WhirrProposal

Whirr, a library of cloud services

Abstract

Whirr will be a set of libraries for running cloud services.

Proposal

Whirr will provide code for running a variety of software services on cloud infrastructure. It will provide bindings in several languages (e.g. Python and Java) for popular cloud providers to make it easy to start and stop services like Hadoop clusters. The project will not be limited to a particular set of services, rather it will be expected that a range of services are developed, as determined by the project contributors. Possible services include Hadoop, HBase, ZooKeeper, Cassandra.

Background

The ability to run services on cloud providers is very useful, particularly for proofs of concept, testing, and also ad hoc production work. Bringing up clusters in the cloud is non-trivial, since careful choreography is required. (Designing an interface that is convenient as well as secure is also a challenge in a cloud context.) Making services that runs on a variety of cloud providers is harder, even with the availability of libraries like libcloud and jclouds, since each platform's quirks and extra features must be considered (and either worked around, or possibly taken advantage of, as appropriate) . Whirr will facilitate sharing of best practices, both for a particular service (such as Hadoop configuration on a particular provider), and for common cloud operations (such as installation of dependencies across cloud providers). It will provide a space to share good configurations and will encode service-specific knowledge.

Rationale

There are already scripts in the Hadoop project that allow users to run Hadoop clusters on Amazon EC2 and other cloud providers. While users have found these scripts useful, their current home as a Hadoop Common contrib project has the following limitations:

- Tying the scripts' release cycle to Hadoop's means that it is difficult to distribute updates to the scripts which are changing fast (new features and bugfixes).
- The scripts support multiple versions of Hadoop, so it makes more sense to distribute them separately from Hadoop itself.
- They are general: people want to contribute code for non-Hadoop services like Cassandra (for example: http://github.com/johanoskarsson/cassandra-ec2).
- Having a uniform approach to running services in the cloud, hosted in one project, makes launching sets of complementary services easier for the user. Today, the scripts and libraries hosted within each project (e.g. in Hadoop, HBase, Cassandra) have slightly different conventions and semantics, and are likely to diverge over time. Building a community around cloud infrastructure services will help enforce a common approach to running services in the cloud.

Initial Goals

- Provide a new home for the existing Hadoop cloud scripts.
- Add more services (e.g. HBase)
- Develop Java libraries for Hadoop clusters
- Add new cloud providers by taking advantage of libcloud and jclouds.
- (Future) Run on own hardware, so users can take advantage of the same interface to control services running locally or in the cloud.

Current Status

Meritocracy

The Hadoop scripts were originally created by Tom White, and have had a substantial number of contributions from members of the Hadoop community. By becoming its own project, significant contributors to Whirr would become committers, and allow the project to grow.

Community

The community interested in cloud service infrastructure is currently spread across many smaller projects, and one of the main goals of this project is to build a vibrant community to share best practices and build common infrastructure. For example, this project would provide a home to facilitate collaboration between the groups of Hadoop and HBase developers who are building cloud services.

Core developers

Tom White wrote most of the original code and is familiar with open source and Apache-style development, being a Hadoop committer and an ASF member. There have been a number of contributors who have provided patches to these scripts over time. Andrew Purtell who created the HBase cloud scripts is a HBase committer. Johan Oskarsson (Hadoop and Cassandra committer) ported the scripts to Cassandra.

Alignment

Whirr complements libcloud, currently in the Incubator. Libcloud provides multi-cloud provider support, while Whirr will provide multi-service support in the cloud. Whirr will build cloud components for several Apache projects, such as Hadoop, HBase, ZooKeeper, Cassandra, and hopefully more.

Known Risks

Orphaned products

There is a risk that Whirr will not gain adoption. However, the current Hadoop scripts seem to be fairly widely used. The small number of initial committers is also a risk, although by starting the project it is expected that new contributors will quickly be attracted to the project and help it grow.

Inexperience with Open Source

The initial code comes from Hadoop where it was developed in an open-source, collaborative way. All the initial committers are committers on other Apache projects, and are experienced in working with new contributors.

Homogenous Developers

The initial set of committers is from a diverse set of organizations, and geographic locations. They are all experienced with developing in a distributed environment.

Reliance on Salaried Developers

It is expected that Whirr will be developed on salaried and volunteer time, although all of the initial developers will work on it mainly on salaried time.

Relationships with Other Apache Products

Whirr will depend on many other Apache Projects as already mentioned above (e.g. Hadoop, ZooKeeper). If the project develops some common infrastructure then it is possible that it becomes a dependency on a project that wishes to use that infrastructure for running in the cloud.

A Excessive Fascination with the Apache Brand

We think that Whirr will benefit from the community sharing ideas and best practices for running cloud services. The ASF does a great job at building communities, which is why we want to build Whirr at Apache.

Documentation

Information on the current scripts and general background can be found at

- http://wiki.apache.org/hadoop/AmazonEC2
- http://archive.cloudera.com/docs/ec2.html
- http://hbase.s3.amazonaws.com/hbase/HBase-EC2-HUG9.pdf
- http://www.slideshare.net/steve_l/new-roles-for-the-cloud

Initial Source

- http://svn.apache.org/viewvc/hadoop/common/trunk/src/contrib/cloud/
- http://github.com/tomwhite/whirr

Source and Intellectual Property Submission Plan

The initial source is already in an Apache project's SVN repository (Hadoop), so there should be no action required here.

External Dependencies

The existing external dependencies all have Apache compatible licenses: boto (MIT), libcloud (Apache 2.0), simplejson (MIT). Jclouds is not a dependency of the current source, but it is Apache 2.0 licensed, so it will be possible to use it in the future if required.

Cryptography

Whirr uses standard APIs and tools for SSH and SSL.

Required Resources

Mailing lists

- whirr-private (with moderated subscriptions)
- whirr-dev
- whirr-commits
- whirr-user

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/whirr

Issue Tracking

- JIRA Whirr (WHIRR)

Other Resources

The existing code already has unit and integration 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

- Tom White (tomwhite at apache dot org)
- Andrew Purtell (apurtell at apache dot org)
- Johan Oskarsson (johan at apache dot org)
- Steve Loughran (stevel at apache dot org)
- Patrick Hunt (phunt at apache dot org)

Affiliations

- Tom White, Cloudera
- Andrew Purtell, Trend Micro
- Johan Oskarsson, Twitter
- Steve Loughran, HP Labs
- Patrick Hunt, Yahoo!

Sponsors

Champion

- Tom White

Nominated Mentors

- Doug Cutting
- Tom White
- Steve Loughran

Sponsoring Entity

- Incubator PMC