

PENNSTATE[®]



Enabling NGS analysis with(out) the infrastructure

Enis Afgan, the Galaxy Team, Anton Nekrutenko,
James Taylor

Bioinformatics Open Source Conference, July 16, 2011, Vienna



EMORY

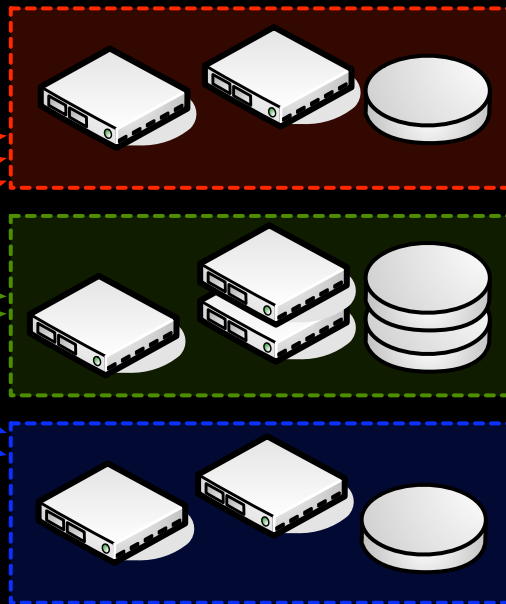
CloudMan-as-a-Bridge

A. Users in different labs

