cloudera

Cloudera Data Science Workbench Installation from Scratch on AWS

Partner Presales @Cloudera:
Filippo Lambiente
Alvin Heib

(greetings to Toby Fergusson)

cloudera

I. Table of Contents

cloudera

cloudera

I. Introduction

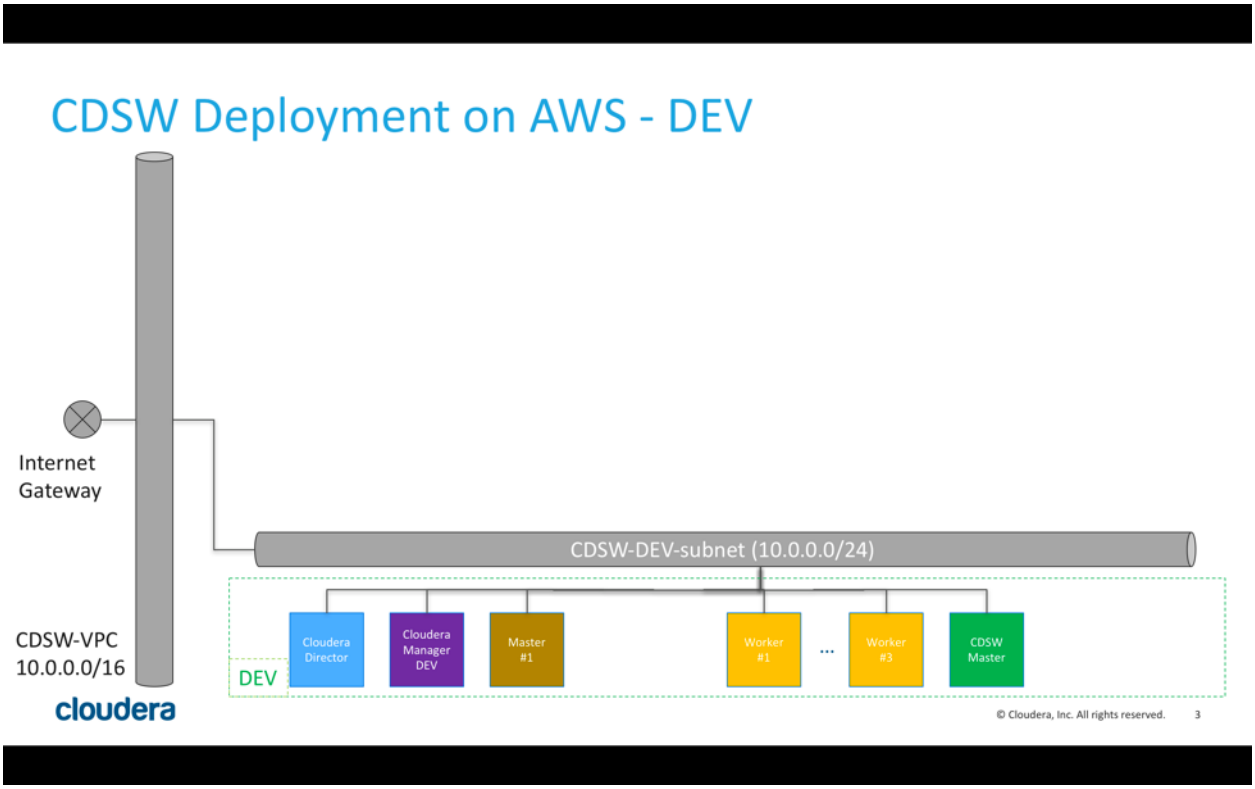
Cloudera Data Science and Engineering is the comprehensive offering for exploratory data science and machine learning at scale. It lives where your datas lives, on-premise and across public clouds. Cloudera gives data science users better access to Hadoop data with familiar and performant tools that address every stage of the modern predictive analytics workflow.

This Enablement Session will focus on creating a fully functional Cloudera Data Science Workbench service based on AWS Public Cloud. The Cloudera Data Engineering cluster will be created through Cloudera Director.

The Operations, Industrialisation and Production/Security readiness will also be covered.

The architecture of the deployment for the Enablement Session:

cloudera



There will be a Virtual Private Cloud (VPC) with a CDSW-DEV-subnet deployed for the Network Layer. This subnet is made public, and will have access to the internet.
An instance of Cloudera Director and Cloudera Manager will be deployed.
The Cloudera Cluster will be a minimal one based on 1 Master and 3 Worker Nodes.
There will be a single Cloudera Data Science Workbench node. (no edge nodes will be deployed).

Here will be the associated sizing for the AWS Deployment:

DEV					SYSTEM DISK				DATA DISK			
NODE TYPE	QTY	INSTANCE TYPE	CPU	RAM	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)
Cloudera Director	1	t2.medium	2	4	1	50	EBS	50				0
Cloudera Manager	1	t2.large	2	8	1	50	EBS	50				0
Master	1	t2.large	2	8	1	50	EBS	50				0
Worker	3	t2.xlarge	4	16	1	50	EBS	50	1	500	EBS	500
CDSW-Master	1	t2.2xlarge	8	32	1	100	GP2	100	2	500	GP2	1 000
			18	68				300				1 500

cloudera

II. Prerequisites

A. AWS Account / Quotas

Please make sur with your AWS administrator that you have an AWS account (with credentials) and sufficient quotas. We will use different Virtual Machine sizing for the different environments DEV / STG / PRD.

The minimum requirement are:
* 1 Virtual Private Cloud (VPC)
* 1 Security Group
* 7 Instances (from t2.medium to t2.2xlarge)
* 18 vCPU total
* 68 GB RAM total
* 300 GB DISK (EBS / GP2) total for OS
* 1.5 TB DISK (EBS / GP2) total for Data Store

B. AWS Command Line Interface installed (CLI)

Please got through this tutorial to correctly install your AWS CLI environment.
Installing the AWS Command Line Interface

C. Knowledge on SSH with Keypairs into a linux terminal

During the session, we will provide a keypairs to ease the Cloudera Staff debuggability. You could find private / public keypairs (**cdsw-admin** / **cdsw-admin.pub**) under this link:

Cloudera Data Science Workbench Installation from scratch on AWS - Github

Special attention for windows users, please make sure sure that you have:

- Installed putty. Putty installation link
- And converted the **cdsw-admin** private key. Putty user guide

cloudera

III. Lab #1: AWS Environment Setup

A. Prepare a directory with files for the Enablement Session

- Identify a storage location to clone or unzip all the scripts & files for our Enablement session (for ex: ~/Github/CDSW)

- Use clone (using `git clone <URL>`) or download (using the Top-Right green download button) the current git repo:
  https://github.com/heibalvin/Cloudera-Data-Science-Workbench-Installation-from-scratch-on-AWS
  You should have a directory named `Cloudera-Data-Science-Workbench-Installation-from-scratch-on-AWS`. Please rename it to `AWS-DEV`.

- You shoud have a directory hierarchy (for ex: ~/Github/CDSW/AWS-DEV) very similar to

```
drwxr-xr-x@ 12 alvinheib  staff     408 Sep 25 22:02 .
drwxr-xr-x   3 alvinheib  staff     102 Sep 25 22:11 ..
-rw-r--r--@  1 alvinheib  staff  627400 Sep 25 22:02 CDSW-Installation-from-scratch-on-AWS.pdf
-rw-r--r--@  1 alvinheib  staff    2382 Sep 25 22:02 README.md
-rw-r--r--@  1 alvinheib  staff      52 Sep 25 22:02 SECRET.properties
-rw-r--r--@  1 alvinheib  staff    3243 Sep 25 22:02 cdsw-admin
-rw-r--r--@  1 alvinheib  staff     745 Sep 25 22:02 cdsw-admin.pub
-rw-r--r--@  1 alvinheib  staff     923 Sep 25 22:02 cloudera-CDH-bootstrap-script.sh
-rw-r--r--@  1 alvinheib  staff     610 Sep 25 22:02 cloudera-director-install-script.sh
-rw-r--r--@  1 alvinheib  staff      22 Sep 25 22:02 owner_tag.properties
-rw-r--r--@  1 alvinheib  staff     170 Sep 25 22:02 provider.properties
-rw-r--r--@  1 alvinheib  staff     658 Sep 25 22:02 ssh.properties
```

- Special attention is required on the cdsw-admin & cdsw-admin.pub key pairs' permissions. We must protect it from being read by other, and from accidental deletion (no more access to your VMs). On linux, please issue a command `chmod 400 cdsw-admin*` .

- During the Enablement Session we will be intensivelly using linux environment variable. Most of them needs to be updated in the corresponding files:

cloudera

owner_tag.properties:
[YOUR-USERNAME]

provider.properties:
[YOUR-AWS-ACCESS-KEY-ID]
[YOUR-AWS-DEV-SECURITY-GROUP-ID]
[YOUR-AWS-DEV-SUBNET-ID]

SECRET.properties:
[YOUR-AWS-SECRET-ACCESS-KEY]

As well as URL for the Different Services:
[CLOUDERA-DIRECTOR-PUBLIC-DNS]
[CLOUDERA-MANAGER-PUBLIC-DNS]
[CLOUDERA-CDSW-PUBLIC-DNS]
[CLOUDERA-CDSW-PRIVATE-IP]

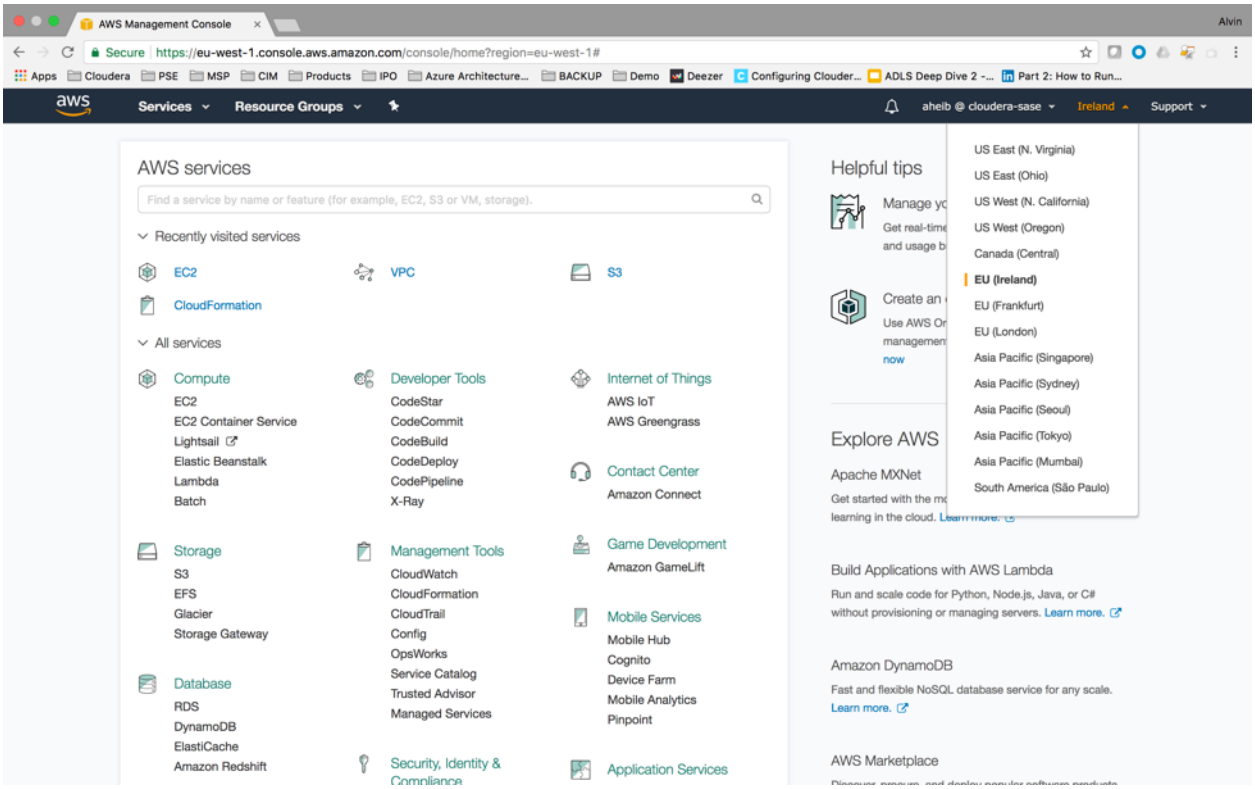
B. Log on to AWS Console

- First step is to open your favorite web browser and use the url. Then click on the button **Sign in to the Console**. You will be able to login with your AWS user/password.

  https://aws.amazon.com/console/

- For the Enablement Session, we will be using Ireland data-center (eu-west-1). Please select the **EU(Ireland)** datacenter on the menu Top-Right (next to Support)

cloudera

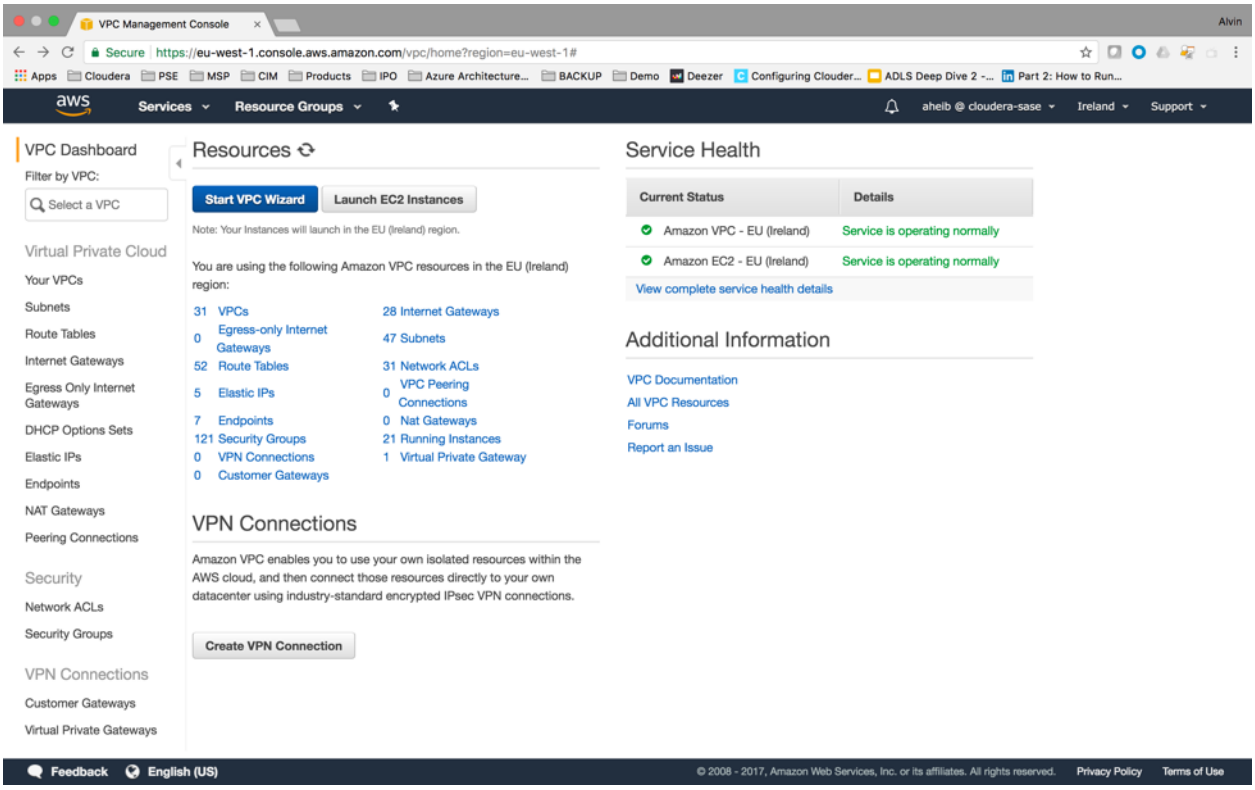


C. Create a new VPC

Click on Top-Left AWS Search Box, and search for **VPC Services**.

We will use the **VPC Wizard**, please click on the button to proceed.

cloudera



For ease of the Enablement Session, we will be creating a single DEV subnet for this VPC. Please select option **VPC with a Single Public Subnet**.

- VPC name = [YOUR-USERNAME]-CDSW-VPC for ex: aheib-CDSW-VPC.
- Subnet name = [YOUR-USERNAME]-CDSW-DEV-subnet for ex: aheib-CDSW-DEV-subnet.

