on their clusters.

PROPOSAL

Dr. Elephant was open sourced by LinkedIn in 2016 and is currently hosted on GitHub. We believe that being a part of the Apache Software Foundation will improve the diversity and help form a strong community around the project.

LinkedIn submits this proposal to donate the code base to the Apache Software Foundation. The code is already under Apache License 2.0. Both the source code and documentation are hosted on Github.

- Code: http://github.com/linkedin/dr-elephant
- Documentation: https://github.com/linkedin/dr-elephant/wiki

Background

Dr. Elephant is a service that helps users of Apache Hadoop and Apache Spark understand, analyze, and improve the performance of jobs and workflows running on their clusters. It automatically gathers metrics, performs analysis, and presents the results along with actionable advice. The goal of the project is to improve developer productivity and increase cluster efficiency by reducing the time and domain expertise required to diagnose and treat sick jobs. It analyzes Hadoop and Spark jobs using a set of configurable, extensible, rule-based heuristics that provide insights on job performance, and then uses this information to provide recommendations about how to tune jobs to make them run more efficiently.

Dr. Elephant was open sourced in 2016 after two years of successful production use at Linkedin. In the time since many new features have been added including support for the Oozie and Airflow workflow schedulers, improved metrics, and enhancements to the Spark history fetcher and Spark heuristics. It is also important to note that many of these contributions came from developers outside of LinkedIn. We have also been happy to see that many people have been able to benefit from running Dr. Elephant including companies like Airbnb, Foursquare, Hulu, and Pinterest.

RATIONALE

Dr. Elephant's entry to the ASF will be beneficial to both the Dr. Elephant and Apache communities. Dr. Elephant has greatly benefited from its open source roots. Its community and adoption has grown greatly as a result. More importantly, the feedback from the community whether through interactions at meetups or through the mailing list have allowed for a rich exchange of ideas. We believe a partnership with the Apache Foundation is the logical next step. The Dr. Elephant community will greatly benefit from the established development and consensus processes that have worked well for other projects. The Apache process has served many other open source projects well and we believe that the Dr. Elephant community will greatly benefit from these practices as well.

CURRENT STATUS

Dr. Elephant is currently open sourced under the Apache License Version 2.0 and is available at github.com/linkedin/dr-elephant. All of the development is done using GitHub Pull Requests.

We are aware of at least 10 organizations that are running Dr. Elephant, and many of these organizations have also contributed code. Dr. Elephant has also been integrated into commercial products such as Pepperdata's Application Profiler.

INITIAL GOALS

Our initial goals are as follows:

- Migrate the existing codebase to Apache
- Study and integrate with the Apache development process
- Ensure all dependencies are compliant with Apache License version 2.0
- Incremental development and releases per Apache guidelines
- Diversify the set of core developers and committers

MERITOCRACY

Following the Apache meritocracy model, we intend to build an open and diverse community around Dr. Elephant. We will encourage the community to contribute to discussions and the codebase.

COMMUNITY

The need for a simple and understandable performance monitoring and tuning service for Hadoop and Spark is tremendous. Dr. Elephant is currently being used by at least 10 organizations worldwide (some examples are listed here). We hope to extend the contributor base significantly by bringing Dr. Elephant into Apache.

CORE DEVELOPERS

Dr. Elephant was started by engineers at LinkedIn. Many other individuals and organizations have contributed to the project, and this diversity is reflected in the list of initial committers.

ALIGNMENT

Apache is the most natural home for Dr. Elephant because of its close relationship to Apache Hadoop and Apache Spark, and its integration with Apache Oozie and Apache Airflow (incubating).

KNOWN RISKS

Orphaned products

The risk of the Dr. Elephant project being abandoned is minimal. As noted earlier, there are many organizations that have benefitted from Dr. Elephant, and which are thus incentivized to continue development. In addition, the software vendor PepperData has integrated Dr. Elephant into their Application Profiler product.

Inexperience with Open Source

Dr. Elephant has existed as a healthy open source project since 2016. Any risks that we foresee are ones associated with scaling our open source communication and operation process rather than with inherent inexperience in operating as an open source project.