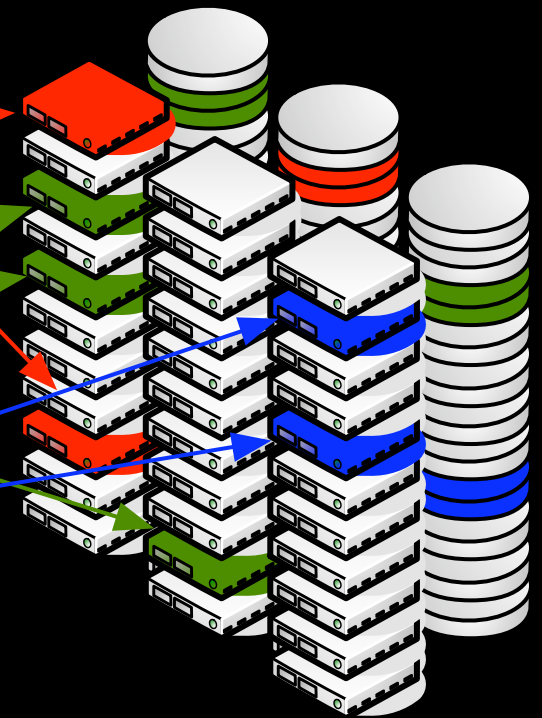


Internet

B. Isolated Galaxy instance(s)



C. Dense data center

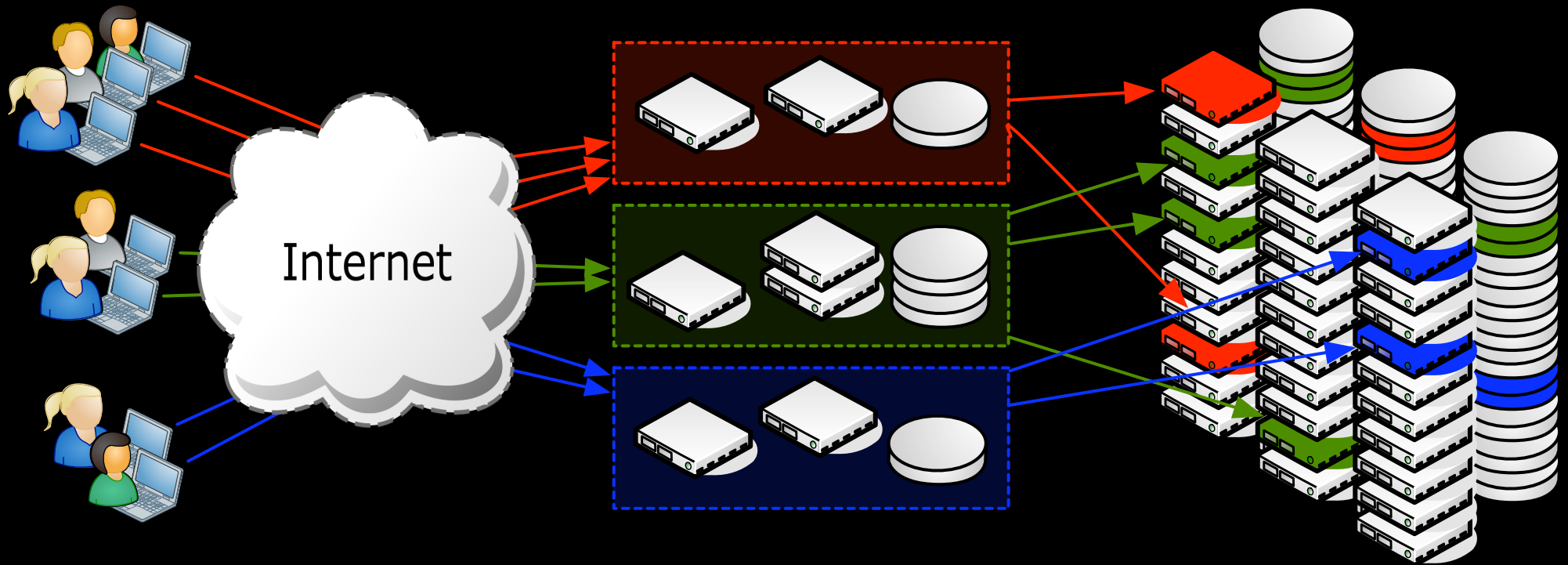


CloudMan-as-a-Bridge

A. Users in different labs

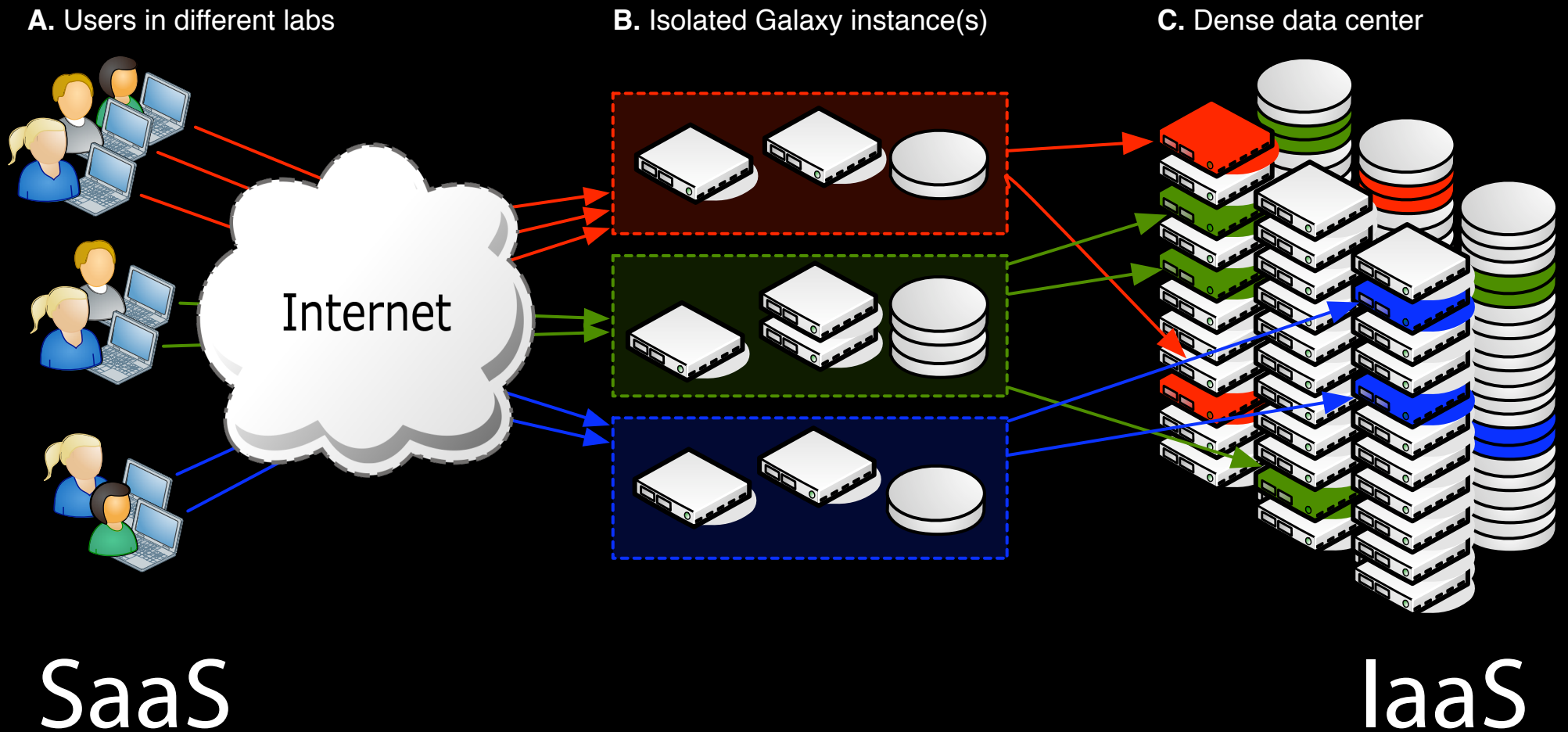
B. Isolated Galaxy instance(s)

