How-to: Create a CDH Cluster on Amazon EC2 via Cloudera Manager

[Cloudera Manager](http://blog.cloudera.com/content/cloudera/en/products/cloudera-manager.html) includes a new express installation wizard for Amazon Web Services (AWS) EC2. Its goal is to enable Cloudera Manager users to provision CDH clusters and Cloudera Impala (the open source distributed query engine for Apache Hadoop) on EC2 as easily as possible (for testing and development purposes only, not supported for production workloads) – and thus is currently the fastest way to provision a Cloudera Manager-managed cluster in EC2.

The new distinguishing feature introduced in version 4.5 is that Cloudera Manager can now launch and configure the instances for you, so you don’t have to worry about launching the instances, authorizing SSH keys, and configuring a firewall. All this can now be done from within Cloudera Manager!

Since Cloudera Manager and the nodes running CDH use internal hostnames to communicate, the Cloudera Manager server must run on EC2 as well. In fact, the Cloud Express Wizard only appears when installing Cloudera Manager on EC2.

Here’s what you can do with Cloud Express Wizard:

* Provision new EC2 instances (AWS credentials required)
* Choose between CentOS and Ubuntu images (or a custom AMI)
* Choose your EC2 instance type
* Install the most recently released CDH, Cloudera Impala, and Cloudera Manager agent packages on them

And here’s what you cannot do:

* Use pre-existing EC2 instances
* Install older (earlier ) versions of CDH and Cloudera Manager, or use Parcels

I am excited to show you how this feature works. These instructions will set up a fully configured CDH cluster (all services with embedded PostgreSQL) from scratch in less than 15 minutes.

Step 1: Install Cloudera Manager Server on EC2

First, you will need to  launch an EC2 instance for the Cloudera Manager server, which will require an AWS Access Key ID and AWS Secret Key — please follow [these instructions](http://docs.aws.amazon.com/fws/1.1/GettingStartedGuide/index.html?AWSCredentials.html) if you need help getting them.

To launch the EC2 instance, go to “EC2” in the AWS web console and select “Instances” in the left menu. Before you provision the instance, select the EC2 region you want your instance to be in (dropdown in top right corner of the web console). For his demo, you can simply use the default “N. Virginia (us-east-1)” region. Click on “Launch Instance” and select the Classic Wizard. On the next page, pick the “Ubuntu Server 12.04 LTS” 64-bit image. You need one instance of type “m1.large.” You can keep the default values of other settings and proceed to the “Create Key Pair” page.

If you don’t have an SSH key imported to EC2 already, select “Create a new Key Pair.” Enter the name of your new key pair, and click “Create and Download your key pair.” This will download a .pem file to your computer. (Important: AWS does not store the private SSH keys, so save this file or you won’t be able to SSH into the instance we’re about to launch.)

It is very important to configure the EC2 firewall correctly. On the “Configure Firewall” page choose “Create a new Security Group,” and authorize all the ports listed below:

|  |  |  |
| --- | --- | --- |
| TCP | 22 | SSH |
| TCP | 7180 | Cloudera Manager web console |
| TCP | 7182 | Agent heartbeat |
| TCP | 7183 | (optional, Cloudera Manager web console with TLS) |
| TCP | 7432 | Embedded PostgreSQL |
| icmp | -1 | ping echo |

Next, go to the last page of the wizard and launch the instance!

**How to Install the Latest Version of Cloudera Manager**  
Once the state of the instance is “running” (provisioning takes usually less than 5 minutes), you  can SSH in and install Cloudera Manager 4.5. The public hostname of the instance is listed in the instance details in the AWS console.

|  |  |
| --- | --- |
| 1 | $ ssh -i your-key.pem ubuntu@ec2-xx-xx-xx-xx.compute-1.amazonaws.com |