Homogenous Developers

Apart from Linkedin's developers, Dr. Elephant has developers from Airbnb, Pepperdata, Flipkart, Hulu, Foursquare, Altiscale, PayPal, Evariant, Didi, Trivago, and Cardlytics.

A lot of effort has been put for efficient communication between all the developers. We have set up different forums for communication like github issues, google groups mailing list, gitter chat, weekly hangouts, and frequent meetups.

Reliance on Salaried Developers

It is expected that Dr. Elephant development will occur on both salaried time and on volunteer time, after hours. Many of the 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.

A Excessive Fascination with the Apache Brand

While we respect the reputation of the Apache brand and have no doubts that it will attract contributors and users, we believe the ASF is the right home for Dr. Elephant to foster a great community that will lead to a better outcome in the long term.

Documentation

Dr Elephant's developer wiki: https://github.com/linkedin/dr-elephant/wiki

Initial Source

Dr Elephant's initial source contribution will come from https://github.com/linkedin/dr-elephant

The code is licensed under the Apache License V2.

Source and Intellectual Property Submission Plan

The Dr. Elephant codebase is currently hosted on Github. This is the exact codebase that we would migrate to the Apache Software Foundation. The Dr. Elephant source code is already licensed under Apache License Version 2.0. Going forward, we will continue to have all the contributions licensed directly to the Apache Software Foundation through our signed Individual Contributor License Agreements for all of the committers on the project.

External Dependencies

To the best of our knowledge all of Dr. Elephant's dependencies are distributed under Apache Software Foundation compatible licenses. Upon acceptance to the incubator, we will begin a thorough analysis of all transitive dependencies to verify this fact and introduce license checking into the build and release process.

Cryptography

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

Required Resources

Mailing lists

- private@drelephant.incubator.apache.org (moderated subscriptions)
- commits@drelephant.incubator.apache.org
- dev@drelephant.incubator.apache.org
- issues@drelephant.incubator.apache.org
- user@drelephant.incubator.apache.org

Git Repository

Git is the preferred source control system:
git://git.apache.org/dr-elephant

Issue Tracking

JIRA project DOCTOR

Other Resources

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

Initial Committers

- Akshay Rai <akshayrai09 at gmail dot com>
- Anant Nag <nntnag17 at gmail dot com>
- Chetna Chaudhari <chetnachaudhari at gmail dot com>
- Clemens Valiente <clemens dot valiente at gmail dot com>
- Fangshi Li <shengzhixia at gmail dot com>
- George Wu <georgieewuu at gmail dot com>
- Krishna Puttaswamy <krishnaprasad dot pn at gmail dot com>
- Maxime Kestemont <maxkestemont at hotmail dot com>
- Noam Shaish <noamshaish at gmail dot com>
- Paul Reed Bramsen <prb at paulbramsen dot com>
- Ragesh K R <ragesh dot rajagopalan at gmail dot com>
- Shankar Manian <shankar37 at gmail dot com>
- Shahrukh Khan <shahrukhkhan489 at gmail dot com>
- Shekhar Gupta <shkhrgptat gmail dot com>
- Shida Li <lishid at gmail dot com>

Affiliations

- Akshay Rai - Linkedin
- Anant Nag - Linkedin
- Chetna Chaudhari - SkyTv New Zealand
- Clemens Valiente - trivago GmbH
- Fangshi Li - Linkedin
- George Wu - Pinterest

- Krishna Puttaswamy - Airbnb
- Mark Wagner - Linkedin
- Maxime Kestemont - Criteo
- Noam Shaish - Nordea Bank
- Ragesh K R - Linkedin
- Shankar Manian - Linkedin
- Shahrukh Khan - Hortonworks
- Shekhar Gupta - Pepperdata
- Shida Li - Dynalist Inc.

Sponsors

Champion

- Carl Steinbach

Nominated Mentors

- Carl Steinbach (Linkedin)
- Timothy Chen (HyperPilot)

Sponsoring Entity

The Apache Incubator