PredictionIO

PredictionIO Proposal

Abstract

PredictionIO is an open source Machine Learning Server built on top of state-of-the-art open source stack, that enables developers to manage and deploy production-ready predictive services for various kinds of machine learning tasks.

Proposal

The PredictionIO platform consists of the following components:

- PredictionIO framework - provides the machine learning stack for building, evaluating and deploying engines with machine learning algorithms. It uses Apache Spark for processing.
- Event Server - the machine learning analytics layer for unifying events from multiple platforms. It can use Apache HBase or any JDBC backends as its data store.

The PredictionIO community also maintains a Template Gallery, a place to publish and download (free or proprietary) engine templates for different types of machine learning applications, and is a complemental part of the project. At this point we exclude the Template Gallery from the proposal, as it has a separate set of contributors and we're not familiar with an Apache approved mechanism to maintain such a gallery.

Background

PredictionIO was started with a mission to democratize and bring machine learning to the masses.

Machine learning has traditionally been a luxury for big companies like Google, Facebook, and Netflix. There are ML libraries and tools lying around the internet but the effort of putting them all together as a production-ready infrastructure is a very resource-intensive task that is remotely reachable by individuals or small businesses.

PredictionIO is a production-ready, full stack machine learning system that allows organizations of any scale to quickly deploy machine learning capabilities. It comes with official and community-contributed machine learning engine templates that are easy to customize.

Rationale

As usage and number of contributors to PredictionIO has grown bigger and more diverse, we have sought for an independent framework for the project to keep thriving. We believe the Apache foundation is a great fit. Joining Apache would ensure that tried and true processes and procedures are in place for the growing number of organizations interested in contributing to PredictionIO. PredictionIO is also a good fit for the Apache foundation. PredictionIO was built on top of several Apache projects (HBase, Spark, Hadoop). We are familiar with the Apache process and believe that the democratic and meritocratic nature of the foundation aligns with the project goals.

Initial Goals

The initial milestones will be to move the existing codebase to Apache and integrate with the Apache development process. Once this is accomplished, we plan for incremental development and releases that follow the Apache guidelines, as well as growing our developer and user communities.

Current Status

PredictionIO has undergone nine minor releases and many patches. PredictionIO is being used in production by Salesforce.com as well as many other organizations and apps. The PredictionIO codebase is currently hosted at GitHub, which will form the basis of the Apache git repository.

Meritocracy

We plan to invest in supporting a meritocracy. We will discuss the requirements in an open forum. 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

Acceptance into the Apache foundation would bolster the already strong user and developer community around PredictionIO. That community includes many contributors from various other companies, and an active mailing list composed of hundreds of users.

Core Developers

The core developers of our project are listed in our contributors and initial PPMC below. Though many are employed at Salesforce.com, there are also engineers from ActionML, and independent developers.

Alignment

The ASF is the natural choice to host the PredictionIO project as its goal is democratizing Machine Learning by making it more easily accessible to every user/developer. PredictionIO is built on top of several top level Apache projects as outlined above.

Known Risks

Orphaned products

PredictionIO has a solid and growing community. It is deployed on production environments by companies of all sizes to run various kinds of predictive engines.

In addition to the community contribution to PredictionIO framework, the community is also actively contributing new engines to the Template Gallery as well as SDKs and documentation for the project. Salesforce is committed to utilize and advance the PredictionIO code base and support its user community.

Inexperience with Open Source

PredictionIO has existed as a healthy open source project for almost two years and is the most starred Scala project on GitHub. All of the proposed committers have contributed to ASF and Linux Foundation open source projects. Several current committers on Apache projects and Apache Members are involved in this proposal and intend to provide mentorship.

Homogeneous Developers

The initial list of committers includes developers from several institutions, including Salesforce, ActionML, Channel4, USC as well as unaffiliated developers.

Reliance on Salaried Developers

Like most open source projects, PredictionIO receives substantial support from salaried developers. PredictionIO development is partially supported by Salesforce.com, but there are many contributors from various other companies, and an active mailing list composed of hundreds of users. We will continue our efforts to ensure stewardship of the project to be independent of salaried developers by meritocratically promoting those contributors to committers.

