



Cloudera Data Science Workbench Installation from Scratch on AWS

Partner Presales @Cloudera:

Filippo Lambiente
Alvin Heib

(greetings to Toby Fergusson)



I. Table of Contents

I.	INTRODUCTION	4
II.	PREREQUISITES	6
A.	AWS ACCOUNT / QUOTAS.....	6
B.	AWS COMMAND LINE INTERFACE INSTALLED (CLI)	6
C.	KNOWLEDGE ON SSH WITH KEYPAIRS INTO A LINUX TERMINAL	6
III.	LAB #1: AWS ENVIRONMENT SETUP	7
A.	PREPARE A DIRECTORY WITH FILES FOR THE ENABLEMENT SESSION.....	7
B.	LOG ON TO AWS CONSOLE	8
C.	CREATE A NEW VPC.....	9
D.	RETRIEVE SUBNET INFORMATION	10
E.	CREATE A NEW SECURITY GROUP	11
F.	IMPORT YOUR KEYPAIR.....	13
IV.	LAB #2: CLOUDERA DIRECTOR SETUP	15
A.	DEPLOY YOUR CLOUDERA DIRECTOR INSTANCE	15
B.	IDENTIFY YOUR CLOUDERA DIRECTOR INSTANCE	17
C.	LOG ONTO CLOUDERA DIRECTOR	18
D.	MODIFY YOUR DEFAULT PASSWORD	18
V.	LAB #3: CLOUDERA DATA ENGINEERING CLUSTER SETUP	20
A.	CREATE YOUR AWS-DEV ENVIRONMENT	20
B.	CREATE NODE TEMPLATES	21
C.	DEPLOY YOUR FIRST CLOUDERA MANAGER INSTANCE	27
D.	DEPLOY YOUR FIRST CLOUDERA CLUSTER	28
E.	IDENTIFY YOUR CLOUDERA MANAGER.....	29
VI.	LAB #4: UPGRADE CLUSTER TO SPARK 2	31
A.	INSTALL SPARK2 ADD-ON SERVICE ON CLOUDERA MANAGER.....	31
B.	RESTART CLOUDERA MANAGEMENT SERVICES.....	31
C.	INSTALL SPARK2 / ANACONDA PARCELS	32
VII.	LAB #5: CLOUDERA DATA SCIENCE WORKBENCH SETUP	36
A.	DEACTIVATE FIREWALLING ON CDSW.....	36
B.	UNMOUNT BOTH DISK (/DEV/XVDF AND /DEV/XVDG)	37
C.	ENSURE IPV6 IS DISABLED.....	38
D.	ENABLE/ACTIVATE RPCBIND ON START-UP	39
E.	REMOVE IPTABLES SERVICE BLACKLISTING.....	39



F. DOWNLOAD AND INSTALL CLOUDERA DATA SCIENCE WORKBENCH	39
G. CONFIGURE CLOUDERA DATA SCIENCE WORKBENCH	40
H. LAUNCH CLOUDERA DATA SCIENCE WORKBENCH	40
VIII. LAB#6: CLOUDERA DATA SCIENCE WORKBENCH OPERATIONS.....	42
A. LOGIN AS ADMINISTRATOR.....	42
B. SETUP HADOOP_USER_NAME.....	43
C. SETUP DOCKER CONTAINER TYPES.....	44
D. ADD USER AND CDSW LIVE DEMO.....	45
E. BLOCK EXTERNAL SIGN-UPS.....	46
IX. LAB #7: INDUSTRIALISATION: CLOUDERA DIRECTOR CLI.....	47
A. PREPARE YOUR CDSW AWS-DEV CONFIG FILES	47
B. LOG ONTO CLOUDERA DIRECTOR.....	48
X. PRODUCTION AND SECURITY CONSIDERATIONS.....	50
A. PERIMETRICAL SECURITY	50
B. CLOUDERA SERVICE DATABASE.....	51
C. AWS REGIONS AND AMI	51
D. STAGING AND PRODUCTION DEPLOYMENT SIZING	51



I. Introduction

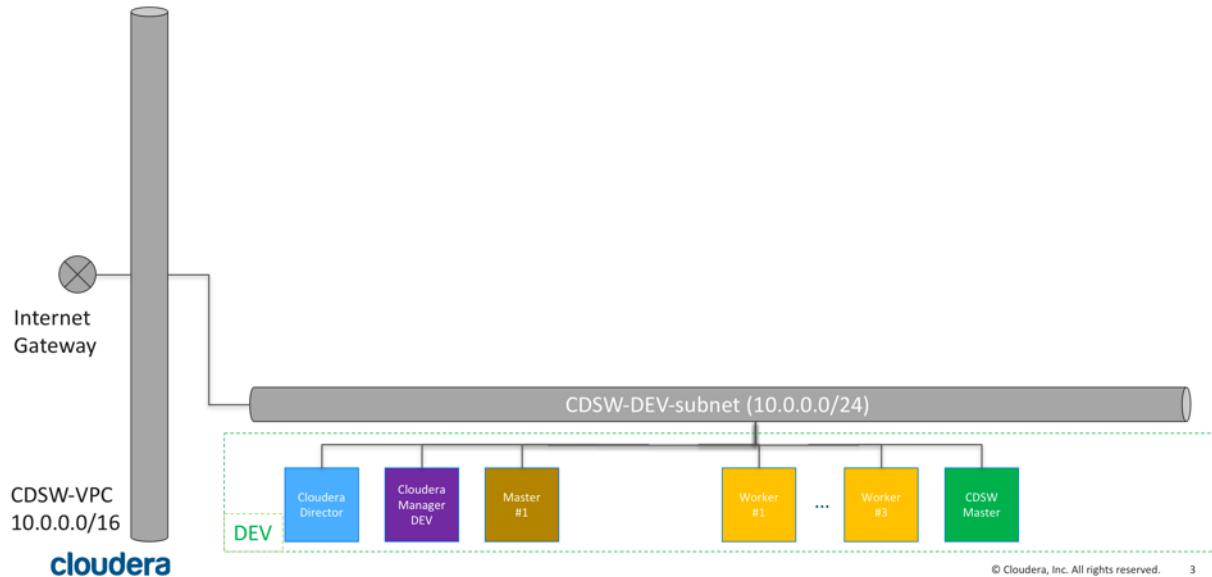
Cloudera Data Science and Engineering is the comprehensive offering for exploratory data science and machine learning at scale. It lives where your data lives, on-premise and across public clouds. Cloudera gives data science users better access to Hadoop data with familiar and performant tools that address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

CDSW Deployment on AWS - DEV



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	QTY	INSTANCE TYPE	CPU	RAM	SYSTEM DISK				DATA DISK			
					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



II. Prerequisites

A. AWS Account / Quotas

Please make sure 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 go 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 that you have:

- Installed putty. [Putty installation link](#)
- And converted the **cdsw-admin** private key. [Putty user guide](#)



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 cdsweb-admin
-rw-r--r--@  1 alvinheib staff    745 Sep 25 22:02 cdsweb-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 intensively using linux environment variable. Most of them needs to be updated in the corresponding files:



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]

