Getting started with Cloud-Bio-Linux

Bela Tiwari

July 21, 2010

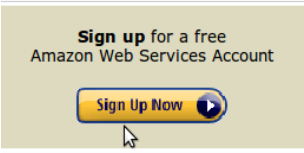

Figure 1: Signing up for an AWS account starts with the click of a button.

1 Get ready to get on the Amazon cloud

1.1 Get an Amazon AWS account

Anyone can set up an account with Amazon to access their computer cloud. Just go to http://aws.amazon.com and sign up for an account.

The rest of this document assumes you have an AWS account and you are logged into it.

1.2 Get an Amazon EC2 Account

There are various ways you can access the power of the Amazon cloud. In this document, we describe using EC2.

If you do not already have one, you need to sign up for an EC2 account. This is in addition to the general Amazon aws account you have if you followed the instructions above.

To get your EC2 account ,

1. click on the Products tab on the Amazon aws page,
2. click on the Amazon Elastic Compute Cloud (EC2) link that appears in the Compute section of the listing, and
3. click on the button in the right hand pane that says Sign up for Amazon EC2.
4. Complete the registration process[1].

1.3 Get an EC2 key pair

After you've signed up for your account, Amazon will send you an email with a link in it to the Access Identifiers section of your account. Amazon provides a list of which credentials you need to do particular tasks.

[1]Signing up for Amazon EC2 also automatically signs you up for Amazon Simple Storage Service and Amazon Virtual Private Cloud. You will not be charged for any service unless you use it.

If all you will be doing is starting up Bio-Linux using the Amazon (graphical) console, then you only need your Amazon EC2 Key Pair.

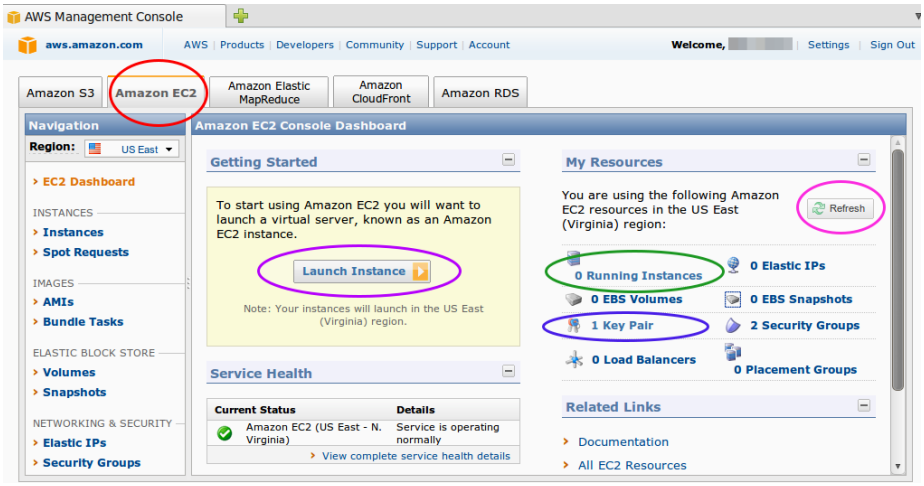

Figure 2: Your AWS home are will look something like this. Here, we are looking at the information under the EC2 tab (Red Circle). Blue Circle: your Key Pairs - if you don't have any, click on the link to create some. Purple Circle: You can easily launch any publicly available EC2 image from Amazon by clicking on this button. Green Circle: The number of running instances you have. If you have any, you can find out more about them by clicking on this link. Pink Cicle: If you add Key Pairs, or start up instances, you may need to hit the refresh button to see changes on your EC2 Dashboard.

To create a key pair:

1. Go to the the EC2 area on Amazon https://console.aws.amazon.com/ec2/home.
2. Click on the *Key Pairs* link under My Resources in the right hand area of the window. See the blue circle in figure .
3. Click on the *Create Key Pair* button near the top of the Key Pairs section of the window.
4. Give your key pair a memorable name when prompted. Save your private key to a safe location. See the further information below about this.
5. Click on the link in the left hand pane to go back to the *EC2 Dashboard* and then click on the *Refresh button* at the far right hand side of the window (see figure).