Relationships with Other Apache Product

PredictionIO relies heavily on top level apache projects such as Apache Spark, HBase and Hadoop. However it brings a distinguished functionality, rather than just an abstraction - Machine Learning in a plug-and-play fashion.

Compared to Apache Mahout, which focuses on the development of a wide variety of algorithms, PredictionIO offers a platform to manage the whole machine learning workflow, including data collection, data preparation, modeling, deployment and management of predictive services in production environments.

An Excessive Fascination with the Apache Brand

PredictionIO is already a widely 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

PredictionIO boasts rich and live documentation, included in the code repo (docs/manual directory), is built with Middleman, and publicly hosted at https://docs.prediction.io

Initial Source and Intellectual Property Submission Plan

Currently, the PredictionIO codebase is distributed under the Apache 2.0 License and hosted on GitHub: https://github.com/PredictionIO/PredictionIO

External Dependencies

PredictionIO has the following external dependencies:

- Apache Hadoop 2.4.0 (optional, required only if YARN and HDFS are needed)
- Apache Spark 1.3.0 for Hadoop 2.4
- Java SE Development Kit 8
- and one of the following sets:
    - PostgreSQL 9.1
      or
    - MySQL 5.1
      or
    - Apache HBase 0.98.6
    - Elasticsearch 1.4.0

Upon acceptance to the incubator, we would begin a thorough analysis of all transitive dependencies to verify this information and introduce license checking into the build and release process by integrating with Apache RAT.

Cryptography

PredictionIO does not include cryptographic code. We utilize standard JCE and JSSE APIs provided by the Java Runtime Environment.

Required Resources

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

Mailing lists

predictionio-private@incubator.apache.org (with moderated subscriptions)

predictionio-dev

predictionio-user

predictionio-commits

We will migrate the existing PredictionIO mailing lists.

Git repository

The PredictionIO team would like to use Git for source control, due to our current use of GitHub.

git://git.apache.org/incubator-predictionio

Documentation

https://predictionio.incubator.apache.org/docs/

JIRA instance

PredictionIO currently uses the GitHub issue tracking system associated with its repository: https://github.com/PredictionIO/PredictionIO/issues. We will migrate to Apache JIRA.

JIRA PREDICTIONIO https://issues.apache.org/jira/browse/PREDICTIONIO

Other Resources

- TravisCI for builds and test running.
- PredictionIO's documentation, included in the code repo (docs/manual directory), is built with Middleman and publicly hosted https://docs.prediction.io
- A blog to drive adoption and excitement at https://blog.prediction.io

Initial Committers

- Pat Ferrell
- Tamas Jambor
- Justin Yip
- Xusen Yin
- Lee Moon Soo
- Donald Szeto
- Kenneth Chan
- Tom Chan
- Simon Chan
- Marco Vivero
- Matthew Tovbin
- Yevgeny Khodorkovsky
- Felipe Oliveira
- Vitaly Gordon
- Alex Merritt

Affiliations

- Pat Ferrell - ActionML
- Tamas Jambor - Channel4
- Justin Yip - independent
- Xusen Yin - USC
- Lee Moon Soo - NFLabs
- Donald Szeto - Salesforce
- Kenneth Chan - Salesforce
- Tom Chan - Salesforce
- Simon Chan - Salesforce
- Marco Vivero - Salesforce
- Matthew Tovbin - Salesforce
- Yevgeny Khodorkovsky - Salesforce
- Felipe Oliveira - Salesforce
- Vitaly Gordon - Salesforce
- Alex Merritt - ActionML

Sponsors

Champion

Andrew Purtell <apurtell at apache dot org>

Nominated Mentors

- Andrew Purtell <apurtell at apache dot org>
- James Taylor <jtaylor at apache dot org>
- Lars Hofhansl <larsh at apache dot org>
- Suneel Marthi <smarthi at apache dot org>
- Xiangrui Meng <meng at apache dot org>
- Luciano Resende <lresende at apache dot org>

Sponsoring Entity

Apache Incubator PMC