D. Retrieve Subnet Information

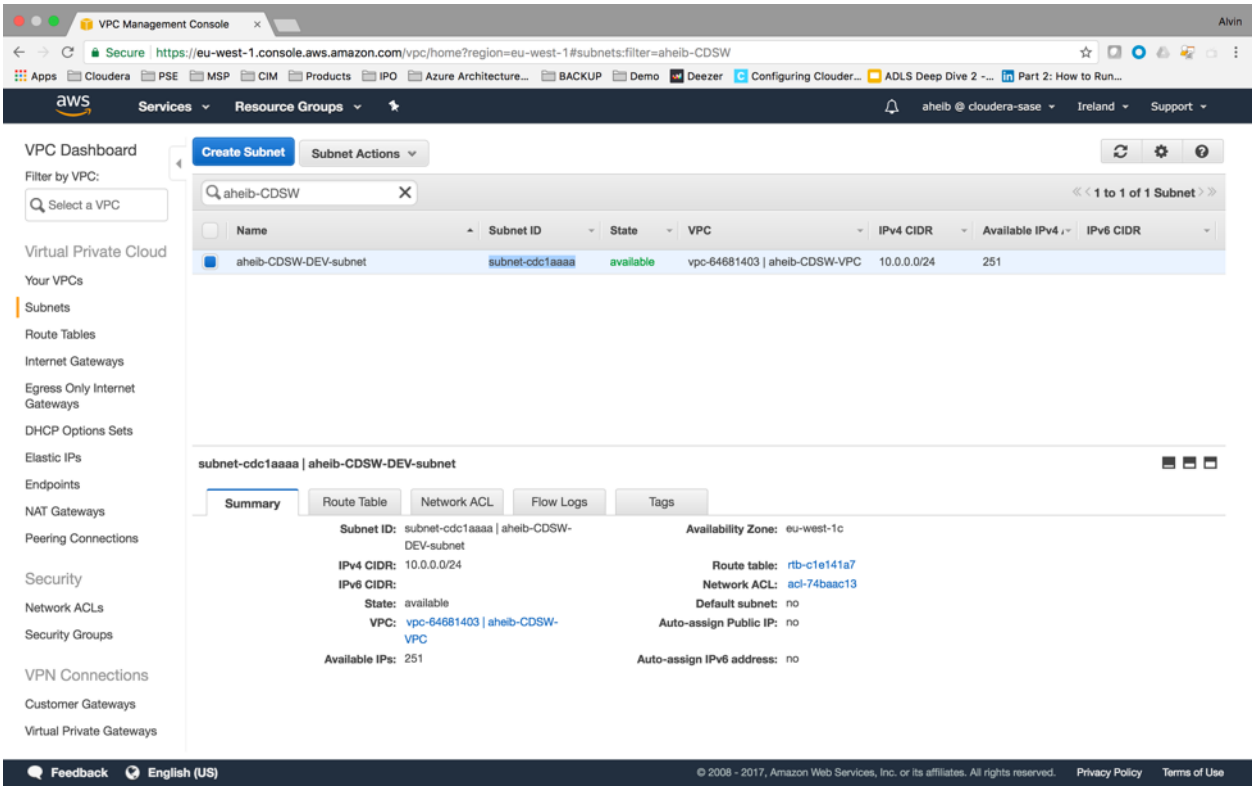
Make sure you are still in the **VPC Services** page.

Please select **Subnets** on the left menu.
And search for your newly created subnet [YOUR-USERNAME]-CDSW-DEV-subnet (for ex: aheib-CDSW-DEV-subnet). Under **subnet-id** field, you should find a **subnet-id**.

This is [YOUR-AWS-DEV-SUBNET-ID] value (for ex: subnet-cdc1aaaa)

cloudera

Please update your **provider.properties** file with the new value.



E. Create a new Security Group

Make sure you are still in the **VPC Services** page.

Please select **Security Groups** on the left menu. And click on the **Create Security Group** button.

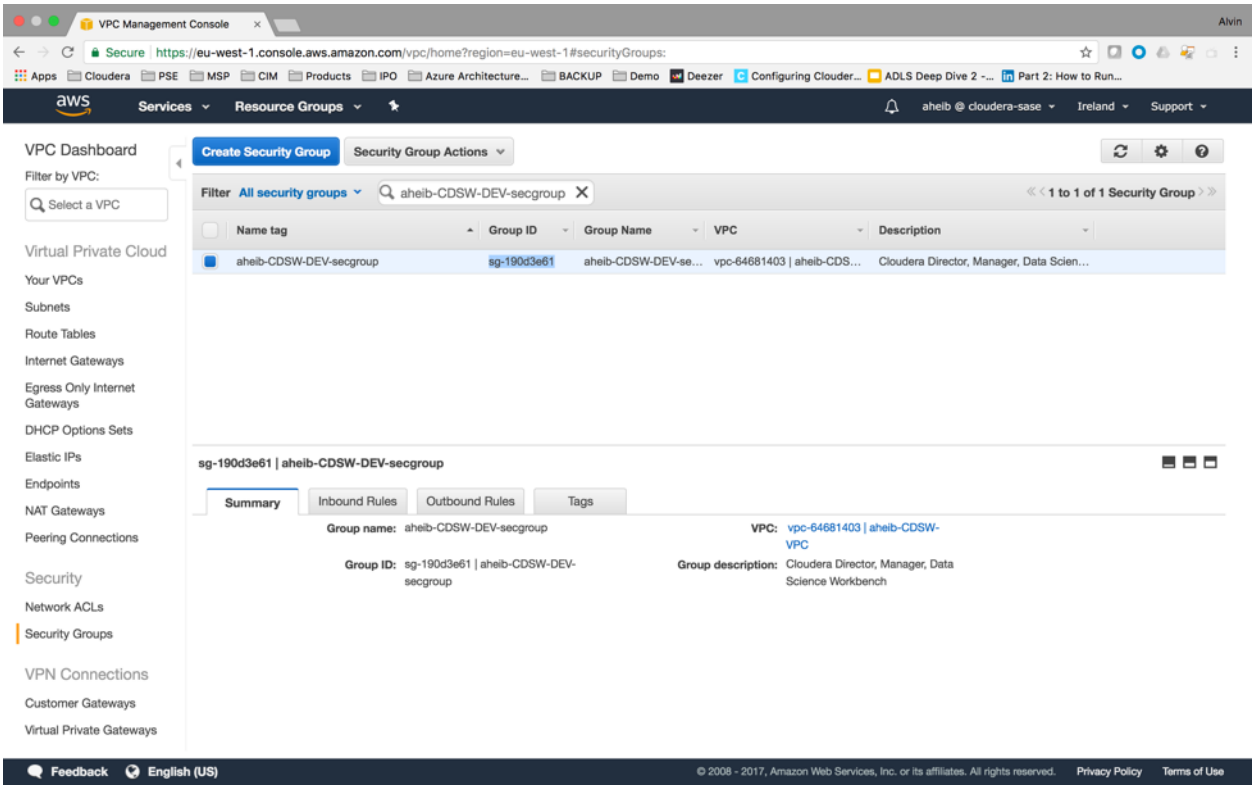
- Name tag = [YOUR-USERNAME]-CDSW-DEV-secgroup for ex: aheib-CDSW-DEV-secgroup
- Group name = [YOUR-USERNAME]-CDSW-DEV-secgroup for ex: aheib-CDSW-DEV-secgroup
- Description = Cloudera Director, Manager, Data Science Workbench
- VPC = identify your newly created VPC for ex: vpc-64681403 | aheib-CDSW-VPC

cloudera

And search for your newly created security group [YOUR-USERNAME]-CDSW-DEV-secgroup (for ex: aheib-CDSW-DEV-secgroup). Under **group-id** field, you should find a security group id.

This is [YOUR-AWS-DEV-SECGROUP-ID] value (for ex: sg-190d3e61)

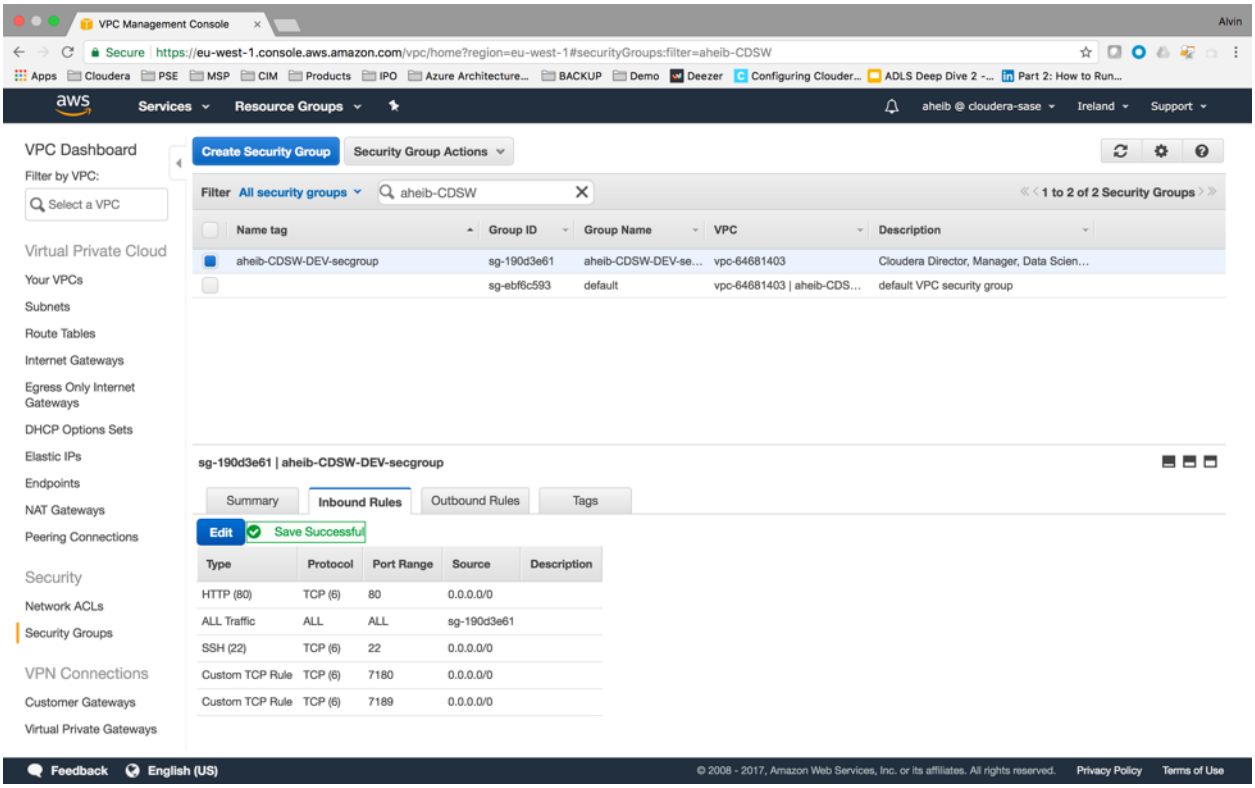
Please update your **provider.properties** file with the new value.



Then, you will need to add Security Group Rules for inbound traffic. Please select **Inbound** Tab, Edit and add rules for each entry below:

- Type = SSH, Port = 22, Source = 0.0.0.0/0 (for SSH services)

- Type = HTTP, Port = 80, Source = 0.0.0.0/0 (for Cloudera Data Science Workbench Services)
- Type = Custom TCP, Port = 7189, Source = 0.0.0.0/0 (for Cloudera Director Services)
- Type = Custom TCP, Port = 7180, Source = 0.0.0.0/0 (for Cloudera Manager Services)
- Type = All Traffic, Port = ALL, Source = [YOUR-AWS-DEV-SECGROUP-ID] (for ex: sg-190d3e61, for CDH services, no firewall on local network)

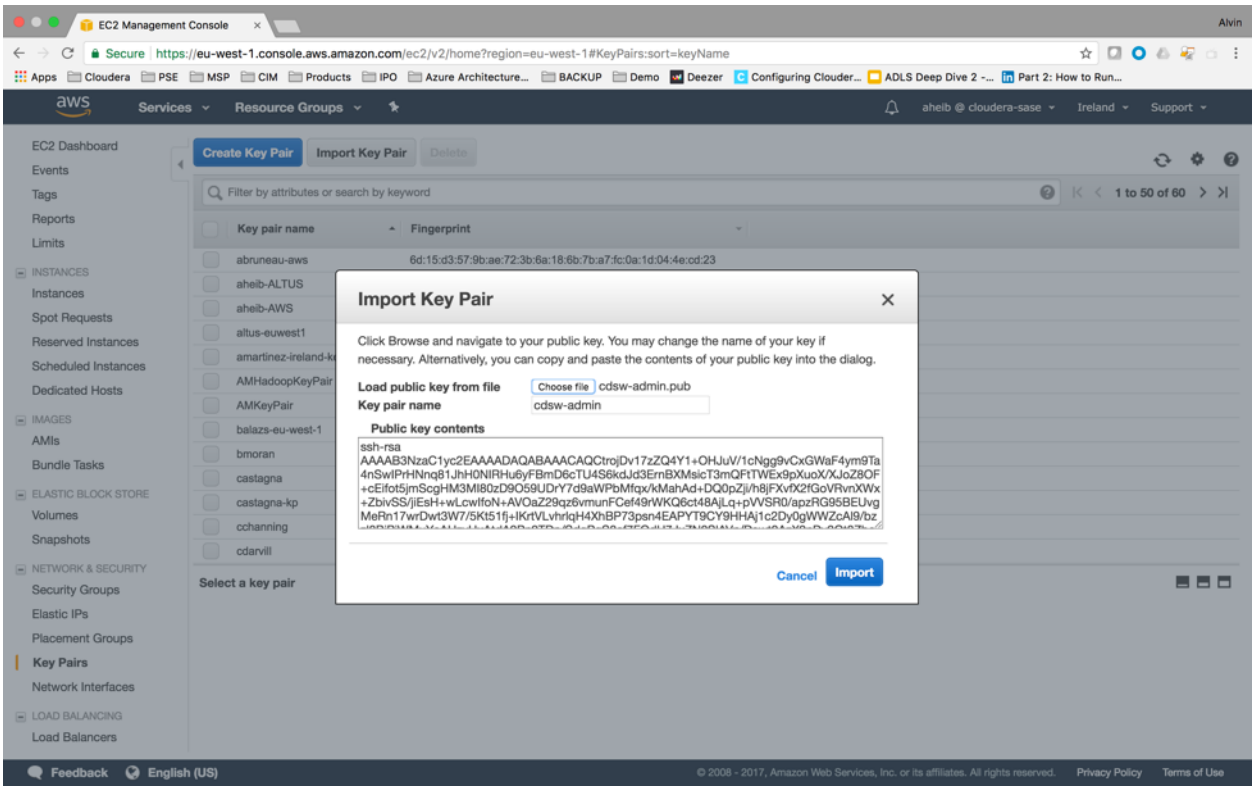


F. Import your KeyPair

Click on Top-Left AWS Search Box, and search for **EC2 Services**.

- Please select **Key Pairs** on the left menu. And click on the **Import Key Pair** button.
  Load public key from file = ~/Github/CDSW/AWS-DEV/**cdsw-admin.pub** which is in your folder.
- Key Pair name = **cdsw-admin**

cloudera


EC2 Management Console
Alvin
Secure https://eu-west-1.console.aws.amazon.com/ec2/v2/home?region=eu-west-1#KeyPairs:sort=keyName
Apps Cloudera PSE MSP CIM Products IPO Azure Architecture... BACKUP Demo Deezer Configuring Clouder... ADLS Deep Dive 2 -... Part 2: How to Run...
aws Services Resource Groups aheib @ cloudera-sase Ireland Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Spot Requests
Reserved Instances
Scheduled Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Network Interfaces
LOAD BALANCING
Load Balancers
Create Key Pair
Import Key Pair
Delete
Filter by attributes or search by keyword
1 to 50 of 60
Key pair name
Fingerprint
abruneau-aws
6d:15:d3:57:9b:ae:72:3b:6a:18:6b:7b:a7:fc:0a:1d:04:4e:cd:23
aheib-ALTUS
aheib-AWS
altus-euwest1
amartinez-ireland-k
AMHadoopKeyPair
AMKeyPair
balazs-eu-west-1
bmoran
castagna
castagna-kp
cchanning
cdarvill
Select a key pair
Import Key Pair
Click Browse and navigate to your public key. You may change the name of your key if necessary. Alternatively, you can copy and paste the contents of your public key into the dialog.
Load public key from file
Choose file
cdsw-admin.pub
Key pair name
cdsw-admin
Public key contents
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAACAQCtrojDv17zZQ4Y1+OHJuV/1cNgg9vCxGWaF4ym9Ta
4nSwIPrHNnq81JhH0NIRHu6yFBmD6cTU4S6kdJd3ErnBXMsicT3mQFtTWEx9pXuoX/XJoZ8OF
+cEifot5jmScgHM3MI80zD9O59UDrY7d9aWPbMfqx/kMahAd+DQ0pZji/h8jFXvfX2fGoVRvnXWx
+ZbivSS/jiEsH+wLcwIfoN+AVOaZ29qz6vmunFCef49rWKQ6ct48AjLq+pVVSR0/apzRG95BEUvg
MeRn17wrDwt3W7/5Kt51fj+IKrtVLvhrlqH4XhBP73psn4EAPYT9CY9HHAj1c2Dy0gWWZcAl9/bz
Cancel
Import
Feedback English (US)
© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

cloudera

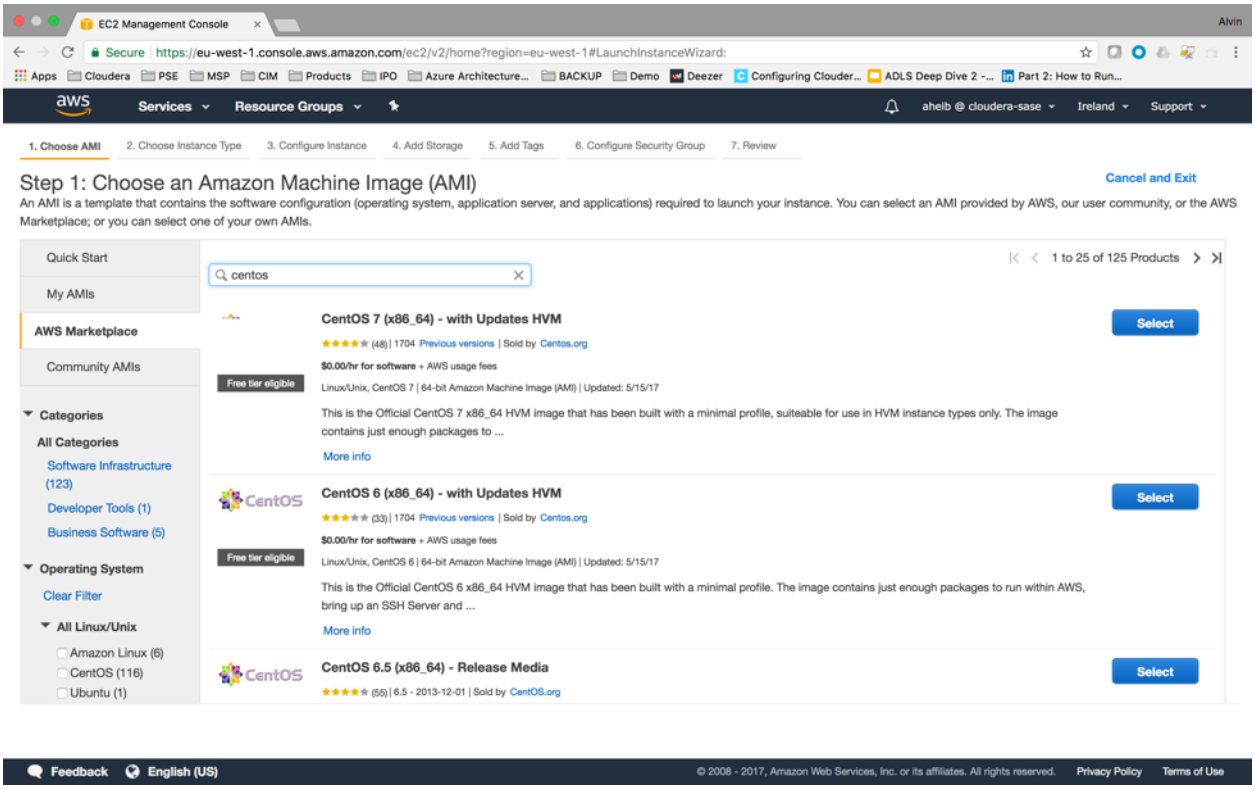
IV. Lab #2: Cloudera Director Setup

A. Deploy your Cloudera Director Instance

Make sure you are still in the **EC2 Services** page.

