DataLabProposal

DataLab Proposal

Abstract

DataLab is a platform for creating self-service, exploratory data science environments in the cloud using best-of-breed data science tools.

DataLab includes a self-service web console, used to create and manage exploratory environments. It allows teams to spin up analytical environments with just a single click of a mouse. Once established, the environment can be managed by an analytical team itself, leveraging simple and easy-to-use web-based interface.

Proposal

In order to work effectively, data scientists rely on a varying suite of analytics tools that are readily available. However, many of those tools are non-trivial to set up in terms of hardware provisioning, software installation, configuration, and deployment. Setting up a collaborative, multi-tenant development environment for data scientists consumes substantial IT and DevOps resources, as well as time. These factors often combine to hinder the agility and effectiveness of data science teams within an organization. Current solutions are largely closed source and/or proprietary, and committing to a given solution introduces the potential for vendor lock-in.

EPAM Systems developed Data in response to the lack of open source, permissibly licensed solutions to better enable data science workflows. The ALv2 was selected to encourage open development and user adoption. DataLab was open sourced on Dec 29, 2016 and is under active development with support from EPAM Systems.

We believe DataLab is a unique solution with no current open source equivalent. Our primary goals of incubation are to grow and diversify the DataLab community to ensure its long-term sustainability.

Rationale

DataLab is a platform that provides data scientists with the ability to self-provision, without IT support, exploratory and production environments with their preferred set of tools installed and pre-configured. Tool options include, but are not limited to:

- Apache Spark
- Apache Flink (planned)
- Apache Zeppelin
- Jupyter
- TensorFlow + Jupyter
- Deep Learning + Jupyter

DataLab leverages cloud computing providers for virtual hardware provisioning and currently supports the following:

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Compute Platform (GCP) (under development)

DataLab offers git-based collaboration tools for data scientists and developers and integrates with the following git service providers:

- GItHub
- GitLab
- BitBucket

Additionally, DataLab includes the option to configure the UnGit tool in an environment to facilitate collaboration. Finally, DataLab integrates closely with many security and SSO offerings, including:

- LDAP
- Microsoft Active Directory
- AWS Identity Access Management service

DataLab was designed from the ground up to be highly configurable, flexible, and extensible platform. We believe these qualities will encourage community growth by enabling contributors to easily add new integrations and extensions.

Initial Goals

The initial goal will be to move the existing codebase to Apache and integrate with the Apache development process and infrastructure. A primary goal of incubation will be to grow and diversify the DataLab PPMC. We are well aware that the project community is comprised of individuals from a single company. We aim to change that during incubation.

Current Status

As previously mentioned, DataLab is under active development at EPAM Systems, and is being used in a number of production deployments:

- [An investment company] is using DataLab as an AWS-based analytics platform for their data scientists to provide a convenient way to perform multi-tenant data analytics. This enables data scientists to easily provision work environments with integrated data sources based on Elasticsearch, Apache HBase, and Neo4j, and utilizing Apache Spark. This enabled a “one click”, self service option for users to provision an environment with the necessary tools and data.

- [An electronics manufacturing company] leverages DataLab for data quality, data exploration, and analytics. The company’s data scientists leverage DataLab to work with data sources that have been transferred to the cloud in order to find new insights on the data, and help the implementation team define requirements for data engineering. The main goal is to increase the utilization of various tools by decreasing time to deployment.

- [A retail company] is using DataLab as an image recognition framework, to enable automated restocking of inventory.

- [A travel company] is using DataLab to create recommendation engine that will allow end users to find more relevant accommodations faster and at a lower cost.

Meritocracy

We value meritocracy and we understand that it is the basis for an open community that encourages multiple companies and individuals to contribute and be invested in the project’s future. We will encourage and monitor participation and make sure to extend privileges and responsibilities to all contributors.

Community

DataLab is currently being used by developers at EPAM and a gowing number of customers are actively using it in production environments. By bringing DataLab to Apache we hope to broaden and diversity the user and developer community through open collaboration.

Core Developers

DataLab was initially developed at EPAM Systems and is under active development. We believe DataLab will be of interest to a broad range of users and devlopers and that incubating the project at the ASF will help us build a diverse, sustainable community.

Alignment

DataLab utilizes other Apache projects such as Apache Spark, Apache Toree (incubating), and Apache Zeppelin, along with a number of other Apache libraries. We anticipate integration with additional Apache projects as the DataLab community and interest in the project grows.

Known Risks

Orphaned products

EPAM Systems is committed to the future development of DataLab and understands that graduation to a TLP, while preferable, is not the only positive outcome of incubation.

Should the DataLab project be accepted by the Incubator, the prospective PPMC would be willing to agree to a target incubation period of 2 years or less, knowing that every Incubator project incurs a certain cost in terms of ASF infrastructure and volunteer time.

Inexperience with Open Source

Many DataLab contributors are already familiar with open source processes and several of them are committers on other Apache projects. We will be actively working with experienced Apache community members to improve our project.

Homogenous Developers

The initial committers of DataLab all come from EPAM Systems, though we are committed to recruiting and developing additional committers from a wide spectrum of industries and backgrounds.