C. Dense data center

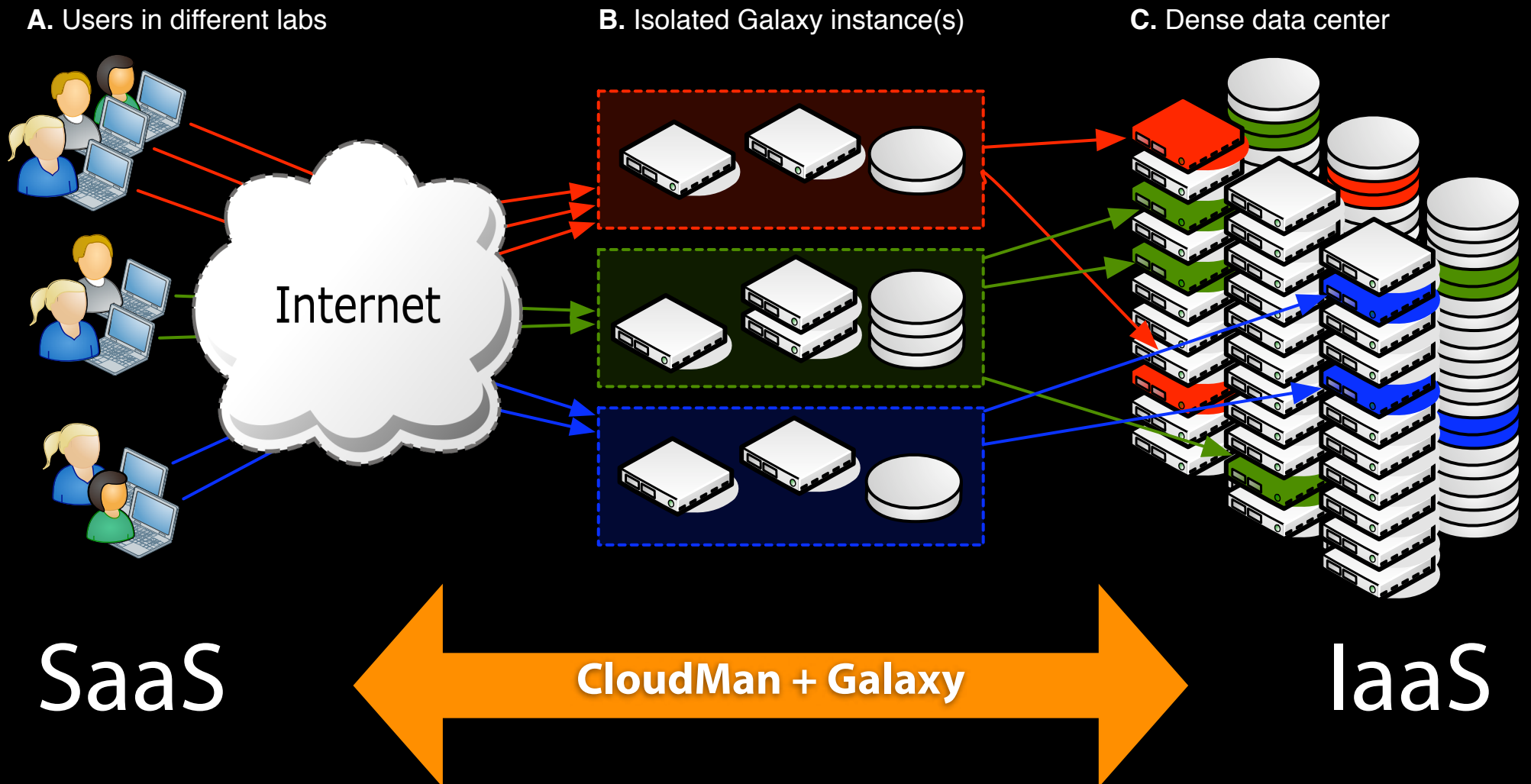


SaaS

CloudMan-as-a-Bridge



CloudMan-as-a-Bridge



CloudMan Platform

A complete solution for instantiating and managing cloud resources

With automatically configured Galaxy (if desired)

Scope of tools and reference datasets **exceed Galaxy Main**

Deploy a (Galaxy) cluster in minutes!