Please select **EC2 Dashboard** on the left menu. And click on the **Launch Instance** button.

- Choose AMI.
  We will be using CentOS 7 type of images for this Enablement Session.
  Please select **AWS Marketplace** on left-menu.
  Then search for **centos** images.
  For this Enablement Session we will be using exclusively **CentOS 7 (x86_64) - with Updates HVM**.



cloudera

- Choose Instance
  Type. For instance type, you will need to select suitable virtual machine for Cloudera Director. Please select **t2.medium** (for a DEV environment)

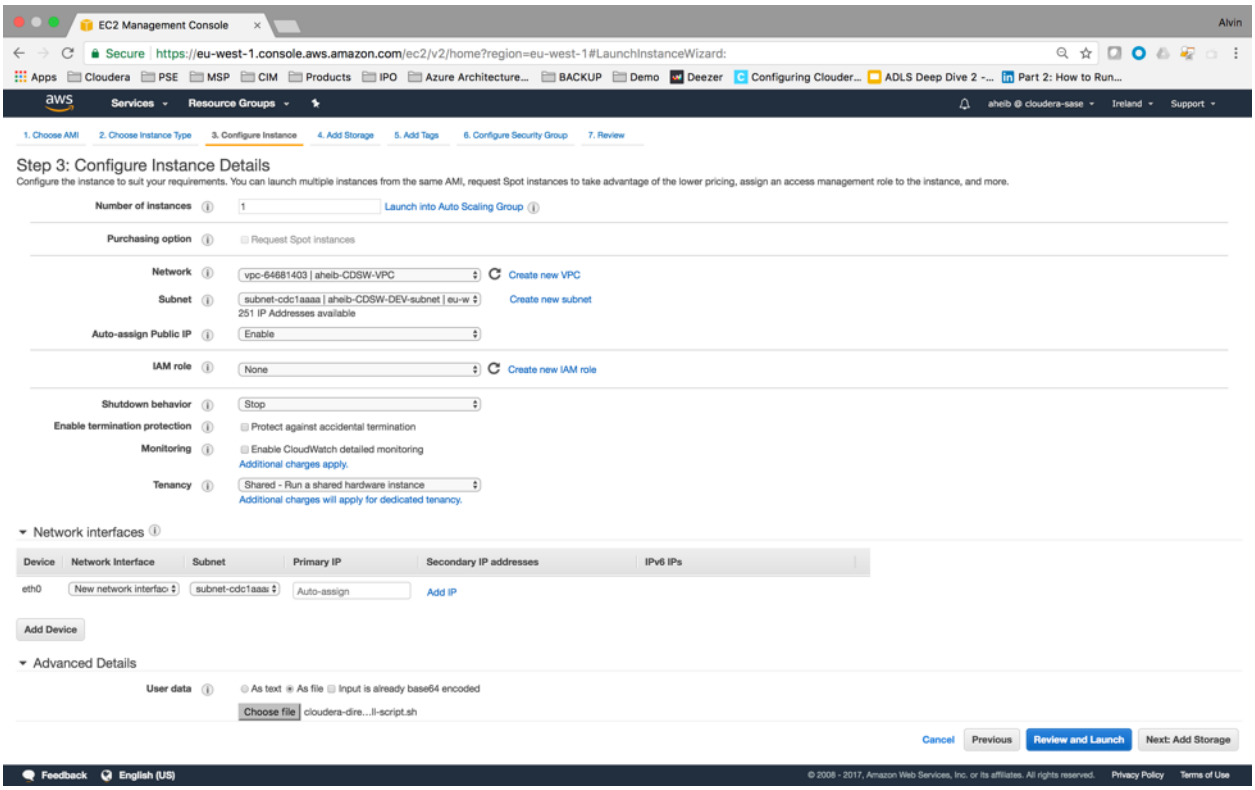
- Configure Instance.

  Select the correct Network (for ex: vpc-64681403 | aheib-CDSW-VPC).

  Select the correct Subnet (for ex: subnet- cdc1aaaa I aheib-CDSW-DEV-subnet I eu-west-1).

  /!\ It is important to set to Enable the flag Auto-assign Public IP /!

  Before moving to the next step, please make sure to add the Cloudera Director install scripts. To do so, please open up Advanced Details and import file: **~/Github/CDSW/AWS-DEV/cloudera-director-install.sh**



cloudera

- Add Storage
  Please modify the storage capacity from 8 -> 30 GB.

- Add Tags.
  We will be adding 2 tags, which represents the Name of the Instance and the Owner of the instance.
  - Name = [YOUR-USERNAME]-CDSW-DEV-Director for ex: aheib-CDSW-DEV-Director
  - owner = [YOUR-USERNAME] for ex: aheib

- Configure Security Group
  Please select and existing Security Group.
  And select the Security Group `[YOUR-USERNAME]-CDSW-DEV-secgroup`.

- Review.
  You will then see a pop-up window asking you the correct Key Pair to be used. Please select the previously imported **cdsw-admin** keypair.

B. Identify your Cloudera Director Instance

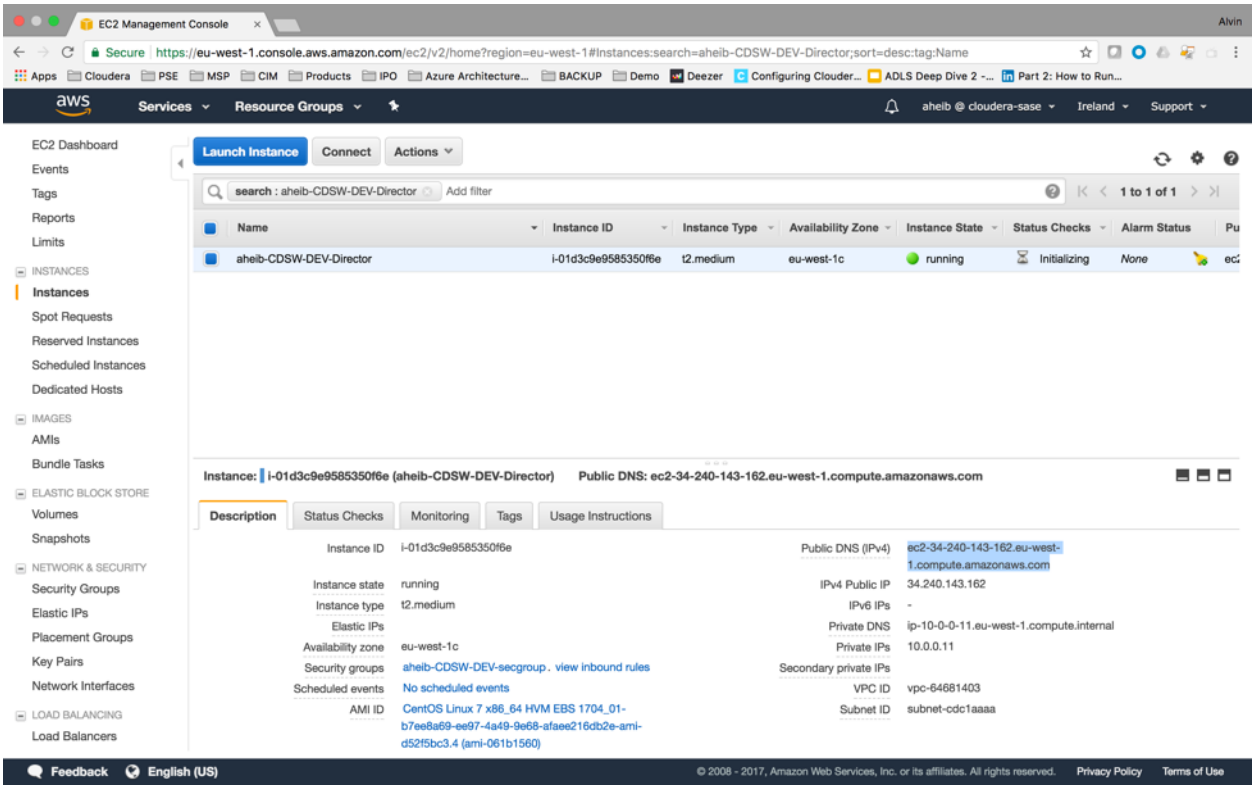
Make sure you are still in the **EC2 Services page** and select **Instances** on the left menu.

Search for your newly created Director Instance **[YOUR-USERNAME]-CDSW-DEV-Director** (for ex: aheib-CDSW-DEV-Director)

Once you have clicked on the instance, you will find in bottom view an entry called Public DNS (IPv4).

This is your [CLOUDERA-DIRECTOR-PUBLIC-DNS] environment variable (for ex: ec2-34-240-143-162.eu-west-1.compute.amazonaws.com).

cloudera



C. Log onto Cloudera Director

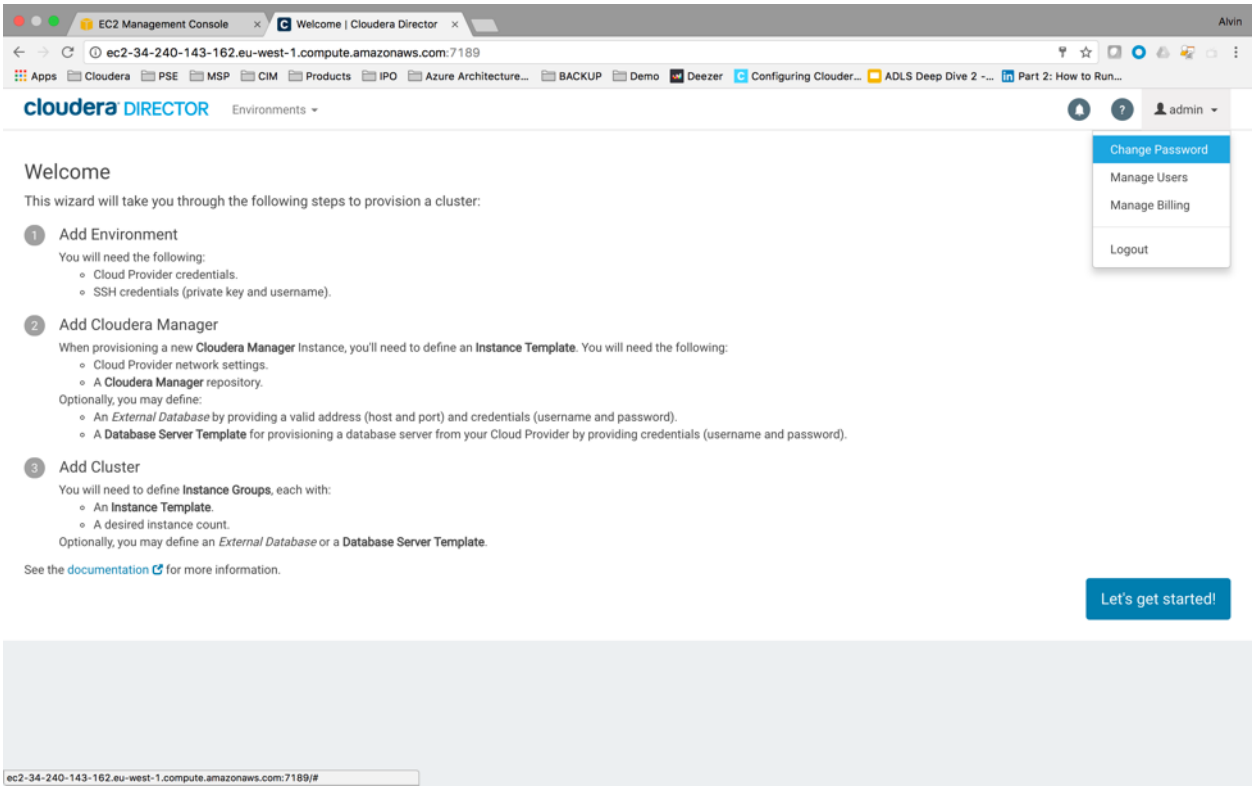
- First step is to open your favourite web browser and use the url: http://ec2-34-240-143-162.eu-west-1.compute.amazonaws.com:7189/
- Then, you will then be asked to accept the Cloudera Director End User License T&C. Please proceed.
- Finally, you will be able to login with the default login/password (**admin/admin**).

D. Modify your Default Password

On the welcome page, you will have a complete access to Cloudera Director options. Please click on **Admin Menu** on top-right and select **Change Password**.

cloudera

For the Enablement debug ability purpose, please use this password:
password = Cloudera_123



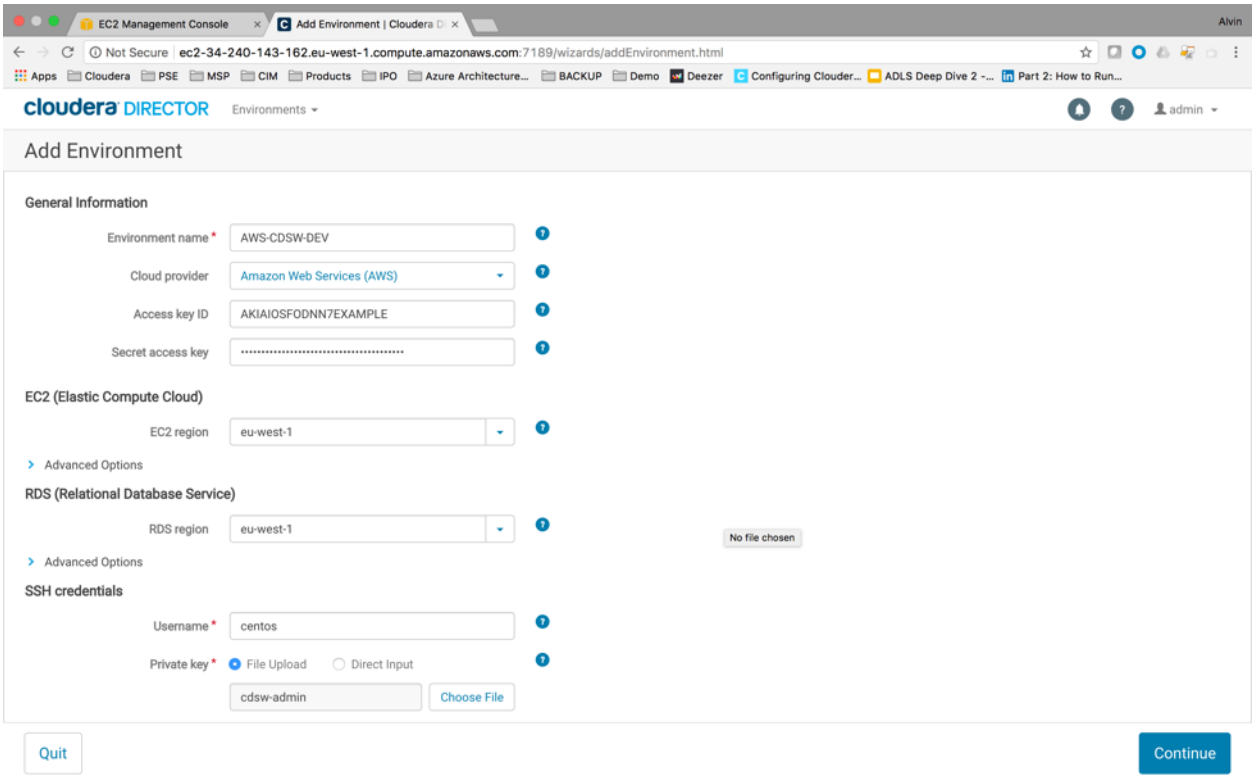
cloudera

V. Lab #3: Cloudera Data Engineering Cluster Setup

A. Create your AWS-DEV Environment

Please click on **Environment** and select **Add Environment**.

We will then enter information related to our AWS account.



