SupersetProposal

Superset

Abstract

Superset is an enterprise-ready web application for data exploration, data visualization and dashboarding.

Proposal

Superset is business intelligence (BI) software that helps modern organizations visualize and interact with their data. Superset enables users explore data from a variety of databases, assemble beautiful dashboards and share their findings. Superset works neatly with all modern SQL-speaking databases, and integrates with Druid.io to provide real-time, interactive, blazing fast data access to large datasets.

Background

Data is mission critical. To succeed in this era, organizations need to provide low-friction, intuitive and interactive access to data. It is paramount for knowledge workers to be capable of answering their own questions by querying, exploring and visualizing data.

The entire business intelligence industry has pivoted from a model of centralized top-down platforms driven by IT organizations to self-service analytics and agile workflows by any user. This shift unblocks centralized service bottlenecks for creating data visualizations while also creating an environment that is iterative and fast-moving. This means that business intelligence software must also be easy and delightful to use. Self-service analytics doesn't mean that admin and governance features are not needed. Modern BI tools provide fine-grain access controls and auditing capabilities to understand how data is being used. Superset is a solution that delivers on all of these vectors.

The technology stack is also constantly morphing - vendors are struggling to provide cheap, quick and easy solutions to access data. Business intelligence users are finding existing solutions lacking as these software products either disregard or react slowly to recent game-changing technologies like Druid.io, PrestoDB, Apache Drill, Apache Kylin, d3.js, React.js and iPython's Jupyter for instance.

Rationale

Business intelligence is more relevant today than at any other point in history. Organizations are currently very limited in options for open source data visualization solutions, especially solutions that are both self-service and enterprise-ready. Every company informing their decisions with data needs a BI tool.

We believe that Superset will be a strong compliment to existing Apache Software Foundation technologies by offering scalable user interactions to distributed storage and computation solutions. Users will often find that Superset can act as a catalyst for tooling that can visualize the byproduct of data and computation infrastructure.

Superset has many key design elements that help fill a gap in current solutions for organizations:

- Easy, low friction access to data through a simple, web-based data exploration interface. Composing charts and dashboards are intuitive. Eliminating the need to write code or SQL empowers anyone to use it.
- Access to a wide array of rich, interactive data visualization types.
- Enterprise-ready: Integration with different authentication mechanisms and granular permissions centered around actions and data access.
- Realtime & fast: Superset provides realtime analytics at the speed of thought on very large datasets when integrated with Druid.io.
- Broad data access: Consume data out of any SQL-speaking relational database.
- Extensible: Can be extended to talk to many noSQL databases like Apache Drill, Elastic Search, and other popular database engines.
- Fast loading dashboards with configurable web-scale caching.
- Plug-in framework that enables organizations to build custom analytical applications with new UI/UX interfaces.
- SQL Lab, a state-of-the-art SQL IDE that empowers SQL-speaking users with more flexibility. SQL Lab integrates with the visualization engine seamlessly.

Initial Goals

The initial goals of the Superset project are several-fold:

- Move the existing codebase to Apache and integrate with the Apache development process.
- Redesign the user interface and interaction model for creating visualizations/dashboards and connecting to data sources
- Build robust support for security and governance of the tool including popular authorization modules (including Apache Ranger and Apache Sentry) and a more sophisticated permissions system
- Grow the extensibility of the project both in terms of enhanced connectivity to NoSQL-based data sources and creating a plug-in framework that enables organizations to build custom analytical applications which require a new UI/UX

Current Status

By many standards, Superset is already a successful open source project. As of March 2017, Superset is officially used in production at about a dozen companies, has received contributions from over one hundred contributors on Github, 1500+ forks, and 12k+ stars.

Sizeable companies like Airbnb, Yahoo! and Hortonworks have made significant contributions, and expressed their commitment to the project. The product is feature complete and has been viable for months. It already serves as the main interface for consuming data at many companies of different sizes.

While the product is usable, there's room for improvement across the board, starting with providing a smoother user experience around content creation, making sure all features work out-of-the-box on more platforms and databases, providing better user training guides and videos, having a predictable release process, and increasing the overall quality of the Superset releases.

