AmaterasuProposal

Apache Amaterasu

Abstract

Apache Amaterasu is a framework providing continuous deployment for Big Data pipelines.

It provides the following capabilities:

- **Continuous integration** tools to **package pipelines and run tests**.
- A repository to store those packaged applications: the **applications repository**.
- A repository to store the pipelines, and engine configuration (for instance, location of the Spark master, etc.): per environment - the **configuration repository**.
- A **dashboard** to monitor the pipelines.
- A **DSL and integration hooks** allowing third parties to easily integrate.

Proposal

Amaterasu is a simple and powerful framework to build and dispense pipelines. It aims to help data engineers and data scientists to compose, configure, test, package, deploy and execute data pipelines written using multiple tools, languages and frameworks. Amaterasu provides a standard repo structure to package big data pipelines, a YAML based Domain Specific Languages (DSL) for data engineers, data scientists and operations engineers to manage complex pipelines throughout their entire lifecycle (Dev, UAT, Prod, etc.).

Background

Amaterasu is a relatively new project that was created to deal with some of the issues that as Consultants, we have seen recurring at different client sites. Mainly the need to continuously deploy complex pipelines built in multiple tools and languages. Amaterasu started as a pet project and is currently being evaluated by a couple of organizations, supported by the contributors, on a personal time and voluntary bases.

Rational

As software engineers working on big data projects we have straggled for a long time to apply the same CI/CD practices that have become the standard in the software industry for the last few years. While some of them are possible, for example Apache Spark is easy to unit test. However large scale pipelines are more complex and often use data, which might be un-structured as integration point, which requires heavy integration tests.

To automate such tests and complex deployments, we have found the need to often handcraft scripts and use a mixture tools, so we have decided to finally build a tool we can apply in a general way and not on a project by project basis.

Another issue Amaterasu is trying to tackle is the Integrating between the work of software engineers, data scientists, and sometimes operations engineers. The approach Amaterasu takes to integrate between those three schools of thought it to provide a simple YAML based DSL that provides a simple way to integrate different pipeline written in the native tools for each task (R, Spark in different languages, etc.).

Initial Goals

Our initial goals are to bring Amaterasu into the ASF, transition internal engineering processes into the open, and foster a collaborative development model according to the "Apache Way".

In addition, we intend to continue the development of Amaterasu, add new features as well as integrate better with other frameworks, including:

- Apache Arrow
- Apache Hive
- Apache Drill
- Apache Beam
- Apache YARN
- Farther and more complete integration with Apache Spark

Other frameworks will be evaluated after those initial goals are reached.

Current Status

Amaterasu is preview state but provide a large set of features. We plan to stabilize and head to a first production ready release during the incubation process. The current license is already Apache 2.0.

Meritocracy

We intend to radically expand the initial developer and user community by running the project in accordance with the "Apache Way". Users and new contributors will be treated with respect and welcomed. By participating in the community and providing quality patches/support that move the project forward, they will earn merit. They also will be encouraged to provide non-code contributions (documentation, events, community management, etc.) and will gain merit for doing so. Those with a proven support and quality track record will be encouraged to become committers.

Community

As a relatively new project, Amaterasu has a small, but growing community. Amaterasu is an open project, not just with it’s source code but also with our discussions which are held openly in our slack https://shintoio.slack.com which contains channels for design, tech and future directions discussions.

If Amaterasu is accepted for incubation, the primary initial goal is to build a large and strong community. We are confident that Amaterasu can become a key project for big data operations, which hopefully will create a large community of users and developers.

Known Risks

Development has been sponsored mostly by a one company. For the project to fully transition to the Apache Way governance model, development must shift towards the meritocracy-centric model of growing a community of contributors balanced with the needs for extreme stability and core implementation coherency.

Orphaned products

We are fully committed on Amaterasu. A few organizations have expressed their interest in using Amaterasu.

Inexperience with Open Source

We have been developing and using open source software for a long time. Additionally, several ASF veterans have agreed to mentor the project and they are listed in this proposal. The project will rely on their guidance and collective wisdom to quickly transition the entire team of initial committers towards practicing the Apache Way.

Reliance on Salaried Developers

Most of the current contributors are employed in the Big Data space. While they might wander from their current employers, they are unlikely to venture far from their core expertises and thus will continue to be engaged with the project regardless of their current employers.

An Excessive Fascination with the Apache Brand

While we intend to leverage the Apache ‘branding’ when talking to other projects as testament of our project’s ‘neutrality’, we have no plans for making use of Apache brand in press releases nor posting billboards advertising acceptance of Amaterasu into Apache Incubator.

The main purpose in applying for Apache incubation is due to the fact that Amaterasu is built with integration already in mind for many tools which are Apache projects, and we see Amaterasu as an extension of these projects. We hope that by being an Apache project, we can integrate better, and collaborate more effectively with the relevant projects. As Amaterasu matures, we see mutual benefits for all involved.

Initial Source

https://github.com/shintoio/amaterasu

External Dependencies

All external dependencies are licensed under an Apache 2.0 license or Apache-compatible license. As we grow the Amaterasu community we will configure our build process to require and validate all contributions and dependencies are licensed under the Apache 2.0 license or are under an Apache-compatible license.

- Apache Spark
- Apache Hadoop
- Apache Maven (maven-core)
- Apache Commons
- Apache Log4j
- Apache Mesos
- Apache Zookeeper
- Apache Curator
- Scala
- Junit
- Py4j

Future versions are planned to integrate with:

- Apache YARN
- Apache Hive
- Apache Drill

Required Resources

Mailing lists

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

Git Repository

- https://git-wip-us.apache.org/repos/asf/incubator-amaterasu.git

Issue Tracking

- JIRA Project Amaterasu

Initial Committers

- Yaniv Rodenski
- Jean-Baptiste Onofré
- Eyal Ben Ivri
- Karel Alfonso
- Kirupagaran (Kirupa) Devarajan
- Nadav Har Tzvi

Affiliations

- Yaniv Rodenski - Shinto
- Jean-Baptiste Onofré - Talend
- Olivier Lamy - Webtide

Sponsors

Champion

- Jean-Baptiste Onofré

Mentors

- Jean-Baptiste Onofré
- Olivier Lamy
- Davor Bonaci

Sponsoring Entity

The Apache Incubator