Managing a Hadoop cluster with Cloudera Manager

Mark Stetzer

@stetzer

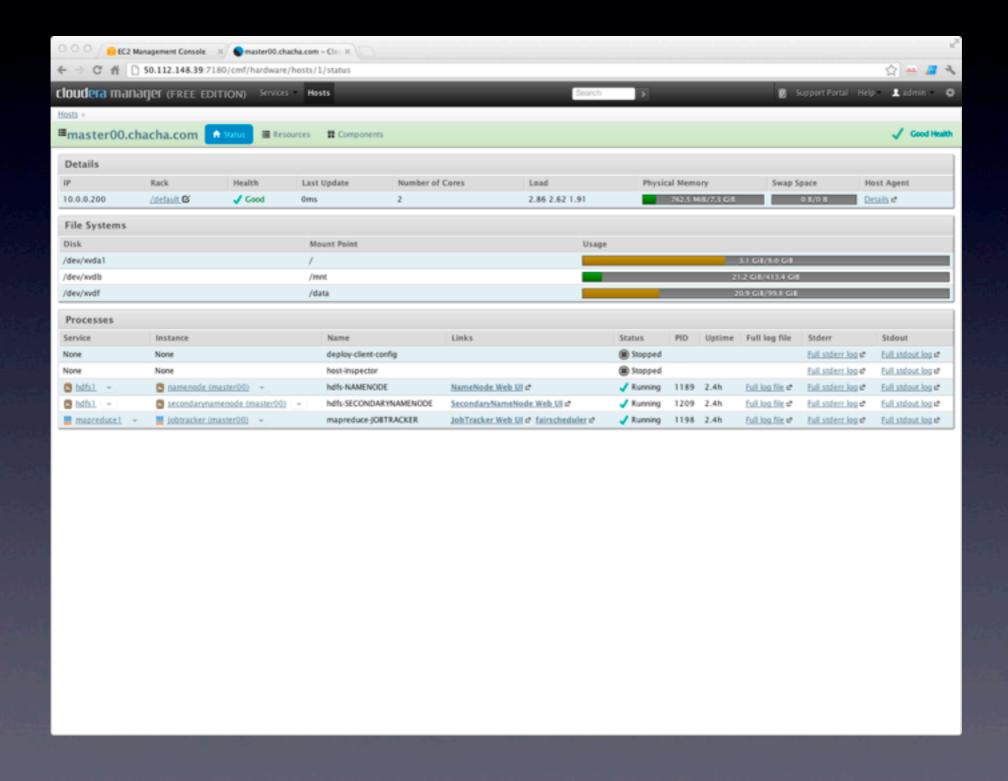
Why use CM?

- Point & click management
- Central location to configure all nodes & services

Capabilities Overview

- Hosts dashboard showing load avg, disk usage, etc.
- Bundled config downloads
- CDH3 & CDH4 support
- HDFS HA (CDH4)
- Lucid, Precise, & Squeeze support
- API
- Gateway role (a.k.a. client box)
- ...and much more!

What does it look like?



First, some AVVS setup...

Setting up a VPC

Create an Amazon Virtual Private Cloud

Cancel X

Select a VPC configuration below:

VPC with a Single Public Subnet Only

Your instances run in a private, isolated section of the AWS cloud with direct access to the Internet. Network access control lists and security groups can be used to provide strict control over inbound and outbound network traffic to your instances.

VPC with Public and Private Subnets

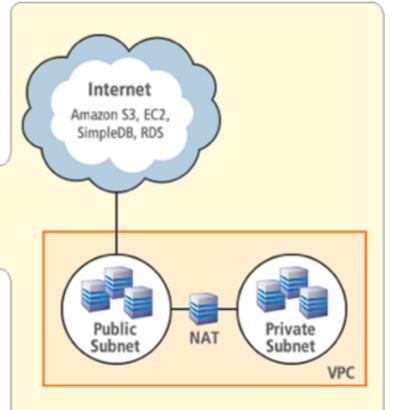
In addition to containing a public subnet, this configuration adds a private subnet whose instances are not addressable from the Internet. Instances in the private subnet can establish outbound connections to the Internet via the public subnet using Network Address Translation.

VPC with Public and Private Subnets and Hardware VPN Access

This configuration adds an IPsec Virtual Private Network (VPN) connection between your Amazon VPC and your datacenter - effectively extending your datacenter to the cloud while also providing direct access to the Internet for public subnet instances in your Amazon VPC.

OVPC with a Private Subnet Only and Hardware VPN Access

Your instances run in a private, isolated section of the AWS cloud with a private subnet whose instances are not addressable from the Internet. You can connect this private subnet to your corporate datacenter via an IPsec Virtual Private Network (VPN) tunnel.



Creates: a /16 network with two /24 subnets. Public subnet instances use Elastic IPs to access the Internet. Private subnet instances access the Internet via a Network Address Translation (NAT) instance in the public subnet. (Hourly charges for NAT instances apply)



Configuring private IPs

 Set up DHCP to give instances same internal IP after restart

Assigning elastic IPs

- Instances in private subnet use NAT to get to Internet
- Instances in public subnet need elastic IPs

Mount EBS volume(s)

- mkfs.ext4 /dev/xvdf
- mkdir -m 000 /data
- echo "/dev/xvdf /data auto noatime 0 0" |
 sudo tee -a /etc/fstab
- mount /data

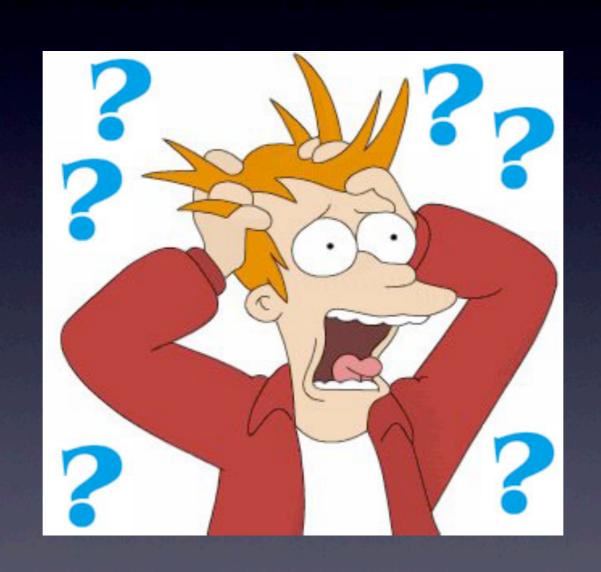
DNS

- Must configure DNS so hosts can address each other & reverse DNS works
- If instances can't identify themselves, Cloudera agents won't work correctly

Now take me through:

- Examining cluster health
- Adding a node to a cluster
- Configuring services on nodes
- Adding a new service

Questions?



Want to work with Hadoop at your day job?

http://about.chacha.com/about/careers/