SentryProposal

Sentry - A fine-grained Authorization System for the Hadoop ecosystem

Abstract

Sentry is a highly modular system for providing fine grained role based authorization to both data and metadata stored on an Apache Hadoop cluster. Sentry can be used to enforce various access policy rules when accessing data stored on Hadoop Distributed File System through various Hadoop ecosystem components such as Apache Hive, Apache Pig or others.

Proposal

Traditionally, user access control in Apache Hadoop has been implemented using file based permissions on HDFS. Following the UNIX permissions model, HDFS offers all or nothing semantics allowing administrator to configure system to allow certain users or user groups read, write or perform both operations on files. This system does not enable more fine grained permissions that allow access policies for logical parts within one file. Furthermore, this model can't be used to restrict access to the rich set of objects in the metadata catalog that are stored outside HDFS.

Sentry will provide true role-based fine-grained user access control for Apache Hadoop and its ecosystem components such as Hive, Pig or HBase. This includes providing fine- grained role based access to both data as well as the metadata, which provides a rich object based abstraction such as databases, tables or columns.

Background

Sentry was initially developed by Cloudera to allow users fine grained access to data as well as the metadata in Apache Hadoop.

Sentry has been maintained as an open source project on Cloudera’s github. Sentry was previously called “Access”. All code in Sentry is open source and has been made publicly available under the Apache 2 license. During this time, Sentry has been formally released two times as versions 1.0.0 and 1.1.0.

Rationale

Currently, users don't have a way to achieve fine grained enforceable user access control to data stored in HDFS and their associated metadata. While users can use file based permissions to control access to specific directories and files, it is insufficient because access can't be restricted to file parts i.e., to specific lines or logical columns. In the absence of such support, users have to resort to duplicating data. Furthermore, file based permissions are insufficient to provide any form of access control to the metadata that provides an object abstraction such as databases, tables, columns or partitions over the data stored in HDFS.

Current Sentry developers subscribe to the mission of ASF and are familiar with the open source development process. Several members are already committers and PMC members of various other Apache projects.

Initial Goals

Sentry is currently in its first major release with a considerable number of enhancement requests, tasks, and issues recorded towards its future development. The initial goal of this project will be to continue to build community in the spirit of the "Apache Way", and to address the highly requested features and bug-fixes towards the next dot release.

Current Status

Meritocracy

Intent of the proposal is to build a diverse community of developers around Sentry. Sentry started as a open source project on Github, driven in the spirit of open source and we would like to continue in this spirit by, for example, encouraging contributors from a variety of organizations.

Community

Sentry stakeholders desire to expand the user and developer base of Sentry further in the future. The current sets of developers in Sentry are committed to building a strong user base and open source community around the project. Development discussions within the current team have been on a public mailing list.

Core Developers

The core developers for the Sentry project are Brock Noland, Shreepadma Venugopalan, Prasad Mujumdar and Jarek Jarcec Cecho. Other contributors include Arvind Prabhakar and Xuefu Zhang. All engineers have deep expertise in Hadoop and various other ecosystem components.

Alignment

Sentry complements the access control feature of some projects in the Apache Hadoop ecosystem, such as HDFS file permissions, by providing finer grained access control to data and metadata. It supersedes the access control capabilities of some other projects such as Apache Hive by providing stronger guarantees against malicious access. Currently, Sentry integrates with Apache Hive, however we are planning to provide support for other components such as Apache Pig.

While projects such as Apache Knox aim to provide perimeter security, the goal of Sentry is to implement a fine-grained role-based access control policy. Thus Sentry complements Apache Knox.

Known Risks

Orphaned Products

Sentry is already deployed in production at a few well established companies and they are actively sharing feature requests. The risks of it being orphaned is negligible.

Inexperience with Open Source

All committers of the Sentry project are intimately familiar with the Apache model for open-source development and are experienced with working with various Apache open -source communities.

Homogeneous Developers

The initial set of committers includes developers from several organizations - Cloudera, Oracle, Lab41, Nvidia and Wibidata. We expect that once approved for incubation, the project will further attract new contributors.

Reliance on Salaried Developers

It is expected that Sentry will be developed on both salaried and volunteer time, although all of the initial developers will work on it mainly on salaried time.

Relationships with Other Apache Products

Sentry depends on other Apache Projects: Apache Hadoop, Apache Log4J, Apache Hive, Apache Shiro, multiple Apache Commons components. Build is orchestrated by Apache Maven. Sentry complements Apache Knox.

