Sahara Project Update November 2019

Jeremy Freudberg

Background - what is Sahara?

Sahara provides a scalable data processing (Hadoop, Spark, etc.) stack and associated management interfaces: deploy and scale clusters, and run jobs on them, with minimal effort.

- Deploy onto virtual machines or bare metal (Ironic)
- Use upstream (Apache) distribution or vendor (Hortonworks, Cloudera, MapR) distributions
- Run jobs against data in object store (Swift, Ceph RGW), shared filesystem (Manila), or HDFS

Background - project history

- Project began in 2013
- First official release, Juno, was in 2014
- Most contributors have been from Mirantis, Red Hat, and EasyStack
- Currently most of the work is done by 3 contributors from Red Hat

“Train” cycle - what was done?

- Python 3 support
- PDF documentation
- Minor fixes

… not much is happening.

(“Stein” was more productive: out-of-tree plugins, and a stable APIv2 were finally finished after a long wait.)

Some bad news

- Vendors - longevity?
  - Hortonworks: merged with Cloudera
  - Cloudera: summer of 2019 marked by low earnings and departure of CEO
  - MapR: assets and IP acquired by HPE
- Maintainers - who?

Some good news

You can help!

Some even better news

You can start as soon as tomorrow! (Thursday PTG)

Kilo table
before lunch: onboarding, dependent upon interest
after lunch: project discussions, dependent upon interest

Please view details and ***indicate interest*** on the etherpad:
https://etherpad.openstack.org/p/shanghai-ptg-sahara
(accessible from ptg.openstack.org)

Why you should contribute to Sahara

- Get your feet wet in OpenStack
- A small project, your voice would be heard
- Maybe you or your organization find it interesting or valuable

What you might contribute to Sahara

Continuing with existing roadmap:

- Fix broken features:
  - Castellan/Barbican integration
  - Support for Spark jobs written in Python
- Finish transition to sahara-image-pack
- Rehome sahara-extra code to Apache
- Upgrade plugins
- Improve CI

Improve the following in order to encourage (re)adoption:

- Documentation
- Efficiency/performance
- Stability
- Monitoring/maintenance of clusters

Summary and contact

- Sahara deploys and manages data processing clusters and jobs — cool!
- “Train” cycle saw little progress
- There’s lots to do!
- YOU can contribute
- YOU can contribute tomorrow
  - See etherpad @ https://etherpad.openstack.org/p/shanghai-ptg-sahara
  - 

Contact me: #openstack-sahara (ping jeremy__bouncer)
openstack-discuss ML: [sahara]