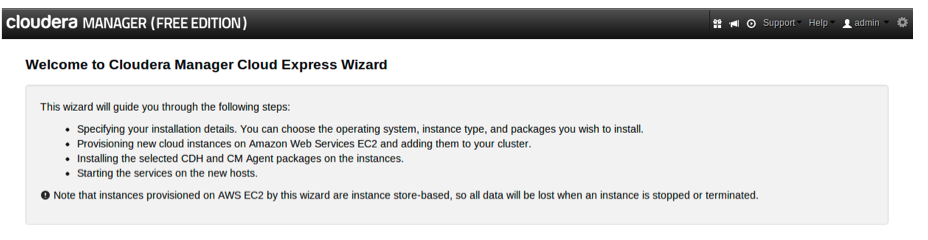
Download the Cloudera Manager 4.5 installer and execute it on the remote instance:

|  |  |
| --- | --- |
| 1  2  3 | $ wget http://archive.cloudera.com/cm4/installer/latest/cloudera-manager-installer.bin  $ chmod +x cloudera-manager-installer.bin  $ sudo ./cloudera-manager-installer.bin |

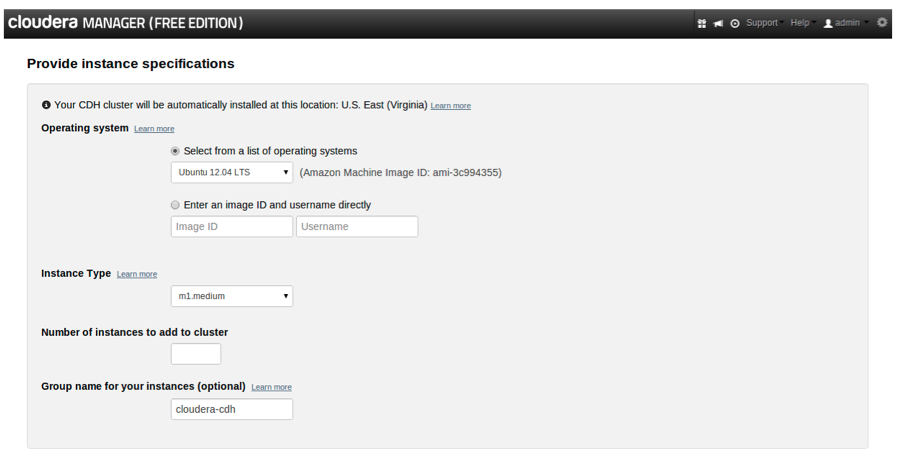
Once the installer finishes, use the public hostname of your server instance to navigate in your browser to http://ec2-xx-xx-xx-xx.compute-1.amazonaws.com:7180, and then log into the web console (the default username and password are both “admin”). If you’re successfully logged in, congratulations!

Step 2: Installing a CDH Cluster with Cloud Express Wizard

After logging in, Cloudera Manager will detect that it runs on EC2, and it will greet you with the welcome screen of the new wizard (see below). There is a warning that the instances started by this installer are instance store-based, which implies that stopping or terminating these instances results in losing all data stored on them. Remember to back-up  important data from the cluster before terminating the instances!

[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek1.png)**Figure 1: Cloud Express Wizard**

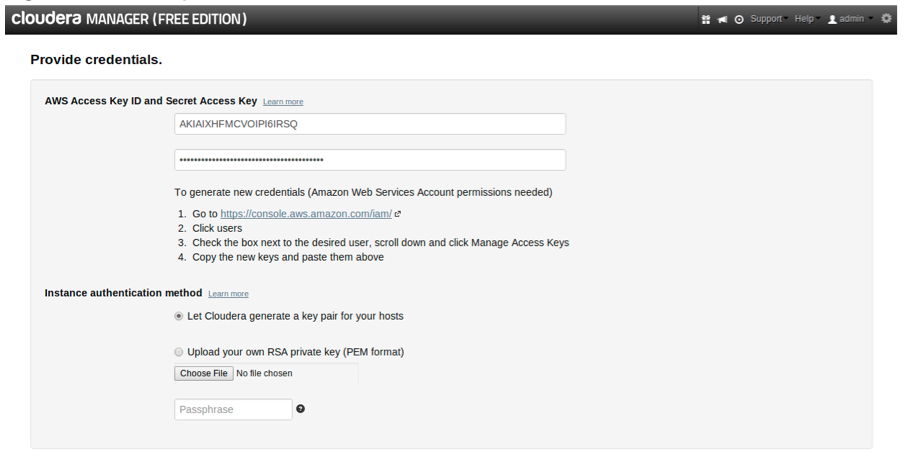
Why does Cloudera Manager prefer [instance store-backed over EBS-backed AMIs](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ComponentsAMIs.html)? Although EBS volumes offer persistent storage, they are network-attached and charge per I/O request, so they are not suitable for Hadoop deployments. If you wish to experiment with EBS-backed instances, you can always use a custom EBS AMI.