Reliance on Salaried Developers

It is expected that DataLab development will occur on both salaried time and on volunteer time, after hours. All of the initial committers are paid by EPAM Systems to contribute to this project. However, they are all passionate about the project, and we are both confident and hopeful that the project will continue even if no salaried developers contribute to the project.

Relationships with Other Apache Products

As mentioned in the Rationale section, DataLab utilizes a number of existing Apache projects (Spark, Toree, Zeppelin, et. al.), and we expect that list to expand as the community grows and diversifies. Any Apache project in the big data, data science, and/or analytics space would be potentially relevant.

A Excessive Fascination with the Apache Brand

We are applying to the Incubator process because we think it is the next logical step for the DataLab project after open-sourcing the code. This proposal is not for the purpose of generating publicity. Rather, we want to make sure to create a very inclusive and meritocratic community, outside the umbrella of a single company. EPAM has a long history of contributing to Apache projects and the DataLab developers and contributors understand the implication of making it an Apache project.

Required Resources

Mailing lists

- dev@DataLab.incubator.apache.org
- commits@DataLab.incubator.apache.org
- private@DataLab.incubator.apache.org

Source control

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

Issue tracking

- JIRA DataLab: https://issues.apache.org/jira/projects/DataLab/

Documentation

- DataLab Website: https://datalab.apache.org/
- DataLab code base: https://github.com/apache/incubator-datalab
- DataLab Overview: https://github.com/apache/incubator-datalab/blob/master/README.md
- DataLab User Guide: https://github.com/apache/incubator-datalab/blob/master/USER_GUIDE.md

Initial Source

The DataLab codebase is currently hosted on Github: https://github.com/apache/incubator-datalab

Source and Intellectual Property Submission Plan

The DataLab source code in Github is currently licensed under Apache License v2.0 and the copyright is assigned to EPAM Systems. If DataLab becomes an Incubator project at the ASF, EPAM Systems will transfer the source code and trademark ownership to the Apache Software Foundation via a Software Grant Agreement.

External Dependencies

To the best of our knowledge, all of DataLab dependencies are distributed under Apache compatible licenses.

DataLab was designed to be highly extensible, and we expect and encourage the development of third-party extensions and plug-ins. We also understand that any such component, if it requires a dependency forbidden by Apache license policy, would not be eligible for inclusion in an Apache release, and would have to be hosted, supported, etc. outside of ASF infrastructure and labeled appropriately.

External dependencies licensed under Apache License 2.0:

MongoDB Java Driver - org.mongodb:mongo-java-driver (http://mongodb.github.io/mongo-java-driver/3.2/driver)

Dropwizard (https://github.com/dropwizard/dropwizard)

Dropwizard Template Config (https://github.com/tkrille/dropwizard-template-config)

Apache Directory Server (https://github.com/apache/directory-server)

Jackson (https://github.com/FasterXML/jackson)

AWS Java SDK (https://github.com/aws/aws-sdk-java)

Boto3 (https://github.com/boto/boto3)

External dependencies licensed under the MIT License:

angular2-app (https://www.npmjs.com/package/angular2-app)

angular2-seed (https://www.npmjs.com/package/angular2-seed)

angular2-seed-advanced (https://www.npmjs.org/package/angular2-seed-advanced)

angular2-seed-n3UX (https://www.npmjs.com/package/angular2-seed-n3UX)

http-status-enum (https://www.npmjs.com/package/http-status-enum) Mockito (https://github.com/mockito/mockito)

ng2-translate (https://www.npmjs.com/package/ng2-translate)

SLF4J (http://www.slf4j.org/)

External dependencies licensed under the CDDL License:

Jersey (https://github.com/jersey/jersey)

External dependencies licensed under the Python Software License Version 2:

jython (https://github.com/jythontools/jython)

ASF Projects:

Apache Spark, Apache Toree (incubating), Apache Zeppelin

Cryptography

Not applicable.

Initial Committers

- Dmytro Liaskovskyi dmytro_liaskovskyi@epam.com
- Volodymyr Veres Volodymyr_Veres@epam.com
- Oleh Hrynets Oleh_Hrynets@epam.com
- Oleh Martushevskyi Oleh_Martushevskyi@epam.com
- Oleh Moskovych Oleh_Moskovych@epam.com
- Vadym Kuznetsov Vadym_Kuznetsov@epam.com
- Bohdan Hliva Bohdan_Hliva@epam.com
- Vira Vitanska Vira_Vitanska@epam.com
- Andriana Kovalyshyn Andriana_Kovalyshyn@epam.com
- Oleksandr Chaparin Oleksandr_Chaparin@epam.com
- Denys Shliakhov Denys_Shliakhov@epam.com
- Nazar Barabash Nazar_Barabash@epam.com

Sponsors

Champion

- P. Taylor Goetz ptgoetz@apache.org

Nominated Mentors

- P. Taylor Goetz ptgoetz@apache.org
- Henry Saputra hsaputra@apache.org

Interested Contributors

- Debo Dutta ddutta@apache.org

Sponsoring Entity

- The Apache Incubator