ZeppelinProposal

Abstract

Zeppelin is a collaborative data analytics and visualization tool for distributed, general-purpose data processing systems such as Apache Spark, Apache Flink, etc.

Proposal

Zeppelin is a modern web-based tool for the data scientists to collaborate over large-scale data exploration and visualization projects. It is a notebook style interpreter that enable collaborative analysis sessions sharing between users. Zeppelin is independent of the execution framework itself. Current version runs on top of Apache Spark but it has pluggable interpreter APIs to support other data processing systems. More execution frameworks could be added at a later date i.e Apache Flink, Crunch as well as SQL-like backends such as Hive, Tajo, MRQL.

We have a strong preference for the project to be called Zeppelin. In case that may not be feasible, alternative names could be: “Mir”, “Yuga” or “Sora”.

Background

Large scale data analysis workflow includes multiple steps like data acquisition, pre-processing, visualization, etc and may include inter-operation of multiple different tools and technologies. With the widespread of the open source general-purpose data processing systems like Spark there is a lack of open source, modern user-friendly tools that combine strengths of interpreted language for data analysis with new in-browser visualization libraries and collaborative capabilities.

Zeppelin initially started as a GUI tool for diverse set of SQL-over-Hadoop systems like Hive, Presto, Shark, etc. It was open source since its inception in Sep 2013. Later, it became clear that there was a need for a greater web-based tool for data scientists to collaborate on data exploration over the large-scale projects, not limited to SQL. So Zeppelin integrated full support of Apache Spark while adding a collaborative environment with the ability to run and share interpreter sessions in-browser

Rationale

There are no open source alternatives for a collaborative notebook-based interpreter with support of multiple distributed data processing systems.

As a number of companies adopting and contributing back to Zeppelin is growing, we think that having a long-term home at Apache foundation would be a great fit for the project ensuring that processes and procedures are in place to keep project and community “healthy” and free of any commercial, political or legal faults.

Initial Goals

The initial goals will be to move the existing codebase to Apache and integrate with the Apache development process. This includes moving all infrastructure that we currently maintain, such as: a website, a mailing list, an issues tracker and a Jenkins CI, as mentioned in “Required Resources” section of current proposal. Once this is accomplished, we plan for incremental development and releases that follow the Apache guidelines. To increase adoption the major goal for the project would be to provide integration with as much projects from Apache data ecosystem as possible, including new interpreters for Apache Hive, Apache Drill and adding Zeppelin distribution to Apache Bigtop. On the community building side the main goal is to attract a diverse set of contributors by promoting Zeppelin to wide variety of engineers, starting a Zeppelin user groups around the globe and by engaging with other existing Apache projects communities online.

Current Status

Currently, Zeppelin has 4 released versions and is used in production at a number of companies across the globe mentioned in Affiliation section. Current implementation status is pre-release with public API not being finalized yet. Current main and default backend processing engine is Apache Spark with consistent support of SparkSQL. Zeppelin is distributed as a binary package which includes an embedded webserver, application itself, a set of libraries and startup/shutdown scripts. No platform-specific installation packages are provided yet but it is something we are looking to provide as part of Apache Bigtop integration. Project codebase is currently hosted at github.com, which will form the basis of the Apache git repository.

Meritocracy

Zeppelin is an open source project that already leverages meritocracy principles. It was started by a handfull of people and now it has multiple contributors, although as the number of contribution grows we want to build a diverse developer and user community that is governed by the "Apache way". Users and new contributors will be treated with respect and welcomed; they will earn merit in the project by tendering quality patches and support that move the project forward. Those with a proven support and quality patch track record will be encouraged to become committers.

Community

Zeppelin already has a burgeoning community of users spread across the world that leverage and contributes to the code base and mailing list. We hope that being part of Apache Foundation will help to grow it more and convert some of the users into active contributors to the project.

Core Developers

The core developers of Zeppelin are listed in our contributors and initial PPMC below. It is a diverse group of people from two companies, NFLabs and Between, as mentioned in Affiliations section including at least one Apache committer and PPMC member, Lee Moon Soo, of Apache MRQL project.

Alignment

Zeppelin is already integrated with Apache Spark. Integration with Apache Tajo and Apache MRQL is something that has been currently worked on. Apache Flink is a potential next integration step. We also plan to add a binary distribution of Zeppelin to Apache Bigtop to align it with whole ASF Hadoop data stack.

Known Risks

We feel that for Zeppelin to become as successful as it can be, it needs to be picked up by as many back-end systems as possible, not only Apache Spark.

Orphaned Products

Initial code contributors were from the same company but in last few months we see signs of the global adoption, at least 2 more companies in Europe and US have products based on a Zeppelin codebase. Other companies use Zeppelin in production for their data analytics workflows. We believe that this, plus the fact that Zeppelin already have contributors from different companies mitigates this risk well.