Meritocracy

We plan to invest in supporting a meritocracy. We will discuss the requirements in an open forum. Several companies have expressed interest in this project, and 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

The need for an enterprise-ready data visualization and exploration platform in the open source community is tremendous. While Superset is fairly well known, recognized and used within the Druid.io community, adoption is currently limited outside of that niche. There is a huge opportunity to grow the community to hundreds if not thousands of organizations, and we are hoping that embracing "the Apache way" will accelerate the growth of our community.

We have already been active at seeking and inviting contributions, and are planning to scale the project by investing time and growing the support structure to grow the community.

Core Developers

The initial committers for Superset include experienced full stack, front-end and data engineers:

- Maxime Beauchemin (Airbnb)
- Alanna Scott (Airbnb)
- Bogdan Kyryliuk (Airbnb)
- Vera Liu (Airbnb)
- Jeff Feng (Airbnb)
- Ashutosh Chauhan (Hortonworks)
- Nishant Bangarwa (Hortonworks)
- Slim Bouguerra (Hortonworks)
- Priyank Shah (Hortonworks)
- Sriharsha Chintalapani (Hortonworks)
- Daniel Dai (Hortonworks)

We realize that additional employer diversity is needed, and we will work aggressively to recruit developers from additional companies.

Alignment

The initial committers strongly believe that a system for interactive visualization of data will gain broader adoption as an open source, community driven project, where the community can contribute not only to the core components, but also to a growing collection of connectors, visualizations and improving integration a all potential data sources. Superset already integrates closely with Apache Hive, the Hive metastore, as well as most SQL-speaking databases found in modern data ecosystems.

Known Risks

Orphaned Products

Superset is a vital component for both visualizing, accessing and democratizing data at Airbnb. Also at Hortonworks, Superset is a core component of the DataFlow product offering. Thus, the risk of the project being orphaned is relatively low. The project could be at risk if Airbnb changes their approach for democratizing data or if Hortonworks changes their strategy in the market. In such an event, the committers plan to continue working on the project on their own time, thought the progress will likely be slower. We plan to mitigate this risk by recruiting additional committers.

Inexperience with Open Source

The initial committers include veteran Apache members (committers and PPMC members) and other developers who have varying degrees of experience with open source projects. All have been involved with source code that has been released under an open source license, and several also have experience developing code with an open source development process.

Homogenous Developers

The initial committers are employed by Airbnb Inc. and Hortonworks. We are committed to recruiting additional committers from other companies.

Reliance on Salaried Developers

It is expected that Superset development will occur on both salaried time and on volunteer time, after hours. The majority of initial committers are paid by their employer to contribute to this project. However, they are all passionate about the project, and we are confident that the project will continue even if no salaried developers contribute to the project. We are committed to recruiting additional committers including non-salaried developers.

Relationships with Other Apache Products

To the knowledge of the Initial Committers, there are no direct competitors to Superset within the Apache Software Foundation. That said, Apache Zeppelin is an indirect competitor, but it solves a different use case.

Apache Zeppelin is a web-based notebook that enables interactive data analytics. It enables the creation of beautiful data-driven, interactive and collaborative documents with SQL, Scala and more. Although a user can create data visualizations using this project, it leverages a notebook style user interfaces and it is geared towards the Spark community where Scala and SQL co-exist

We look forward to collaborating with those communities, as well as other Apache communities.

An Excessive Fascination with the Apache Brand

Superset is solving two huge challenges:
The challenge of enabling every knowledge worker to make data informed decisions, particularly those who are not deeply skilled at writing SQL.
The challenge of visualizing huge amounts of data interactively and in real-time

Superset was first developed as a data visualization solution for Druid.io as a way to visualize billions of rows of data. Since then, usage of Superset has expanded to address data visualization use cases across SQL speaking data sources as well.

Our rationale for developing Superset as an Apache project is detailed in the Rationale Section. We believe that the Apache brand and community process will help us attract more contributors to this project, and help grow the footprint of the project through usage at other organizations and within other applications. Establishing consensus among users and developers will result in a more valuable tool for everyone.

Documentation

References to further reading material:

- Superset Documentation
- Blog Post: Superset: Airbnb's Data Exploration Platform
- Blog Post: Superset: Scaling Data Access & Visual Insights at Airbnb

Initial Source

The origin of the proposed code base can be found at https://github.com/airbnb/superset. The code base is primarily in Python.