Key	Value	Comments
Environment Name	AWS-CDSW-DEV	in our example

cloudera

Key	Value	Comments
Cloud Provider	AWS	in our example (could be also Azure / GCE)
Access Key Id	[YOUR-AWS-ACCESS-KEY-ID]	for ex: AKIAIOSFODNN7EXAMPLE
Secret Access Key	[YOUR-AWS-SECRET-ACCESS-KEY]	for ex: wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY
EC2 Region	eu-west-1	in our example
RDS Region	eu-west-1	in case we are using a managed AWS DB for Cloudera Manager, not covered in our example
Username	centos	username for the centos 7 image on AWS with root priviliedges **"/!\ It is not your AWS username /!"**
Private Key	cdsw-admin	Link to file ~/Github/CDSW/AWS/cdsw-admin

B. Create Node Templates

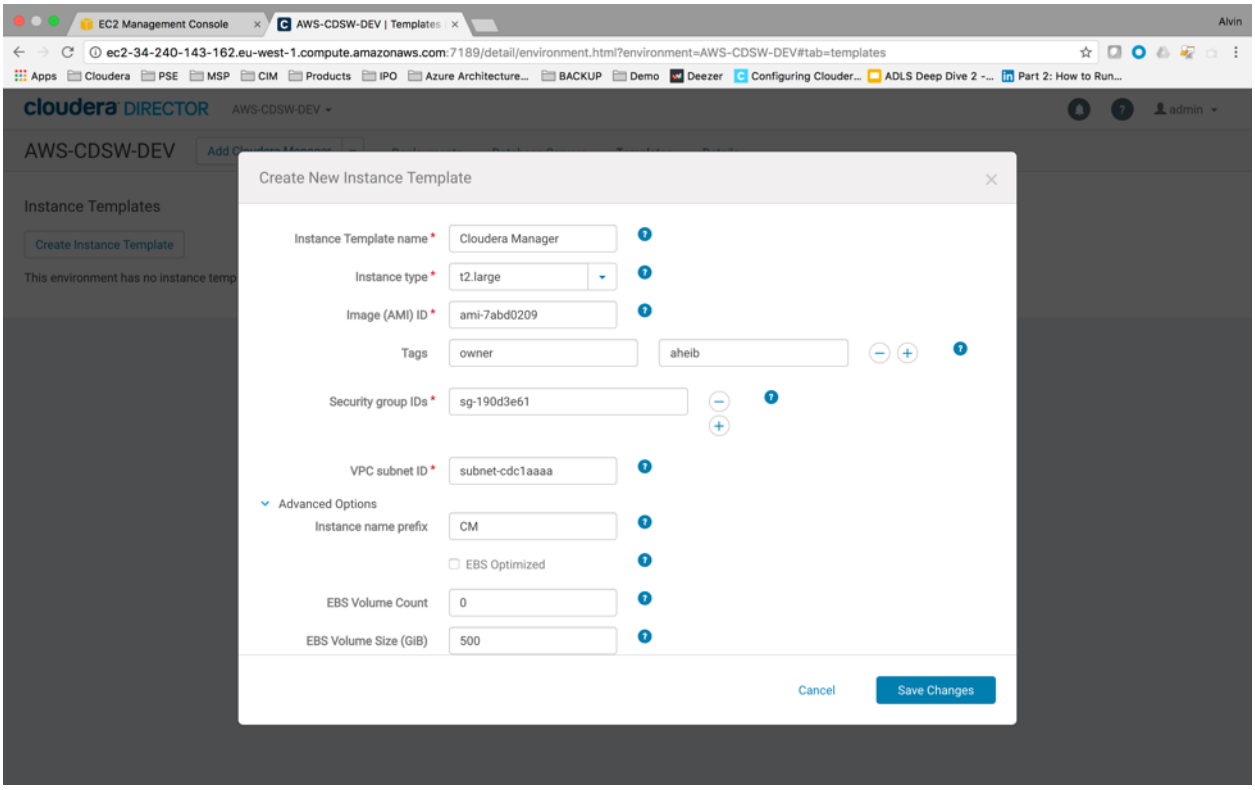
We will not proceed with the Cloudera Manager creation directly, instead we will create templates for Cloudera Manager / Master / Worker and CDSW-Master nodes.

Select the newly created environment in menu **Environment** and select **AWS-CDSW-DEV**. If requested, please Leave the current web page.

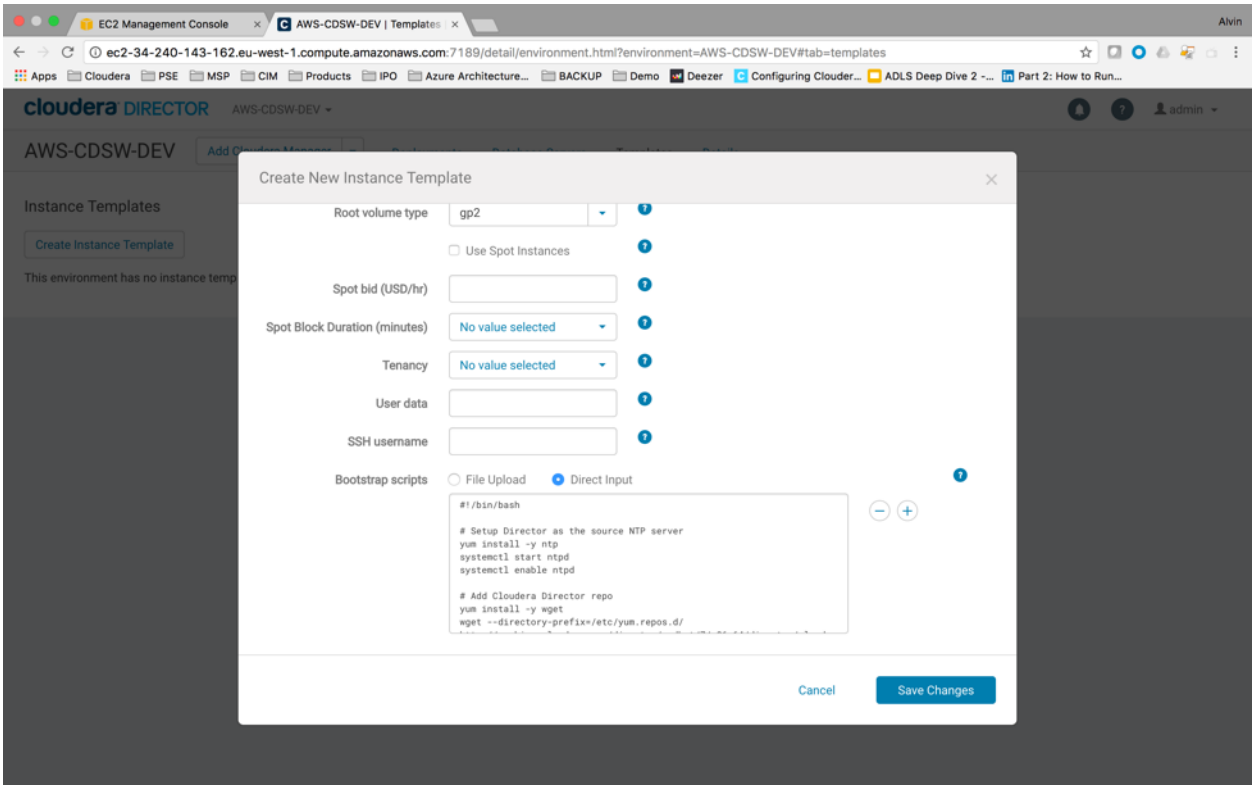
Select **Templates** menu, on top menu row. Then click on **Create Instance Template**.

- Create a **Cloudera Manager** Node template.
  Which is the longest part, since we will use Cloudera Directors Template copy feature to create all the others (Master, Worker, etc ...)

cloudera


EC2 Management Console
AWS-CDSW-DEV | Templates
Alvin
ec2-34-240-143-162.eu-west-1.compute.amazonaws.com:7189/detail/environment.html?environment=AWS-CDSW-DEV#tab=templates
Apps
Cloudera
PSE
MSP
CIM
Products
IPO
Azure Architecture...
BACKUP
Demo
Deezer
Configuring Clouder...
ADLS Deep Dive 2 -...
Part 2: How to Run...
cloudera DIRECTOR
AWS-CDSW-DEV
admin
AWS-CDSW-DEV
Instance Templates
Create Instance Template
This environment has no instance temp
Create New Instance Template
Instance Template name *
Cloudera Manager
Instance type *
t2.large
Image (AMI) ID *
ami-7abd0209
Tags
owner
aheib
Security group IDs *
sg-190d3e61
VPC subnet ID *
subnet-cdc1aaaa
Advanced Options
Instance name prefix
CM
EBS Optimized
EBS Volume Count
0
EBS Volume Size (GiB)
500
Cancel
Save Changes

cloudera



Key	Value	Comments
Template Name	Cloudera Manager	in our example
Instance Type	t2.large	for a DEV environment
Image (AMI ID)	ami-7abd0209	which is the AMI ID for eu-west-1 region for CentOS 7
Tags	(owner, [YOUR-USERNAME])	for ex: (owner, aheib)
Security Group Id	[YOUR-AWS-DEV-SECURITY-GROUP-ID]	for ex: sg-190d3e61
Subnet Id	[YOUR-AWS-DEV-	for ex: subnet-cdc1aaaa

cloudera

Key	Value	Comments
	SUBNET-ID]	
Instance Name Prefix	CM	this will help in identifying the node type under AWS Console
Bootstrap scripts	cloudera-CDH-bootstrap-script.sh	Link to file ~/Github/CDSW/AWS-DEV/ cloudera-CDH-bootstrap-script.sh

Once completed you will obtain a first node template. And then, we could start copying this template and create the following ones.

- Create a **Master Node** template.
  Copy Template from the Cloudera Manager Node template, and proceed to the specific updates.

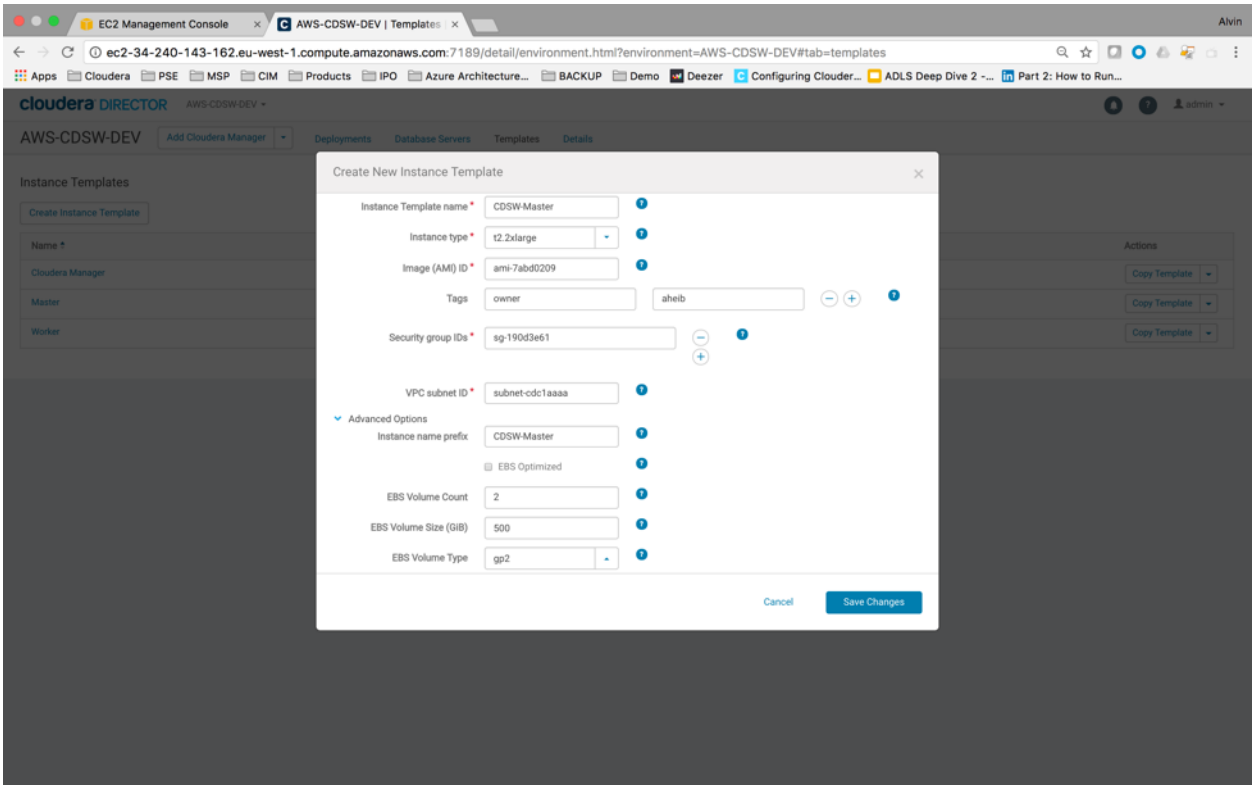
Key	Value	Comments
Template Name	Master	in our example
Instance Name Prefix	Master	this will help in identifying the node type under AWS Console

- Create a **Worker Node** template.
  Copy Template from the Cloudera Manager Node template, and proceed to the specific updates.

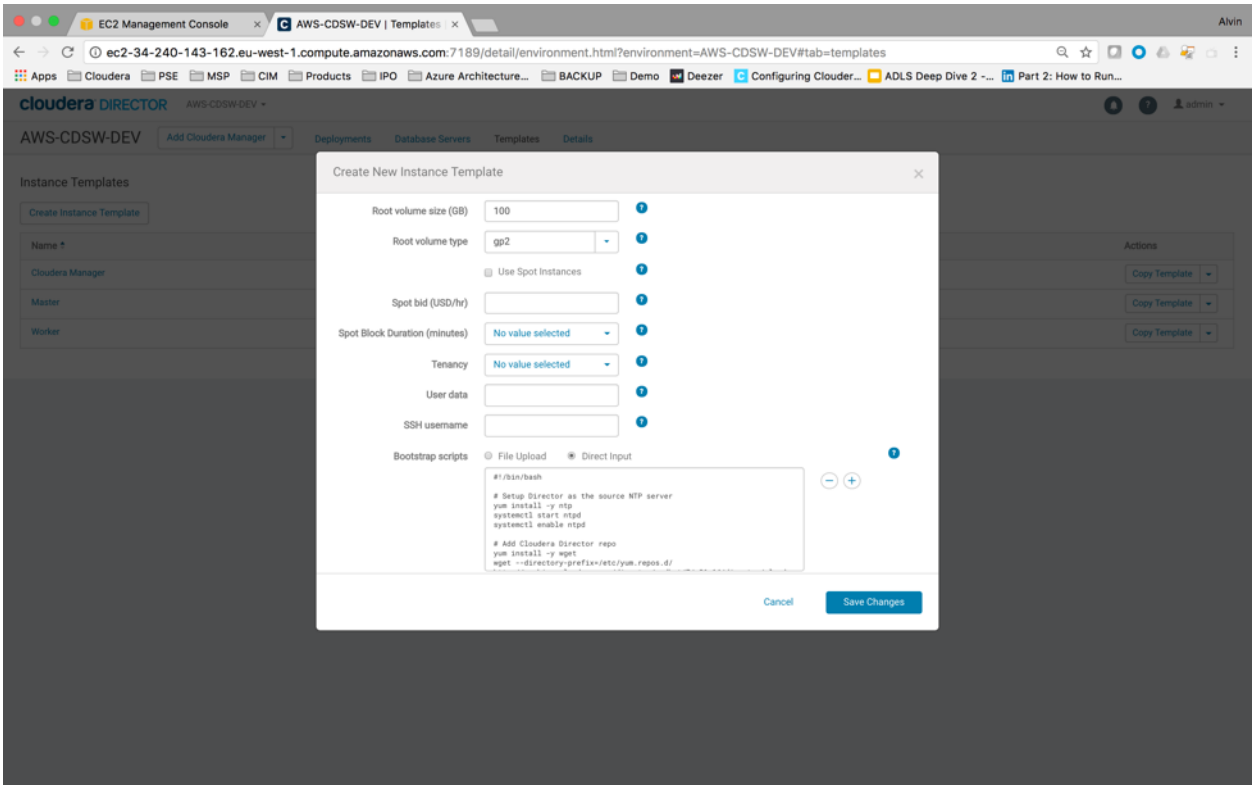
Key	Value	Comments
Template Name	Worker	in our example
Instance Type	t2.xlarge	for a DEV environment
Instance Name Prefix	Worker	this will help in identifying the node type under AWS Console
EBS Volume Count	1	we will need a 1 * 500 GiB HDFS disk
EBS Volume Size	500	we will need a 1 * 500 GiB HDFS disk

cloudera

- Create a **CDSW-Master** Node template.
  Copy Template from the Cloudera Manager Node template, and proceed to the specific updates.



cloudera

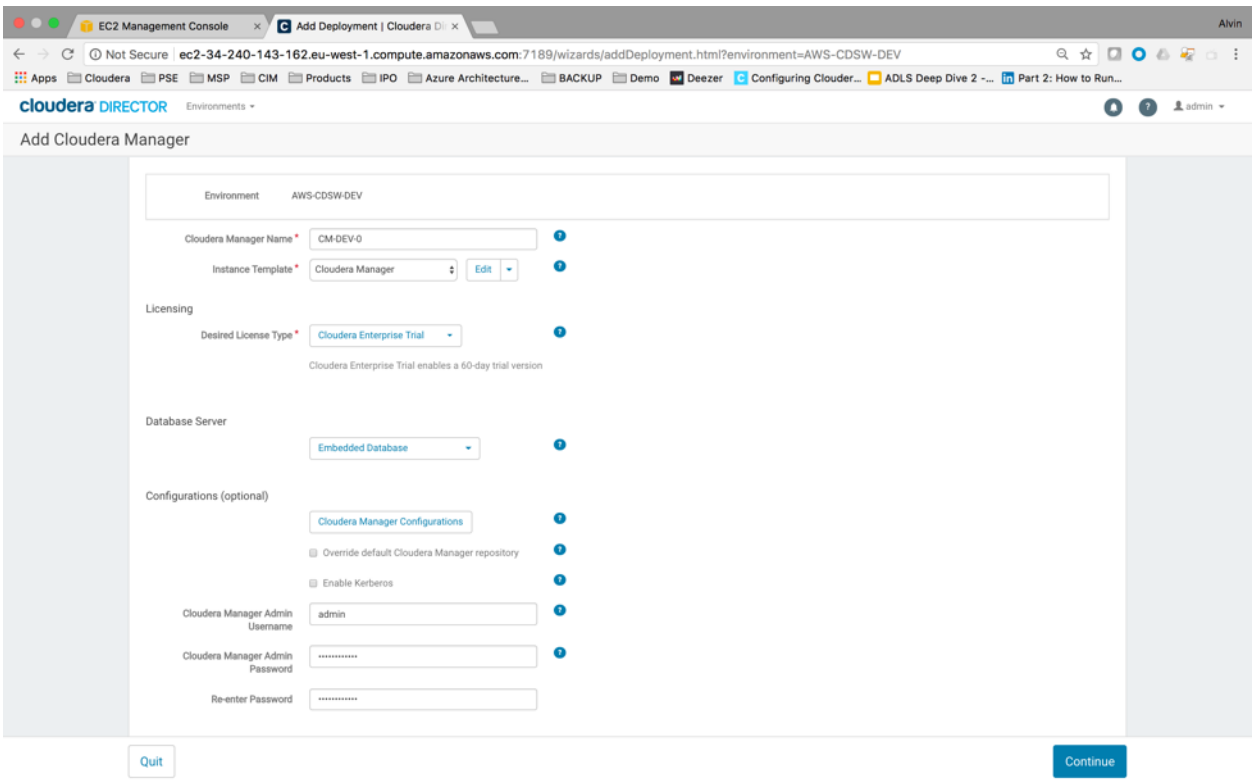


Key	Value	Comments
Template Name	CDSW-Master	in our example
Instance Type	t2.2xlarge	for a DEV environment
Instance Name Prefix	CDSW-Master	this will help in identifying the node type under AWS Console
EBS Volume Count	2	we will need a 2 * 500 GiB HDFS disk
EBS Volume Size	500	we will need a 2 * 500 GiB HDFS disk
EBS Volume Type	gp2	in our example
Root Volume Size	100	in our example

cloudera

C. Deploy your first Cloudera Manager Instance

Make sure you are still in the **AWS-CDSW-DEV** environment, then select **Add Cloudera Manager**.