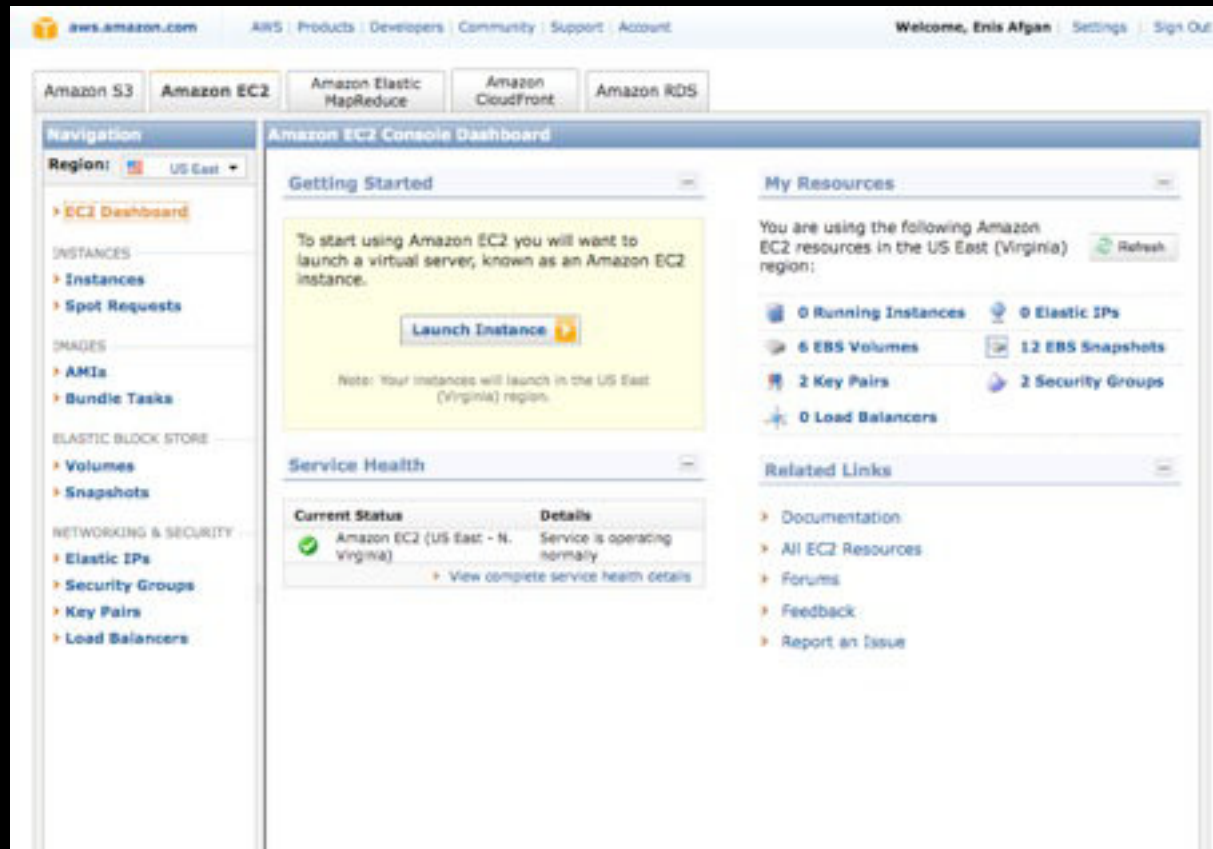
CloudMan features

- Deployment on Amazon Web Services Cloud
 - **Wizard-guided setup**: requires no computational expertise, no infrastructure, no software
- **Automated** (thus reproducible) configuration for machine image, tools, and data
- **Four modes** of cluster type setup
- **Dynamic persistent storage**
- **Elastic resource scaling**: manual or automatic based on workload
- **Standalone** deployment, requiring no external dependencies or services
- **Customizable** by individual users
- **Sharing** of derived cluster instances -> even the customized ones!

CloudMan features

- Deployment on Amazon Web Services Cloud
- **Wizard-guided setup**: requires no computational expertise, no infrastructure, no software
- **Automated** (thus reproducible) configuration for machine image, tools, and data
- **Four modes** of cluster type setup
- **Dynamic persistent storage**
- **Elastic resource scaling**: manual or automatic based on workload
- **Standalone** deployment, requiring no external dependencies or services
- **Customizable** by individual users
- **Sharing** of derived cluster instances -> even the customized ones!

Deploying a cluster on AWS



1.

Deploying a cluster on AWS

1.

The screenshot shows the AWS Management Console with the Amazon EC2 Console Dashboard selected. The Galaxy Cloudman application is open, displaying the 'Initial Cluster Configuration' dialog. The dialog has a title bar with 'Galaxy Cloudman' and links for 'Admin', 'Report bugs', 'Wiki', and 'Screencast'. The main content area is titled 'Initial Cluster Configuration' and includes a welcome message: 'Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.' There are four radio button options: 1. 'Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)' with a text input field set to '1' and a 'GB' label. 2. 'Share-an-instance' with a text input field for 'Shared instance bucket path'. 3. 'Data volume and SGE only. Specify initial storage size (in Gigabytes)' with a text input field set to '1' and a 'GB' label. 4. 'SGE Only. No persistent storage created.' Below these options is a link 'Hide extra options' and a 'Start Cluster' button at the bottom.

aws.amazon.com AWS Products Developers Community Support Account Welcome, Enis Afgan Settings Sign Out

Amazon S3 Amazon EC2 Amazon Elastic MapReduce Amazon CloudFront Amazon RDS

Navigation Amazon EC2 Console Dashboard

Region: us-east-1

Galaxy Cloudman Admin | Report bugs | Wiki | Screencast

Galaxy Cloudman

Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)

1 GB

Share-an-instance

Shared instance bucket path

Data volume and SGE only. Specify initial storage size (in Gigabytes)

1 GB

SGE Only. No persistent storage created.

