

OpenStack 101 and services integration for supporting Data Processing toolkits

ccamacho, dmellado

Red Hat

ccamacho@redhat.com, dmellado@redhat.com

January 25, 2017

Agenda

Available in:

<https://github.com/ccamacho/openstack-presentations/tree/master/2017-01-25-meetup-openstack101-bigdata>

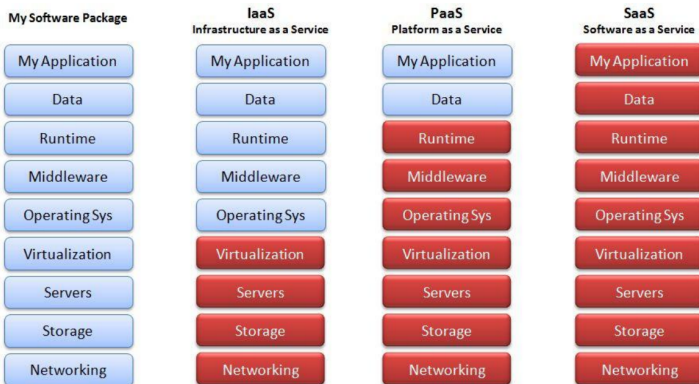
Cloud computing has been defined by the U.S. National Institute of Standards and Technology (NIST) as "...a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

Cloud computing characteristics

Cloud computing has several essential characteristics:

- Self-service: Allows cloud consumers to provision instances with computing resources.
- Global network access: Access the applications on the instance from the Internet.
- Multitenancy: Allows multiple cloud consumers to share the underlying hardware.
- Elasticity: Scales out (or scales in) instances to satisfy demand.
- Telemetry: Resources can be monitored and metered by the service provider as well as the cloud consumer.

Infra service levels



Legends:

Managed by Me

Managed by the Vendor

What is OpenStack

OpenStack is a set of software tools for building and managing cloud computing platforms for public and private clouds.



- Compute project
- Provision & manage virtual machines
- Multi-hypervisor support, included KVM & Xen

- Networking project
- Manage virtual networks (L2 & L3)
- Multi-backend support: Linux Bridge, OVS, etc

- Image project
- Catalog & manage library of server images
- Backends: Swift, Amazon, Ceph, GlusterFS, etc

- Object storage project
- Redundant and scalable
- Long-term storage system for large amounts of data
- HTTP API (RESTFull)
- Similar to Amazon S3

- Block storage project
- Manage volumes, pluggable to virtual machines
- Backends: Ceph, NFS, iSCSI, etc
- Similar to Amazon Elastic storage

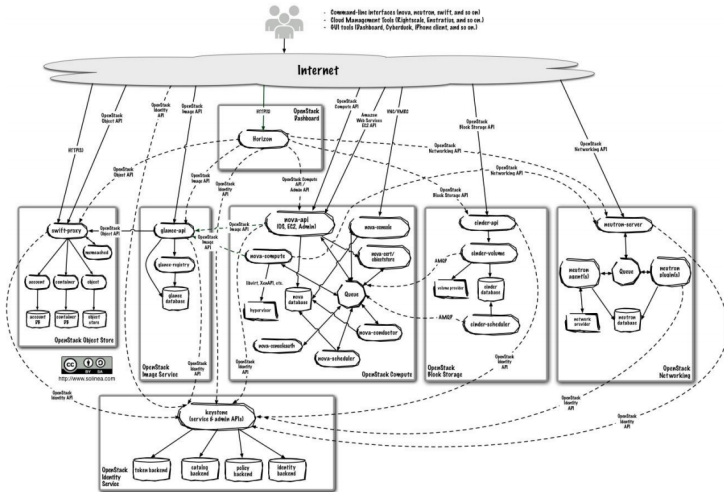
- Identity service
- Provide unified authentication for OpenStack projects
- Also manage services endpoints catalog
- Concepts of User, Tenant, Role
- Backends: MySQL, LDAP

- Telemetry project
- Provide collection of metering data (CPU usage, network costs, etc) used by virtual machines
- Custom data by plugins

- Orchestration project
- Provide a template-based for describing an application
- Integrated with OpenStack projects
- Auto-scaling and High-Availability for VMs
- Compatible with AWS CloudFormation

How does it look like?