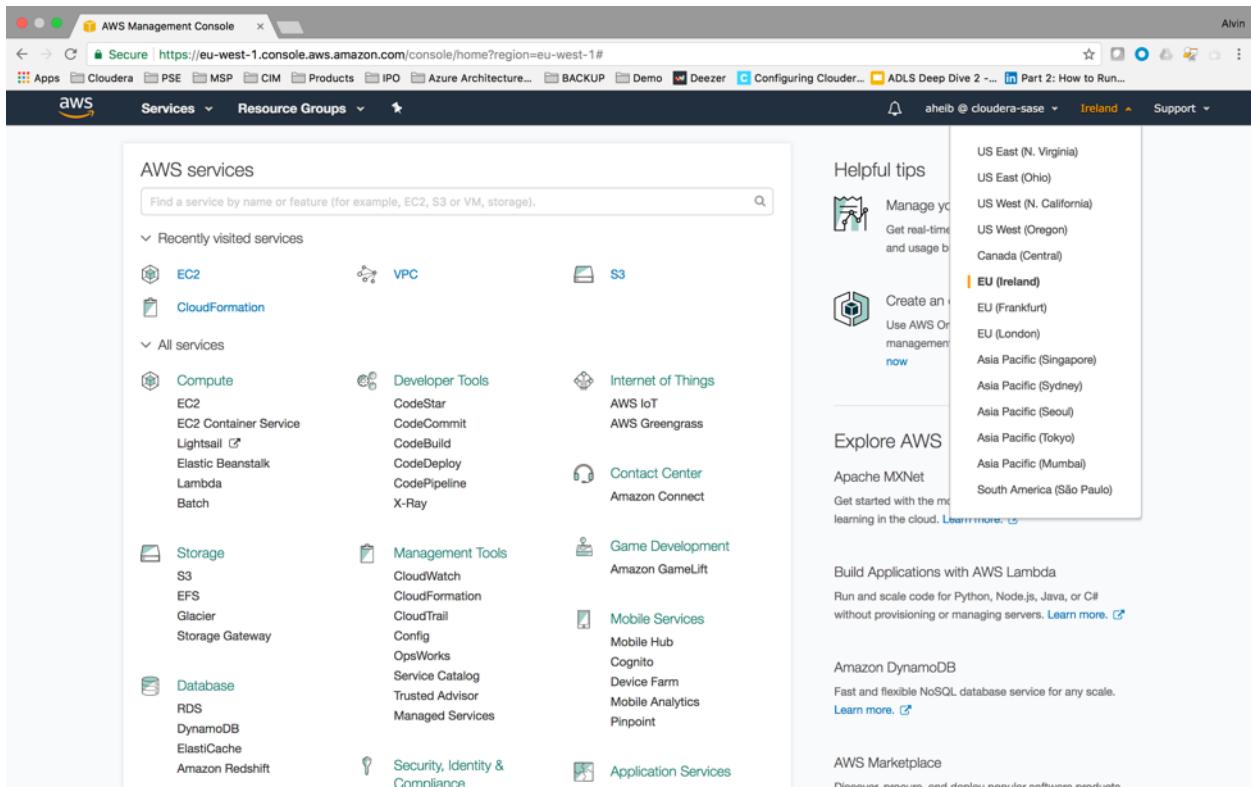
B. Log on to AWS Console

- 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

VPC Dashboard Resources ↗
Filter by VPC:
Select a VPC
Start VPC Wizard Launch EC2 Instances

Virtual Private Cloud
Your VPCs
Subnets
Route Tables
Internet Gateways
Egress Only Internet Gateways
DHCP Options Sets
Elastic IPs
Endpoints
NAT Gateways
Peering Connections
Security
Network ACLs
Security Groups
VPN Connections
Customer Gateways
Virtual Private Gateways

31 VPCs 28 Internet Gateways
0 Egress-only Internet Gateways 47 Subnets
52 Route Tables 31 Network ACLs
5 Elastic IPs 0 VPC Peering Connections
7 Endpoints 0 Nat Gateways
121 Security Groups 21 Running Instances
0 VPN Connections 1 Virtual Private Gateway
0 Customer Gateways

Service Health

Current Status	Details
Amazon VPC - EU (Ireland)	Service is operating normally
Amazon EC2 - EU (Ireland)	Service is operating normally

View complete service health details

Additional Information

VPC Documentation All VPC Resources Forums Report an Issue

VPN Connections

Create VPN Connection

Feedback English (US) © 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

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.

The screenshot shows the AWS VPC Management Console interface. On the left, there's a sidebar with various VPC-related options like Subnets, Route Tables, Internet Gateways, etc. The main area is titled 'Create Subnet' and shows a table of subnets. One subnet is selected: 'aheib-CDSW-DEV-subnet'. The table includes columns for Name, Subnet ID, State, VPC, IPv4 CIDR, and Available IPv4. Below the table, there's a detailed view of the selected subnet ('subnet-cdc1aaaa | aheib-CDSW-DEV-subnet'). This view shows details such as Subnet ID, Availability Zone (eu-west-1c), IPv4 CIDR (10.0.0.0/24), and VPC (vpc-64681403 | aheib-CDSW-VPC). It also lists Route table (rtb-c1e141a7) and Network ACL (acl-74baac13).

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.

Name tag	Group ID	Group Name	VPC	Description
aheib-CDSW-DEV-secgroup	sg-190d3e61	aheib-CDSW-DEV-se...	vpc-64681403 aheib-CDS...	Cloudera Director, Manager, Data Scien...

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)

cloudera

- 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)

The screenshot shows the AWS VPC Management Console interface. On the left, there's a sidebar with various VPC-related options like Virtual Private Cloud, Route Tables, Internet Gateways, etc. The main area is titled 'Create Security Group' and shows a list of security groups. One group, 'ahelb-CDSW-DEV-secgroup', is selected and detailed below. This group has a description of 'Cloudera Director, Manager, Data Scien...' and is associated with VPC 'vpc-64681403'. It also lists a 'default' security group. Below this, the specific configuration for 'ahelb-CDSW-DEV-secgroup' is shown, including its inbound rules:

Type	Protocol	Port Range	Source	Description
HTTP (80)	TCP (6)	80	0.0.0.0/0	
All Traffic	ALL	ALL	sg-190d3e61	
SSH (22)	TCP (6)	22	0.0.0.0/0	
Custom TCP Rule	TCP (6)	7180	0.0.0.0/0	
Custom TCP Rule	TCP (6)	7189	0.0.0.0/0	

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

The screenshot shows the AWS EC2 Management Console interface. On the left, there's a sidebar with various navigation links like EC2 Dashboard, Instances, and Key Pairs. The main area shows a list of existing key pairs. A modal dialog box titled "Import Key Pair" is open in the center. It contains instructions to click "Browse" and navigate to a public key file, or to copy and paste its contents. A "Choose file" button is highlighted, and the file path "cdsw-admin.pub" is shown. Below it, a "Key pair name" input field contains "cdsw-admin". Underneath, a "Public key contents" section displays a long string of base64-encoded RSA public key data:

```
ssh-rsa
AAAAB3NzaC1yc2EAAAQABAAQACQCrqjPv17zZQ4Y1+OHJu/V/1cNg9vCxGwaf4ym9Ta
4nSwIPRHnng81Jh+10lRh6yFBmD6cTU4S6kdJd3ErnBXMeicT3mQFTWEx9pXuoX/XJoZ8OF
+cEfot5jmScgHM3Ml80zD9t059UDY7d9aWPbmffx/kMahAd+DQOp2jl/n8FxvrX2fGvVrnxWx
+ZbiSS/jlEsH+wLcwifN+AVOaz29qz6muFnCeF49rWKQ6ct48Ajl-q-pVSRI/0apzRG95BEUvg
MefRn17wrDw3W7/5K51fj+lKrtVLvhrlqH4XhBP73pan4EAPYT8CY9HHA1c2Dy0gWWZcA9/bz
... (truncated)
```

At the bottom of the dialog, there are "Cancel" and "Import" buttons.

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**.