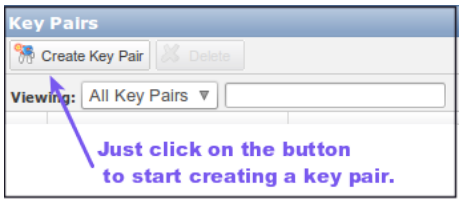

Figure 3: Keypair creation is simple - just click on the button and follow the instructions on screen.

You should now see that you have a key pair registered in the *My Resources* section. You may need to log out of Amazon at this point and log back in for the key pairs to be noticed by the system when you try to start up Cloud-Bio-Linux.

Each EC2 key pair includes a private key file and a public key file. **Save your private key to a secure and memorable location**. Don't lose it or share it. (Amazon does not make a copy of it.) If you're working on Linux, adjust the permissions on your key file so it is readable only by you.

If you plan to use the command line tools to start up an instance, you will also need to get your X509 certificates. This document assumes that you will only be using the graphical console, so this is not covered further here.

How the EC2 key pairs work

When you launch a Cloud-Bio-Linux instance from Amazon, you will specify a particular EC2 key pair name. The Amazon systeem puts a copy of your public key, which it has a record of, on the instance. You, as the (only!) holder of the private key will be the only one able to access the Bio-Linux instance you just started up.

2 Working on Cloud-Bio-Linux - the basics

2.1 The process in a nutshell

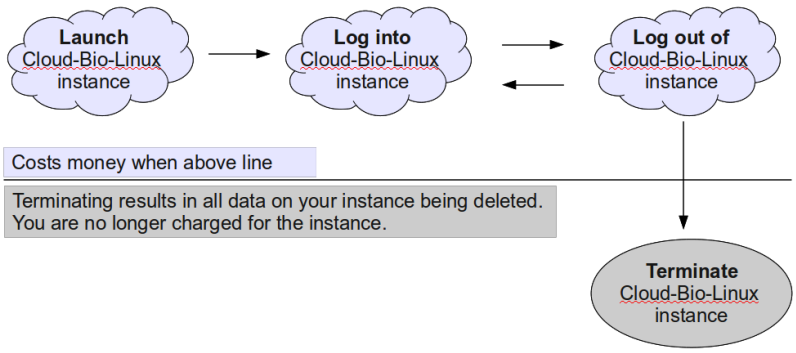

Figure 4: The most basic process of working on Cloud-Bio-Linux on Amazon EC2 is depicted here. While you can log in and out of your instance as often as you like, you continue to be charged for it whether you are logged in or not. You can also stop and re-start instances (not shown here). When you are finished with an instance, you should terminate it. Once terminated, the system and all data on it are deleted. This is the simplest setup. There are, however, easy ways to store data and even whole systems, at a fraction of the price of a running image.

The general process you will follow when working with Cloud-Bio-Linux is outlined in figure :

1. Start up a Cloud-Bio-Linux instance
2. Log into your Cloud-Bio-Linux instance
3. Log out of the Cloud-Bio-Linux instance
4. Still want to work on this instance? You can log into it and out of it as often as you like, or you can stop and start the instance, which can work out slightly cheaper.

5. When you're really finished, and don't need the Cloud-Bio-Linux instance anymore, terminate the instance. You will stop being charged for this instance when it is terminated. Stopping and terminating are different.

This chapter focusses on starting and logging into a full graphical Bio-Linux desktop on the cloud.

Of course, there are other things that you may wish to do, like save your system and data for use again later, or share it with others. These things and more are covered in the http://aws.amazon.com/ebs/official documentation for Elastic Block Storage.

A note about charging:

The charging structure for Amazon EC2 is well defined, if quite detailed. It is important to understand what you are being charged for, so you can make good decisions about when using the cloud is a cost effective option, and when it is not. You will be charged for running instances, and also for things like bandwidth when transferring data on and off Amazon systems, and data volumes you wish to use later. Please read the Amazon pricing documentation so you don't get surprised when you next see your credit card bill.