[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek2.png)  
**Figure 2: Cloud Express Wizard – instance specifications**

Go to the second page of the wizard (Figure 2) to specify the details about the hosts we are about to launch. Cloudera Manager detects the region it runs in, and the new instances will be installed there as well. The following attributes can be specified:

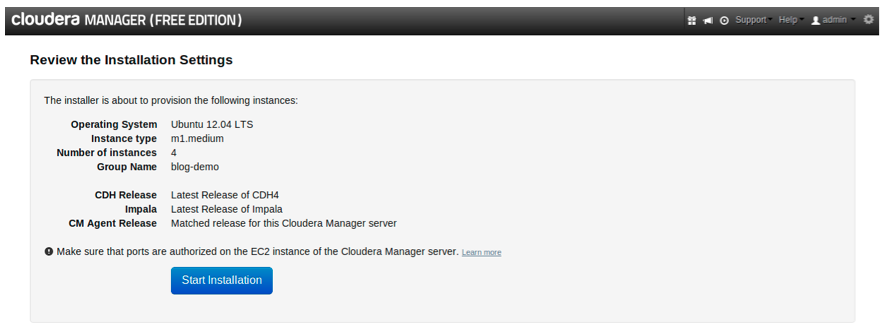
* OS (Amazon Machine Image, AMI): Cloudera supports Ubuntu 12.04 and CentOS 6.3 images. Cloudera Manager knows which AMI to use for the specified region. If you choose to use a custom AMI (this is especially handy if you want to pre-install some tools or authorize SSH keys on your hosts), make sure the AMI is available in the specified region.
* Instance Type: Only instance types matching the minimum requirements for CDH hosts are available. m1.medium will be sufficient for this demo. The high-storage instances (hs1.8xlarge) are not yet available but will be included in a future release of Cloudera Manager .
* Number of Instances: You will create four instances for this demo. Although there is no limit on the number of instances, you’re likely to exceed the EC2 API request limit  if you try to create more than ~20 instances at once.
* Group name: The optional “group name” is there to help you identify the instances launched by the wizard, and it will be used as suffix for the name, Security Group, and Key Pair of the instances.

The next page (Figure 3) shows you the credentials page. You need to paste in the AWS Access ID and AWS Secret Key. Then you can choose an SSH key for the hosts; in this demo I will let Cloudera Manager generate a new key pair for my instances, and the private key will be available for download on the next page once the instances are launched. If you upload an existing private SSH key, Cloudera Manager will extract the public part and authorize it in your AWS account.

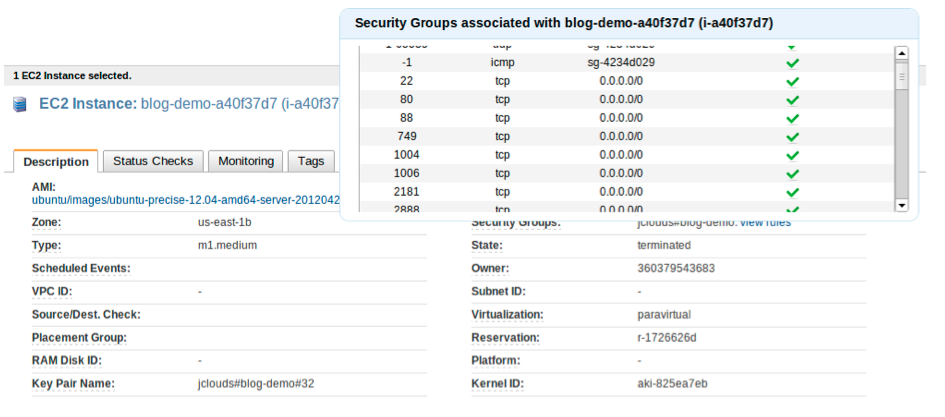
[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek3.png)  
**Figure 3: Cloud Express Wizard – Credentials**

Proceed to the review page (Figure 4), where you can double-check your installation settings. You can easily go back to modify the settings. However, once the instances are provisioned, you must terminate  them in order to make changes.