The screenshot shows the AWS EC2 Management Console interface. The user is on the 'Choose an Amazon Machine Image (AMI)' step of the instance creation wizard. A search bar at the top right contains the text 'centos'. Below it, a list of AMIs is displayed:

- CentOS 7 (x86_64) - with Updates HVM**: Selected. It has a rating of ★★★★☆ (48) and 1704 previous versions. It is sold by Centos.org. It is a Linux/Unix image. It is marked as 'Free tier eligible'. The description states it is the official CentOS 7 x86_64 HVM image built with a minimal profile for HVM instance types. A 'Select' button is visible.
- CentOS 6 (x86_64) - with Updates HVM**: It has a rating of ★★★☆☆ (32) and 1704 previous versions. It is sold by Centos.org. It is a Linux/Unix image. It is marked as 'Free tier eligible'. The description states it is the official CentOS 6 x86_64 HVM image built with a minimal profile for HVM instance types. A 'Select' button is visible.
- CentOS 6.5 (x86_64) - Release Media**: It has a rating of ★★★★☆ (55) and 6.5 - 2013-12-01. It is sold by Centos.org. It is a Linux/Unix image. A 'Select' button is visible.

On the left sidebar, there are filters for 'Quick Start', 'My AMIs', 'AWS Marketplace', 'Community AMIs', 'Categories' (All Categories, Software Infrastructure (123), Developer Tools (1), Business Software (5)), 'Operating System' (All Linux/Unix, Amazon Linux (6), CentOS (116), Ubuntu (1)), and 'All Linux/Unix' (Amazon Linux (6), CentOS (116), Ubuntu (1)).

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 | aheib-CDSW-DEV-subnet | 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**

The screenshot shows the AWS EC2 Management Console Launch Instance Wizard. The current step is "Step 3: Configure Instance Details". The configuration includes:

- Number of instances: 1
- Purchasing option: Request Spot Instances
- Network: vpc-64681403 | aheib-CDSW-VPC
- Subnet: subnet-cdc1aaaa | aheib-CDSW-DEV-subnet | eu-w
- Auto-assign Public IP: Enabled
- IAM role: None
- Shutdown behavior: Stop
- Enable termination protection: Protect against accidental termination
- Monitoring: Enable CloudWatch detailed monitoring (Additional charges apply)
- Tenancy: Shared - Run a shared hardware instance

Under "Advanced Details", User data is set to "As text" and the file "cloudera-director-install.sh" is selected. The "Review and Launch" button is visible at the bottom right.



- 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 an 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

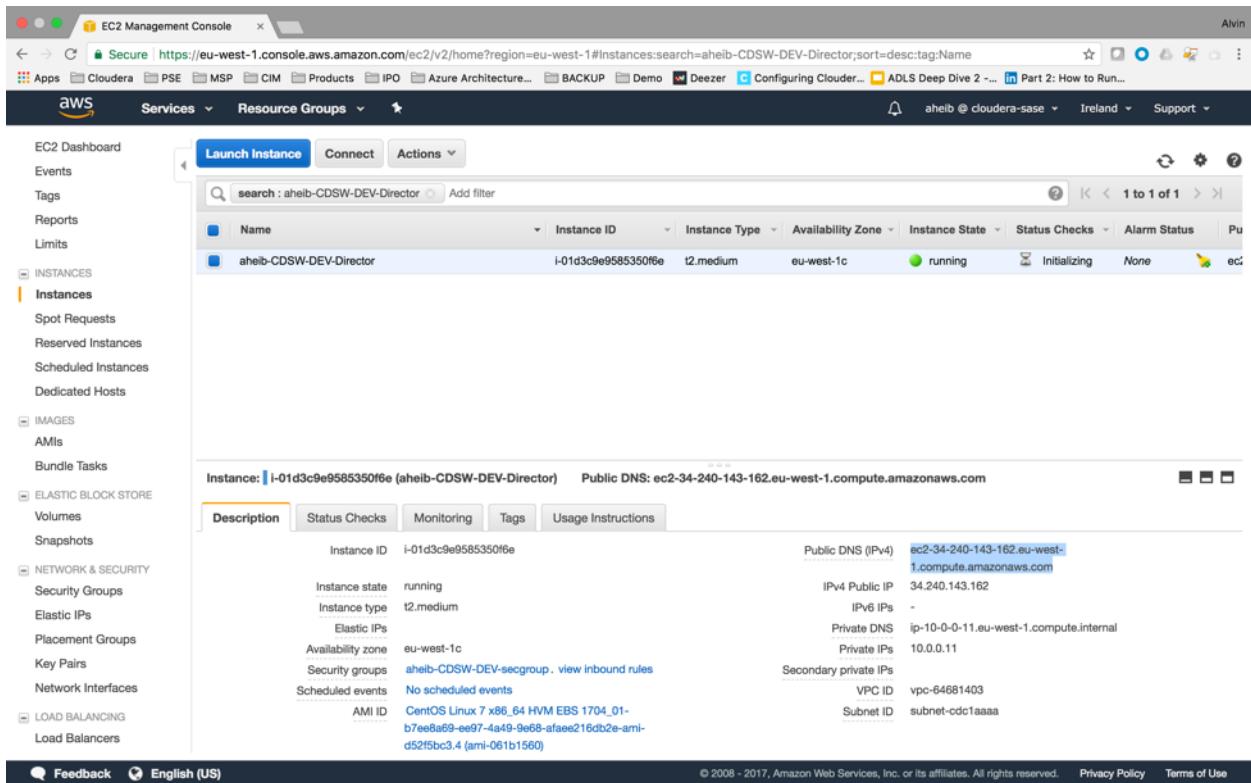
Make sure you are still in the **EC2 Services page** and select **Instances** on the left menu.

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

The screenshot shows the Cloudera Director interface. At the top, there's a navigation bar with tabs like EC2 Management Console, Welcome | Cloudera Director, and a search bar. Below the bar, a menu bar includes Apps, Cloudera, PSE, MSP, CIM, Products, IPO, Azure Architecture..., BACKUP, Demo, Deezer, Configuring Cloudera, ADLS Deep Dive 2, Part 2: How to Run..., and Part 2: How to Run... (repeated). A sidebar on the left lists steps: 1. Add Environment, 2. Add Cloudera Manager, and 3. Add Cluster. Step 1 is expanded, showing requirements for Cloud Provider credentials and SSH keys. Step 2 shows requirements for an Instance Template and Cloudera Manager repository, with options for External Database and Database Server Template. Step 3 shows requirements for Instance Groups, Instance Template, and instance count. A note at the bottom says to see the documentation for more information. On the right, a user menu for 'admin' is open, showing options for Change Password, Manage Users, Manage Billing, and Logout. A large blue button at the bottom right says 'Let's get started!'. The URL in the address bar is ec2-34-240-143-162.eu-west-1.compute.amazonaws.com:7189/#.



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.

The screenshot shows the 'Add Environment' wizard in the Cloudera Director interface. The 'General Information' step is active. The form fields are as follows:

- Environment name: AWS-CDSW-DEV
- Cloud provider: Amazon Web Services (AWS)
- Access key ID: AKIAIOSFODNN7EXAMPLE
- Secret access key: (redacted)
- EC2 region: eu-west-1
- RDS region: eu-west-1
- Username: centos
- Private key: cdsw-admin (File Upload selected)

At the bottom right are 'Quit' and 'Continue' buttons.