An Excessive Fascination with the Apache Brand

We would like Sentry to become an Apache project to further foster a healthy community of users and developers around it. Since Sentry solves an important problem faced by Apache Hadoop users and interacts with other components of the Apache Hadoop ecosystem, we believe that Apache is the right home for Sentry.

Documentation

- Cloudera provides documentation specific to its distribution of Sentry at: http://www.cloudera.com/content/cloudera-content/cloudera-docs/Sentry/Sentry.pdf
- Sentry jira at Cloudera: https://issues.cloudera.org/browse/access

Initial Source

https://github.com/cloudera/access

Source and Intellectual Property Submission Plan

All of Sentry's code is under Apache 2 license already.

External Dependencies

All dependencies have licenses compatible with ASL. Dependencies that are not directly using ASL are,

- Junit - Eclipse Public License

Cryptography

Sentry currently doesn't directly use any cryptographic libraries. However, Sentry uses Apache Shiro, which provides support for cryptography features such as hash, cipher etc.

Required Resources

Mailing Lists

- private@sentry.incubator.apache.org for private PMC discussions (with moderated subscriptions)
- security@sentry.incubator.apache.org for private security related discussions
- dev@sentry.incubator.apache.org
- commits@sentry.incubator.apache.org

Source code repository

Git repository running at http://git-wip-us.apache.org/.

Issue Tracking

JIRA Sentry (SENTRY)

Other Resources

The existing code already has unit and integration tests so we would like a Jenkins CI instance that would run the tests on reference environment. We would also like to use Jenkins to run tests for every newly submitted patch (so called pre-commit hook), however this can be added after project creation.

Initial Committers

- Ali Rizvi (ali.rizvi at oracle.com)
- Arvind Prabhakar (arvind at apache.org)
- Brock Noland (brock at apache.org)
- Chaoyu Tang (ctang at cloudera.com)
- Daisy Zhou (daisy at wibidata.com)
- David Nalley (ke4qqq at apache.org)
- Erick Tryzelaar(etryzelaar at iqt.org)
- Greg Chanan (gchanan at apache.org)
- Hadi Nahari (hnahari at nvidia.com)
- Jarek Jarcec Cecho (jarcec at apache.org)
- Johnny Zhang (xiaoyuz at cloudera.com)
- Karthik Ramachandran (kramachandran at iqt.org)
- Mark Grover (mgrover at cloudera.com)
- Milo Polte (milo at wibidata.com)
- Lenni Kuff (lskuff at cloudera.com)
- Patrick Daly (daly at cloudera.com)
- Patrick Hunt (phunt at apache.org)
- Prasad Mujumdar (prasadm at apache.org)
- Raghu Mani (raghu.mani at oracle.com)
- Sean Mackrory (sean at cloudera.com)
- Shreepadma Venugopalan (shreepadma at cloudera.com)
- Sravya Tirukkovalur (sravya at cloudera.com)
- Tom White (tomwhite at apache.org)
- Xuefu Zhang (xuefu at apache.org)

Affiliations

- Ali Rizvi (Oracle)
- Arvind Prabhakar (Cloudera)
- Brock Noland (Cloudera)
- Chaoyu Tang (Cloudera)
- Daisy Zhou (Wibidata)
- David Nalley (Citrix)
- Erick Tryzelaar (Lab41)
- Greg Chanan (Cloudera)
- Hadi Nahari (Nvidia)
- Jarek Jarcec Cecho (Cloudera)
- Johnny Zhang (Cloudera)
- Karthik Ramachandran (Lab41)
- Mark Grover (Cloudera)
- Milo Polte (Wibidata)
- Lenni Kuff (Cloudera)
- Patrick Daly (Cloudera)
- Patrick Hunt (Cloudera)
- Prasad Mujumdar (Cloudera)
- Raghu Mani (Oracle)
- Sean Mackrory (Cloudera)
- Shreepadma Venugopalan (Cloudera)
- Sravya Tirukkovalur (Cloudera)

- Tom White (Cloudera)
- Xuefu Zhang (Cloudera)

Sponsors

Champion

- Arvind Prabhakar (Cloudera)

Nominated Mentors

- Arvind Prabhakar (Cloudera)
- David Nalley (Citrix)
- Joe Brockmeier (Citrix)
- Olivier Lamy (Ecetera)
- Patrick Hunt (Cloudera)
- Tom White (Cloudera)

Sponsoring Entity

We are requesting the Incubator to sponsor this project.