A couple of things to note when starting out:

- **You will be charged for the time your instance is running.** It's not about when you're logged into it that counts. Charging for the instance terminates when you terminate the instance. When you do this, all your data and files will be deleted, along with the running instance. To avoid losing all your work, you can just transfer your files off the system onto your local machine - but be aware that you will be charged for the bandwidth you use. Alternatively, you could consider taking a snapshot. See the official documentation on Elastic Block Storage for more on this.
- **You are charged by the time-hour.** This means that if you start up an instance at 1:55pm and use it until 2:05pm, you are charged for two hours - because your instance was running in two different hours of the clock.

2.2 Starting up a Cloud-Bio-Linux instance

This document focusses on using the AWS Management Console, a web-interface, for starting up Cloud-Bio-Linux.

1. Go to the EC2 Management Console URL: http://console.aws.amazon.com/ec2/home
2. You should see a button saying **Launch instance**. Click on this.

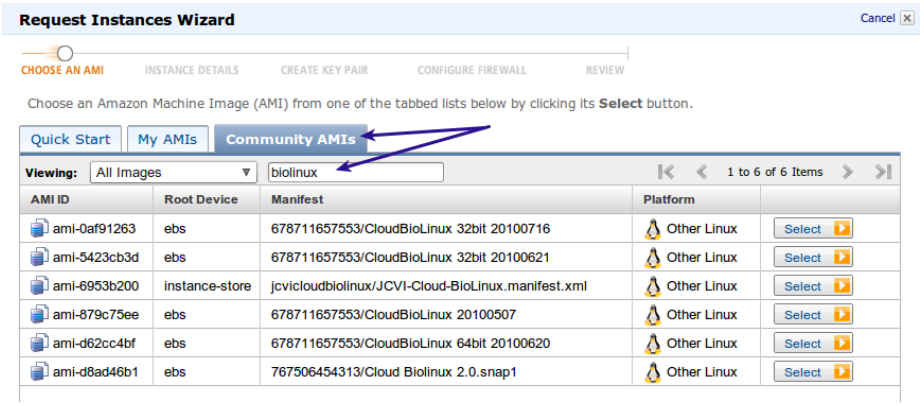

Figure 5: Search for the term "biolinux" in the Community instances. You are likely to find a number of different images available. Those listed here were available on July 20, 2010. We recommend you pick the one with the system you want (e.g. 32 bit or 64 bit) that has the most recent date included in the description.

3. You are presented with a window called **Request Instances Wizard**.

4. To start up Cloud-Bio-Linux, go to the **Community AMIs tab** and search All Images for the term **biolinux**. This will bring up a list of available Cloud-Bio-Linux images. See figure .

5. Once you have chosen your instance, leave the selection **Launch Instances** chosen in the next window presented to you.

6. Click on the Continue Button at the bottom of the window. *(Can't see a Continue button? Check out the FAQ.)*

7. Leave the Advanced options on the next page alone this time. *Note that you will not be able to change any of these things in a running instance.*

8. In the next window, you'll need to provide the name of your Key Pair. *If you created a key pair earlier, but are not offered the option of using it, and if you created your keys in the same session you are currently logged into, try logging out of Amazon and logging back in again.*

9. Once you have provided a key pair name, the next window will ask about your preferred security settings. This is analagous to setting up your firewall. At a minimum you will need to enable ssh access - ssh is how you are going to connect, whether you do so via the command line or via a graphical NX connection. See figure for more information about how to do this. If you want to access web pages provided by your instance, then you also need to open a port for http. You will want to do this if, say, you wish to refer to the Bio-Linux documentation pages on your instance.

If you will be running MySQL or postgreSQL for example, you'll need to enable access to these also.

10. Once you've done all this, you should be able to review the information you've provided, and if you're happy click on the **Launch instance** button.

If you go back to your Amazon EC2 home area and click on the Instances link in the left hand pane, you should see your Cloud-Bio-Linux instance starting up. When you see a green icon with the word running beside it, your instance is ready to log into.