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



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 priviledges "/!\ It is not your AWS username !/"
Private Key	cdsw-admin	Link to file ~/Github/CDSW/AWS/cdsw-admin

B. Create Node Templates

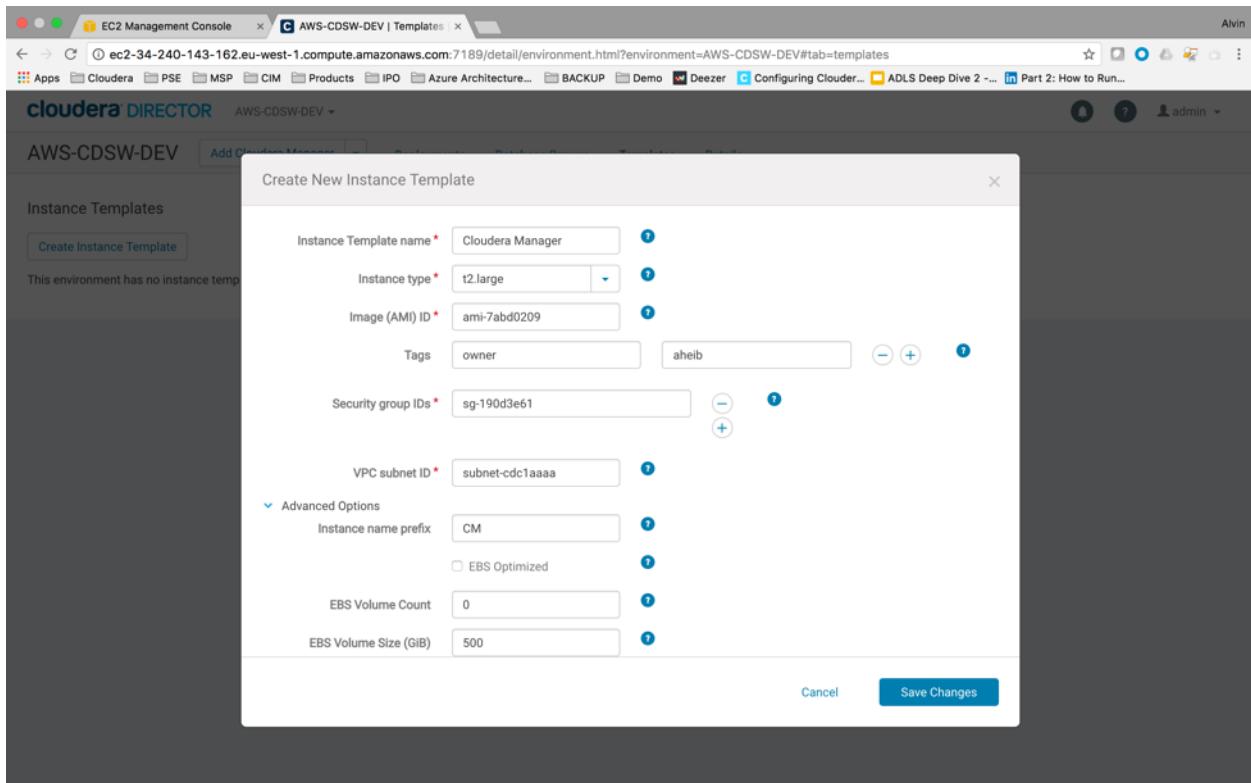
We will not proceed with the Cloudera Manager creation directly, instead we will create templates for Cloudera Manager / Master / Worker and CDSW-Master nodes.

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

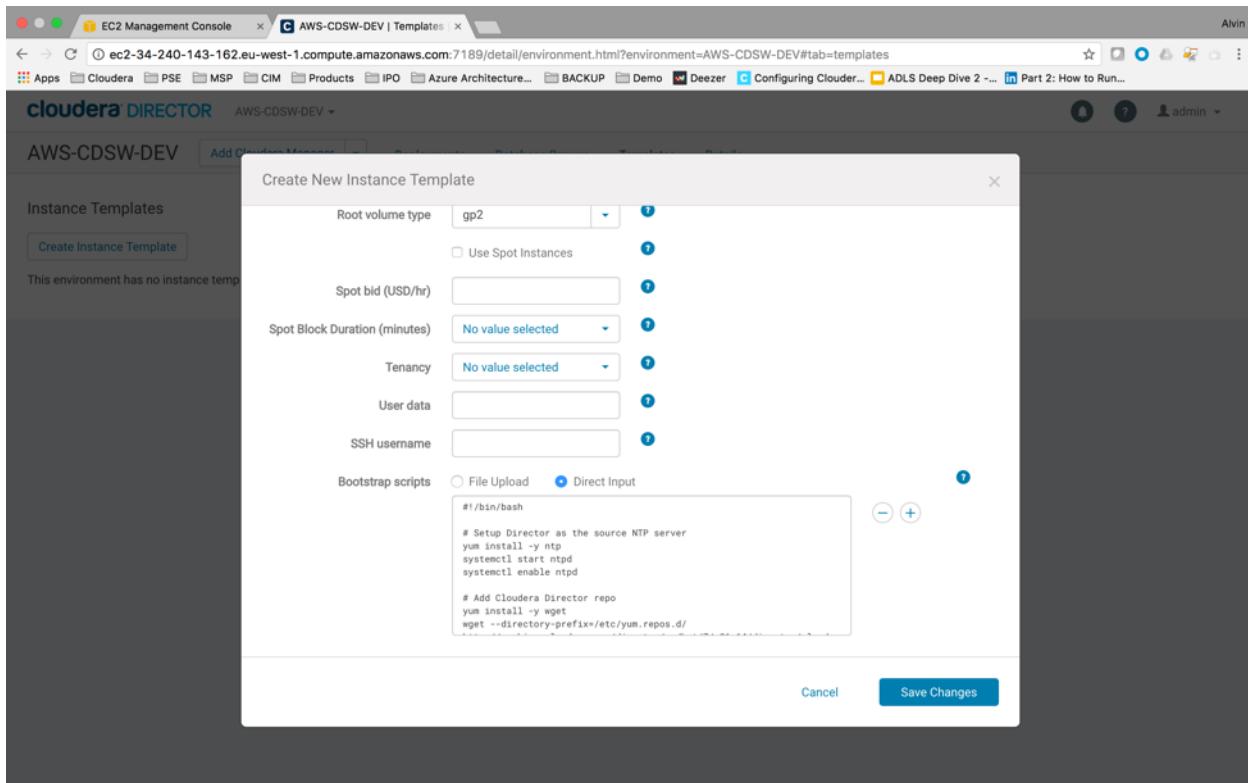
Select **Templates** menu, on top menu row. Then click on **Create Instance Template**.

- 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



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.

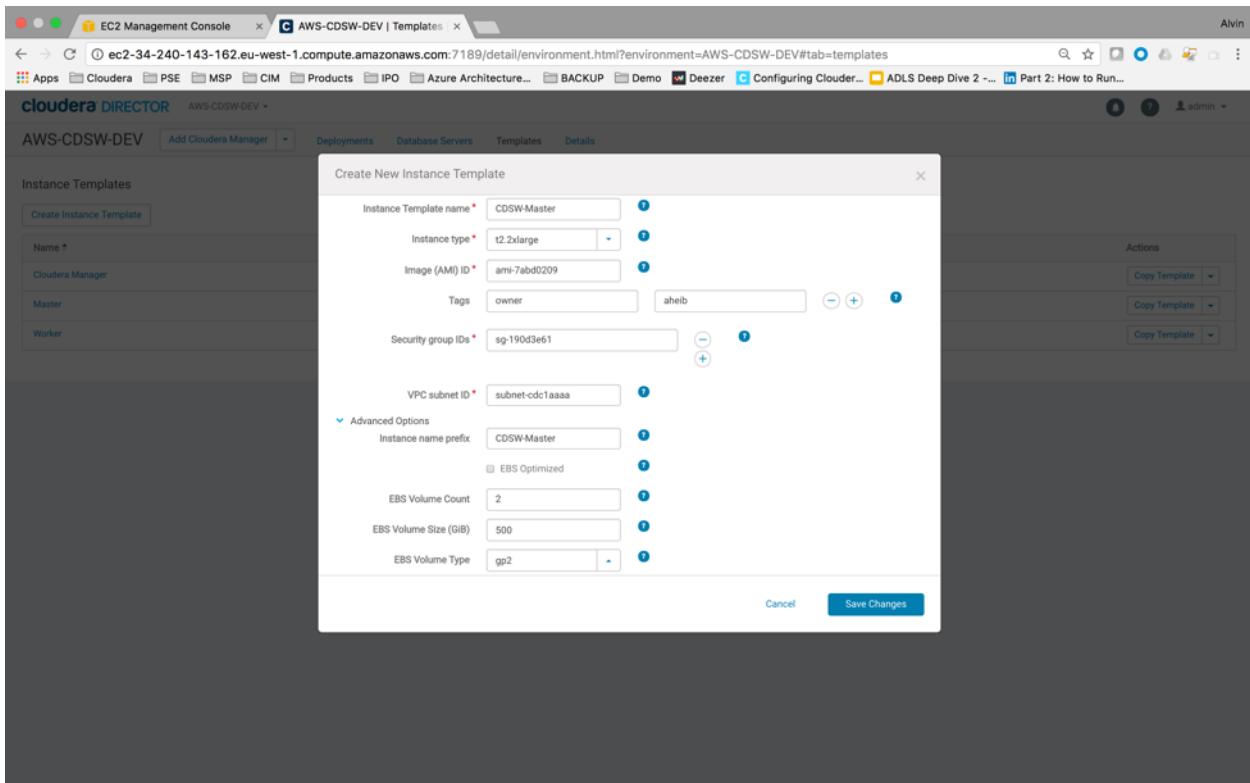
Key	Value	Comments
Template Name	Master	in our example
Instance Name Prefix	Master	this will help in identifying the node type under AWS Console

