SparkProposal

Abstract

Spark is an open source system for large-scale data analysis on clusters.

Proposal

Spark is an open source system for fast and flexible large-scale data analysis. Spark provides a general purpose runtime that supports low-latency execution in several forms. These include interactive exploration of very large datasets, near real-time stream processing, and ad-hoc SQL analytics (through higher layer extensions). Spark interfaces with HDFS, HBase, Cassandra and several other storage storage layers, and exposes APIs in Scala, Java and Python. Background
Spark started as U.C. Berkeley research project, designed to efficiently run machine learning algorithms on large datasets. Over time, it has evolved into a general computing engine as outlined above. Spark's developer community has also grown to include additional institutions, such as universities, research labs, and corporations. Funding has been provided by various institutions including the U.S. National Science Foundation, DARPA, and a number of industry sponsors. See: https://amplab.cs.berkeley.edu/sponsors/ for full details.

Rationale

As the number of contributors to Spark has grown, we have sought for a long-term home for the project, and we believe the Apache foundation would be a great fit. Spark is a natural fit for the Apache foundation: Spark already interoperates with several existing Apache projects (HDFS, HBase, Hive, Cassandra, Avro and Flume to name a few). The Spark team is familiar with the Apache process and and subscribes to the Apache mission - the team includes multiple Apache committers already. Finally, joining Apache will help coordinate the development effort of the growing number of organizations which contribute to Spark.

Initial Goals

The initial goals will most likely be to move the existing codebase to Apache and integrate with the Apache development process. Furthermore, we plan for incremental development, and releases along with the Apache guidelines.

Current Status

Meritocracy

The Spark project already operates on meritocratic principles. Today, Spark has several developers and has accepted multiple major patches from outside of U.C. Berkeley. While this process has remained mostly informal (we do not have an official committer list), an implicit organization exists in which individuals who contribute major components act as maintainers for those modules. If accepted, the Spark project would include several of these participants as committers from the onset. We will work to identify all committers and PPMC members for the project and to operate under the ASF meritocratic principles.

Community

Acceptance into the Apache foundation would bolster the already strong user and developer community around Spark. That community includes dozens of contributors from several institutions, a meetup group with several hundred members, 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 exist at UC Berkeley, there is a representative cross sampling of other organizations including Quantifind, Microsoft, Yahoo!, ClearStory Data, Bizo, Intel, Tagged and Webtrends.

Alignment

Our proposed effort aligns with several ongoing BIGDATA and U.S. National priority funding interests including the NSF and its Expeditions program, and the DARPA XDATA project. Our industry partners and collaborators are well aligned with our code base.

There are also a number of related Apache projects and dependencies, that will be mentioned in the Relationships with Other Apache products section.

Known Risks

Orphaned Products

Given the current level of investment in Spark - the risk of the project being abandoned is minimal. There are several constituents who are highly incentivized to continue development. The U.C. Berkeley AMPLab relies on Spark as a platform for a large number of long-term research projects. Several companies have build verticalized products which are tightly dependent on Spark. Other companies have devoted significant internal infrastructure investment in Spark.

Inexperience with Open Source

Spark has existed as a healthy open source project for several years. During that time, Matei and others have curated an open-source community successfully, attracting developers from a diverse group of companies including Quantifind, Microsoft, Yahoo!, ClearStory Data, Bizo, Intel, and Webtrends.

Homogenous Developers

The initial list of committers includes developers from several institutions, including Quantifind, Microsoft, Yahoo!, ClearStory Data, Bizo, Intel, and Webtrends.

Reliance on Salaried Developers

Like most open source projects, Spark receives a substantial support from salaried developers. A large fraction of Spark development is supported by graduate students at U.C. Berkeley in the course of research degrees - this is more a “volunteer” relationship, since in most cases students contribute vastly more than is necessary to immediately support research. In addition, those working from within corporations often devote “after hours” or spare time in the project - and these come from several organizations. We will work to ensure that the ability for the project to continuously be stewarded and to proceed forward independent of salaried developers is continued.

Relationship with Other Apache Products

Spark inter-operates with several existing Apache products by supporting them as storage layers: Apache Cassandra, Apache HBase, and Apache Hadoop (HDFS). It also uses several Apache components internally including Apache Maven and several Apache Commons libraries. Finally, Shark (a higher layer framework built on Spark) inter-operates with Apache Hive. We will explore the relationship between Spark and Apache Gora, which also provides in-memory object storage (Champion Mattmann was the Champion for Apace Gora so we expect alignment and cross pollination between our efforts).

Spark offers an alternative computation engine to Apache Hadoop (MapReduce). Unlike MapReduce, Spark is designed for lower-latency and interactive workloads. This makes the projects complimentary: many users run MapReduce and Spark side-by-side.

