Tổng quan về Sahara

1. Rationale

Introduction

Apache Hadoop là một tiêu chuẩn công nghiệp và áp dụng rộng rãi thực hiện MapReduce. Mục đích của dự án này là cho phép người dùng dễ dàng cung cấp và quản lý cụm Hadoop trên OpenStack. Điều đáng nói rằng Amazon cung cấp Hadoop trong nhiều năm như dịch vụ Amazon Elastic MapReduce (EMR).

Sahara nhằm cung cấp cho người sử dụng với phương tiện đơn giản để cung cấp cụm Hadoop bằng cách xác định một số thông số như phiên bản Hadoop, cụm cấu trúc liên kết, chi tiết các nút phần cứng và một vài chi tiết. Sau khi người dùng điền vào tất cả các thông số, Sahara triển khai cụm trong một vài phút. Sahara cũng cung cấp phương tiện để mở rộng cụm đã được cung cấp bằng cách thêm/gỡ bỏ các nút làm việc theo yêu cầu.

Các giải pháp sau đây sẽ giải quyết trường hợp sử dụng:

- Cung cấp nhanh chóng các cụm cluster Hadoop trên Opensatck cho Dev và QA
- Tận dụng tài nguyên không sử dụng của OpenStack IaaS Cloud.
- "Analytics như một dịch vụ" cho ad-hoc hoặc khối lượng công việc phân tích bùng phát (tương tự như AWS EMR).

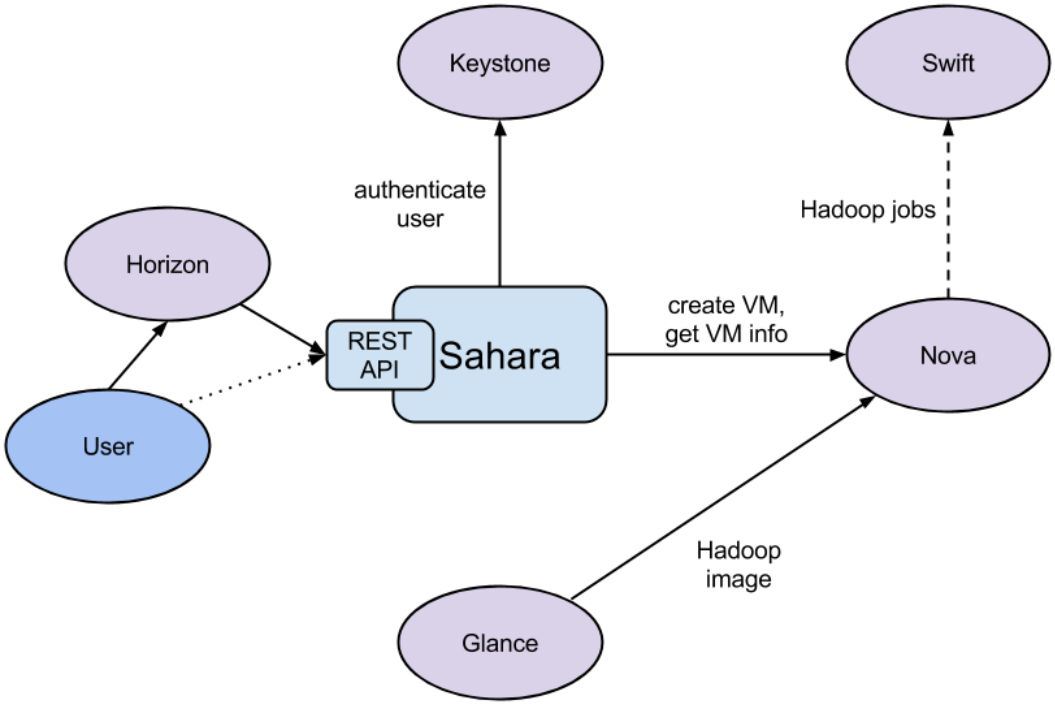
Các tính năng chính:

- Thiết kế như là một thành phần OpenStack
- Quản lý thông qua REST API với giao diện người dùng có sẵn như là một phần của OpenStack Dashboard.
- Hỗ trợ cho các bản phân phối Hadoop khác nhau:
  - Hệ thống cắm của động cơ cài đặt Hadoop
  - Tích hợp với các nhà cung cấp công cụ quản lý cụ thể, chẳng hạn như Apache Ambari hoặc Cloudera Management Console;
- Các mẫu định sẵn các cấu hình Hadoop với khả năng sửa đổi các thông số.

Details

Sahara giao tiếp với các thành phần khác của Openstack như sau:

- Horizon: Cung cấp giao diện đồ họa với khả năng sử dụng tất cả các tính năng Sahara.
- Keystone: Xác thực người dùng và cung cấp tính bảo mật , được sử dụng để làm việc với Openstack. Giới hạn quyền người sử dụng Sahara trong Openstack.
- Nova: Được sử dụng để cung cấp các máy ảo cho các cụm hadoop.
- Glance: File image máy ảo hadoop được lưu ở đây. Mỗi image đều đã được cài hệ điều hành và hadoop.
- Swift: Được sử dụng như một bộ lưu trữ dữ liệu sẽ được xử lý bằng hadoop jobs



General workflow

Sahara sẽ cung cấp 2 mức độ cho API và UI dựa trên các trường hợp sử dụng: cluster provisioning và analytics as a service.