Source and Intellectual Property Submission Plan

Airbnb will submit a Software Grant Agreement (SGA) as Superset joins the incubator. We do not expect any complications for the submission of the Superset code base. Our code is already in Github and there is only a single code base.

External Dependencies

List of Python packages, from the Python Package Index (Pypi):

- boto3
- celery
- cryptography
- flask-appbuilder
- flask-cache
- flask-migrate
- flask-script
- flask-sqlalchemy
- flask-testing
- humanize
- gunicorn
- markdown
- pandas
- parsedatetime
- pydruid
- PyHive
- python-dateutil
- requests
- simplejson
- six
- sqlalchemy
- sqlalchemy-utils
- sqlparse
- thrift
- thrift-sasl
- werkzeug

List of Javascript packages, from NPM:

- autobind-decorator
- bootstrap
- bootstrap-datepicker
- brace
- brfs
- cal-heatmap
- classnames
- d3
- d3-cloud
- d3-sankey
- d3-scale
- d3-tip
- datamaps
- datatables-bootstrap3-plugin
- datatables.net-bs
- font-awesome
- gridster
- immutability-helper
- immutable
- jquery
- lodash.throttle
- mapbox-gl
- moment
- moments
- mustache
- nvd3
- react
- react-ace
- react-bootstrap
- react-bootstrap-table
- react-dom
- react-draggable
- react-gravatar
- react-grid-layout
- react-map-gl
- react-redux
- react-resizable
- react-select
- react-syntax-highlighter
- reactable
- redux
- redux-localstorage
- redux-thunk
- shortid
- style-loader
- supercluster
- topojson
- victory
- viewport-mercator-project

Cryptography

The proposal does not include cryptographic code.

Required Resources

Mailing List

There is a current mailing list as a Google Group “airbnb_superset” that we are planning on deprecating as the Apache.org become ready to serve our community.

- superset-private
- superset-dev
- superset-user

Subversion Directory

Git is the preferred source control system. http://svn.apache.org/repos/asf/incubator/superset

Git Repository

Git is the preferred source control system, we’re assuming https://github.com/apache/incubator-superset based on the naming scheme

Issue Tracking

JIRA Superset (SUPERSET). If possible, we'd like to use Github issues & PRs to manage our project as much as possible. It's been said that there are ways to keep Github's issues in sync with Jira, allowing us to get best of both worlds. If that is not possible, we will comply to using Jira.

Other Resources

We currently use a set of Github integrated services that are free to the open source community, like Travis-ci, Code Climate, Coveralls, Landscape.io, Requires.io, david-dm and Gitter. We would like to keep using these services as they allow us to scale contributions and optimize our development flows. These services require some elevated rights on the Github repository in order to set up or tune and we would like for the committers to have the required rights.

Initial Committers

- Maxime Beauchemin <maxime.beauchemin@airbnb.com> - PPMC & Committer
- Alanna Scott <alanna.scott@airbnb.com> - PPMC & Committer
- Bogdan Kyryliuk <b.kyryliuk@gmail.com> - PPMC & Committer
- Vera Liu <vera.liu@airbnb.com> - Committer
- Jeff Feng <jeff.feng@airbnb.com> - PPMC & Committer
- Ashutosh Chauhan <hashutosh@apache.org> - Mentor & Committer
- Nishant Bangarwa <nbangarwa@hortonworks.com> - PPMC & Committer
- Slim Bouguerra <sbouguerra@hortonworks.com> - Committer
- Priyank Shah <pshah@hortonworks.com> - Committer
- Harsha Chintalapani <schintalapani@hortonworks.com> - Committer
- Daniel Dai <daijy@apache.org> - Champion & Committer
- Luke Han <luke.han@apache.org> - Mentor
- Jim Jagielski <jim@jaguNET.com> - Mentor

Affiliations

The initial committers are employees of Airbnb Inc. and Hortonworks.

Sponsors

Champion

Daniel Dai <daijy@apache.org>

Nominated Mentors

- Ashutosh Chauhan <hashutosh@apache.org>
- Luke Han <luke.han@apache.org>
- Jim Jagielski <jim@jaguNET.com>

Sponsoring Entity

Incubator PMC