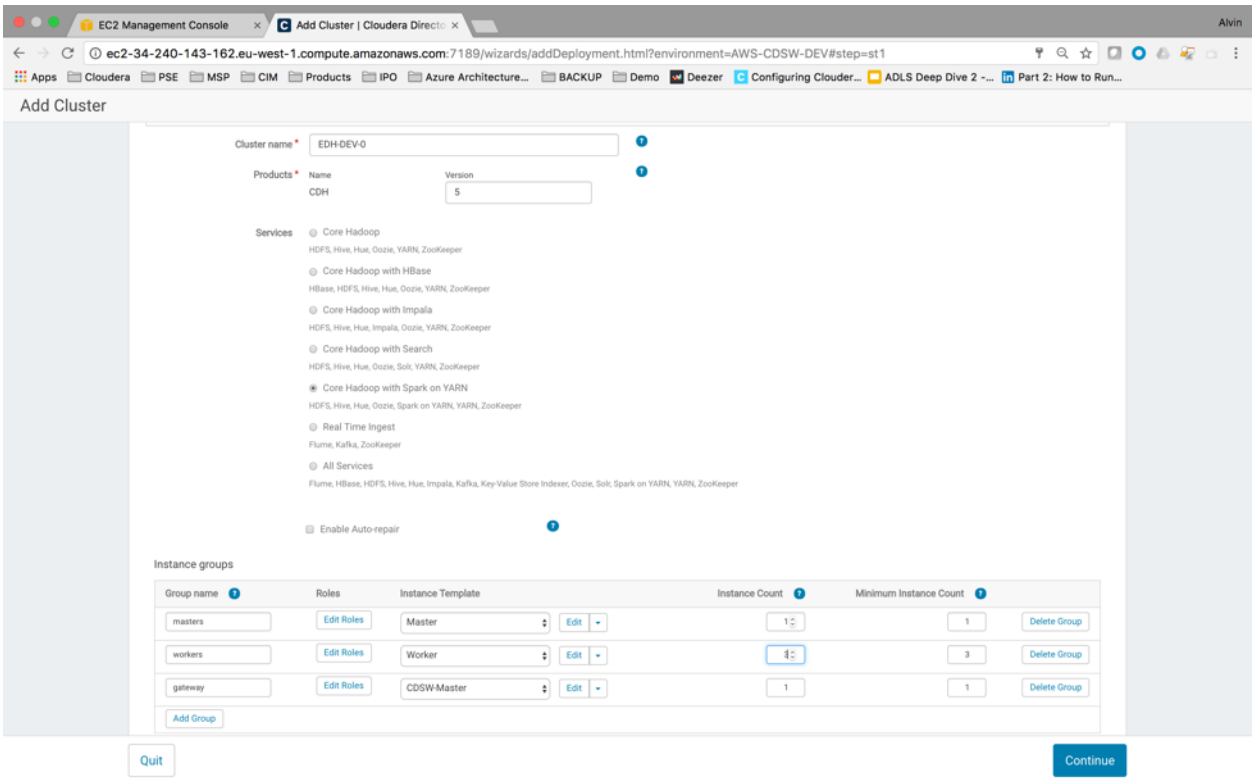
Key	Value	Comments
Cloudera Manager Name	CM-DEV-0	in our example
Instance Template	Cloudera Manager	the one we created on previous steps
License Type	Cloudera Enterprise Trial	in our example, and expires within 30 days

cloudera

Key	Value	Comments
Database Server	Embedded Database	in our example, should be an external DB for production
Cloudera Manager Admin Username	admin	please use this username for enablement session debugging ease
Cloudera Manager Admin Username	Cloudera_123	please use this password for enablement session debugging ease

D. Deploy your first Cloudera Cluster

You will automatically reach the wizard to create your first cluster.



cloudera

Key	Value	Comments
Cluster Name	EDH-DEV-0	
Services	Core Hadoop with Spark on Yarn	Spark is needed for CDSW, all Services is also possible

On the same page, you will need to define the cluster architecture (number of master, slave, CDSW-master nodes).

- for **masters** group name, choose **Master** node template and set the instance count to **1**.
- for **workers** group name, choose **Worker** node template and set the instance count to **3**
- for **gateways** group name, choose **CDSW-Master** node template and set the instance count to **1**.

Your cluster (Cloudera Manager, Master, Workers and CDSW-Master) should be deploying in background. You should observe after 8 min, that the entire cluster is deployed.

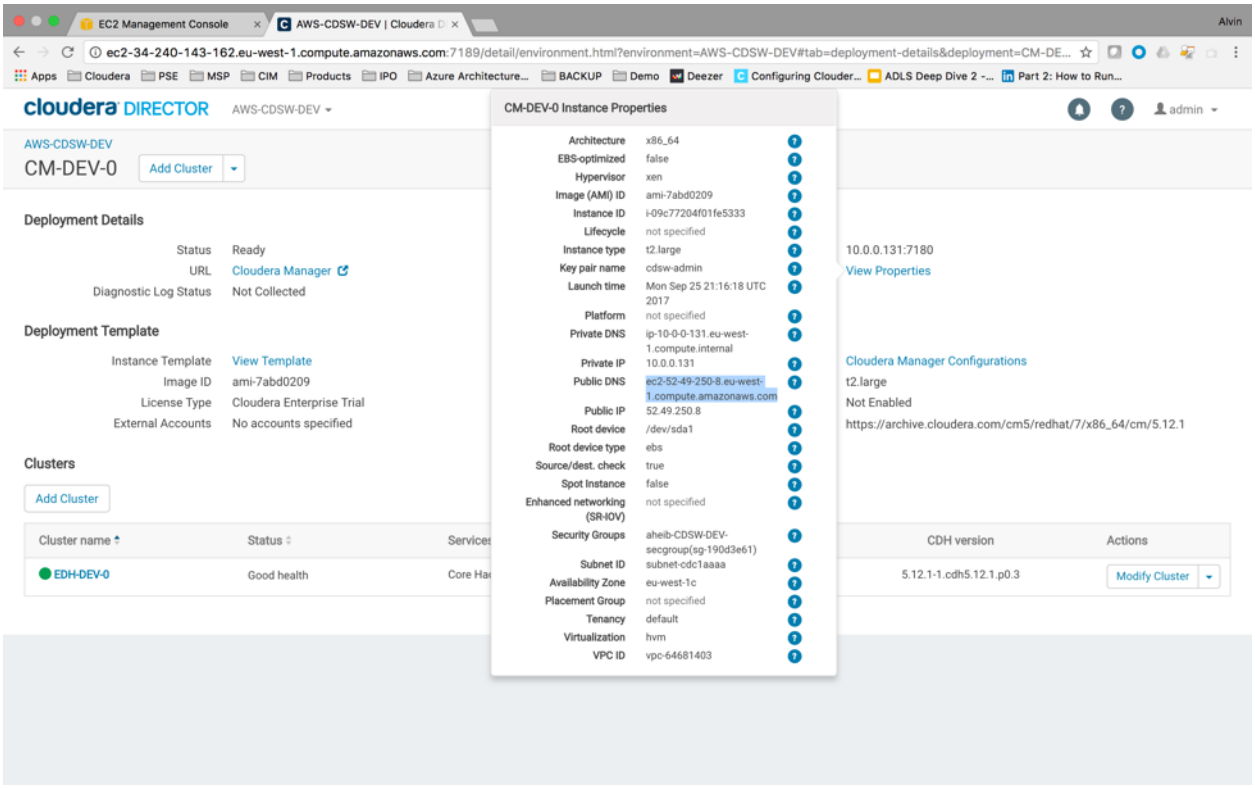
E. Identify your Cloudera Manager

You have a complete view of your cluster deployed from Cloudera Director.

Click on your newly deployed Cloudera Manager called **CM-DEV-0**.

When you extend the **View Properties**, you will identify your **[CLOUDERA-MANAGER-PUBLIC-DNS]** (for ex: ec2-52-49-250-8.eu-west-1.compute.amazonaws.com)

cloudera


EC2 Management Console
AWS-CDSW-DEV | Cloudera
Alvin
ec2-34-240-143-162.eu-west-1.compute.amazonaws.com:7189/detail/environment.html?environment=AWS-CDSW-DEV#tab=deployment-details&deployment=CM-DE...
Apps
Cloudera
PSE
MSP
CIM
Products
IPO
Azure Architecture...
BACKUP
Demo
Deezer
Configuring Clouder...
ADLS Deep Dive 2 -...
Part 2: How to Run...
cloudera DIRECTOR
AWS-CDSW-DEV
admin
AWS-CDSW-DEV
CM-DEV-0
Add Cluster
Deployment Details
Status
Ready
URL
Cloudera Manager
Diagnostic Log Status
Not Collected
Deployment Template
Instance Template
View Template
Image ID
ami-7abd0209
License Type
Cloudera Enterprise Trial
External Accounts
No accounts specified
Clusters
Add Cluster
Cluster name
Status
Services
CDH version
Actions
EDH-DEV-0
Good health
Core Ha
5.12.1-1.cdh5.12.1.p0.3
Modify Cluster
CM-DEV-0 Instance Properties
Architecture
x86_64
EBS-optimized
false
Hypervisor
xen
Image (AMI) ID
ami-7abd0209
Instance ID
i-09c77204f01fe5333
Lifecycle
not specified
Instance type
t2.large
Key pair name
cdsw-admin
Launch time
Mon Sep 25 21:16:18 UTC 2017
Platform
not specified
Private DNS
ip-10-0-0-131.eu-west-1.compute.internal
Private IP
10.0.0.131
Public DNS
ec2-52-49-250-8.eu-west-1.compute.amazonaws.com
Public IP
52.49.250.8
Root device
/dev/sda1
Root device type
ebs
Source/dest. check
true
Spot Instance
false
Enhanced networking (SR-IOV)
not specified
Security Groups
aheib-CDSW-DEV-secgroup(sg-190d3e61)
Subnet ID
subnet-cdc1aaaa
Availability Zone
eu-west-1c
Placement Group
not specified
Tenancy
default
Virtualization
hvm
VPC ID
vpc-64681403
10.0.0.131:7180
View Properties
Cloudera Manager Configurations
t2.large
Not Enabled
https://archive.cloudera.com/cm5/redhat/7/x86_64/cm/5.12.1

cloudera

VI. Lab #4: Upgrade Cluster to Spark 2

A. Install Spark2 Add-On Service on Cloudera Manager

- First, we need to install new packages on Cloudera Manager for Spark 2 libraries. Connect to your Cloudera Manager instance through a terminal window, using command:

```
ssh -i cdsw-admin centos@[CLOUDERA-MANAGER-PUBLIC-DNS]
```

- Update the JAVA_HOME environment variable for cloudera manager server config file:

```
sudo sh -c "echo export JAVA_HOME=/usr/java/jdk1.8.0_121-cloudera >>
/etc/default/cloudera-scm-server"
```

- Then you will be able to download appropriate the official CSD file:

```
sudo wget --directory-prefix=/opt/cloudera/csd/
http://archive.cloudera.com/spark2/csd/SPARK2_ON_YARN-
2.2.0.cloudera1.jar

sudo chmod 644 /opt/cloudera/csd/SPARK2_ON_YARN-2.2.0.cloudera1.jar

sudo chown cloudera-scm:cloudera-scm /opt/cloudera/csd/SPARK2_ON_YARN-
2.2.0.cloudera1.jar

sudo systemctl restart cloudera-scm-server
```

B. Restart Cloudera Management Services

- First step is to open your favourite web browser and use the below url. you will be able to login with the default login/password (**admin/Cloudera_123**).

  http://[CLOUDERA-MANAGER-PUBLIC-DNS]:7180/

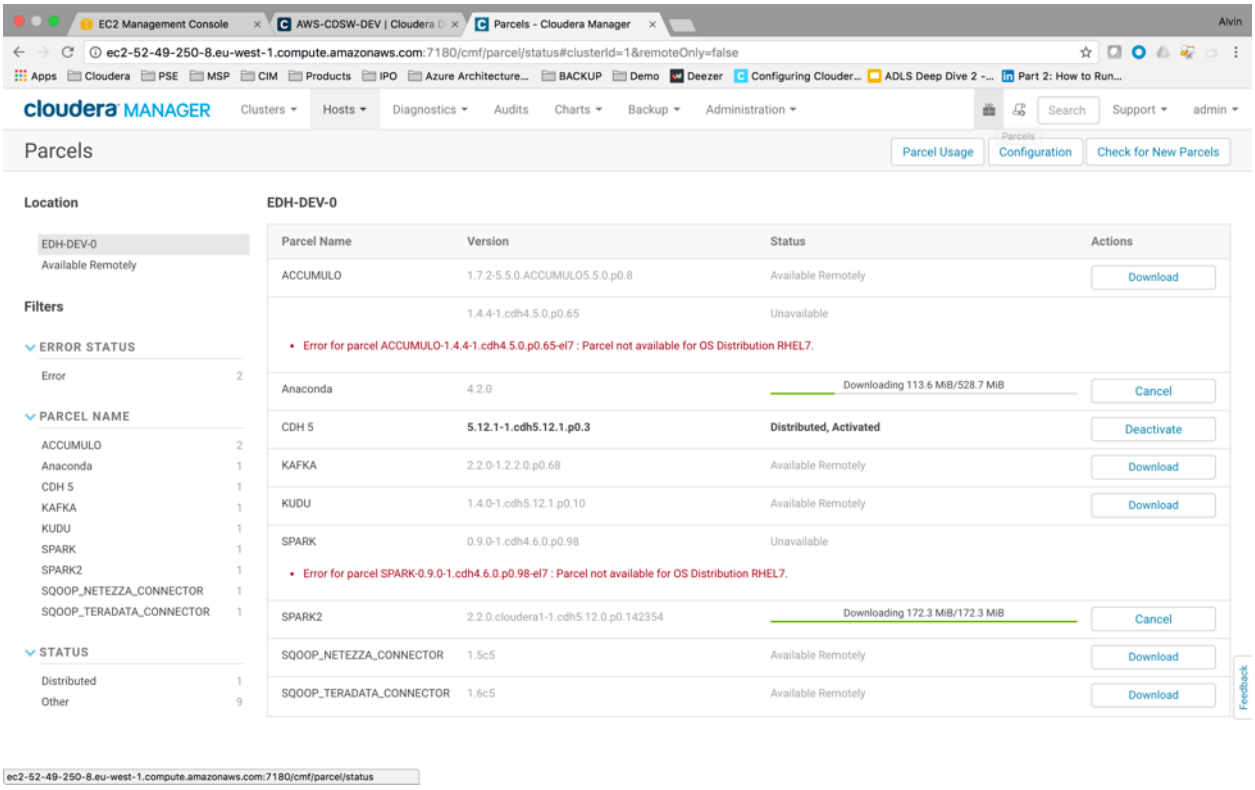
- We will need to setup JAVA_HOME environment variable. On the Cloudera Manager search bar please enter JAVA_HOME and set the environment variable to `/usr/java/jdk1.8.0_121-cloudera`

cloudera

- Click on Top-Left Button **Cloudera Manager** to return to dashboard. We can now restart Cloudera Management Service. And, also restart the EDH-DEV-0 Cluster.