Inexperience with Open Source

Zeppelin was born as an open source project from scratch. Majority of the current core contributors have experience working on other open source projects. We also expect that as we grow the community further based on meritocracy and with the guidance of more experienced mentors this will have a positive influence on the project in the long term.

Homogenous Developers

The initial committers are from same region but there are already 2 companies in the Europe that contribute to Zeppelin and others in US also reviewing it and being active on the mailing list. We are committed to create diverse mix of developers from all over the world.

Reliance on Salaried Developers

Most of the Zeppelin contributors use it as tool of choice either in their own companies internally or distribute it as part of the product. Backend agnostic design helps to keep it as tool of choice for diverse community of data analysts even if they move from one employee to another. There also is at least one university in US with students who potentially might use Zeppelin for R'n'D projects.

Relationship with Other Apache Products

Right now Zeppelin relies on Apache Spark to run distributed task across a cluster of machines, but it's abstract interpreter design allows it to work with other systems like Apache MRQL, Apache Crunch as well as SQL-based systems like Apache Tajo, Apache Hive

A Excessive Fascination with the Apache Brand

We believe that joining Apache will help us attract more contributors to Zeppelin, by giving us a well-defined, transparent development and governance process under a known brand. The reason for this proposal is not to gain publicity, but to further strengthen the longevity of the project without affiliation with any particular company. There are no plans to use of Apache brand in press releases nor posting advertising of acceptance it into Apache Incubator.

Documentation

Additional documentation on Zeppelin may be found on its github website:

- Zeppelin overview: https://github.com/NFLabs/zeppelin/blob/master/README.md
- Zeppelin docs: http://zeppelin-project.org/docs/index.html
- Zeppelin road map: https://github.com/NFLabs/zeppelin/blob/master/Roadmap.md
- Zeppelin issue tracking: https://zeppelin-project.atlassian.net/browse/ZEPPELIN
- Zeppelin codebase: https://github.com/NFLabs/zeppelin
- User group: https://groups.google.com/group/zeppelin-developers

Initial Source

Zeppelin codebase is currently hosted on Github: https://github.com/NFLabs/zeppelin

Source and Intellectual Property Submission Plan

Currently, the Zeppleing codebase is distributed under an Apache 2.0 License.

External Dependencies

To the best of our knowledge, all other dependencies of Zeppelin are distributed under Apache compatible licenses (e.g. junit is EPL, Eclipse Public License v1.0, atmosphere-jersey is CDDL1.0 and dom4j:dom4 is BSD licensed, org.slf4j and org.java-websocket:Java-WebSocket are MIT). Only org.reflections:reflections https://github.com/ronmamo/reflections is WTFPL 2.0, which should not be a problem as of https://issues.apache.org/jira/browse/LEGAL-135 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.

Required Resources

Mailing list

We will migrate the existing Zeppelin mailing lists as follows:

- zeppelin-developers@googlegroups.com --> dev@zeppelin.incubator.apache.org
- users@zeppelin.incubator.apache.org
- private@zeppelin.incubator.apache.org for PPMC members
- commits@zeppelin.incubator.apache.org

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

Source control

Zeppelin team would like to use Git for source control, as it already uses Git. We request a writeable Git repo for Zeppelin, and mirroring to be set up to Github through INFRA. https://git-wip-us.apache.org/repos/asf/incubator-zeppelin.git

Issue Tracking

Zeppelin currently uses the Jira tracking system https://zeppelin-project.atlassian.net/browse/ZEPPELIN. We will migrate to the Apache JIRA: http://issues.apache.org/jira/browse/ZEPPELIN

Other Resources

- Jenkins/Hudson for builds and test running.
- Wiki for documentation purposes
- Blog to improve project dissemination

Initial Committers

- Lee Moon Soo <moon at apache dot org>
- Anthony Corbacho <corbacho.anthony at gmail dot com>, CLA submitted
- Damien Corneau <corneadoug at gmail dot com>, CLA submitted
- Alexander Bezzubov <abezzubov at nflabs dot com>, CLA confirmed
- Kevin Sangwoo Kim <sangwookim dot me at gmail dot us>, CLA confirmed

Affiliations

- Lee Moon Soo: NFLabs
- Anthony Corbacho: NFLabs
- Damien Corneau: NFLabs
- Alexander Bezzubov: NFLabs
- Kevin Sangwoo Kim: VCNC (a.k.a Between)

Sponsors

Champion

- Roman Shaposhnik

Nominated Mentors

- Konstantin Boudnik
- Ted Dunning
- Henry Saputra
- Roman Shaposhnik
- Hyunsik Choi

Sponsoring Entity

The Apache Incubator