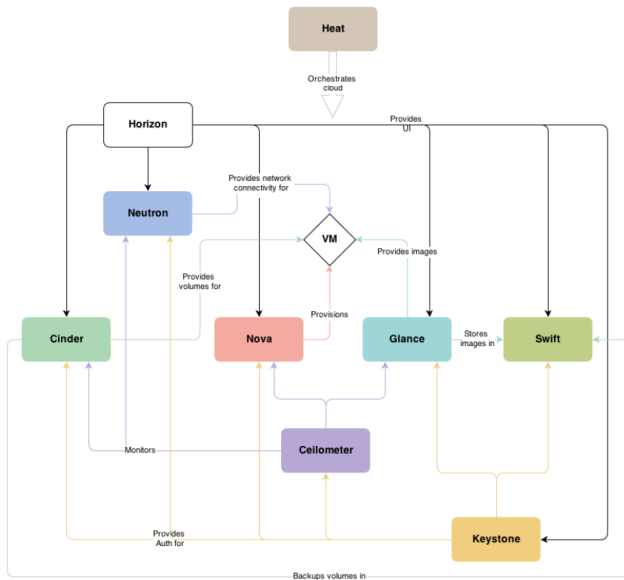
True story



Really???



A simpler view

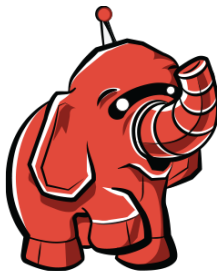
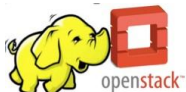


- TripleO/OSP
- Packstack
- Fuel

Big data is a term that describes the analysis and processing of large volumes of data.

Apache Hadoop is an open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Hadoop services provide for data storage, data processing, data access, data governance, security, and operations.

The intersection of Hadoop and OpenStack



Sahara

Saharas mission is to provide a scalable data processing stack and associated management interfaces. Sahara delivers on that mission by providing the ability to rapidly create and manage Apache Hadoop clusters and easily run workloads across them. All on OpenStack managed infrastructure, without having to deal with the details of cluster management.

With full cluster lifecycle management, provisioning, scaling and termination, Sahara allows the user to select different Hadoop versions, cluster topology and node hardware details.

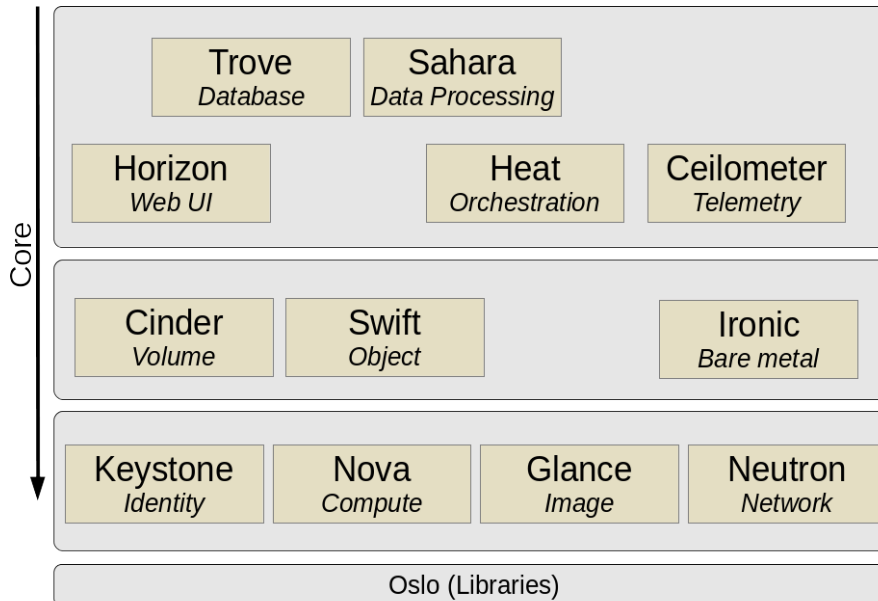
Sahara key features and use cases

- Fast and agile Hadoop cluster deployment
- An extensible framework for management and provisioning components
- Run Hadoop workloads in few clicks without expertise in Hadoop operations
- Analytics as a Service utilization of unused compute capacity for ad-hoc or bursty analytic workloads
- Sahara supports different types of jobs: MapReduce, Hive, Pig and Oozie workflows. The data could be taken from various sources: Swift, HDFS, NoSQL and SQL databases. It also supports various provisioning plugins.
- The intersection of two of the largest open source movements
- OpenStack provides the foundation and hub of innovation for cleanly managing infrastructure resources. While Apache Hadoop serves as the core and innovation driver for storing and processing data.