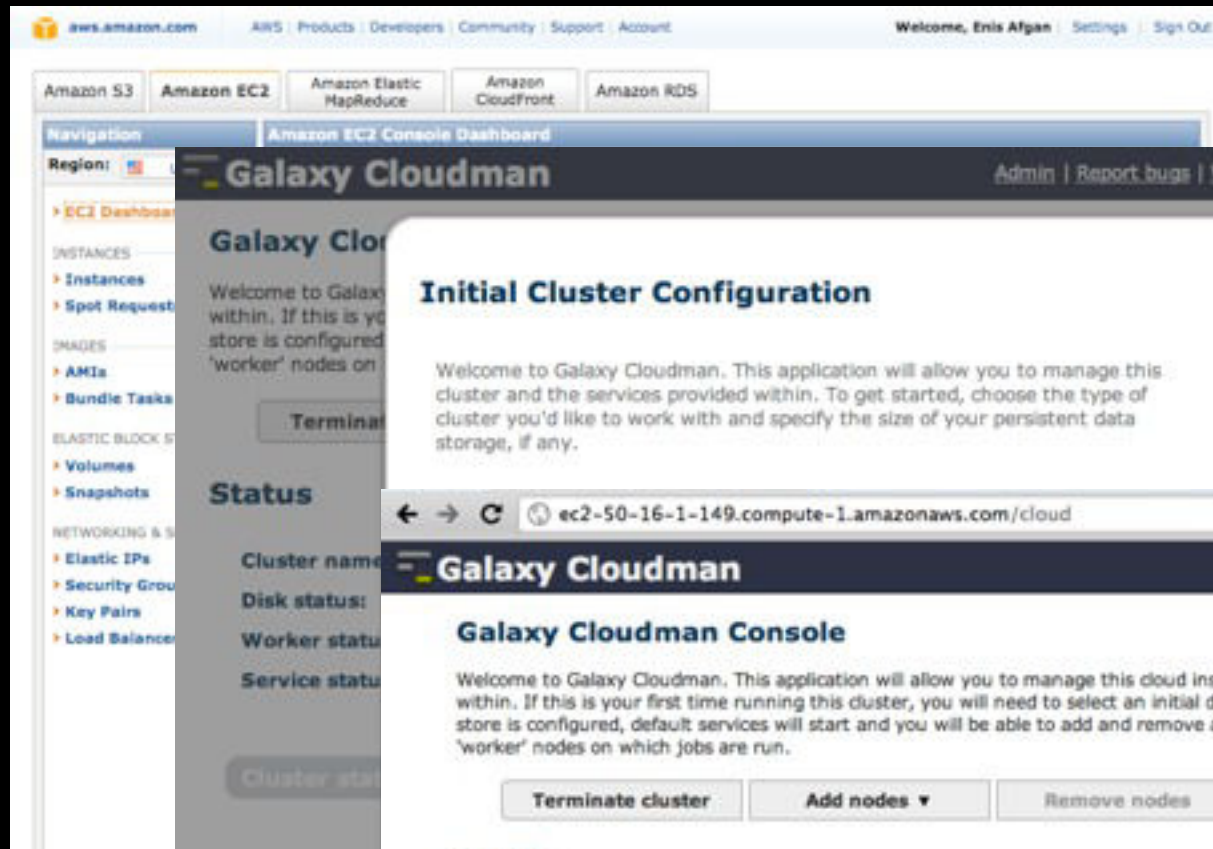
[Hide extra options](#)

