ImpalaProposal

Abstract

Impala is a high-performance C++ and Java SQL query engine for data stored in Apache Hadoop-based clusters.

Proposal

We propose to contribute the Impala codebase and associated artifacts (e.g. documentation, web-site content etc.) to the Apache Software Foundation with the intent of forming a productive, meritocratic and open community around Impala's continued development, according to the 'Apache Way'.

Cloudera owns several trademarks regarding Impala, and proposes to transfer ownership of those trademarks in full to the ASF.

Background

Engineers at Cloudera developed Impala and released it as an Apache-licensed open-source project in Fall 2012. Impala was written as a brand-new, modern C++ SQL engine targeted from the start for data stored in Apache Hadoop clusters.

Impala's most important benefit to users is high-performance, making it extremely appropriate for common enterprise analytic and business intelligence workloads. This is achieved by a number of software techniques, including: native support for data stored in HDFS and related filesystems, just-in-time compilation and optimization of individual query plans, high-performance C++ codebase and massively-parallel distributed architecture. In benchmarks, Impala is routinely amongst the very highest performing SQL query engines.

Rationale

Despite the exciting innovation in the so-called 'big-data' space, SQL remains by far the most common interface for interacting with data in both traditional warehouses and modern 'big-data' clusters. There is clearly a need, as evidenced by the eager adoption of Impala and other SQL engines in enterprise contexts, for a query engine that offers the familiar SQL interface, but that has been specifically designed to operate in massive, distributed clusters rather than in traditional, fixed-hardware, warehouse-specific deployments. Impala is one such query engine.

We believe that the ASF is the right venue to foster an open-source community around Impala's development. We expect that Impala will benefit from more productive collaboration with related Apache projects, and under the auspices of the ASF will attract talented contributors who will push Impala's development forward at pace.

We believe that the timing is right for Impala's development to move wholesale to the ASF: Impala is well-established, has been Apache-licensed open-source for more than three years, and the core project is relatively stable. We are excited to see where an ASF-based community can take Impala from this strong starting point.

Initial Goals

Our initial goals are as follows:

- Establish ASF-compatible engineering practices and workflows
- Refactor and publish existing internal build scripts and test infrastructure, in order to make them usable by any community member.
- Transfer source code, documentation and associated artifacts to the ASF.
- Grow the user and developer communities

Current Status

Impala is developed as an Apache-licensed open-source project. The source code is available at http://github.com/cloudera/Impala, and developer documentation is at https://github.com/cloudera/Impala/wiki. The majority of commits to the project have come from Cloudera-employed developers, but we have accepted some contributions from individuals from other organizations.

All code reviews are done via a public instance of the Gerrit review tool at http://gerrit.cloudera.org:8080/, and discussed on a public mailing list. All patches must be reviewed before they are accepted into the codebase, via a voting mechanism that is similar to that used on Apache projects such as Hadoop and HBase.

Before a patch is committed, it must pass a suite of pre-commit tests. These tests are currently run on Cloudera's internal infrastructure. One of our initial goals will be to work with the ASF Infrastructure team to find a way to run these tests in an acceptable way on publicly accessible machines.

Issues are tracked in JIRA at https://issues.cloudera.org/projects/IMPALA, in a way that is extremely similar to existing practices at other ASF projects.

Meritocracy

We understand the central importance of meritocracy to the Apache Way. We will work to establish a welcoming, fair and meritocratic community, in part by expanding the set of committers on the project. Although Impala's committer list will initially be dominated by members of the Impala engineering team at Cloudera, we look forward to growing a rich user and developer community.

Community

Impala has a strong user community (see https://groups.google.com/a/cloudera.org/forum/#!forum/impala-user), and a growing developer community (see https://groups.google.com/a/cloudera.org/forum/#!forum/impala-dev). We wish to attract more developers to the project, and we believe that the ASF's open and meritocratic philosophy will help us with this. We note the success of other, similar projects already part of the ASF.

Core Developers

Most - but not all - of Impala's core developers are not currently affiliated with the ASF, and will require new ICLAs.

Alignment

Impala is related to several other Apache projects:

- Data that is read by Impala is very often stored in Apache Hadoop clusters powered by the HDFS filesystem.
- Impala can also read data stored in Apache HBase
- Metadata for databases, tables and so on is read by Impala from Apache Hive.
- The preferred data format for HDFS-based tables is Apache Parquet, and Apache Avro is also a supported data format.
- Impala is closely integrated with Kudu, which is also being proposed to the Incubator.
- Impala uses Apache Thrift as its RPC and serialization framework of choice.