Cung cấp nhanh các cụm cluster, quy trình công việc chung sẽ như sau:

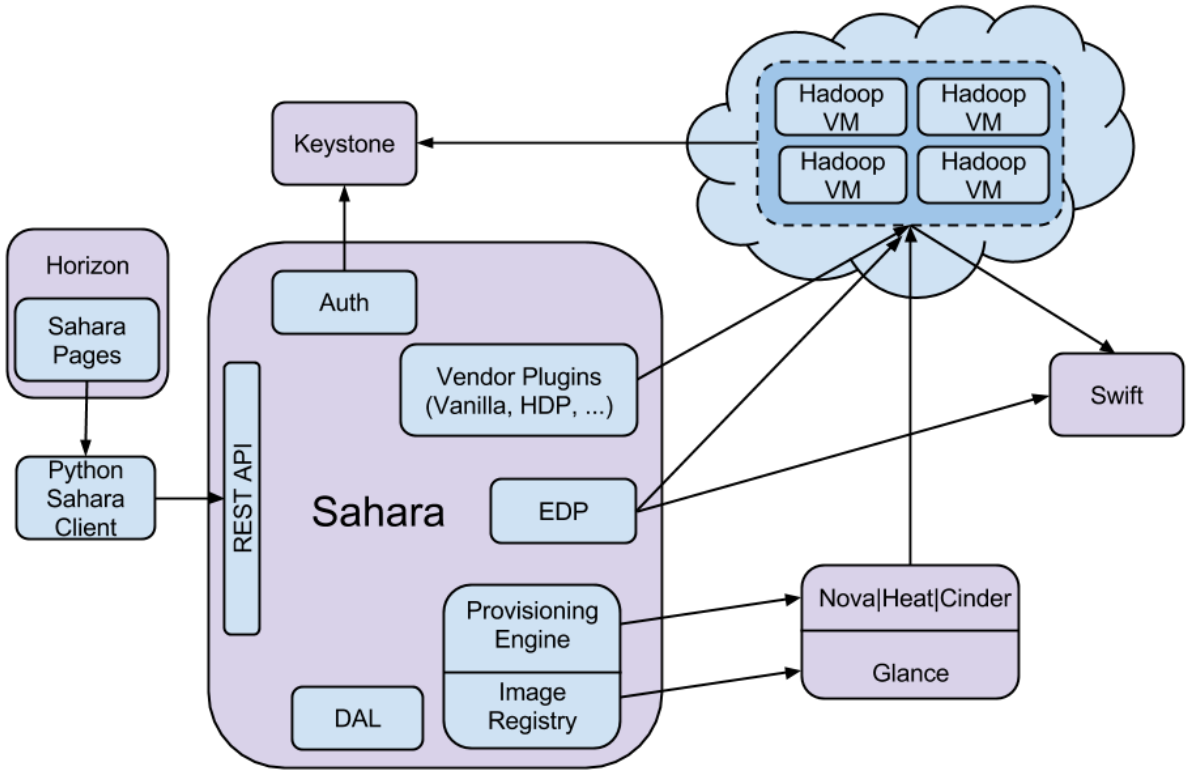
- Chọn phiên bản Hadoop.
- Chọn images đã hoặc chưa cài sẵn hadoop.
  Các file images không có hadoop cài sẵn sahara sẽ hỗ trợ các công cụ triển khai tích hợp với các công cụ của nhà cung cấp.
- Xác định cấu hình cluster, bao gồm kích thước và cấu trúc liên kết của các cụm cluster và thiết lập các kiểu khác nhau của các thông số hadoop.
  Để dễ dàng cấu hình các thông số như cơ chế của mẫu cấu hình sẽ bỏ qua.
- Cung cấp các cụm cluster: Sahara sẽ cung cấp máy ảo, cài đặt và cấu hình hadoop.
- Các hoạt động trên các cluster: Thêm/Xóa các nodes.
- Dừng/Xóa các cluster khi không sử dụng nữa.

Đối với analytic as a service, công việc chung sẽ như sau:

- Chọn một trong các phiên bản hadoop được xác định trước.

- Cấu hình job.
  - Chọn kiểu cho job: pig, hive, jar-file,...
  - Cung cấp nguồn job script hoặc vị trí jar.
  - Chọn đầu vào và vị trí đầu ra dữ liệu (ban đầu chỉ Swift sẽ được hỗ trợ).
  - Chọn vị trí cho các file logs.
- Thiết lập giới hạn kích thước cho cluster.
- Thực hiện các công việc:
  - Tất cả các cluster trích lập và thực hiện các công việc sẽ xảy ra một cách trong suốt với người dùng.
  - Cluster sẽ được gỡ bỏ tự động sau khi thực hiện xong công việc.
- Lấy về kết quả tính toán (ví dụ từ swift).

2. Architecture



Kiến trúc của Sahara:

- Auth component (thành phần auth): Chịu trách nghiệm xác thực và ủy quyền phía client và giao tiếp với Keystone.
- Data Access Layer (DAL): Vẫn tồn tại mô hình bên trong DB.
- Provisioning Engine (Công cụ cung cấp): Thành phần này chịu trách nghiệm giao tiếp với Nova, Heat, Cinder và Glance.

- Vendor Plugins (Plugins nhà cung cấp): Chịu trách nghiệm cấu hình và triển khai Hadoop trên các máy ảo được cung cấp. Các giải pháp quản lý hiện tại như Apache Ambari and Cloudera Management Console có thể được sử dụng cho vấn đề này.
- EDP - Elastic Data Processing (Xử lý dữ liệu đàn hồi): Chịu trách nghiệm lập kế hoạch và quản lý công việc Hadoop trên các cụm cluster được cung cấp bởi Sahara
- REST API: Trình bày các chức năng của Sahara thông qua REST
- Python Sahara Client: Tương tự như các thành phần khác của Openstack, Sahara có python client của riêng mình.
- Sahara pages: Giao diện cho Sahara nằm trên Horizon.