Start Cluster

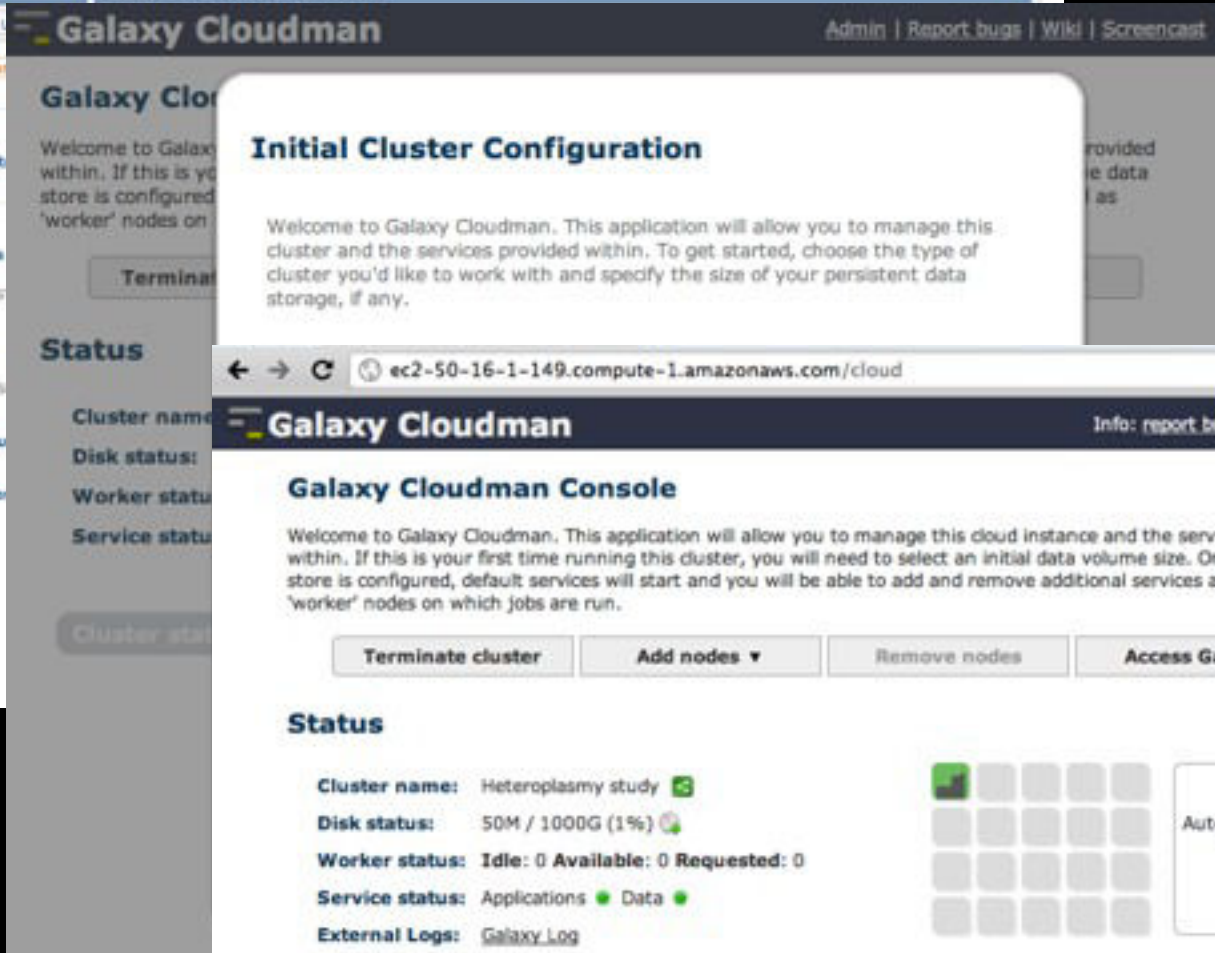
2.