C. Install Spark2 / Anaconda Parcels

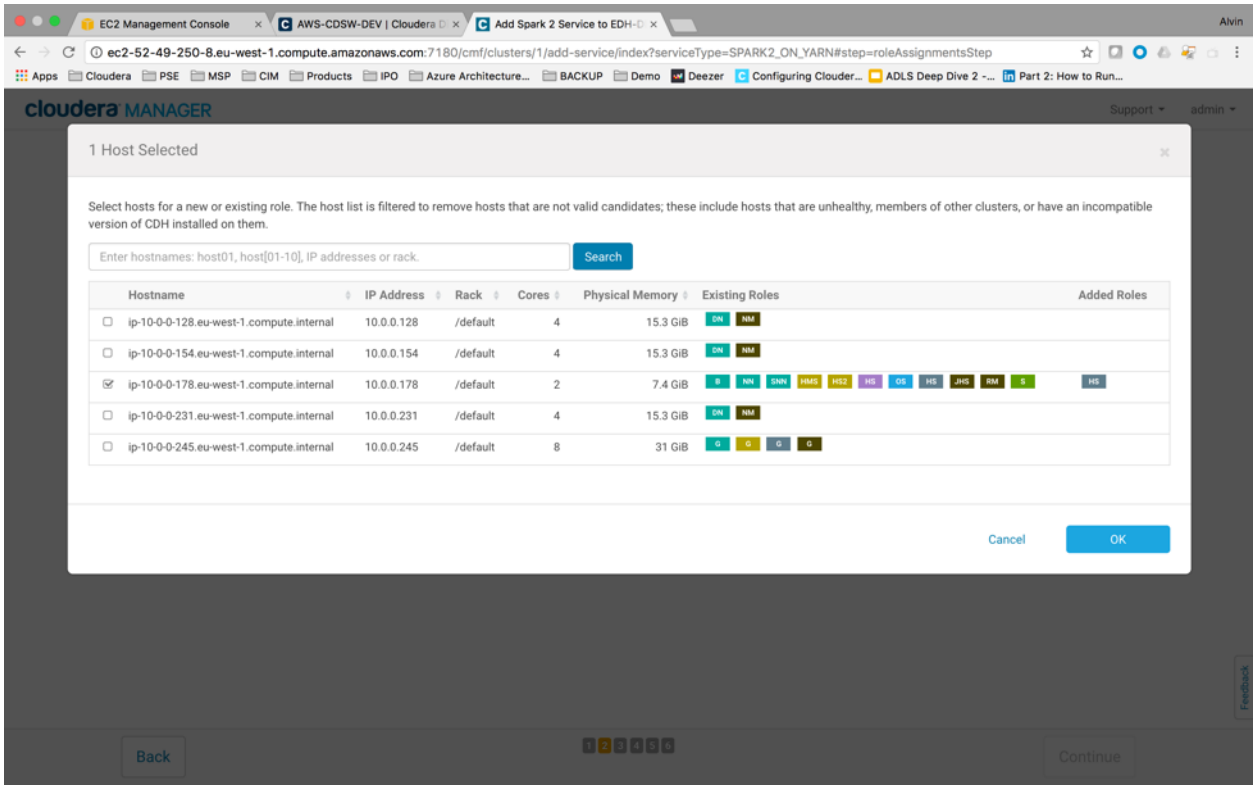
- Add new repos for Anaconda parcels (Spark 2 parcel is already existing).
  Click on the icon top left looking like a delivery package.
  Then, click on **Configure** and add new **Remote Parcel Repository URLs** for Spark2 & Anaconda repos.:
  https://repo.continuum.io/pkgs/misc/parcels/



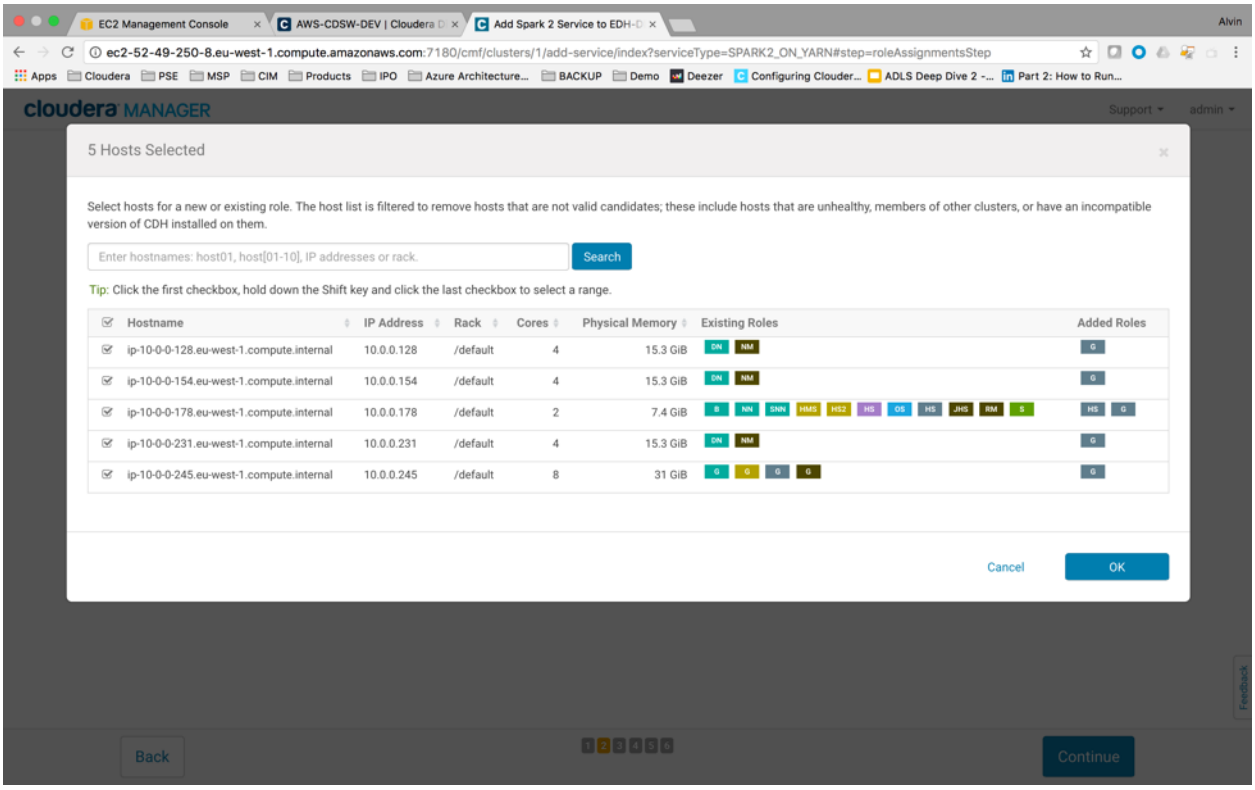
- Sequentially click on Download (to Cloudera Manager), Distribute (to all hosts) and then Activate (on all hosts) both packages.

cloudera

- Deploy new Spark2 services on the Cluster.
  Choose **Add Services** under menu next to the EDH-DEV-0 cluster.
  Choose **Spark2** under the list of deployable services for EDH.
  There are 2 options for the dependencies regarding Spark2. Please choose the option including services for only for **HDFS / YARN / ZOOKEEPER** services (**not HIVE / SPARK-ON-YARN**).
  History Server will be set on the Master Node (the node with most number of services).
  Spark2 Gateway will be set on All Nodes.

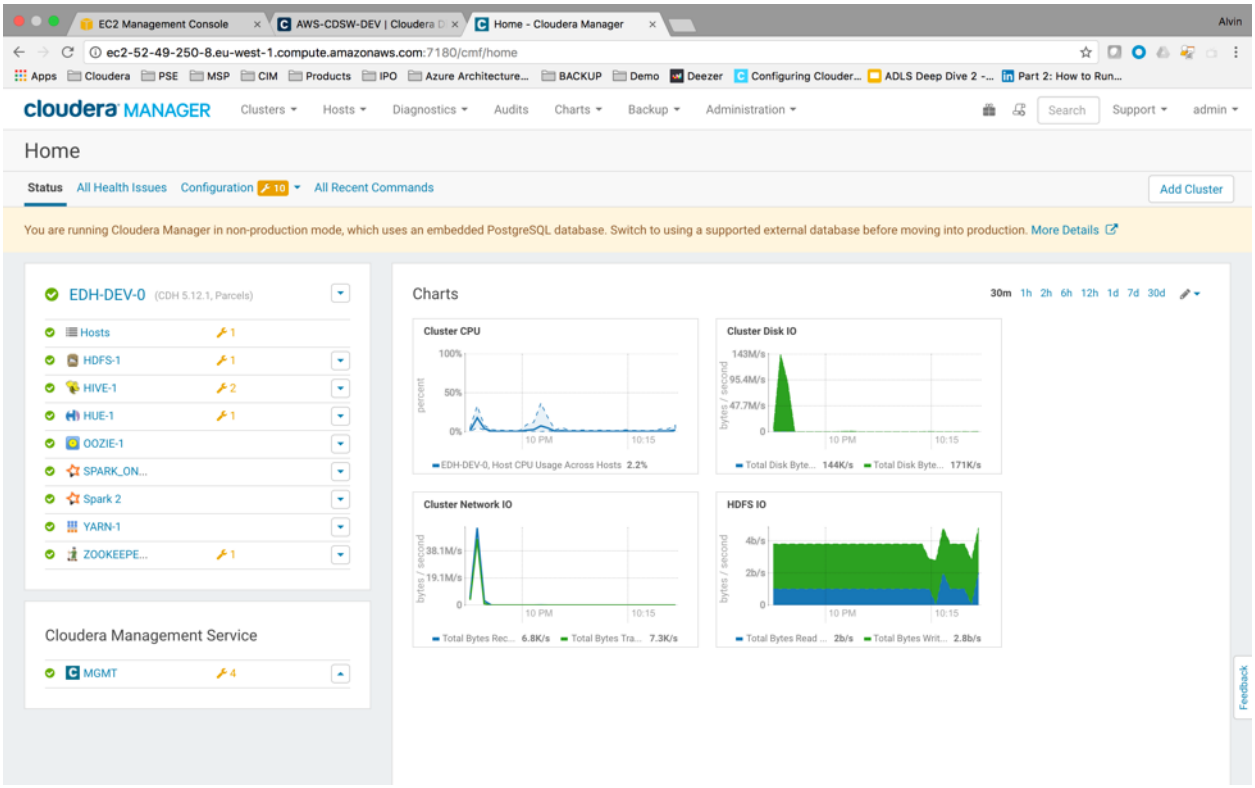


cloudera



- Coming back to Cloudera Manager Dashboard, you will be asked to re-deploy services configurations (Spark services has stalled configurations to be taken into account).
- You can now manually start Spark2 service on drop-down menu on the services right.

cloudera

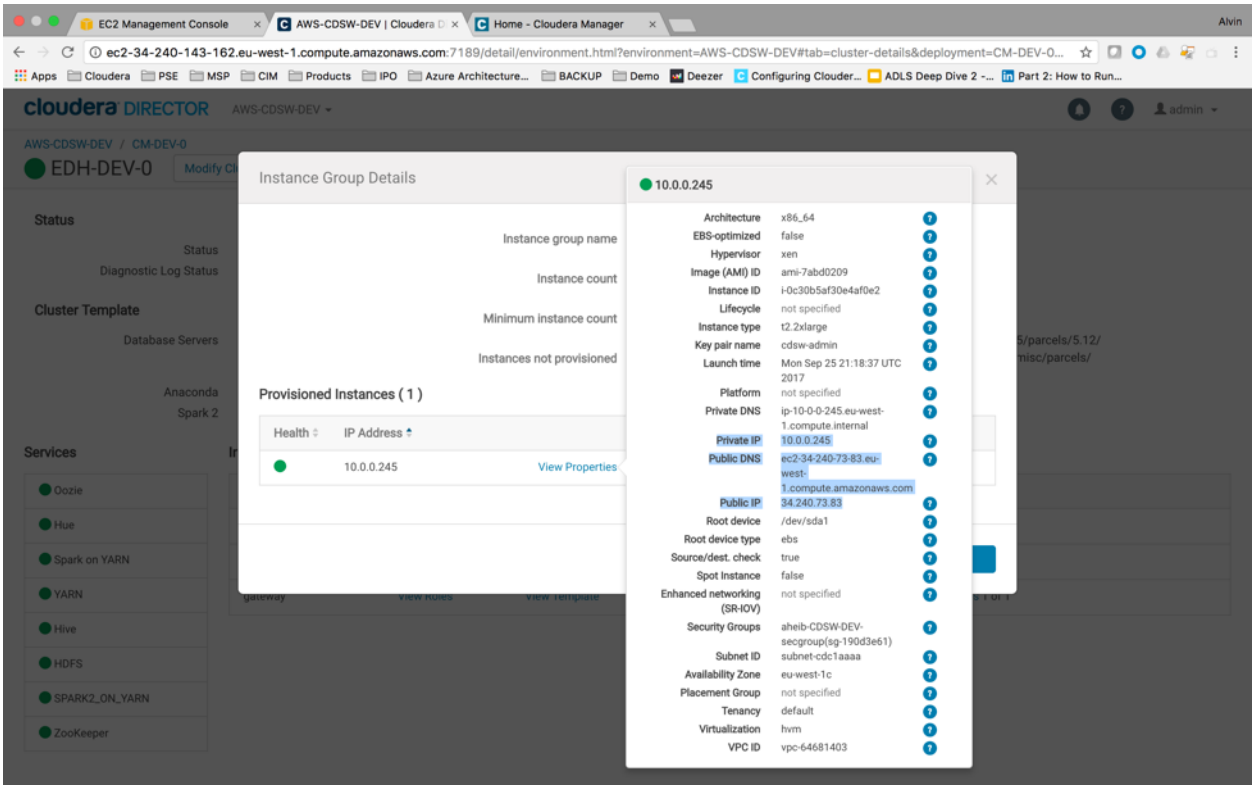

EC2 Management Console
AWS-CDSW-DEV | Cloudera
Home - Cloudera Manager
Alvin
ec2-52-49-250-8.eu-west-1.compute.amazonaws.com:7180/cmf/home
cloudera MANAGER
Clusters
Hosts
Diagnostics
Audits
Charts
Backup
Administration
Search
Support
admin
Home
Status
All Health Issues
Configuration
All Recent Commands
Add Cluster
You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a supported external database before moving into production. More Details
EDH-DEV-0 (CDH 5.12.1, Parcels)
Hosts
HDFS-1
HIVE-1
HUE-1
OOZIE-1
SPARK_ON...
Spark 2
YARN-1
ZOOKEEPE...
Cloudera Management Service
MGMT
Charts
30m 1h 2h 6h 12h 1d 7d 30d
Cluster CPU
Cluster Disk IO
Cluster Network IO
HDFS IO
Feedback

cloudera

VII. Lab #5: Cloudera Data Science Workbench Setup

A. Deactivate Firewalling on CDSW

- To identify your CDSW-Master node IP address, you first need to head back to your Cloudera Director webUI.
  Click on your newly deployed Cloudera Cluster called **EDH-DEV-0**.
  From there onwards, you shall be able to identify the gateway group, click on View Instance.
  When you extend the **View Properties**, you will identify multiple critical informations, like **[CDSW-MASTER-PRIVATE-IP], [CDSW-MASTER-PUBLIC-DNS] and [CDSW-MASTER-PUBLIC-IP].**
-



- You will then need to connect to your CDSW-Master instance using:

cloudera

```
ssh -i cdsw-admin centos@[CLOUDERA-CDSW-PUBLIC-DNS]
```

- Deactivate Permanently the SELinux service.
  Check you SELinux status using command: `grep SELINUX= /etc/selinux/config`. The command should return a similar output.

  ```
  # SELINUX= can take one of these three values:
  SELINUX=disabled
  ```

  If SELinux is enabled in the config file, please update the file.

- Deactivate Temporarily the SELinux service.
  Check you SELinux status using command: `sudo sestatus | grep "SELinux status"`.
  The command should return a similar output.
  ```
  SELinux status:                 enabled
  ```

  Deactivate temporarily the SELinux service using command: `sudo setenforce 0`.

  Verify that you have successfully temporarily deactivate SELinux.

- You could now reboot the CDSW-Master node using command `sudo reboot now`.

- You can wait a couple of minutes before re-connecting to CDSW-Master instance using:

  ```
  ssh -i cdsw-admin centos@[CLOUDERA-CDSW-PUBLIC-DNS]
  ```

B. Unmount both disk (/dev/xvdf and /dev/xvdg)

.

- Un-mount both disk on your CDSW-Master instance:
  Check the number of disks attached to the virtual machine using `sudo mount | grep xvd` command. You should have something similar to:

  ```
  /dev/xvda1 on / type xfs (rw,relatime,seclabel,attr2,inode64,noquota)
  /dev/xvdf on /data1 type ext4 (rw,noatime,seclabel,data=ordered)
  /dev/xvdg on /data2 type ext4 (rw,noatime,seclabel,data=ordered)
  ```

cloudera

(xvda1=Operating System, xvdf=block device for Docker Containers, and xvdg=block device for cdsw user-data)

- Now, you need to un-mount both disks **/dev/xvdf** and **/dev/xvdg**.

```
sudo umount /dev/xvdf /dev/xvdg
```

- Verify that you have successfully un-mounted both disks using `sudo mount | grep xvd` command.

- permanently un-mount both disk on your CDSW-Master instance.
  You could observe that in file **/etc/fstab** you will still find some enrtries for **xvdf** and **xvdg**.
  We will need to remove these entries in case of CDSW-Master reboot operations.
  Check that both entries are in the file **/etc/fstab** using command `sudo cat /etc/fstab | grep xvd`. You should have something similar to:

```
/dev/xvdf /data1 ext4 defaults,noatime 0 0
/dev/xvdg /data1 ext4 defaults,noatime 0 0
```

- Now, please remove both entries in the file. Please backup the file to avoid any manual errors.
  A handy linux command to do it is using `sudo sed -i.bak '/xvd/d' /etc/fstab` command.

- Verify that you have successfully permanently un-mounted both disks using `sudo cat /etc/fstab | grep xvd` command.