2.3 Logging into your Cloud-Bio-Linux instance

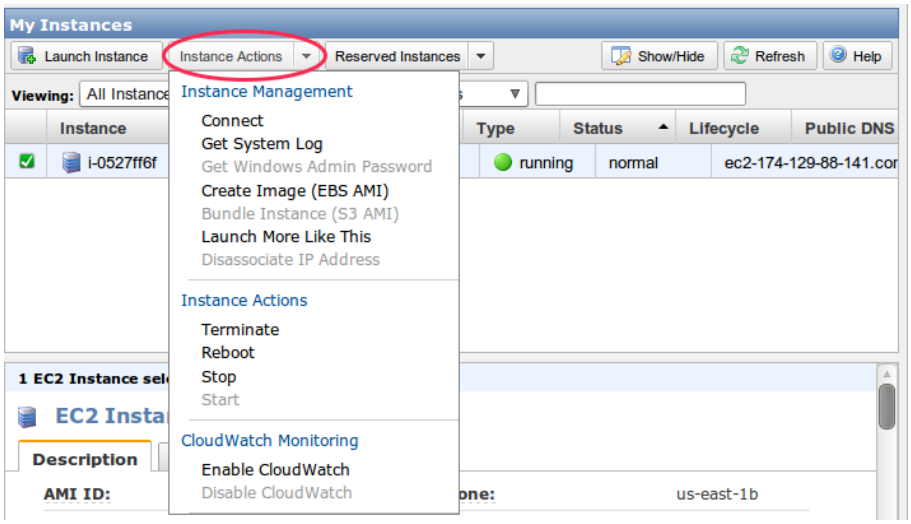

Figure 6: Click the Instance Actions button (pink circle) to bring up a menu with options including connecting to an instance you have already started up, stopping a running instance and terminating an instance.

- Assuming you have already clicked on the Instances link on the left side of your EC2 Dashboard, click on the **Instance Actions** button near the top of the Instances page. See figure .
- Choose **Connect**. A window will open containing directions about how to connect to your Cloud-Bio-Linux instance using ssh. You need to make a couple of changes to the suggested connection instructions - please see the sections below for more on this.

We recommend you log into a full graphical desktop using NX at this point . Instructions on how to log in using command line ssh are given after the NX instructions below.

Logging into graphical desktop using NX

This information will be filled in as soon as it is available.

Logging into a command terminal using ssh

The instructions in the small window you saw after you chose to Connect to your instance suggest something like the following as an ssh command to use:

ssh -i mykey.pem root@ec2-184-72-144-209.compute-1.amazonaws.com

Cloud-Bio-Linux is based on Ubuntu. To log into the instance, you need to use the ubuntu user. The default for most systems on Amazon EC2 is to log in as the root user.

Note: If you get a warning when you try to connect that suggests that your key cannot be found, it may mean that you have saved your key to a non-standard location andor given it a non-standard name. In this case add path information for your key to the command line so that the private key can be found from where you run the ssh connection command. For example, if your key is stored in a subdirectory of your home directory called *keys*, and you want to log in as the *ubuntu* user, you could log in using ssh and the command, you need to

ssh -i /home/mydirectory/keys/mykey.pem ubuntu@ec2-184-72-144-209.compute-1.amazonaws.com

If you are logging in using an ssh command line tool, you just need to type a command like that above into a terminal to log into your Cloud Bio-Linux instance. If you are logging in using Putty on Windows, you will need to enter the relevant information into the Putty system in order to connect.

If you're working from a Linux system, you should be able to run command line or graphical applications via the command line now. For most users, we still recommend connecting to your Cloud-Bio-Linux instance using the NX client, rather then directly using a command line ssh client, as NX connections provide access to a full Bio-Linux desktop, with menus, links, etc.

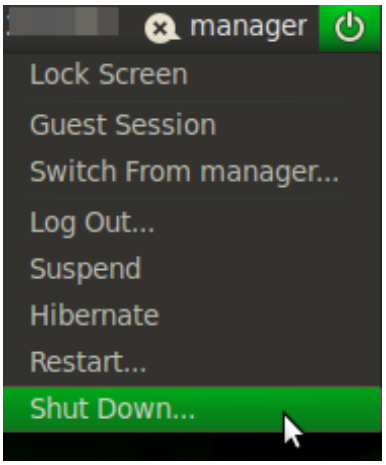

Figure 7: Choosing the **Shut Down...** option in an NX session logs you out. Logging out is not the same as stopping or terminate your Cloud-Bio-Linux instance. You will still be charged while the instance is running - whether you are logged into it or not.