Deploying a cluster on AWS

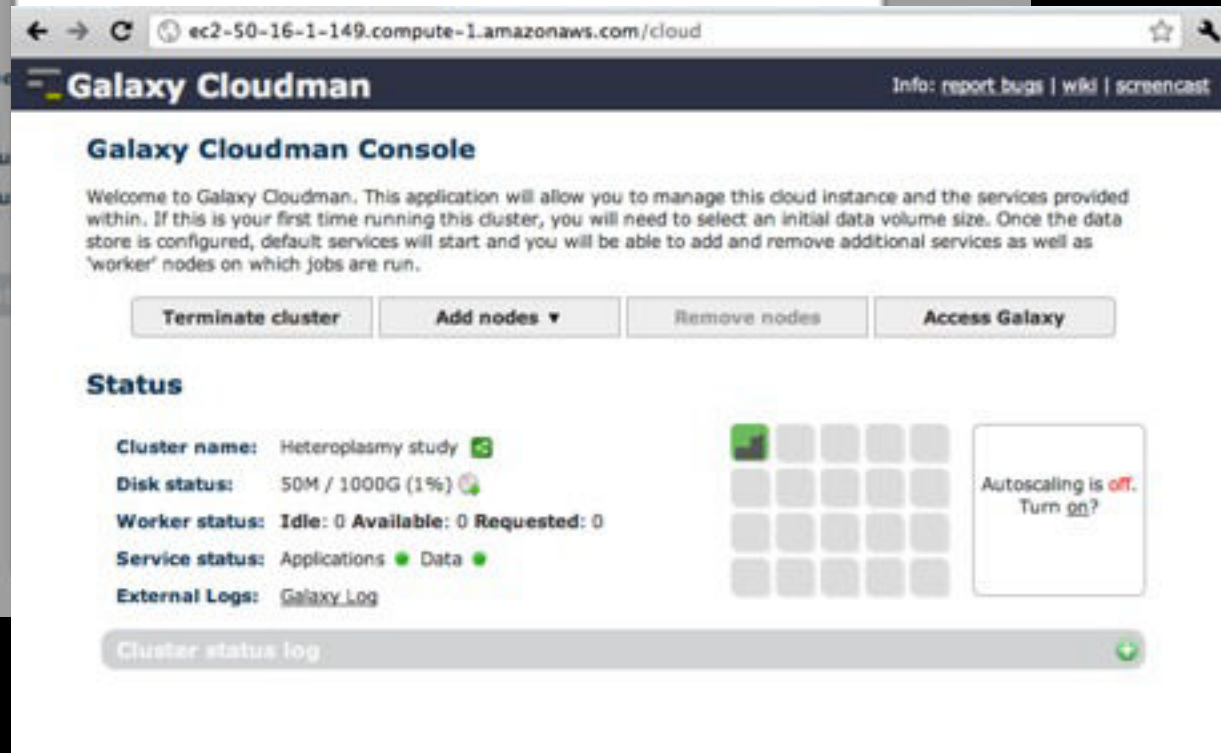
1.



