DataFuProposal

Abstract

DataFu makes it easier to solve data problems using Hadoop and higher level languages based on it.

Proposal

Data{{`Fu provides a collection of Hadoop Map}}Reduce jobs and functions in higher level languages based on it to perform data analysis. It provides functions for common statistics tasks (e.g. quantiles, sampling), PageRank, stream sessionization, and set and bag operations. DataFu also provides Hadoop jobs for incremental data processing in Map`Reduce.

Background

Data{{`Fu began two years ago as set of UDFs developed internally at Linked}}In, coming from our desire to solve common problems with reusable components. Recognizing that the community could benefit from such a library, we added documentation, an extensive suite of unit tests, and open sourced the code. Since then there have been steady contributions to DataFu as we encountered common problems not yet solved by it. Others outside LinkedIn have contributed as well. More recently we recognized the challenges with efficient incremental processing of data in Hadoop and have contributed a set of Hadoop Map`Reduce jobs as a solution.

Data{{`Fu began as a project at Linked}}In, but it has shown itself to be useful to other organizations and developers as well as they have faced similar problems. We would like to share Data`Fu with the ASF and begin developing a community of developers and users within Apache.

Rationale

There is a strong need for well tested libraries that help developers solve common data problems in Hadoop and higher level languages such as Pig, Hive, Crunch, Scalding, etc.

Current Status

Meritocracy

Our intent with this incubator proposal is to start building a diverse developer community around Data{{`Fu following the Apache meritocracy model. Since Data}}Fu was initially open sourced in 2011, it has received contributions from both within and outside Linked`In. We plan to continue support for new contributors and work with those who contribute significantly to the project to make them committers.

Community

Data{{`Fu has been building a community of developers for two years. It began with contributors from Linked}}`In and has received contributions from developers at Cloudera since very early on. It has been included included in Cloudera's Hadoop Distribution and Apache Bigtop. We hope to extend our contributor base significantly and invite all those who are interested in solving large-scale data processing problems to participate.

Core Developers

Data{{`Fu has a strong base of developers at Linked}}In. Matthew Hayes initiated the project in 2011, and aside from continued contributions to DataFu has also contributed the sub-project Hourglass for incremental MapReduce processing. Separate from DataFu he has also open sourced the White Elephant project. Sam Shah contributed a significant portion of the original code and continues to contribute to the project. William Vaughan has been contributing regularly to DataFu for the past two years. Evion Kim has been contributing to DataFu for the past year. Xiangrui Meng recently contributed implementations of scalable sampling algorithms based on research from a paper he published. Chris Lloyd has provided some important bug fixes and unit tests. Mitul Tiwari has also contributed to DataFu. Mathieu Bastian has been developing MapReduce jobs that we hope to include in Data`Fu. In addition he also leads the open source Gephi project.

Alignment

The ASF is the natural choice to host the Data{{`Fu project as its goal of encouraging community-driven open-source projects fits with our vision for Data}}Fu. Additionally, other projects Data`Fu integrates with, such as Apache Pig and Apache Hadoop, and in the future Apache Hive and Apache Crunch, are hosted by the ASF and we will benefit and provide benefit by close proximity to them.

Known Risks

Orphaned Products

The core developers have been contributing to Data{{`Fu for the past two years. There is very little risk of Data}}`Fu being abandoned given its widespread use within Linked`In.

Inexperience with Open Source

Data{{`Fu was started as an open source project in 2011 and has remained so for two years. Matt initiated the project, and additionally is the creator of the open source White Elephant project. He has also contributed patches to Apache Pig. Most recently he has released Hourglass as a sub-project of Data}}`Fu. Sam contributed much of the original code and continues to contribute to the project. Will has been contributing to Data`Fu since it was first open sourced. Evion has been contributing for the past year. Mathieu leads the open source Gephi project. Jakob has been actively involved with the ASF as a full-time Hadoop committer and PMC member.