On Windows, there are additional steps to take if you wish to run graphical programs. Our recommendation is to follow the instructions in the **Logging into graphical desktop using NX** section above.

2.4 Logging out of your Cloud-Bio-Linux instance

From an NX connection you need to go to the menu under the power button in the top right of your Cloud-Bio-Linux desktop and choose the option **Shut Down...**. See figure .

From an ssh command line (or Putty) connection you need to type **exit** at the command prompt.

2.5 Stop your Cloud-Bio-Linux instance

Highlight the instance you wish to stop in the list on your Instances page. Click on the **Instance Actions** button (see figure) and choose **Stop** under the Instance Action section of the menu.

You will still be charged a fee if you have only stopped your instance, as opposed to terminating it, and your data may still be deleted depending on how you have set things up. Stopping and terminating are different.

2.6 Terminating your Cloud-Bio-Linux instance

Highlight the instance you wish to terminate in the list on your Instances page. Click on the **Instance Actions** button (see figure) and choose **Terminate** under the Instance Action section of the menu. In basic terms, terminating results in the system and all the files and data on it being deleted. If you have work you wish to save before terminating, or if you wish to keep a copy of this image such that you can use it later, without paying as much as you would for a

running instance, please check out the Amazon documentation on EBS Volumes and taking snapshots of instances.

3 Working with data on Cloud-Bio-Linux

For many bioinformatics tasks, you will want to work on your own data and files for example, perhaps your own sequence data and blast databases. To do this, you will need to *upload your files onto a machine that your Cloud-Bio-Linux instance can access.* Three options are covered in this chapter:

1. Copy your data directly onto the Cloud-Bio-Linux instance you are running. This would be alright if you were going to use this data only on this running instance and you're happy for it to be deleted when you terminate the instance.

2. Copy your date on a separate EBS volume. This would be useful if you wish to store your files for use in other sessions, but you do not plan to keep the same running instance. (EBS Volumes are cheaper than running instances.)

3. If the data you want to use is already available on Amazon EBS volumes (for example, ENSEMBL data), you can access this easily, with no data transfer costs.

3.1 Copying data onto your Cloud-Bio-Linux instance

If you only need your data for a single Cloud-Bio-Linux instance, then you can just copy your data onto that instance directly.

Once you are logged into your Cloud-Bio-Linux instance, there are a number of ways to do this. For example, there are command line tools like **scp**, for copying files from a machine you have an account on, or **wget** to bring in data from public websites or ftp sites.

Alternatively, if you are logged into the full graphical desktop using NX (see section on page 9), you can use the file browser to connect to a remote site and **drag and drop** your files to your running Bio-Linux instance. This is the method we focus on here.

- Go to the **Places** menu in the top taskbar and open up a file browser, for example by clicking on your Home Folder.

- Now go to the **Go** menu and click on Location... (or just type Ctrl-L).

- If you are going to copy files from a machine that you have login permissions on, then in the box next to the word *Location* that appears in your file browser, type: **ssh://your.machine.com**, replacing your.machine.com with the address of the machine your files are on. Alternatively, if you wanted to copy files from a public ftp server, say, then you would enter something like the following in the Location box:
  ftp://ftp.someother.database.site

As a specific example, if I want to copy fasta files from the EMBL database sections, I would type the following into the Location box:
ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/emblrelease