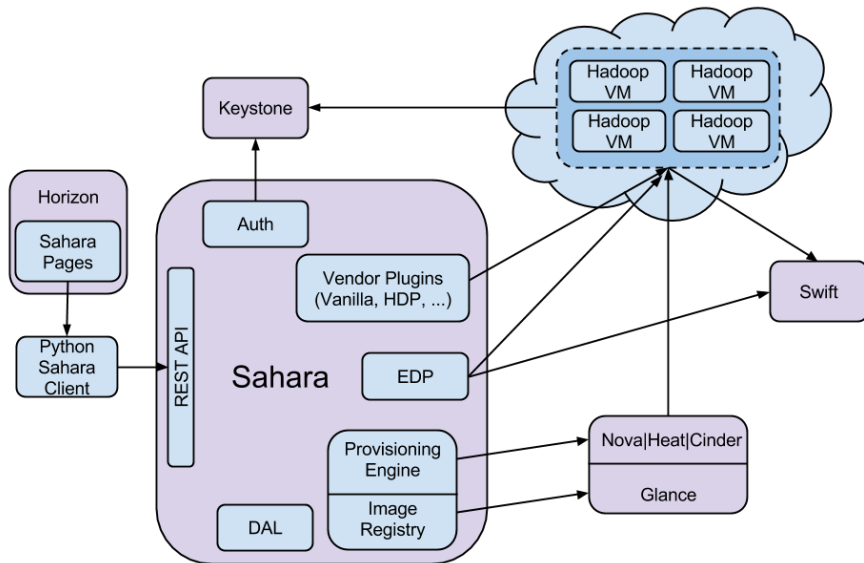
Sahara Functionality

- Bringing up cluster
- Configure it along the way
- Scale cluster
- Terminate cluster
- Job execution (Elastic Data Processing)

OpenStack Layers



Sahara Architecture



OpenStack Integration

The screenshot shows the OpenStack dashboard interface. The top navigation bar includes the OpenStack logo, a user profile dropdown labeled 'demo', and a 'Sign Out' button. The left sidebar contains a navigation menu with categories: Project, Compute, Overview, Instances, Volumes, Images, Access & Security (highlighted), Network, Object Store, Orchestration, Data Processing, and Identity. The main content area is titled 'Manage Security Group Rules: default' and 'Security Group Rules'. It features a '+ Add Rule' button and a 'Delete Rules' button. Below these is a table listing the security group rules.

<input type="checkbox"/>	Direction	Ether Type	IP Protocol	Port Range	Remote	Actions
<input type="checkbox"/>	Ingress	IPv6	Any	-	default	Delete Rule
<input type="checkbox"/>	Egress	IPv6	Any	-	:::0 (CIDR)	Delete Rule
<input type="checkbox"/>	Egress	IPv4	Any	-	0.0.0.0/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	Any	-	default	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	ICMP	-	0.0.0.0/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	TCP	1 - 65535	172.24.4.225/28 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	TCP	22 (SSH)	0.0.0.0/0 (CIDR)	Delete Rule

Displaying 7 items

Upload Sahara image

```
fedora@devstack-sahara ~$ openstack image create --disk-format qcow2 --container-format bare --file sahara-newton-vanilla-2.7.1-centos7.qcow2 centos7-sahara
```

Field	Value
checksum	a7853681c18fe1f611d70e003f3c9639
container_format	bare
created_at	2017-01-24T14:52:47Z
disk_format	qcow2
file	/v2/images/dc751e7c-65d8-4f60-ae72-58abe0a780d4/file
id	dc751e7c-65d8-4f60-ae72-58abe0a780d4
min_disk	0
min_ram	0
name	centos7-sahara
owner	5ff033fe302641eb9ec1317ee7e77c1c
protected	False
schema	/v2/schemas/image
size	1575288832
status	active
tags	
updated_at	2017-01-24T14:53:08Z
virtual_size	None
visibility	shared

Register Sahara image

```
fedora@devstack-sahara ~]$ openstack dataprocessing image register centos7-sahara --username cloud-user
```

Field	Value
Description	
Id	dc751e7c-65d8-4f60-ae72-58abe0a780d4
Name	centos7-sahara
Status	ACTIVE
Tags	
Username	cloud-user

Configure Hadoop plugin

```
fedora@devstack-sahara ~]$ openstack dataprocessing image tags add centos7-sahara --tags vanilla 2.7.1
```

Field	Value
Description	
Id	dc751e7c-65d8-4f60-ae72-58abe0a780d4
Name	centos7-sahara
Status	ACTIVE
Tags	2.7.1, vanilla
Username	cloud-user