Homogeneous Developers

The current core developers are all from Linked{{`In. Data}}`Fu has also received contributions from other corporations such as Cloudera. Two of these developers are among the Initial Committers listed below. We hope to establish a developer community that includes contributors from several other corporations and we are actively encouraging new contributors via presentations and blog posts.

Reliance on Salaried Developers

The current core developers are salaried employees of Linked{{`In, however they are not paid specifically to work on Data}}`Fu. Contributions to DataFu arise from the developers solving problems they encounter in their various projects. The purpose of Data`Fu is to share these solutions so that others may benefit and build a community of developers striving to solve common problems together. Furthermore, once the project has a community built around it, we expect to get committers, developers and contributions from outside the current core developers.

Relationships with Other Apache Products

DataFu is deeply integrated with Apache products. It began as a library of user-defined functions for Apache Pig. It has grown to also include Hadoop jobs for incremental data processing and in the future will include code for other higher level languages built on top of Apache Hadoop.

An Excessive Obsession with the Apache Brand

While we respect the reputation of the Apache brand and have no doubts that it will attract contributors and users, our interest is primarily to give DataFu a solid home as an open source project following an established development model.

Documentation

Information on DataFu can be found at:

https://github.com/LinkedIn/DataFu/blob/master/README.md

Initial Source

The initial source is available at:

https://github.com/LinkedIn/DataFu

Source and Intellectual Property Submission Plan

- The Data{{`Fu library source code, available on Git}}`Hub.

External Dependencies

The initial source has the following external dependencies that are either included in the final DataFu library or required in order to use it:

- fastutil (Apache 2.0)
- joda-time (Apache 2.0)
- commons-math (Apache 2.0)
- guava (Apache 2.0)
- stream (Apache 2.0)
- jsr-305 (BSD)
- log4j (Apache 2.0)
- json (The JSON License)
- avro (Apache 2.0)

In addition, the following external libraries are used either in building, developing, or testing the project:

- pig (Apache 2.0)
- hadoop (Apache 2.0)
- jline (BSD)

- antlr (BSD)
- commons-io (Apache 2.0)
- testng (Apache 2.0)
- maven (Apache 2.0)
- jsr-311 (CDDL-1.0)
- slf4j (MIT)
- eclipse (Eclipse Public License 1.0)
- autojar (GPLv2)
- jarjar (Apache 2.0)

Cryptography

Data{{`Fu has user-defined functions that use MD5 and SHA provided by Java's java.security.Message}}`Digest.

Required Resources

Mailing Lists

Data{{`Fu-private for private PMC discussions (with moderated subscriptions) Data}}`Fu-dev Data`Fu-commits

Subversion Directory

Git is the preferred source control system: git://git.apache.org/DataFu

Issue Tracking

JIRA Data{{`Fu (Data}}`Fu)

Other Resources

The existing code already has unit tests, so we would like a Hudson instance to run them whenever a new patch is submitted. This can be added after project creation.

Initial Committers

- Matthew Hayes
- William Vaughan
- Evion Kim
- Sam Shah
- Xiangrui Meng
- Christopher Lloyd
- Mathieu Bastian
- Mitul Tiwari
- Josh Wills
- Jarek Jarcec Cecho

Affiliations

- Matthew Hayes (LinkedIn)
- William Vaughan (LinkedIn)
- Evion Kim (LinkedIn)
- Sam Shah (LinkedIn)
- Xiangrui Meng (LinkedIn)
- Christopher Lloyd (LinkedIn)
- Mathieu Bastian (LinkedIn)
- Mitul Tiwari (LinkedIn)
- Josh Wills (Cloudera)
- Jarek Jarcec Cecho (Cloudera)

Sponsors

Champion

Jakob Homan (Apache Member)

Nominated Mentors

- Ashutosh Chauhan <hashutosh at apache dot org>

- Roman Shaposhnik <rvs at apache dot org>
- Ted Dunning <tdunning at apache dot org>

Sponsoring Entity

We are requesting the Incubator to sponsor this project.