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

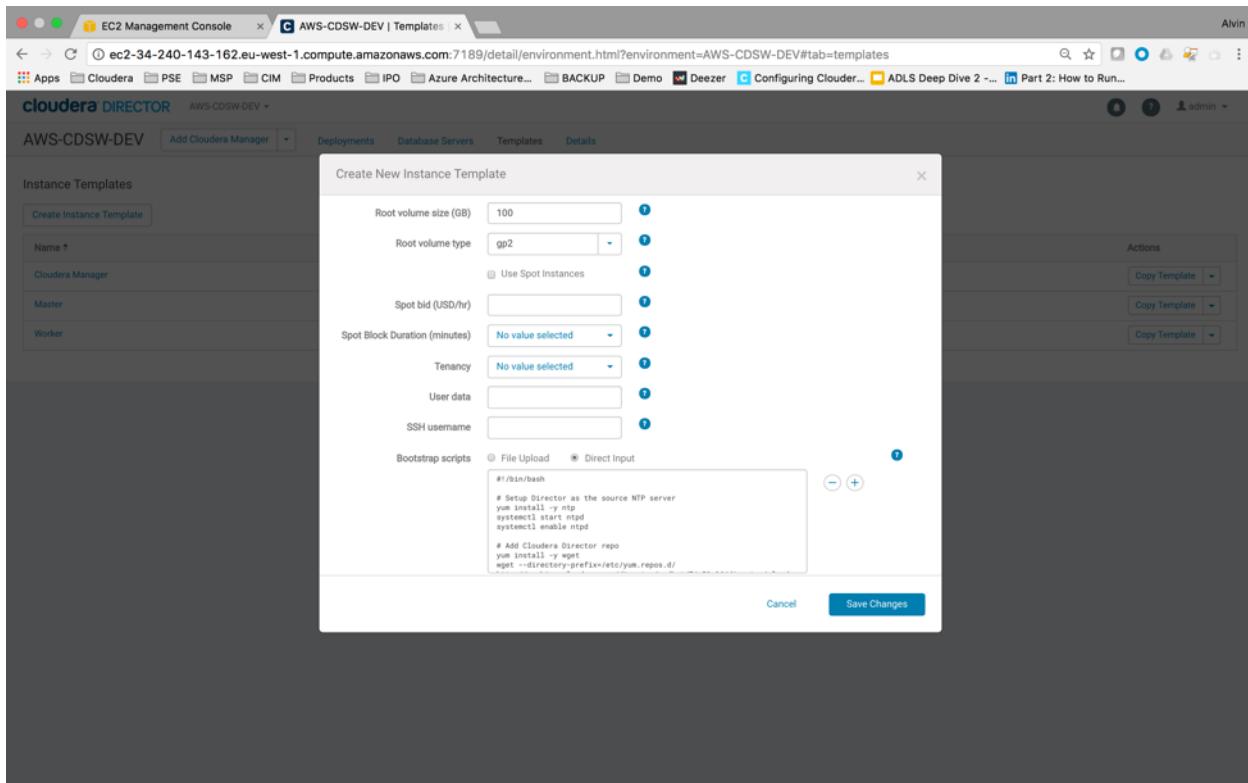
Key	Value	Comments
Template Name	Worker	in our example
Instance Type	t2.xlarge	for a DEV environment
Instance Name Prefix	Worker	this will help in identifying the node type under AWS Console
EBS Volume Count	1	we will need a 1 * 500 GiB HDFS disk
EBS Volume Size	500	we will need a 1 * 500 GiB HDFS disk

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



C. Deploy your first Cloudera Manager Instance

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

The screenshot shows the 'Add Deployment' wizard for Cloudera Director. The 'Environment' dropdown is set to 'AWS-CDSW-DEV'. The 'Cloudera Manager Name' field contains 'CM-DEV-0'. The 'Instance Template' dropdown is set to 'Cloudera Manager'. The 'Desired License Type' dropdown is set to 'Cloudera Enterprise Trial'. The 'Database Server' dropdown is set to 'Embedded Database'. Under 'Configurations (optional)', there is a 'Cloudera Manager Configurations' button, an unchecked checkbox for 'Override default Cloudera Manager repository', and an unchecked checkbox for 'Enable Kerberos'. The 'Cloudera Manager Admin Username' field contains 'admin', and the 'Cloudera Manager Admin Password' and 'Re-enter Password' fields both contain masked text. At the bottom left is a 'Quit' button, and at the bottom right is a 'Continue' button.

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.

The screenshot shows the 'Add Cluster' wizard interface. At the top, the cluster name is set to 'EDH-DEV-0'. Under 'Products', 'Name' is selected and 'CDH' is chosen, with 'Version' set to '5'. In the 'Services' section, 'Core Hadoop' is selected, which includes HDFS, Hive, Hue, Oozie, YARN, ZooKeeper. Other options like 'Core Hadoop with Impala' and 'Core Hadoop with Spark' are also listed. Below the services, there's a checkbox for 'Enable Auto-repair'. The 'Instance groups' section lists three groups: 'masters', 'workers', and 'gateway'. Each group has its 'Group name', 'Roles' (Master, Worker, CDSW-Master), 'Instance Template' (Master, Worker, CDSW-Master), 'Instance Count' (1, 4, 1), and 'Minimum Instance Count' (1, 3, 1). Buttons at the bottom include 'Quit' and 'Continue'.



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...