Note that when provisioning the instance fails on “503 Error: Api Request Limit exceeded”, it’s likely because other applications (or users) are issuing API calls to the same AWS account at the same time, or because you are launching a large number of instances at once. (In testing we successfully spun up as many as 20 instances  simultaneously.) This limitation will be removed in a future Cloudera Manager release.

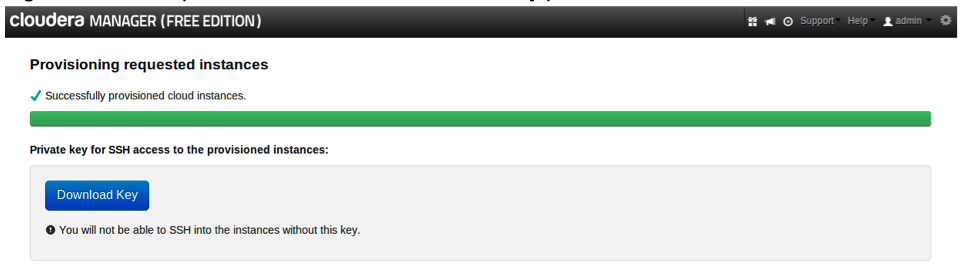
[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek4.png)  
**Figure 4: Cloud Express Wizard – Review Installation**

The review page indicates you are about to install the latest packages of CDH and Impala. Currently this is the only supported option in this installation wizard. If everything looks right, click the “Start Installation” button. (Note: if node installation fails because “CM failed to receive a heartbeat from Agent”, Confirm that port 7182 is authorized in the Security Group of Cloudera Manager server and re-try the installation.)

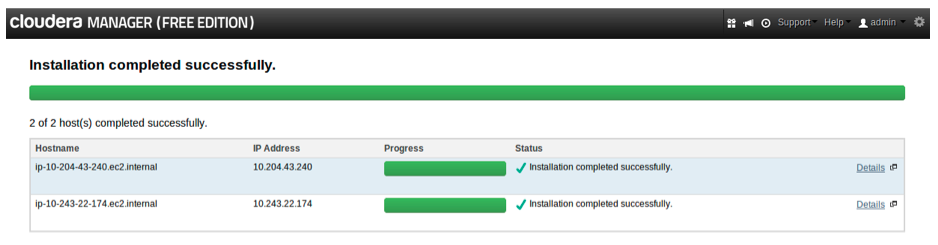
[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek5.png)  
**Figure 5: AWS web console – EC2 instance started by Cloudera Manager**

Cloudera Manager uses [jclouds](http://www.jclouds.org/) to create new key pair and security group, and to launch the EC2 instances. The new instances will also appear in your AWS EC2 console (Figure 5). You can see that the security group and the key pair starts with “jclouds#” prefix. Also, all ports required for CDH have already been enabled. Provisioning new instances takes usually less than five minutes.

Once the instances are successfully provisioned, you can download the private SSH key (Figure 6). It’s a good idea to download the key in case something goes wrong and you need to SSH in to investigate the issue. However, this installation path won’t require us to do anything manually on the remote hosts.

[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek6.png)  
**Figure 6: Cloud Express Wizard – Instances successfully provisioned**

The next screen looks familiar if you’ve used the classic express wizard in Cloudera Manager. It shows the progress of package installation on the newly provisioned hosts (Figure 7).

[](http://blog.cloudera.com/wp-content/uploads/2013/03/buzek7.png)  
**Figure 7: Cloud Express Wizard – Package installation**

After finishing the package installation, you can proceed to the Host Inspector and Services First Run page – you’re done. Congratulations, the CDH cluster is up and running now!

Note: The hosts cannot be terminated from Cloudera Manager, so to do that you’ll need to use EC2 CLI tools or the AWS web console instead. Go to the Instances page in <https://console.aws.amazon.com/ec2>, select the instance you created for the server and all the instances launched by the wizard (hint: use the Group Name string to filter them out), and click “Actions > Terminate”.

*Emanuel Buzek is a Software Engineer on the Enterprise team.*

**Editor’s Note (added Feb. 28, 2014): The instructions above are deprecated for Cloudera Manager releases beyond 4.5. Please refer to**[**this doc**](http://blog.cloudera.com/content/cloudera-content/cloudera-docs/CM4Ent/latest/Cloudera-Manager-Installation-Guide/cmig_install_on_EC2.html)**for instructions pertaining to releases 4.6 and later.**