Known Risks

Orphaned Products

Impala is used by most of Cloudera's customers, and Cloudera remains committed to developing and supporting the project. Cloudera has a strong track record in standing behind projects that were contributed to the ASF by its employees, including Apache Flume, Apache Sqoop, and others. Other companies both ship and support Impala, lending credence to the idea that Impala is not at risk of being suddenly orphaned.

Inexperience with Open Source

Although all committers on the initial list have significant experience with at least one open-source project - namely Impala - fewer have much experience with ASF-based software projects as contributors and community members. However, with the guidance of our mentors, committers who do have ASF experience, and time to learn during Incubation, we are confident that the project can be run in accordance with Apache principles on an ongoing basis.

Homogeneous Developers

The initial committers are employees of Cloudera.

The project has received some contributions from developers outside of Cloudera, from individuals belonging to organizations such as Intel and Google, from hobbyists and from students using Impala to advance their understanding of distributed databases. The project attracted an active user community as well. We hope to continue to encourage contributions from these developers and community members and grow them into committers after they have had time to continue their contributions.

Reliance on Salaried Developers

Many of Impala's initial set of committers work full-time on Impala, and are paid to do so. However, as mentioned elsewhere, we anticipate growth in the developer community which we hope will include hobbyists and academics who have an interested in distributed data systems.

An Excessive Fascination with the Apache Brand

Although we hope that Impala benefits from the Apache Brand, any reflected goodwill to Cloudera as the contributing entity is not the goal of establishing Impala as an Apache project. We will work with the Incubator PMC and the PRC to ensure that the Apache Brand is respected.

Documentation

Impala: A Modern, Open-Source SQL Engine for Hadoop (http://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper28.pdf)

Impala's developer wiki (https://github.com/cloudera/Impala/wiki)

Impala's auto-generated API documentation (http://impala.io/doc/html/index.html)

Initial Source

Impala's initial source contribution will come from http://github.com/cloudera/Impala/.

External Dependencies

Impala depends upon a number of third-party libraries, which we list below. We intend to compile a LICENSE.txt file in the very short term (see https://issues.cloudera.org/browse/IMPALA-2670).

- Google gflags (BSD)
- Google glog (BSD)
- Apache Thrift (Apache Software License v2.0)
- Apache Commons (Apache Software License v2.0)
- Apache Hadoop (Apache Software License v2.0)
- Apache HBase (Apache Software License v2.0)
- Apache Hive (Apache Software License v2.0)
- Boost (Boost Software License)
- OpenLdap (OpenLDAP Software License)
- rapidjson (MIT)
- Google RE2 (BSD-style)
- lz4 (BSD)
- snappy (BSD)
- cyrus-sasl (CMU License)
- Apache Avro (Apache Software License v2.0)
- Cloudera squeasel (Apache Software License v2.0)
- Apache htrace (Incubating) (Apache Software License v2.0)
- Apache Sentry (Incubating) (Apache Software License v2.0)
- Apache Shiro (Apache Software License v2.0)
- Twitter Bootstrap (Apache Software License v2.0)
- d3 (BSD)
- LLVM (BSD-like)

Build and test dependencies:

- ant (Apache Software License v2.0)
- Apache Maven (Apache Software License v2.0)
- cmake (BSD)
- clang (BSD)
- Google gtest (Apache Software License v2.0)

Required Resources

We request that following resources be created for the project to use:

Mailing lists

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

Git repository

https://git.apache.org/impala.git

JIRA instance

JIRA project IMPALA (IMPALA or IMP)

Other Resources

We hope to continue using Gerrit for our code review and commit workflow. We are involved with discussions that the Kudu team at Cloudera have been having with Jake Farrell to start discussions on how Gerrit can fit into the ASF. We know that several other ASF projects or podlings are also interested in Gerrit.

If the Infrastructure team does not have the bandwidth to support gerrit, we will continue to support our own instance of gerrit for Impala, and make the necessary integrations such that commits are properly authenticated and maintain sufficient provenance to uphold the ASF standards (e.g. via the solution adopted by the AsterixDB podling).

Initial Committers

- Tim Armstrong
- Alex Behm
- Taras Bobrovytsky
- Casey Ching
- Martin Grund
- Daniel Hecht
- Michael Ho
- Matthew Jacobs
- Ishaan Joshi
- Lenni Kuff
- Marcel Kornacker
- Sailesh Mukil
- Henry Robinson
- John Russell
- Dimitris Tsirogiannis
- Skye Wanderman-Milne
- Juan Yu

Affiliations

All: Cloudera Inc.

Sponsors

Champion

Tom White

Nominated Mentors

- Tom White (Cloudera)
- Todd Lipcon (Cloudera)
- Carl Steinbach (LinkedIn)
- Brock Noland (StreamSets)

Sponsoring Entity

We ask that the Incubator PMC sponsor this proposal.