A Excessive Fascination with the Apache Brand

Spark is already a healthy and relatively well 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

The reader will find these websites highly relevant:

- Spark website: http://spark-project.org/
- Spark documentation: http://spark-project.org/documentation/
- Issue tracking: https://spark-project.atlassian.net/
- Codebase: https://github.com/mesos/spark
- User group: https://groups.google.com/group/spark-users

Initial Source

The Spark codebase is currently hosted on Github: https://github.com/mesos/spark. This is the exact codebase that we would migrate to the Apache foundation. Source and Intellectual Property Submission Plan Currently, the Spark codebase is distributed under a BSD license. The vast majority of code has copyright held by the University of California. Upon entering Apache, Spark will migrate to an Apache License with all copyright assigned to the Apache Foundation. The University of California will transfer all copyright to the Apache Foundation. In certain cases where individuals hold copyright, we will have individuals sign over copyright to the Apache foundation as well.

Going forward, all commits would assign copyright directly to the Apache foundation through our signed Individual Contributor License Agreements for all initial committers on the project.

External Dependencies

To the best of our knowledge, all dependencies of Spark are distributed under Apache compatible licenses. Upon acceptance to the incubator, we would begin a thorough analysis of all transitive dependencies to verify this fact and introduce license checking into the build and release process (for instance integrating Apache Rat).

Required Resources

Mailing list

We will migrate the existing Spark mailing lists as follows:

- spark-users@googlegroups --> users@spark.incubator.apache.org
- spark-developers@googlegroups --> dev@spark.incubator.apache.org
- spark-commits are hosted on Github, so we would request commits@spark.incubator.apache.org

The latter is to be consistent with the new PIAO naming scheme for podlings.

Source control

The Spark team would like to use Git for source control, due to our current use of Git. We request a writeable Git repo for Spark, and mirroring to be set up to Github through INFRA. Champion Mattmann can assist with creating INFRA tickets for this.

Issue Tracking

Spark currently uses a hosted JIRA deployment for issue tracking. We will migrate to the Apache JIRA. http://issues.apache.org/jira/browse/SPARK

Initial Committers

- Matei Zaharia <matei@apache.org>
- Ankur Dave <ankurdave@gmail.com>
- Tathagata Das <tdas@eecs.berkeley.edu>
- Haoyuan Li <haoyuan@cs.berkeley.edu>
- Josh Rosen <joshrosen@cs.berkeley.edu>
- Reynold Xin <rxin@cs.berkeley.edu>
- Shivaram Venkataraman <shivaram@eecs.berkeley.edu>
- Mosharaf Chowdhury <mosharaf@cs.berkeley.edu>
- Charles Reiss <charles@eecs.berkeley.edu>
- Andy Konwinski <andykonwinski@gmail.com>
- Patrick Wendell <pwendell@eecs.berkeley.edu>
- Imran Rashid <imran@quantifind.com>
- Ryan LeCompte <lecompte@gmail.com>
- Ravi Pandya <ravip@exchange.microsoft.com>
- Ram Sriharsha <harshars@yahoo-inc.com>
- Robert Evans <evans@yahoo-inc.com>
- Mridul Muralidharan <mridulm@yahoo-inc.com>
- Thomas Dudziak <tomdz@clearstorydata.com>
- Mark Hamstra <mark@clearstorydata.com>
- Stephen Haberman <stephen.haberman@gmail.com>
- Jason Dai <jason.dai@intel.com>
- Shane Huang <shannie.huang@gmail.com>
- Andrew xia <xiajunluan@gmail.com>
- Nick Pentreath <nick.pentreath@gmail.com>
- Sean McNamara <sean.mcnamara@webtrends.com>

Affiliations

The initial committers are from nine organizations: UC Berkeley, Quantifind, Microsoft, Yahoo!, ClearStory Data, Bizo, Intel, Mxit and Webtrends.

- Matei Zaharia (UCB)
- Ankur Dave (UCB)
- Tathagata Das (UCB)
- Haoyuan Li (UCB)
- Josh Rosen (UCB)
- Reynold Xin (UCB)
- Shivaram Venkataraman (UCB)
- Mosharaf Chowdhury (UCB)
- Charles Reiss (UCB)
- Andy Konwinski (UCB)
- Patrick Wendell (UCB)
- Imran Rashid (Quantifind)
- Ryan LeCompte (Quantifind)
- Ravi Pandya (Microsoft)
- Ram Sriharsha (Yahoo!)
- Robert Evans (Yahoo!)
- Mridul Muralidharam (Yahoo!)
- Thomas Dudziak (ClearStory)
- Mark Hamstra (ClearStory)
- Stephen Haberman (Bizo)
- Jason Dai (Intel)
- Shane Huang (Intel)
- Andrew Xia (Intel)
- Nick Pentreath (Mxit)
- Sean McNamara (Webtrends)

Sponsors

Champion

- Chris Mattmann

Nominated Mentors

- Chris Mattmann
- Paul Ramirez
- Andrew Hart
- Thomas Dudziak
- Suresh Marru
- Henry Saputra
- Roman Shaposhnik

Sponsoring Entity

The Apache Incubator