AirflowProposal

Abstract

Airflow is a workflow automation and scheduling system that can be used to author and manage data pipelines.

Proposal

Airflow provides a system for authoring and managing workflows a.k.a. data pipelines a.k.a. DAGs (Directed Acyclic Graphs). The developer authors DAGs in Python using an Airflow-provided framework. He/She then executes the DAG using Airflow's scheduler or registers the DAG for event-based execution. A web-based UI provides the developer with a range of options for managing and viewing his/her data pipelines. Background

Airflow was developed at Airbnb to enable easier authorship and management of DAGs than were possible with existing solutions such as Oozie and Azkaban. For starters, both Oozie and Azkaban rely on one or more XML or property files to be bundled together to define a workflow. This separation of code and config can present a challenge to understanding the DAG - in Azkaban, a DAG's structure is reflected by its file system tree and one can find himself/herself traversing the file system when inspecting or changing the structure of the DAG. Airflow workflows, on the other hand, are simply and elegantly defined in Python code, often a single file. Airflow merges the powerful Web-based management aspects of projects like Azkaban and Oozie with the simplicity and elegance of defining workflows in Python. Airflow, less than a year old in terms of its Open Source launch, is currently used in production environments in more than 30 companies and boasts an active contributor list of more than 100 developers, the vast majority of which (>95%) are outside of Airbnb.

We would like to share it with the ASF and begin developing a community of developers and users within Apache.

Rationale

Many organizations (>30) already benefit from running Airflow to manage data pipelines. Our 100+ contributors continue to provide integrations with 3rd party systems through the implementation of new hooks and operators, both of which are used in defining the tasks that compose workflows.

Current Status

Meritocracy

Our intent with this incubator proposal is to start building a diverse developer community around Airflow following the Apache meritocracy model. Since Airflow was open-sourced in mid-2015, we have had fast adoption and contributions by multiple organizations the world over. We plan to continue to support new contributors and we will work to actively promote those who contribute significantly to the project to committers.

Community

Airflow is currently being used in over 30 companies. We hope to extend our contributor base significantly and invite all those who are interested in building large-scale distributed systems to participate.

Core Developers

Airflow is currently being developed by four engineers: Maxime Beauchemin, Siddharth Anand, Bolke de Bruin, and Chris Riccomini. Chris is a member of the Apache Samza PMC and a contributor to various Apache projects, including Apache Kafka and Apache YARN. Maxime, Siddharth, and Bolke have contributed to Airflow.

Alignment

The ASF is the natural choice to host the Airflow project as its goal of encouraging community-driven open-source projects fits with our vision for Airflow.

Known Risks

Orphaned Products

The core developers plan to work part time on the project. There is very little risk of Airflow being abandoned as all of our companies rely on it.

Inexperience with Open Source

All of the core developers have experience with open source development. Chris is a member of the Apache Samza PMC and a contributor to various Apache projects, including Apache Kafka and Apache YARN. Bolke is contributor on multiple open source projects and a few Apache projects as well, including Apache Hive, Apache Hadoop, and Apache Ranger.

Homogeneous Developers

The current core developers are all from different companies. Our community of 100 contributors hail from over 30 different companies from across the world.

Reliance on Salaried Developers

Currently, the only developer paid to work on this project is Maxime.

Relationships with Other Apache Products

Airflow is deeply integrated with Apache products. It currently provides hooks and operators to enable workflows to leverage Apache Pig, Apache Hive, Apache Spark, Apache Sqoop, Apache Hadoop, etc… We plan to add support for other Apache projects in the future.

An Excessive Fascination 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 Airflow a solid home as an open source project following an established development model. We have also given reasons in the Rationale and Alignment sections.

Documentation

http://wiki.apache.org/incubator/AirflowProposal

Initial Source

https://github.com/airbnb/airflow

Source and Intellectual Property Submission Plan

As soon as Airflow is approved to join Apache Incubator, Airbnb will execute a Software Grant Agreement and the source code will be transitioned onto ASF infrastructure. The code is already licensed under the Apache Software License, version 2.0. We know of no legal encumberments that would inhibit the transfer of source code to the ASF.

External Dependencies

The dependencies all have Apache compatible licenses.

- alembic (MIT)
- boto (MIT)
- celery (BSD)
- chartkick (MIT)
- cryptography (Apache 2.0/BSD)
- coverage (Apache 2.0)
- coveralls (MIT)
- croniter (MIT)
- dill (BSD)
- docker-py (Apache 2.0)
- filechunkio (MIT)
- flake8 (MIT)
- flask (BSD)
- flask-admin (BSD)
- flask-cache (BSD)
- flask-login (MIT)
- flower (BSD)
- future (MIT)
- gunicorn (MIT)
- hive-thrift-py (Apache 2.0)
- ipython (BSD)
- jinja2 (BSD)
- markdown (BSD)
- pandas (BSD)
- pygments (BSD)
- pyhive
- pydruid
- PyOpenSSL
- PySmbClient
- python-dateutil
- redis
- requests
- setproctitle
- statsd
- sphinx
- sphinx-argparse

- sphinx_rtd_theme
- Sphinx-PyPI-upload
- sqlalchemy (MIT)
- thrift
- jaydebeapi
- mysqlclient
- unicodecsv
- slackclient
- ldap3
- Flask-WTF
- lxml
- pykerberos (Apache 2.0)
- bcrypt (Apache 2.0)
- flask-bcrypt (BSD)
- mock (BSD)
- hdfs (MIT)

Cryptography

None

Required Resources

Mailing Lists

- private@airflow.incubator.apache.org (moderated)
- dev@airflow.incubator.apache.org
- commits@airflow.incubator.apache.org

Subversion Directory

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

Issue Tracking

JIRA Airflow (Airflow)

Other Resources

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

Initial Committers

- Maxime Beauchemin
- Siddharth Anand
- Chris Riccomini
- Bolke de Bruin
- Arthur Wiedmer
- Dan Davydov
- Jeremiah Lowin
- Patrick Leo Tardif

Affiliations

- Maxime Beauchemin (Airbnb)
- Siddharth Anand (Agari)
- Chris Riccomini (WePay)
- Bolke de Bruin (ING)
- Arthur Wiedmer (Airbnb)
- Dan Davydov (Airbnb)
- Jeremiah Lowin (Kokino)
- Patrick Leo Tardif (Airbnb)

Sponsors

Champion

Chris Riccomini (WePay, Apache Samza PMC)

Nominated Mentors

- Chris Nauroth (HortonWorks, Apache Hadoop Committer/PMC Member, Apache ZooKeeper Committer, Apache Software Foundation Member)
- Hitesh Shah (HortonWorks, Apache Hadoop Committer/PMC Member, Apache Ambari Committer/PMC Member, Apache Tez Committer/PMC Member, Apache Software Foundation Member)
- Jakob Homan (OfferUp, Apache Hadoop Committer/PMC Member, Apache Kafka Committer/PMC Member, Apache Samza Committer/PMC Member, Apache Giraph Committer/PMC Member, Apache Software Foundation Member)

Sponsoring Entity

We are requesting the Incubator to sponsor this project.