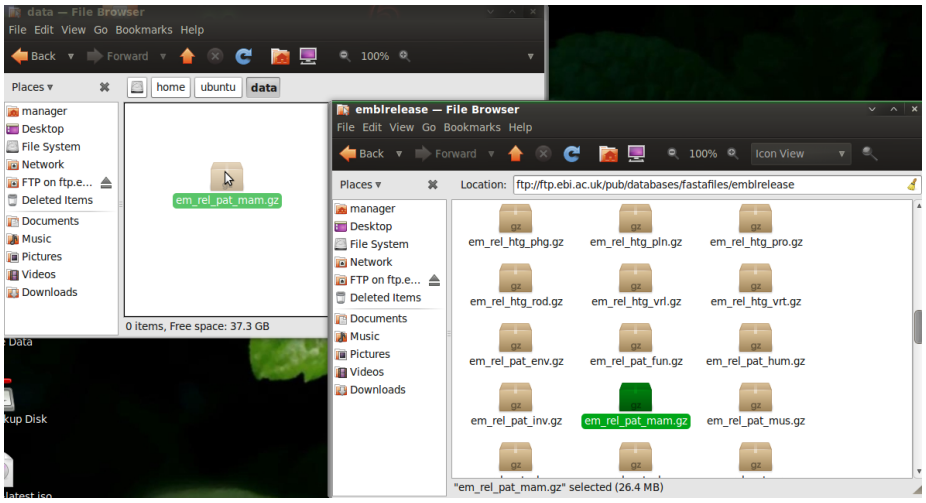

Figure 8: Copying files from remote machine is easy using the graphical File Browser, which can be launched from under the Places menu in the top taskbar. Choosing the Locations option under the Go menu of the file browser will allow you to type in a protocol (e.g. ftp, ssh) and a location. Here, a section of EMBL from the EBI is copied to my system using drag and drop between the two file browser windows.

Now open another file browser by going to the Places menu. Navigate to the folder you wish to store the files in. Now you can just drag and drop your files from the remote machine onto your Cloud-Bio-Linux instance. See figure

This process is simple, and for one-off jobs, is perfectly adequate. Note that you will generally pay for the network traffic you generate in transferring the data[2] So if you are going to use the same dataset numerous times, it is worth considering setting up an EBS volume rather than transferring data onto new instances. Even if this transfer is free, it will still generally take more time than mounting an EBS volume that already has your data on it.

3.2 Using EBS volumes for data

An Amazon EBS volume is what you need if

[2]Until November 1, 2010, data transfer onto Amazon is free. The first Gb per month of transfer off is also free. (Information taken on July 21, 2010, with no guarantees to be correct at the time you are reading this document. Check out the official pricing list.

- you are going to use a dataset a number of times, with gaps in time between uses, or
- you want to store your data such that you can connect to it from different Cloud-Bio-Linux (or other Amazon EC2 images), or
- if you wish to share your data with other people when they are working on an Amazon EC2 system.

This guide presents only a small part of what is possible with EBS volumes. Please check out the EBS volume documentation on the Amazon website for further information.

A note on charging: You will be charged for your Amazon EBS volume as long as it is in existence, and you will be charged for the space you request, not the space you are really using. So if you ask for 1Gb, you are paying for 1Gb, even if you only use 100Kb.

Creating your volume

We suggest that you have started up and have logged into an instance before creating your EBS volume.

To create an EBS Volume:

- In the Navigation pane (left side) of the AWS Management Console, go to the Elastic Block Store area and choose **Volumes**. See figure .
- Click on the **Create Volume** button.
- Choose the same availability zones as the images you plan to use this volume with.
- After changing any other settings in this window, press the **Create** button.
- Wait until the yellow circle beside the word *creating* is replaced by a blue circle beside the word *available*.

You now have an EBS volume - but this is not usable to copy your data onto yet. You first must mount it on your instance - this makes the volume accessible to you when you are logged into your instance. Here we cover how to do this from the Console. It is also easy to do using standard Linux command line tools to mount a volume.

- Click on the **Attach Volume** button on the Volumes page. See figure .
- Fill in the requested information. See figure .

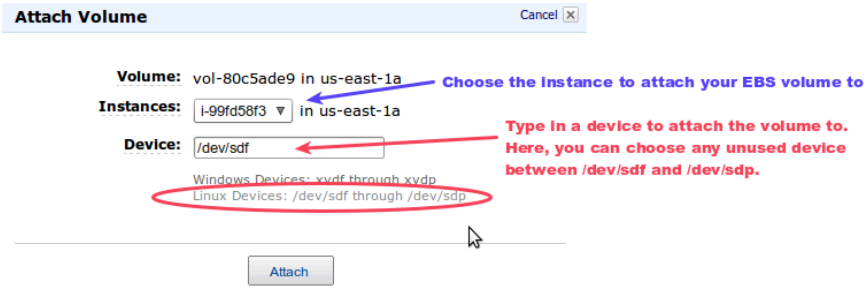

Figure 9: Fill in the information requested. Note that both the instance and the EBS volume are in the same region. You just need to provide an unused device to attach your volume to. If you don't know what this means, and are starting up a Cloud-Bio-Linux instance, then use any of the suggested locations: /dev/sdf, /dev/sdg, /dev/sdh....up to /dev/sdp.