Cloudera DIRECTOR AWS-CDSW-DEV

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: EDH-DEV-0 Status: Good health Services: Core Hadoop

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	ahelb-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

CDH version: 5.12.1-1.cdh5.12.1.p0.3 Actions: Modify Cluster



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.clouderajar
```

```
sudo chmod 644 /opt/cloudera/csd/SPARK2_ON_YARN-2.2.0.clouderajar
```

```
sudo chown cloudera-scm:cloudera-scm /opt/cloudera/csd/SPARK2_ON_YARN-2.2.0.clouderajar
```

```
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/](http://[CLOUDERA-MANAGER-PUBLIC-DNS]:7180/)

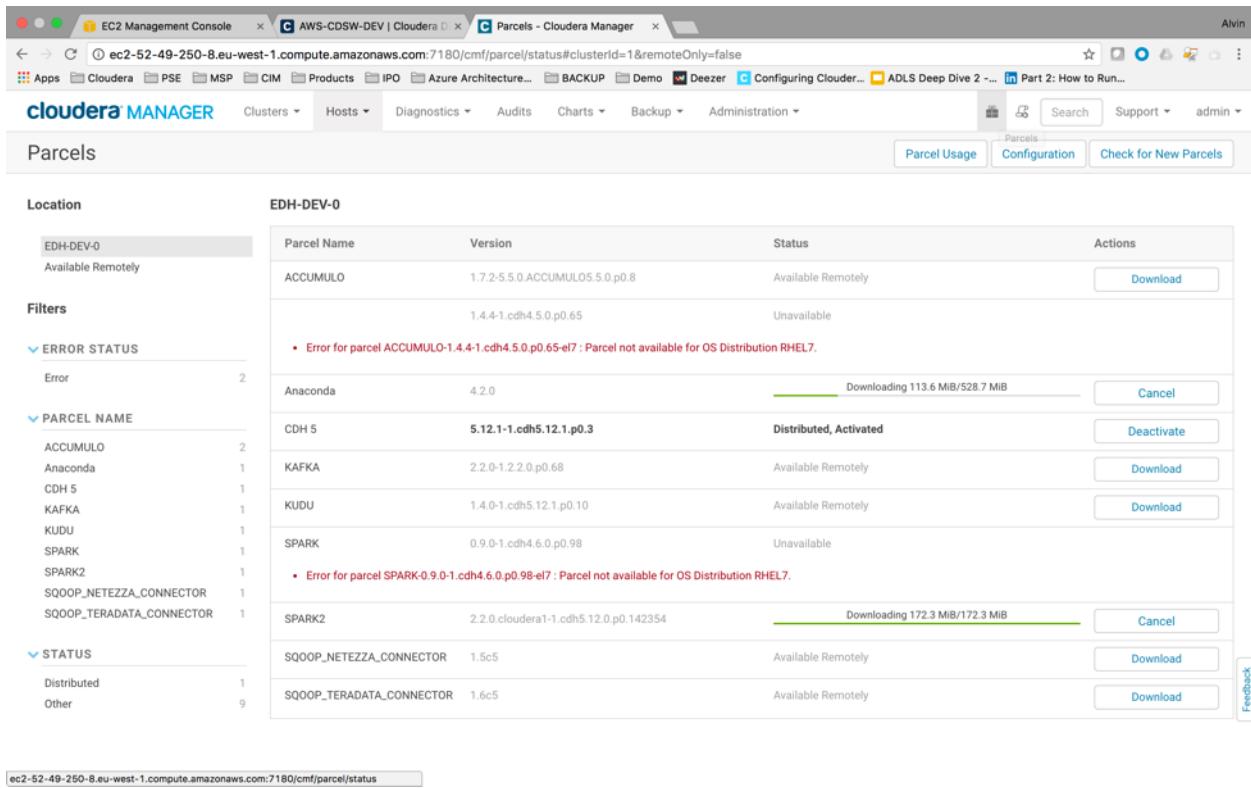
- We will need to setup JAVA_HOME environment variable. On the Cloudera Manager search bar please enter JAVA_HOME and set the environment variable to
`/usr/java/jdk1.8.0_121-cloudera`

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/>



The screenshot shows the Cloudera Manager interface for the EDH-DEV-0 cluster. The 'Parcels' tab is selected. The main table displays the following data:

Parcel Name	Version	Status	Actions
ACCUMULO	1.7.2-5.5.0.ACCUMULOS.5.0.p0.8	Available Remotely	<button>Download</button>
ANACONDA	1.4.4-1.cdh4.5.0.p0.65	Unavailable	
• Error for parcel ACCUMULO-1.4.4-1.cdh4.5.0.p0.65-el7 : Parcel not available for OS Distribution RHEL7.			
CDH 5	5.12.1-1.cdh5.12.1.p0.3	Distributed, Activated	<button>Deactivate</button>
KAFKA	2.2.0-1.2.2.0.p0.68	Available Remotely	<button>Download</button>
KUDU	1.4.0-1.cdh5.12.1.p0.10	Available Remotely	<button>Download</button>
SPARK	0.9.0-1.cdh4.6.0.p0.98	Unavailable	
• Error for parcel SPARK-0.9.0-1.cdh4.6.0.p0.98-el7 : Parcel not available for OS Distribution RHEL7.			
SPARK2	2.2.0.cloudera1-1.cdh5.12.0.p0.142354	Downloading 172.3 MiB/172.3 MiB	<button>Cancel</button>
SQOOP_NETEZZA_CONNECTOR	1.5c5	Available Remotely	<button>Download</button>
SQOOP_TERADATA_CONNECTOR	1.6c5	Available Remotely	<button>Download</button>

On the left, there are filters for ERROR STATUS (Error: 2), PARCEL NAME (e.g., ACCUMULO, ANACONDA, CDH 5, KAFKA, KUDU, SPARK, SPARK2, SQOOP_NETEZZA_CONNECTOR, SQOOP_TERADATA_CONNECTOR), and STATUS (Distributed: 1, Other: 9).

- 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.

1 Host Selected

Select hosts for a new or existing role. The host list is filtered to remove hosts that are not valid candidates; these include hosts that are unhealthy, members of other clusters, or have an incompatible version of CDH installed on them.

Hostname	IP Address	Rack	Cores	Physical Memory	Existing Roles	Added Roles
<input type="checkbox"/> ip-10-0-0-128.eu-west-1.compute.internal	10.0.0.128	/default	4	15.3 GiB	DN NM	
<input type="checkbox"/> ip-10-0-0-154.eu-west-1.compute.internal	10.0.0.154	/default	4	15.3 GiB	DN NM	
<input checked="" type="checkbox"/> ip-10-0-0-178.eu-west-1.compute.internal	10.0.0.178	/default	2	7.4 GiB	B NN SNM HDFS HS2 HS OS RM S	HS
<input type="checkbox"/> ip-10-0-0-231.eu-west-1.compute.internal	10.0.0.231	/default	4	15.3 GiB	DN NM	
<input type="checkbox"/> ip-10-0-0-245.eu-west-1.compute.internal	10.0.0.245	/default	8	31 GiB	G G G G	

Cancel OK

cloudera

5 Hosts Selected

Select hosts for a new or existing role. The host list is filtered to remove hosts that are not valid candidates; these include hosts that are unhealthy, members of other clusters, or have an incompatible version of CDH installed on them.

Enter hostname: host01, host[01-10], IP addresses or rack.

Tip: Click the first checkbox, hold down the Shift key and click the last checkbox to select a range.

Hostname	IP Address	Rack	Cores	Physical Memory	Existing Roles	Added Roles
ip-10-0-0-128.eu-west-1.compute.internal	10.0.0.128	/default	4	15.3 GiB	EN NM	G
ip-10-0-0-154.eu-west-1.compute.internal	10.0.0.154	/default	4	15.3 GiB	EN NM	G
ip-10-0-0-178.eu-west-1.compute.internal	10.0.0.178	/default	2	7.4 GiB	B NN SNN HMS HS2 HS OS HS JHS RM G	HS G
ip-10-0-0-231.eu-west-1.compute.internal	10.0.0.231	/default	4	15.3 GiB	EN NM	G
ip-10-0-0-245.eu-west-1.compute.internal	10.0.0.245	/default	8	31 GiB	G G G G	G

Feedback

- 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

The screenshot shows the Cloudera Manager Home page. At the top left, there's a sidebar for the 'EDH-DEV-0' cluster, listing components like Hosts, HDFS-1, HIVE-1, HUE-1, OOZIE-1, SPARK_ON..., Spark 2, YARN-1, and ZOOKEEPER... with their respective健康状态 (green checkmark). Below this is a section for the 'Cloudera Management Service' with one component listed. The main area features four performance charts:

- Cluster CPU:** A line chart showing host CPU usage across hosts over time. It indicates a peak around 10 PM at approximately 2.2% usage.
- Cluster Disk IO:** A stacked area chart showing total disk bytes per second. It includes segments for total disk bytes, total bytes read, and total bytes written. The total bytes written peak at about 143M/s around 10 PM.
- Cluster Network IO:** A stacked area chart showing bytes per second. It includes segments for total bytes received and total bytes transferred. The total bytes transferred peak at about 98.1M/s around 10 PM.
- HDFS IO:** A stacked area chart showing bytes per second. It includes segments for total bytes read and total bytes written. The total bytes written peak at about 4b/s around 10 PM.