C. Ensure IPv6 is Disabled

- Check that IPv6 is enabled using `sudo cat /etc/sysctl.conf | grep ipv6` command. You should have something similar to:

  net.ipv6.conf.all.disable_ipv6=1

- Now, please disable the IPv6 feature by setting the value to 0. Please backup the file to avoid any manual errors. A handy linux command to do it is using `sudo sed -i.bak`

cloudera

`'s/net.ipv6.conf.all.disable_ipv6=1/net.ipv6.conf.all.disable_ipv6=0/g' /etc/sysctl.conf` command.

- Verify that you have successfully deactivated IPv6 using command `sudo cat /etc/sysctl.conf | grep ipv6` command.

D. Enable/Activate RPCbind on Start-Up

- Check RPCbind service status by using `sudo systemctl status rpcbind` command. You should have something similar to:

  rpcbind.service - RPC bind service
  Loaded: loaded (/usr/lib/systemd/system/rpcbind.service; indirect; vendor preset: enabled)
  Active: inactive (dead)

- Now, please enable the rpcbind service at startup using below commands:

```
sudo systemctl enable rpcbind
sudo systemctl start rpcbind
```

- Verify that you have successfully activate/enabled RPCbind using command `sudo systemctl status rpcbind` command.

E. Remove IPtables service blacklisting

- First you will need to remove the blocklist file using `sudo rm -f /etc/modprobe.d/iptables-blacklist.conf` command.
- Then, you will need to activate iptables module `sudo modprobe ip_tables`.
- Finally, you will need to load iptables filters modules `sudo modprobe iptable_filter`.

F. Download and Install Cloudera Data Science Workbench

- Now that the Instance is ready, we could start downloading the Cloudera Data Science Manager packages and install.

cloudera

```
sudo wget --directory-prefix=/etc/yum.repos.d/
https://archive.cloudera.com/cdsw/1/redhat/7/x86_64/cdsw/cloudera-
cdsw.repo

sudo yum install -y cloudera-data-science-workbench
```

G. Configure Cloudera Data Science Workbench

- Identify the CDSW configuration file at location `/etc/cdsw/config/cdsw.conf` and please edit with your favourite text editor for ex: `sudo vi /etc/cdsw/config/cdsw.conf`

- Update the CDSW-Master Public IP using the XIP trick.
  Update the `DOMAIN="cdsw.company.com"` with specific value `cdsw.[CLOUDERA-CDSW-PUBLIC-IP].xip.io` (for ex: DOMAIN="cdsw.34.240.73.83.xip.io")

- Update the CDSW-Master Private IP.
  Update the `MASTER_IP=""` with specific value `[CLOUDERA-CDSW-PRIVATE-IP]` (for ex: MASTER_IP ="10.0.0.245")

- Update the CDSW Block Device to store Docker Images.
  Update the `DOCKER_BLOCK_DEVICES=""` with specific value `/dev/xvdf` (for ex: DOCKER_BLOCK_DEVICES="/dev/xvdf")

- Update the CDSW Block Device to store Application Data.
  Update the `APPLICATION_BLOCK_DEVICE=""` with specific value `/dev/xvdg` (for ex: APPLICATION_BLOCK_DEVICE="/dev/xvdg")

- Update the JAVA_HOME environment variable.
  Update the `JAVA_HOME=""` with specific value `/usr/java/jdk1.8.0_121-cloudera` (for ex: JAVA_HOME="/usr/java/jdk1.8.0_121-cloudera"

H. Launch Cloudera Data Science Workbench

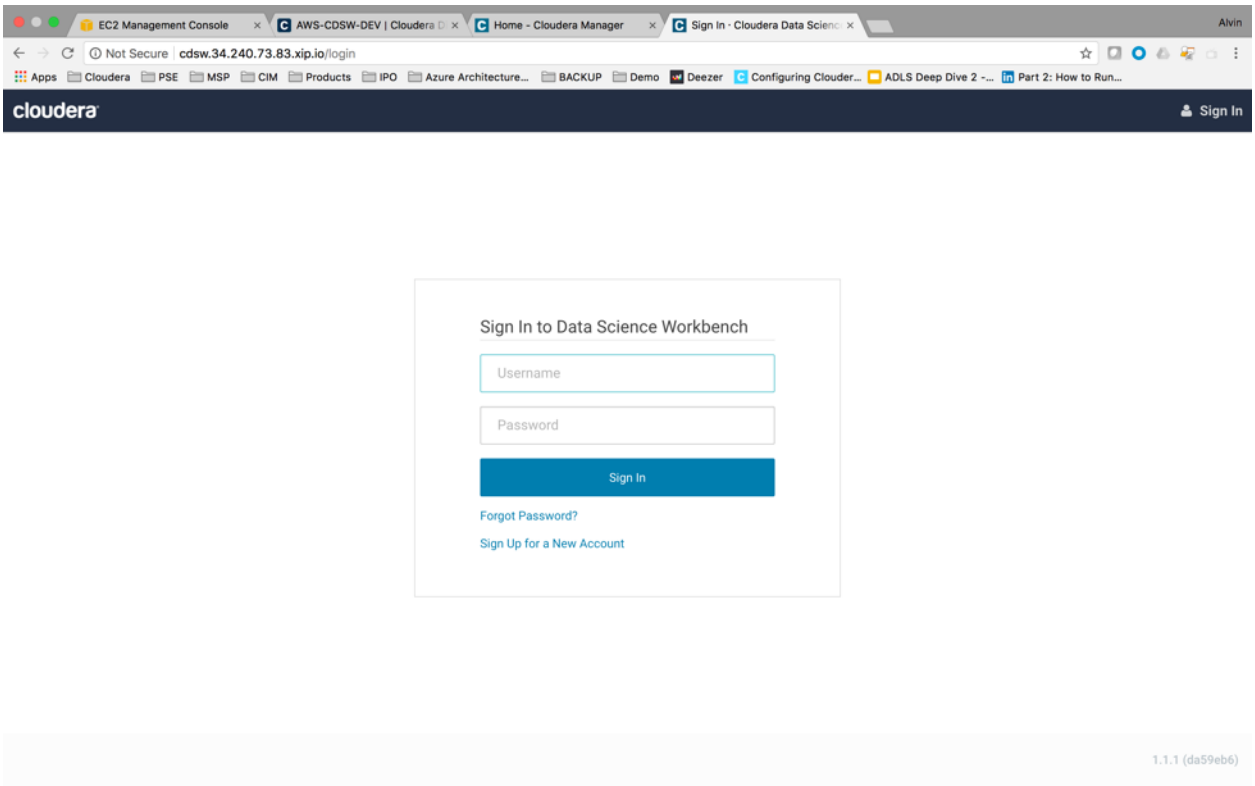
- Start Cloudera Data Science Workbench using command: `sudo cdsw init`

cloudera

- You could check Cloudera Data Science Workbench status using command `sudo cdsw status`.

- You can now use Cloudera Data Science Workbench at URL:

  http://cdsw.34.240.73.83.xip.io/



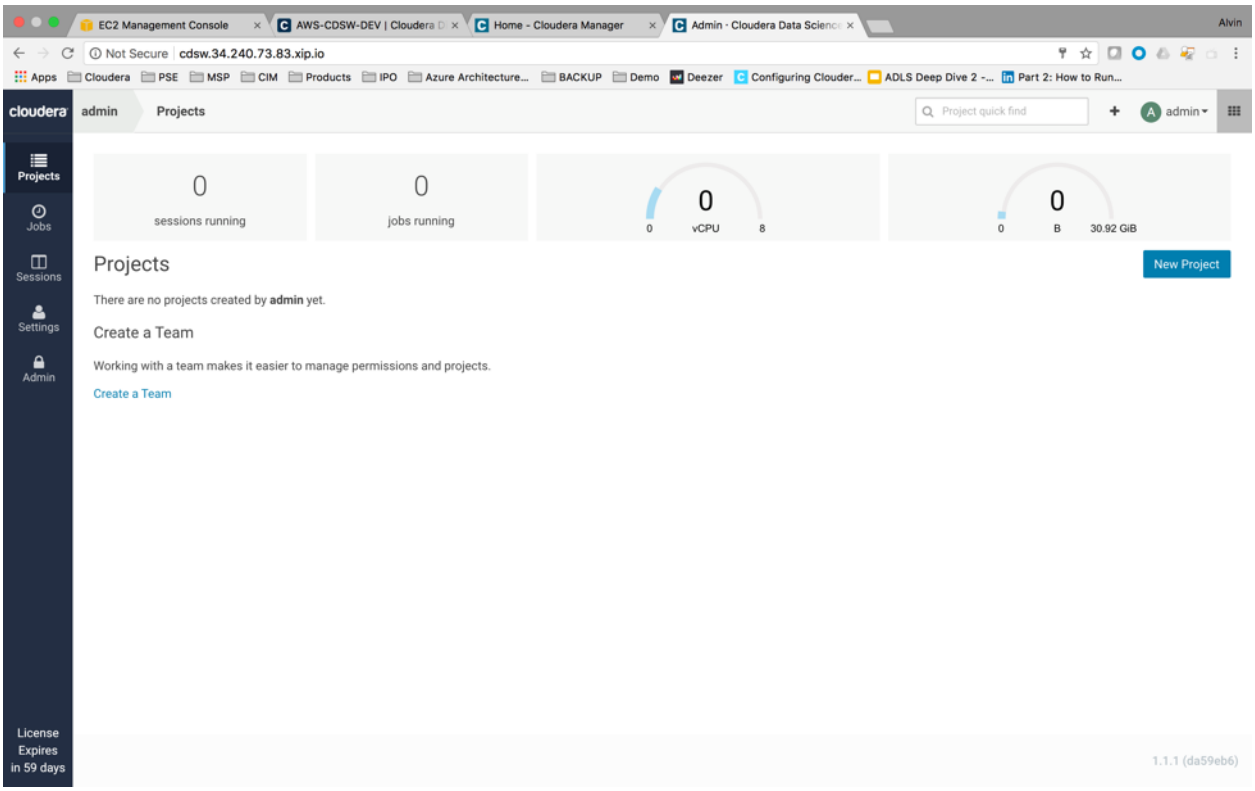
cloudera

VIII. Lab#6: Cloudera Data Science Workbench Operations

A. Login as Administrator

Please note that first account will be the admin for Cloudera Data Science Workbench.
Please **Sign Up for a New Account** with:

- Full Name = Admin
- Username = admin
- Email = any email (please note that CDSW will be identifying user also by email – do not reuse)
- Password = Cloudera_123



cloudera

B. Setup HADOOP_USER_NAME

- Since we did not have setup a LDAP or AD, we will not be having a correspondence between HDFS and CDSW user. Only a single user will be able to launch Spark / HDFS jobs. Please connect to your CDSW-Master node using:

```
ssh -i cdsw-admin centos@[CLOUDERA-CDSW-PUBLIC-DNS]
```

- We will then create manually the HDFS user called hdfs_super.

```
sudo groupadd supergroup
```

```
sudo useradd -G supergroup -u 12354 hdfs_super
```

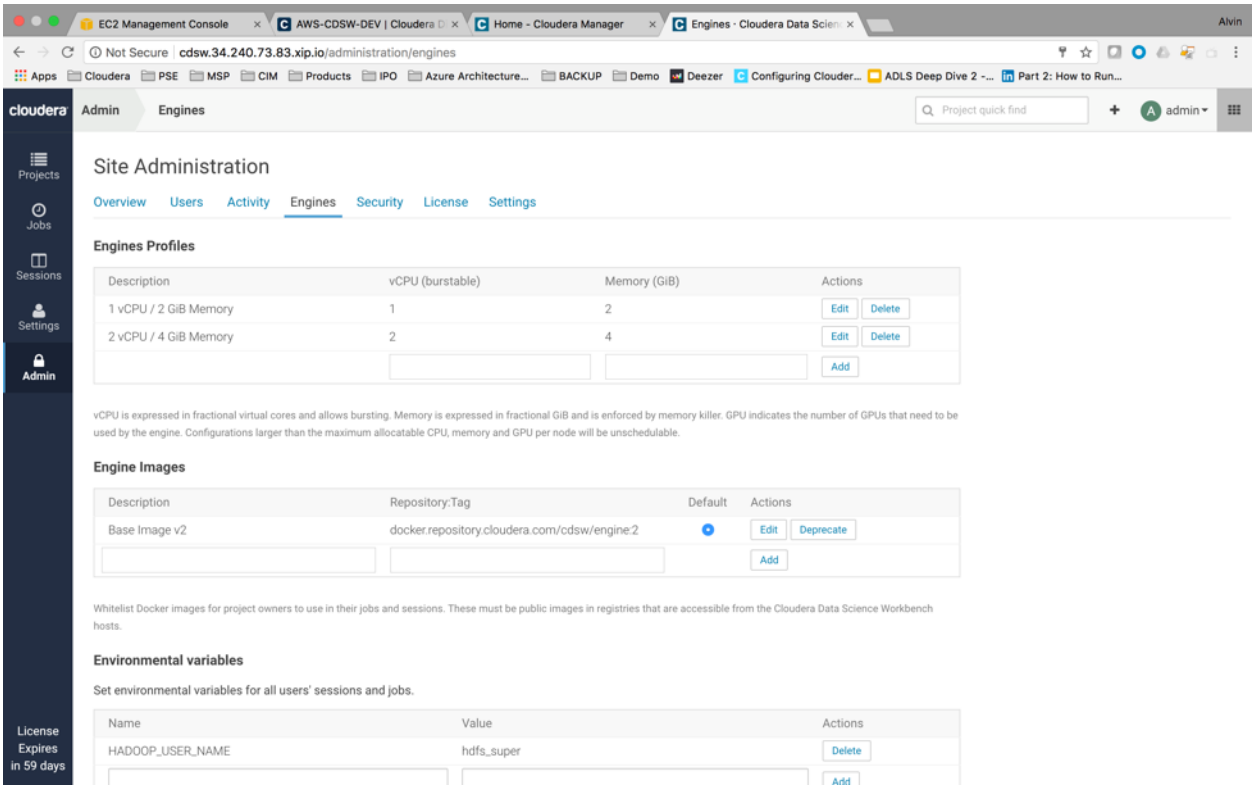
```
sudo su -c "echo Cloudera_123 | passwd --stdin hdfs_super"
```

```
sudo su hdfs -c "hadoop dfs -mkdir /user/hdfs_super"
```

```
sudo su hdfs -c "hadoop dfs -chown hdfs_super:hdfs_super
/user/hdfs_super"
```

- On the CDSW WebUI we will also set an environment variable to hdfs_super.
  Please select **Admin** menu on the left, then **Engines**.
  You will find an **Environmental variables** section, please **add** a variable:
  HADOOP_USER_NAME = hdfs_super