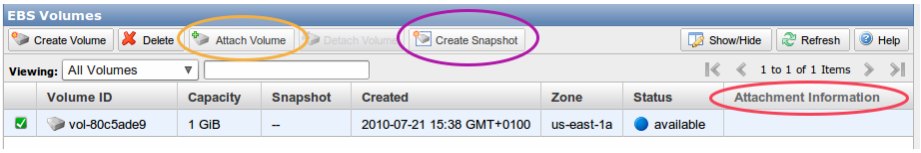

Figure 10: Click on the Attach Volume button (Orange circle) to Attach a volume to an instance.

Selecting any of your volumes in the AWS Management Console will bring up details of that volume at the bottom of the page.

Getting access to your volume

This is where it gets a bit ugly, as you need to log into your machine and use the command line for the next couple of steps. The first of these, **formatting your disk**, you only need to do the first time you use a particular volume. The second step, **mounting the volume**, needs to be done each time you want to access data on your volume from a new instance. There are ways to automate this, but these are not covered here.

1. Log into your instance. If you are logged in using NX, start up a terminal window.

2. **The first time you mount a volume for use only:** Type the following command to create an ext3 filesystem on your volume. Here I assume you have mounted it to devsdf.
   sudo mkfs -t ext3 /dev/sdf

3. Now make a directory that you will mount the data volume onto. For example, this command creates a directory called mntdatasets:
   sudo mkdir /mnt/datasets

4. Now mount your volume:
   sudo mount /dev/sdf /mnt/datasets

You will now be able to put data under the folder mntdatasets. All files under that area are on your EBS volume and will not be lost when your instance terminates. Ensure you read the section below on **unmounting your volume** as failure to do so before detaching your volume or terminating your instance could lead to data corruption.

Putting data on your volume

Unmounting your volume

This is a simple but vital step to avoid the possibility of data corruption. Do this before you detach your volume or terminate your instance.

If you had attached your device to /dev/sdf, then you simply need to type:
umount -d /dev/sdf

Detaching your volume

You can detach your volume from your instance using the AWS Management Console using the Detach Volume button on the console Navigation pane. Alternatively, your volumes will be detached automatically when you terminate your instance.

Backing up or sharing your volume

Check out the Userguide information on creating snapshots and on modifying permisisons on snapshots.

Deleting your volume

You can delete your volume using the Delete button on the console Navigation pane.

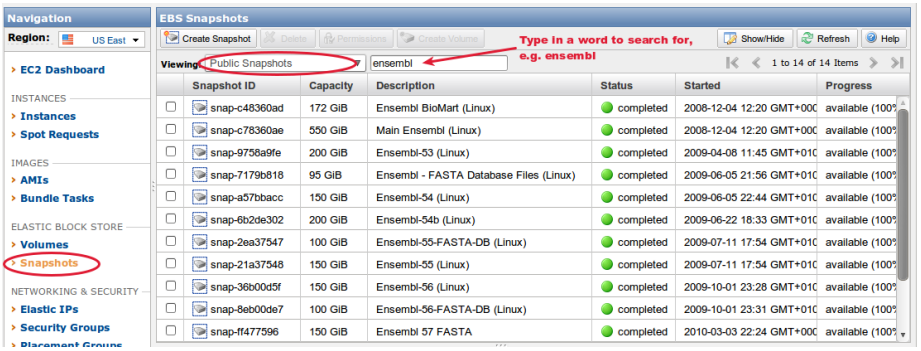

Figure 11: Click the Snapshot link in the Navigation pane. Then select Public Snapshots in the Viewing menu. Searching for a term such as *ensembl* brings up all the public snapshots that contain ensembl in their description.

3.3 Accesing public datasets on Amazon

Amazon makes some public data sets available as snapshots. You can just attach and mount these - no data transfer is necessary. Check out the full public data set listing. Finding datasets is easy: just search through the public snapshots for relevant terms. See figure .

Amazon provides documentation on how to make use of these public data resources.