At the bottom right of the charts, there's a 'Feedback' button. The top right corner of the screen shows the user 'Alvin'.



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]**.

The screenshot shows the Cloudera Director interface for the cluster **EDH-DEV-0**. The left sidebar lists services like Oozie, Hue, Spark on YARN, YARN, Hive, HDFS, SPARK2_ON_YARN, and ZooKeeper. The main panel shows the **Instance Group Details** for the gateway instance. The instance group name is **10.0.0.245**. The modal window provides detailed information about the instance:

Property	Value
Architecture	x86_64
EBS-optimized	false
Hypervisor	xen
Image (AMI) ID	ami-7abbd0209
Instance ID	i-0c30b5af30e4af0e2
Lifecycle	not specified
Instance type	t2.2xlarge
Key pair name	cdsw-admin
Launch time	Mon Sep 25 21:18:37 UTC 2017
Platform	not specified
Private DNS	ip-10-0-0-245.eu-west-1.compute.internal
Private IP	10.0.0.245
Public DNS	ec2-34-240-73-83.eu-west-1.compute.amazonaws.com
Public IP	34.240.73.83
Root device	/dev/sda1
Root device type	ebs
Source/dest. check	true
Spot Instance	false
Enhanced networking (SR-IOV)	not specified
Security Groups	ahelb-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

- 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 xvgd=block device for cdsw user-data)

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

```
sudo umount /dev/xvdf /dev/xvgd
```

- 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 entries for **xvdf** and **xvgd**.
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/xvgd /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"`)

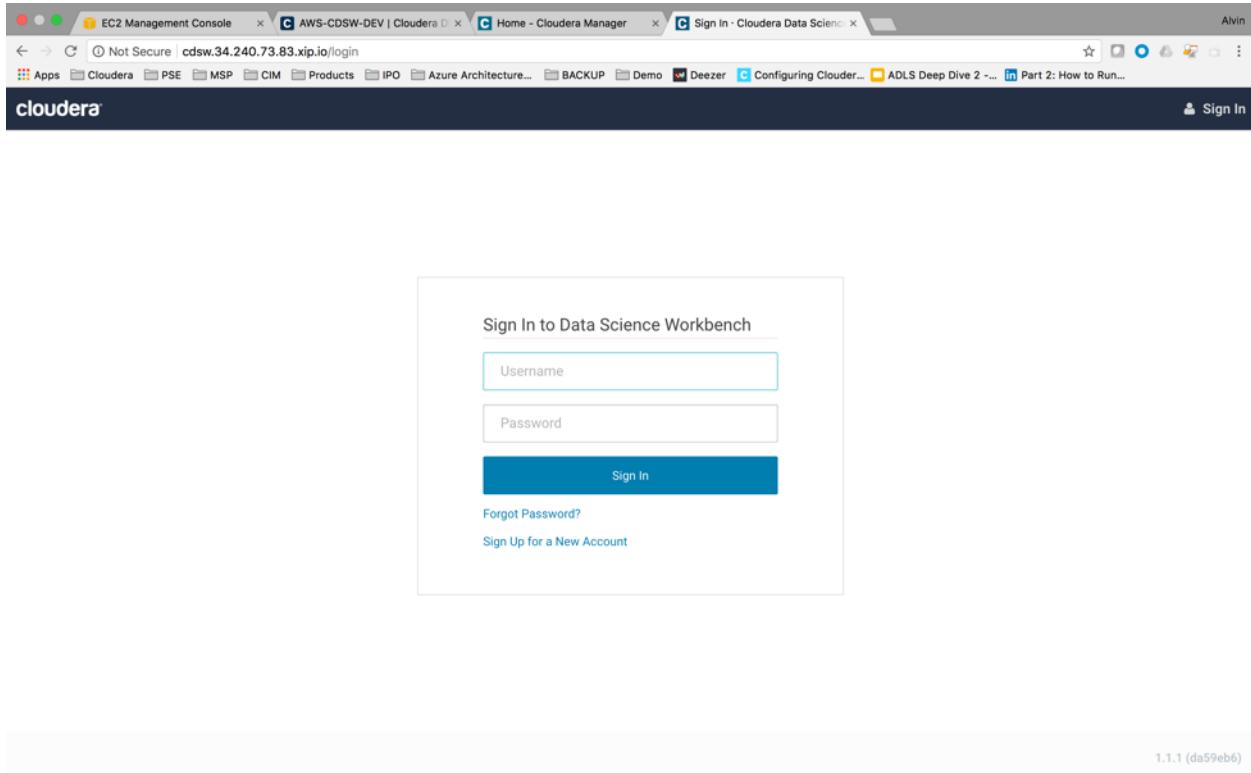
H. Launch Cloudera Data Science Workbench

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

cloudera

- You could check Cloudera Data Science Workbench status using command `sudo cdsweb status`.
- You can now use Cloudera Data Science Workbench at URL:

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





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

The screenshot shows the Cloudera Data Science Workbench (CDSW) Admin interface. The top navigation bar includes tabs for EC2 Management Console, AWS-CDSW-DEV | Cloudera Manager, Home - Cloudera Manager, and Admin - Cloudera Data Science Workbench. The current tab is 'Admin - Cloudera Data Science Workbench'. The top right corner shows the user 'Alvin' and a dropdown menu. Below the header is a toolbar with various icons for Cloudera, PSE, MSP, CIM, Products, IPO, Azure Architecture, BACKUP, Demo, Deezer, Configuring Cloudera, ADLS Deep Dive 2, and Part 2: How to Run... A search bar labeled 'Project quick find' and a 'New Project' button are also present.

The main dashboard has a dark sidebar on the left with icons for Projects, Sessions, and Admin. The 'Projects' section shows 0 sessions running and 0 jobs running. It also features four circular progress indicators: 0 vCPU (out of 8), 0 B (out of 30.92 GiB), and two other unlabelled ones. A 'New Project' button is located in the bottom right of this section. The 'Sessions' section indicates there are no projects created by 'admin' yet. The 'Admin' section provides instructions for working with teams and a 'Create a Team' link. At the bottom left, a 'License Expires in 59 days' message is displayed, and at the bottom right, the version '1.1.1 (da59eb6)' is shown.



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

The screenshot shows the Cloudera Data Science Workbench Admin interface. On the left, there's a sidebar with 'Projects', 'Jobs', 'Sessions', 'Settings', and 'Admin' selected. A message 'License Expires in 59 days' is displayed. The main content area has tabs for 'Overview', 'Users', 'Activity', 'Engines' (selected), 'Security', 'License', and 'Settings'. Under 'Engines Profiles', there are two entries:

Description	vCPU (burstable)	Memory (GiB)	Actions
1 vCPU / 2 GiB Memory	1	2	Edit Delete
2 vCPU / 4 GiB Memory	2	4	Edit Delete

A note below says: "vCPU is expressed in fractional virtual cores and allows bursting. Memory is expressed in fractional GiB and is enforced by memory killer. GPU indicates the number of GPUs that need to be used by the engine. Configurations larger than the maximum allocatable CPU, memory and GPU per node will be unschedulable." Under 'Engine Images', there is one entry:

Description	Repository:Tag	Default	Actions
Base Image v2	docker.repository.cloudera.com/cdsw/engine:2	●	Edit Deprecate Add

A note below says: "Whitelist Docker images for project owners to use in their jobs and sessions. These must be public images in registries that are accessible from the Cloudera Data Science Workbench hosts." Under 'Environmental variables', there is one entry:

Name	Value	Actions
HADOOP_USER_NAME	hdbs_super	Delete Add

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

The screenshot shows the Cloudera Manager interface for Site Administration. On the left, there's a sidebar with 'Admin' selected. The main area has tabs for Overview, Users, Activity, Engines (which is selected), Security, License, and Settings.