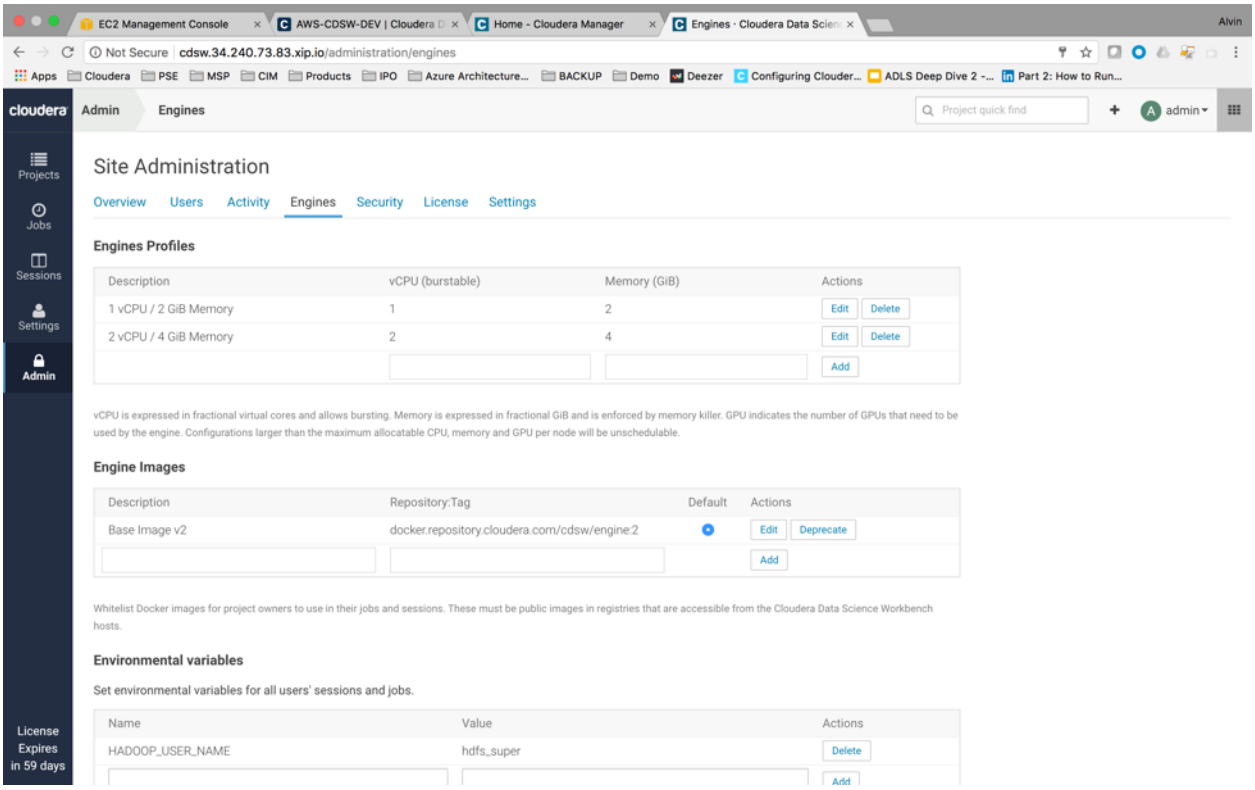
cloudera



C. Setup Docker Container types

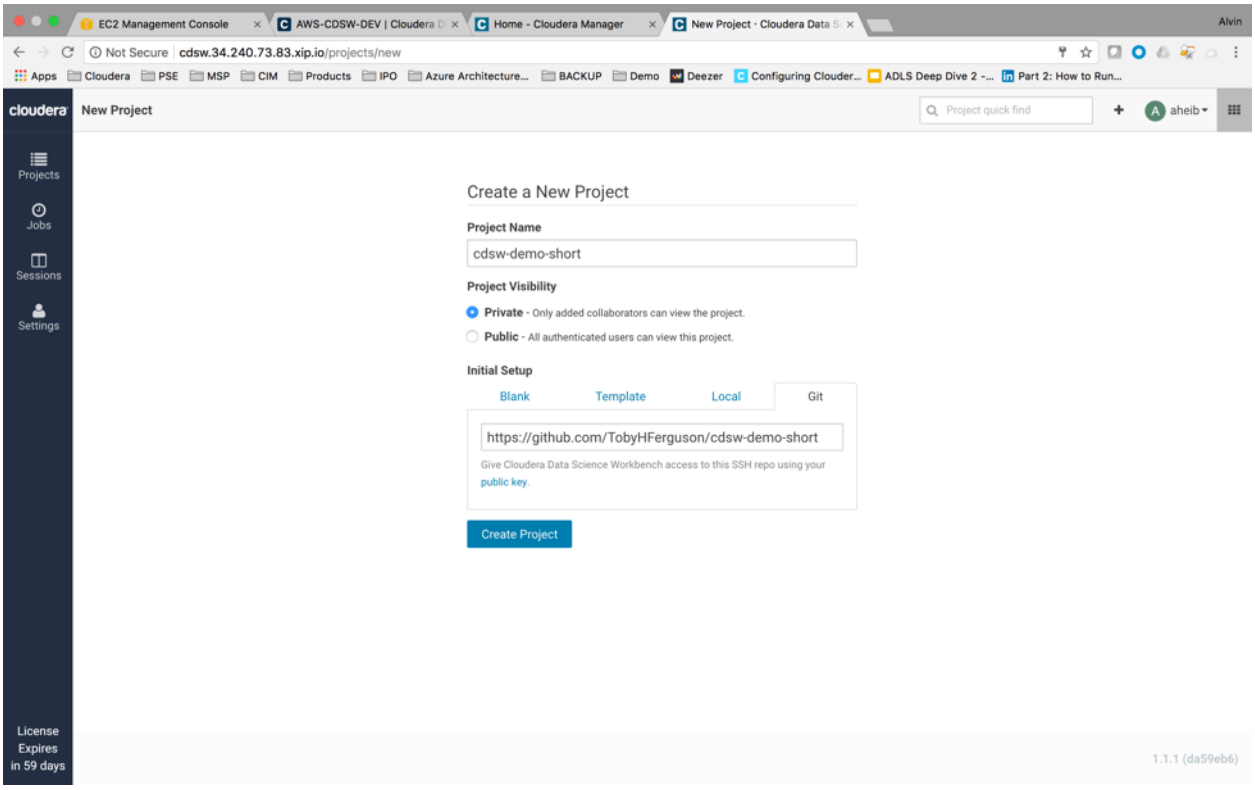
- Under the same menu, you will be also able to add additional container resource type (for ex: 2 vCPU / 4 GB Ram)

cloudera



D. Add user and CDSW Live Demo

- You will need to sign out before creating a new user. From menu Top-Right, please select **admin** and **sign-out**.
- Please re-run step A to create a new user, with distinct email.
- Let's import our cdsw demo source, by clicking on Workbench (button top-left) and Selecting option Git.

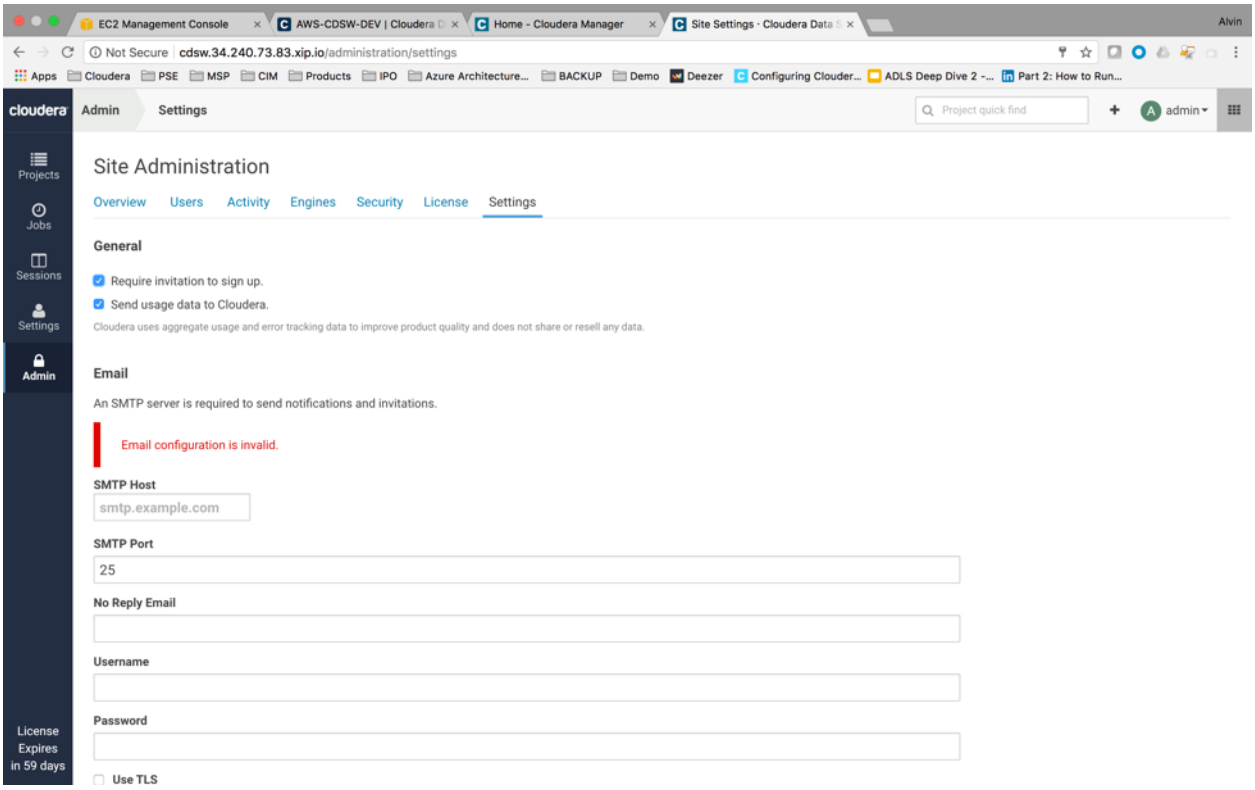


Open Workbench and follow the README.md

E. Block External Sign-Ups

- Quick tips, in the **admin** WebUI, you could deactivate new sign-in.
  Under Admin menu, Settings pane, please check “Require invitation to sign up” field.

cloudera



IX. Lab #7: Industrialisation: Cloudera Director CLI

A. Prepare your CDSW AWS-DEV config files

- Update **~Github/CDSW/AWS-DEV/owner_tag.properties** file.
  Update the `OWNER=[YOUR-USERNAME]` with specific value (for ex: OWNER=aheib)

- Update **~Github/CDSW/AWS-DEV/provider.properties** file.
  Update the `AWS_ACCESS_KEY_ID=[YOUR-AWS-ACCESS-KEY-ID]` with specific value (for ex: AWS_ACCESS_KEY_ID=AKIAIOSFODNN7EXAMPLE)

- Update **~Github/CDSW/AWS-DEV/SECRET.properties** file.
  Update the `AWS_SECRET_ACCESS_KEY=[YOUR-AWS-SECRET-ACCESS-KEY]` with specific

value (for ex: AWS_SECRET_ACCESS_KEY= wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY)

- Verify **~Github/CDSW/AWS-DEV/ssh.properties** file content.
  You should have 2 entries for SSH: `SSH_USERNAME=centos` (for our specific case using CentOS 7) and `SSH_PRIVATEKEY=aws/cdsw-admin` (the path to our cdsw-admin key, in our case it is in the same folder).

- Zip your **~/Github/CDSW/AWS-DEV** folder to be transferred on Cloudera Director instance. The easiest way on Linux will be:

```
cd ~Github/CDSW/

tar -cvf AWS-DEV.tar AWS-DEV

gzip AWS-DEV.tar

scp -i AWS-DEV/cdsw-admin AWS-DEV.tar centos@[CLOUDERA-DIRECTOR-PUBLIC-DNS]:.
```

B. Log onto Cloudera Director

- Connect to your Cloudera Director instance through a terminal window, using command:

```
ssh -i cdsw-admin centos@[CLOUDERA-DIRECTOR-PUBLIC-DNS]
```

- You will need some installation scripts coming from my friend Toby. Use clone (using `git clone <URL>`) or download (using the Top-Right green download button) the current git repo:
  https://github.com/TobyHFerguson/cdsw_install
  You should have a directory named `cdsw_install`.

- UnZip your **AWS-DEV.tar.gz** file on Cloudera Director instance, and move it into **cdsw_install** folder, and cleanup the tar file. The easiest way on Linux will be:

```
tar -xvf AWS-DEV.tar.gz

mv AWS-DEV cdsw_install/
```

cloudera

```
rm AWS-DEV.tar.gz
```

```
cd cdsw_install
```

You should now see both folders `aws` and `rm AWS-DEV.tar.gz`.

- If you browse the **aws** folder content, you will see that there are all the **.properties** files as in your AWS-DEV folder. The aim is to have multiple deployment files.

- Rename the **kerberos.properties** file in **cdsw_install/aws**, since we do not have Kerberos activated.

```
mv aws/kerberos.properties aws/kerberos.properties.bak
```

- Rename the **kerberos.properties** file in **cdsw_install/aws**, since we do not have Kerberos activated.

```
mv aws/kerberos.properties aws/kerberos.properties.bak
```

- Copy the relevant properties and key pair into aws folder.

```
cp AWS-DEV/*.properties aws
```

```
cp AWS-DEV/cdsw-admin* aws
```

```
chmod og+r aws/cdsw-admin*
```

- Launch your Cloudera Cluster with CDSW Node.

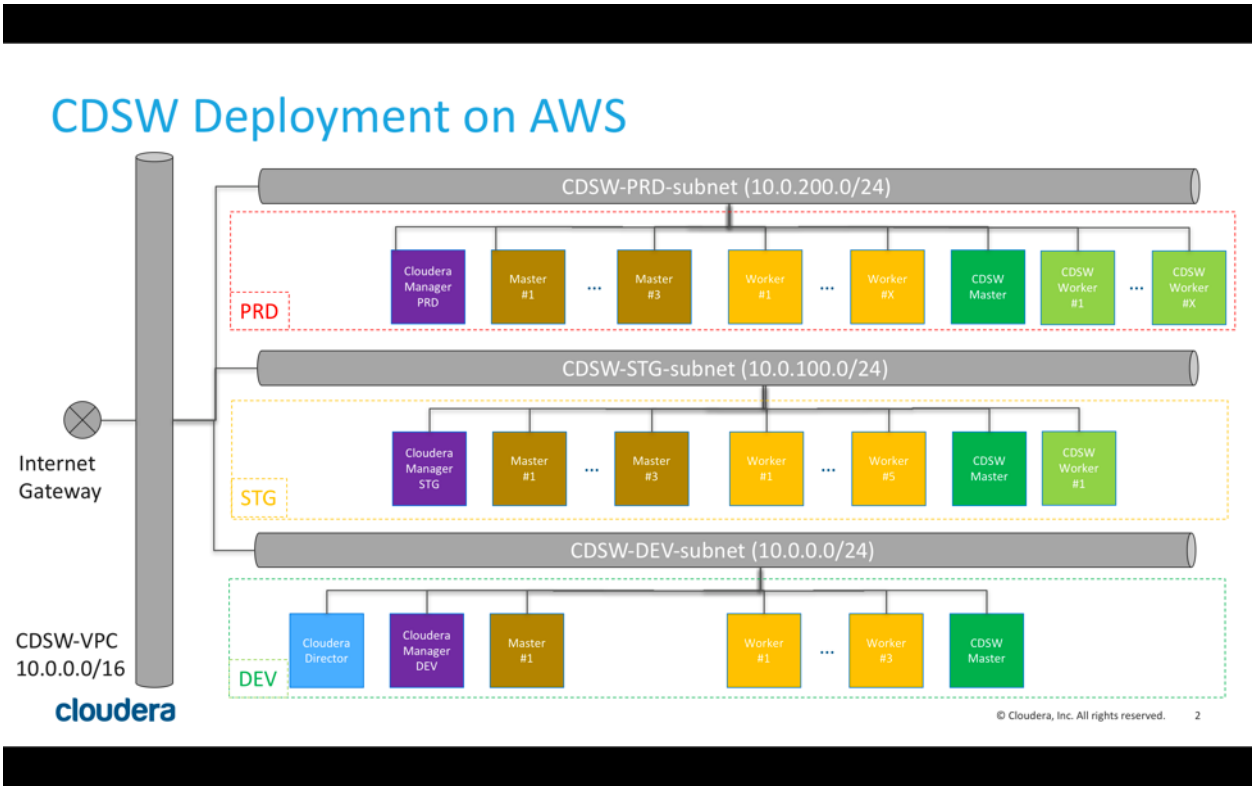
```
cloudera-director bootstrap-remote aws.conf --lp.remote.username=admin
--lp.remote.password=Cloudera_123
```

cloudera

X. Production and Security Considerations

A. Perimetrical Security

In a production environment, we shall not have public IPs for each nodes (Master / Worker / CDSW-Master). Indeed, the entire cluster should be secured within a perimeter security within a private subnet.
Another consideration is concerning Cloudera Director, Cloudera Manager and Cloudera Data Science Workbench. These services need an external connectivity. This is possible either with a Public IP or using a Network Address Translation (NAT) mechanism. In both cases, the AWS SecurityGroup should be filtered on the source of the accessible machines.



cloudera

B. Cloudera Service Database

During the Enablement Session, we are using Unsupported trial version of Cloudera EDH. With this version, only Embedded Postgres Database is installed on Cloudera Director and Cloudera Manager.
These types of Databases are not suited for STG or PRD environments (no resilience, scalability, backup and security).
Please take a special look a current link:
https://www.cloudera.com/documentation/enterprise/5-11-x/topics/cm_ig_installing_configuring_dbs.html

C. AWS Regions and AMI

For enablement ease, we have fixed AWS Region (eu-west-1) and the corresponding AMI. You are able to select another AWS Region (closer to your business). Please take good note that a corresponding AMI need to match your region.
Please take a special look a current link:
https://www.cloudera.com/documentation/director/latest/topics/director_deployment_ami.html

D. Staging and Production Deployment Sizing

For a DEV deployment, minimal resources are requested to provide Data Science Workbench functionality.

STG and PRD deployment should be similar in architecture. Only differences are based on higher instance resources, and number of worker nodes (horizontal scaling).

To provide High-Availability Cluster functionality, both STG / PRD shall have a deployment of 3 Masters (avoid Single Point of Failure). And will also need a minimum of 5 Worker Nodes.

Regarding Cloudera Data Science Workbench, for STG / PRD high-availability needs, we will need to deploy 1 CDSW-Master Node and at least 1 CDSW-Worker Node.
Please note that to install CDSW-Worker, the Lab#5 is identical to CDSW-Master. Only difference resides in step G and H. We will not modify the `/etc/cdsw/config/cdsw.conf`

cloudera

instead, we will copy the one from CDSW-Master. We will not start a new CDSW installation, instead, we will join both nodes through `sudo cdsw init` command.

Please find some sizing guidances for DEV, STG and PRD.

DEV					SYSTEM DISK				DATA DISK			
NODE TYPE	QTY	INSTANCE TYPE	CPU	RAM	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)
Cloudera Director	1	t2.medium	2	4	1	50	EBS	50				0
Cloudera Manager	1	t2.large	2	8	1	50	EBS	50				0
Master	1	t2.large	2	8	1	50	EBS	50				0
Worker	3	t2.xlarge	4	16	1	50	EBS	50	1	500	EBS	500
CDSW-Master	1	t2.2xlarge	8	32	1	100	GP2	100	2	500	GP2	1 000
			18	68				300				1 500

STG					SYSTEM DISK				DATA DISK			
NODE TYPE	QTY	INSTANCE TYPE	CPU	RAM	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)
Cloudera Director	1	c4.large	2	3.75	1	100	EBS	100				0
Cloudera Manager	1	m4.xlarge	4	16	1	100	EBS	100				0
Master	3	m4.xlarge	4	16	1	100	EBS	100				0
Worker	5	m4.2xlarge	8	32	1	100	EBS	100	1	1000	GP2	1000
CDSW-Master	1	m4.4xlarge	16	64	1	100	GP2	100	2	1000	GP2	2000
CDSW-Worker	1	m4.4xlarge	16	64	1	100	GP2	100	1	1000	GP2	1000
			50	192				600				4 000

PRD					SYSTEM DISK				DATA DISK			
NODE TYPE	QTY	INSTANCE TYPE	CPU	RAM	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)	QTY	CAPACITY (GB)	TYPE	TOTAL (GB)
Cloudera Director	1	c4.large	2	3.75	1	500	EBS	500				0
Cloudera Manager	1	m4.xlarge	4	16	1	500	EBS	500				0
Master	3	m4.2xlarge	8	32	1	500	EBS	500				0
Worker	5	m4.4xlarge	16	64	1	500	EBS	500	1	1000	GP2	1000
CDSW-Master	1	m4.16xlarge	64	256	1	500	GP2	500	2	1000	GP2	2000
CDSW-Worker	1	m4.16xlarge	64	256	1	500	GP2	500	1	1000	GP2	1000
			158	624				3 000				4 000