Available Hadoop plugins

```
fedora@devstack-centos: ~$ openstack dataprocessing plugin show vanilla --plugin-version 2.7.1
```

Field	Value
Description	The Apache Vanilla plugin provides the ability to launch upstream Vanilla Apache Hadoop cluster without any management consoles. It can also deploy the Oozie component.
Name	vanilla
Required image tags	2.7.1, vanilla
Title	Vanilla Apache Hadoop
Plugin version 2.7.1: enabled	True
Plugin version 2.7.1: stable	True
Plugin: enabled	True
Plugin: stable	True
Service:	Available processes:
HDFS	datanode, namenode, secondarynamenode
Hadoop	
Hive	hiveserver
JobFlow	oozie
MapReduce	historyserver
Spark	spark history server
YARN	nodemanager, resourcemanager

Create node templates

```
fedora@devstack-sahara ~$ openstack dataprocessing node group template create --name vanilla-default-master --plugin vanilla --plugin-version 2.7.1 --processes namenode,resourcemanager --flavor 2 --auto-security-group
```

Field	Value
Auto security group	True
Availability zone	None
Flavor id	2
Floating ip pool	None
Id	48c6a09f-9228-4b40-a5e8-e6ab079d8505
Is default	False
Is protected	False
Is proxy gateway	False
Is public	False
Name	vanilla-default-master
Node processes	namenode, resourcemanager
Plugin name	vanilla
Plugin version	2.7.1
Security groups	None
Use autoconfig	False
Volumes per node	0

```
fedora@devstack-sahara ~$ openstack dataprocessing node group template create --name vanilla-default-worker --plugin vanilla --plugin-version 2.7.1 --processes datanode,nodemanager --flavor 2 --auto-security-group
```

Field	Value
Auto security group	True
Availability zone	None
Flavor id	2
Floating ip pool	None
Id	835b9b18-92d1-492f-a2e9-a41cd9c2a01e
Is default	False
Is protected	False
Is proxy gateway	False
Is public	False
Name	vanilla-default-worker
Node processes	datanode, nodemanager
Plugin name	vanilla
Plugin version	2.7.1
Security groups	None
Use autoconfig	False
Volumes per node	0

Create cluster template

```
fedora@devstack-sahara ~$ openstack dataprocessing cluster template create --name vanilla-default-cluster --node-groups vanilla-default-master:1 vanilla-default-worker:3
```

Field	Value
Anti affinity	
Description	None
Domain name	None
Id	bl273b68-7ac5-46ad-82af-e6c32b93811a
Is default	False
Is protected	False
Is public	False
Name	vanilla-default-cluster
Node groups	vanilla-default-master:1, vanilla-default-worker:3
Plugin name	vanilla
Plugin version	2.7.1
Use autoconfig	False

Create cluster

```
fedora@devstack:~$ openstack dataprocessing cluster create --name my-cluster-1 --cluster-template vanilla-default-cluster --user-keypair vagrant --neutron-network private --image centos7-sahara
```

Field	Value
Anti affinity	
Cluster template id	b1273b68-7ac5-46ad-82af-e6c32b93811a
Description	None
Id	298cea5f-c1e5-480a-ad21-1cae01d5e04
Image	dc751e7c-65d8-4f60-ae72-58abe0a780d4
Info	{}
Is protected	False
Is public	False
Is transient	False
Name	my-cluster-1
Neutron management network	62a80be4-0dea-45c0-913c-4bef8229a7e0
Node groups	vanilla-default-worker:3, vanilla-default-master:1
Plugin name	vanilla
Plugin version	2.7.1
Status	Validating
Use autoconfig	False
User keypair id	vagrant

Show cluster

```
[fedora@devstack-sahara ~]$ openstack dataprocessing cluster show my-cluster-1
```

Field	Value
Anti affinity	
Cluster template id	b1273b68-7ac5-46ad-82af-e6c32b93811a
Description	None
Id	298cea5f-c1e5-480a-ad21-1caef01d5a04
Image	dc751e7c-65d8-4f60-ae72-58abe0a780d4
Info	{}
Is protected	False
Is public	False
Is transient	False
Name	my-cluster-1
Neutron management network	62a86be4-0dea-45c0-913c-4bef8229a7e0
Node groups	vanilla-default-worker:0, vanilla-default-master:0
Plugin name	vanilla
Plugin version	2.7.1
Status	Spawning
Use autoconfig	False
User keypair id	vagrant