Engines Profiles: A table lists two profiles:

Description	vCPU (burstable)	Memory (GiB)	Actions
1 vCPU / 2 GiB Memory	1	2	Edit Delete
2 vCPU / 4 GiB Memory	2	4	Edit Delete
			Add

A note below the table states: "vCPU is expressed in fractional virtual cores and allows bursting. Memory is expressed in fractional GiB and is enforced by memory killer. GPU indicates the number of GPUs that need to be used by the engine. Configurations larger than the maximum allocatable CPU, memory and GPU per node will be unschedulable."

Engine Images: A table lists one image:

Description	Repository:Tag	Default	Actions
Base Image v2	docker.repository.cloudera.com/cdsw/engine:2	●	Edit Deprecate Add

A note below the table states: "Whitelist Docker images for project owners to use in their jobs and sessions. These must be public images in registries that are accessible from the Cloudera Data Science Workbench hosts."

Environmental variables: A table lists one variable:

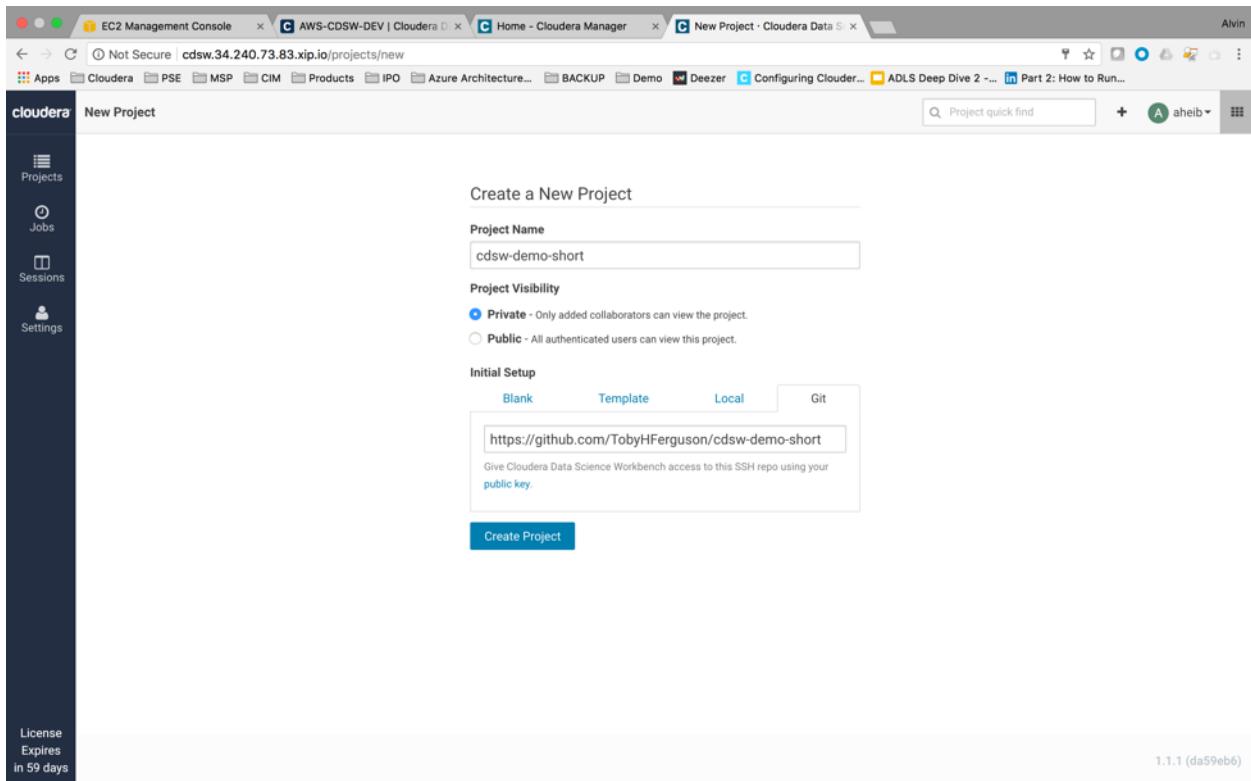
Name	Value	Actions
HADOOP_USER_NAME	hdbs_super	Delete Add

License Expires in 59 days is displayed at the bottom left.

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.

cloudera



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.



The screenshot shows the Cloudera Manager interface for Site Administration. The 'Email' section is active. The 'SMTP Host' field is set to 'smtp.example.com'. The 'SMTP Port' field is set to '25'. An error message 'Email configuration is invalid.' is displayed in a red box above the form. The 'Admin' sidebar on the left shows various management options like Projects, Jobs, Sessions, and Settings.

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

cloudera

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/TobyH Ferguson/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/*.* 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
```

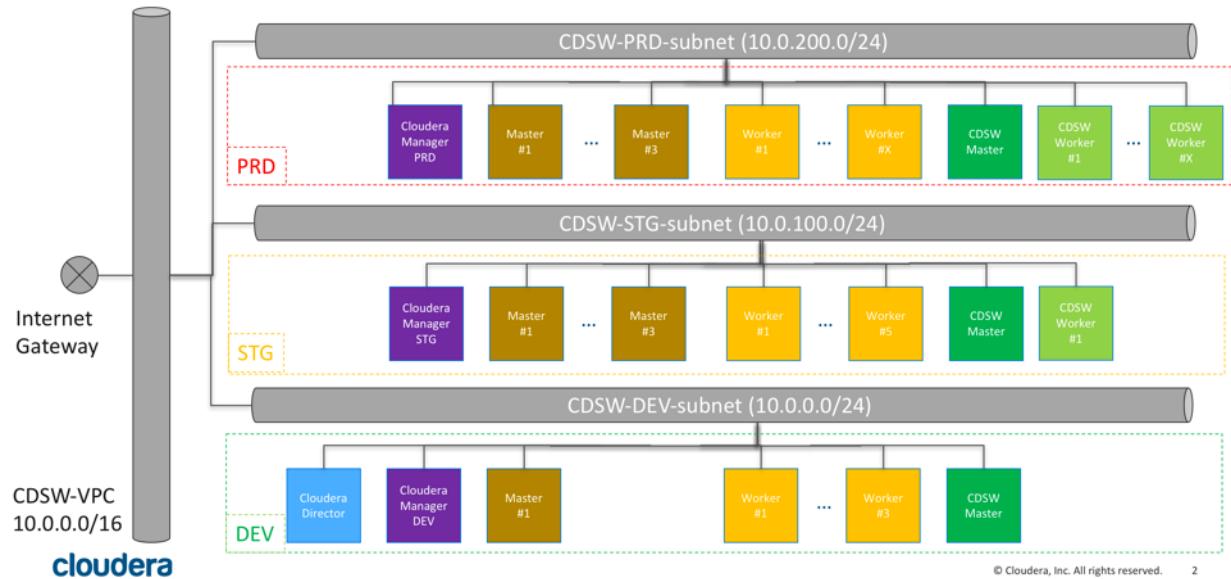
X. Production and Security Considerations

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.

CDSW Deployment on AWS





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/cdsweb/config/cdsweb.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 cdsweb 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 NODE TYPE	SYSTEM DISK							DATA DISK				
	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				4000

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