Installing a virtual Hadoop cluster with Vagrant and Cloudera Manager (http://dandydev.net/blog/installing-virtual-hadoop-cluster)

I recently started as a Big Data Engineer at The New Motion (http://www.thenewmotion.com). While researching our best options for running a Hadoop cluster, I wanted to try out some of the features available in the newest version of Cloudera's Hadoop distribution: CDH5 (http://www.cloudera.com/content/support/en/documentation/cdh5-documentation/cdh5-documentation-v5-latest.html). Of course I could've downloaded their QuickstartVM (http://www.cloudera.com/content/cloudera-content/cloudera-docs/DemoVMs/Cloudera-QuickStart-VM/cloudera_quickstart_vm.html), but I rather wanted to run a virtual cluster, making use of the 16GB of RAM my shiney new 15" Retina Macbook Pro has ;)

There are some tutorials, and repositories available for installing a local virtualized cluster, but none of them did what I wanted to do: install the bare cluster using Vagrant, and install the Hadoop stack using the Cloudera Manager. So I created a simple Vagrant setup myself. You can find it here (https://github.com/DandyDev/virtual-hadoop-cluster).

Setting up the virtual machines

As per the instructions from the Gitub repo:

Depending on the hardware of your computer, installation will probably take between 15 and 25 minutes.

First install VirtualBox (https://www.virtualbox.org/) and Vagrant (http://www.vagrantup.com/).

Install the Vagrant Hostmanager plugin (https://github.com/smdahlen/vagrant-hostmanager)

```
$ vagrant plugin install vagrant-hostmanager
```

Clone this repository.

```
$ git clone https://github.com/DandyDev/virtual-hadoop-cluster.git
```

Provision the bare cluster. It will ask you to enter your password, so it can modify your `/etc/hosts` file for easy access in your browser. It uses the Vagrant Hostmanager plugin to do this.

```
$ cd virtual-hadoop-cluster
$ vagrant up
```

Now we can install the Hadoop stack.

Installing Hadoop and related components

- Surf to: http://vm-cluster-node1:7180
- Login with `admin` / `admin`
- Select **Cloudera Express** and click *Continue* twice
- On the page where you have to specifiy hosts, enter the following: `vm-cluster-node[1-4]` and click *Search*. 4 nodes should pop up and be selected. Click *Continue*.
- On the next page ("Cluster Installation > Select Repository"), leave everything as is and click *Continue*
- On the next page ("Cluster Installation > Configure Java Encryption") I'd advise to tick the box, but only if your country allows it. Click *Continue*
- On this page do the following:
    - **Login To All Hosts As** : *Another user* -> enter `vagrant`
    - In the two password fields enter: `vagrant`
    - Click *Continue*
- wait for CM to install the prerequisites... and click *Continue*
- wait for CM to download and distribute the CDH packages... and click *Continue*
- wait while the installer is inspecting the hosts, and *Run Again* if you encounter any (serious) errors (I got some that went away the second time). After this, click *Finish*
- For now, we'll install everything but HBase. You can add HBase later, but it's quite taxing for the virtual cluster. So on the "Cluster Setup" page, choose *Custom Services* and select the following: **HDFS, Hive, Hue, Impala, Oozie, Solr, Spark, Sqoop2, YARN and ZooKeeper**. Click *Continue*
- On the next page, you can select what services end up on what nodes. Usually Cloudera Manager chooses the best configuration here, but you can change it if you want. For now, click *Continue*

- On the “Database Setup” page, leave it on *Use Embedded Database*. Click *Test Connection* (it says it will skip this step) and click *Continue*
- Click *Continue* on the “Review Changes” step. Cloudera Manager will now try to configure and start all services.

And you're **Done!**. Have fun experimenting with Hadoop!