2.

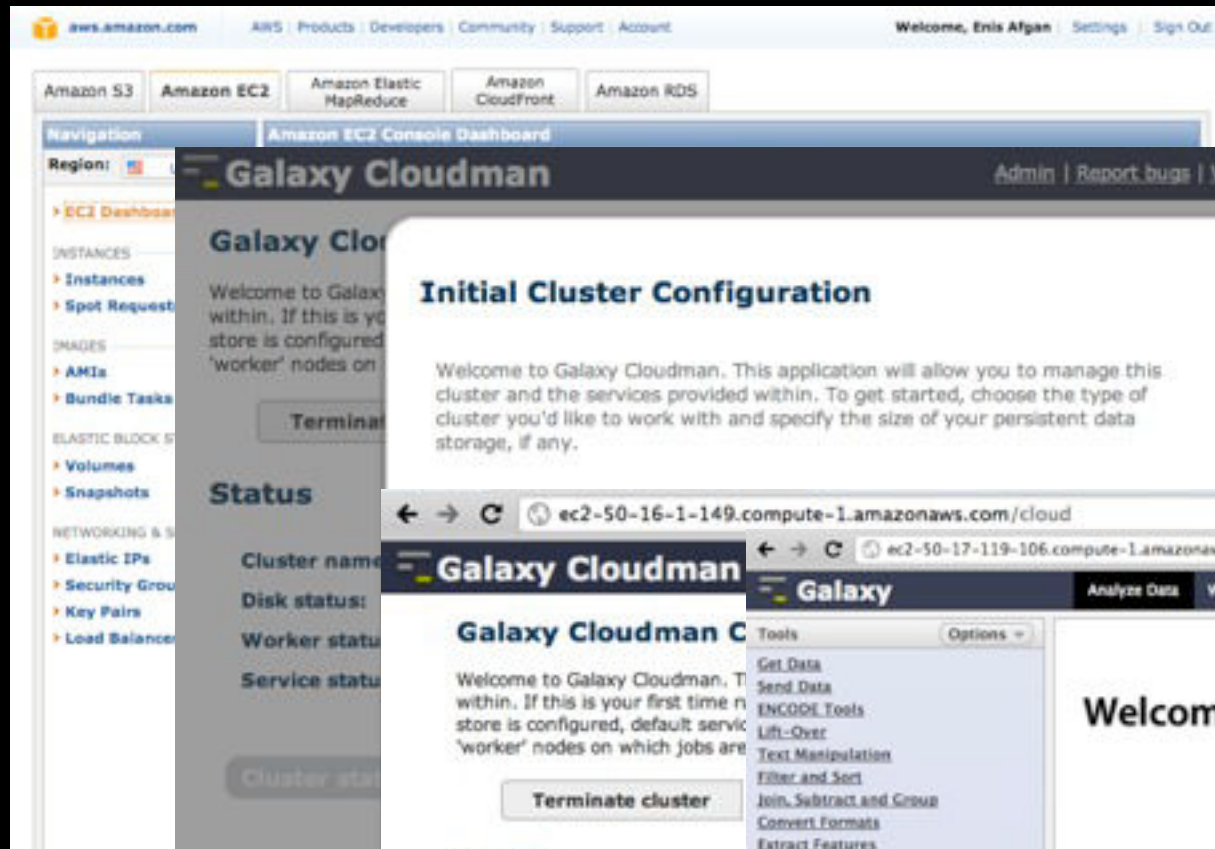


3.

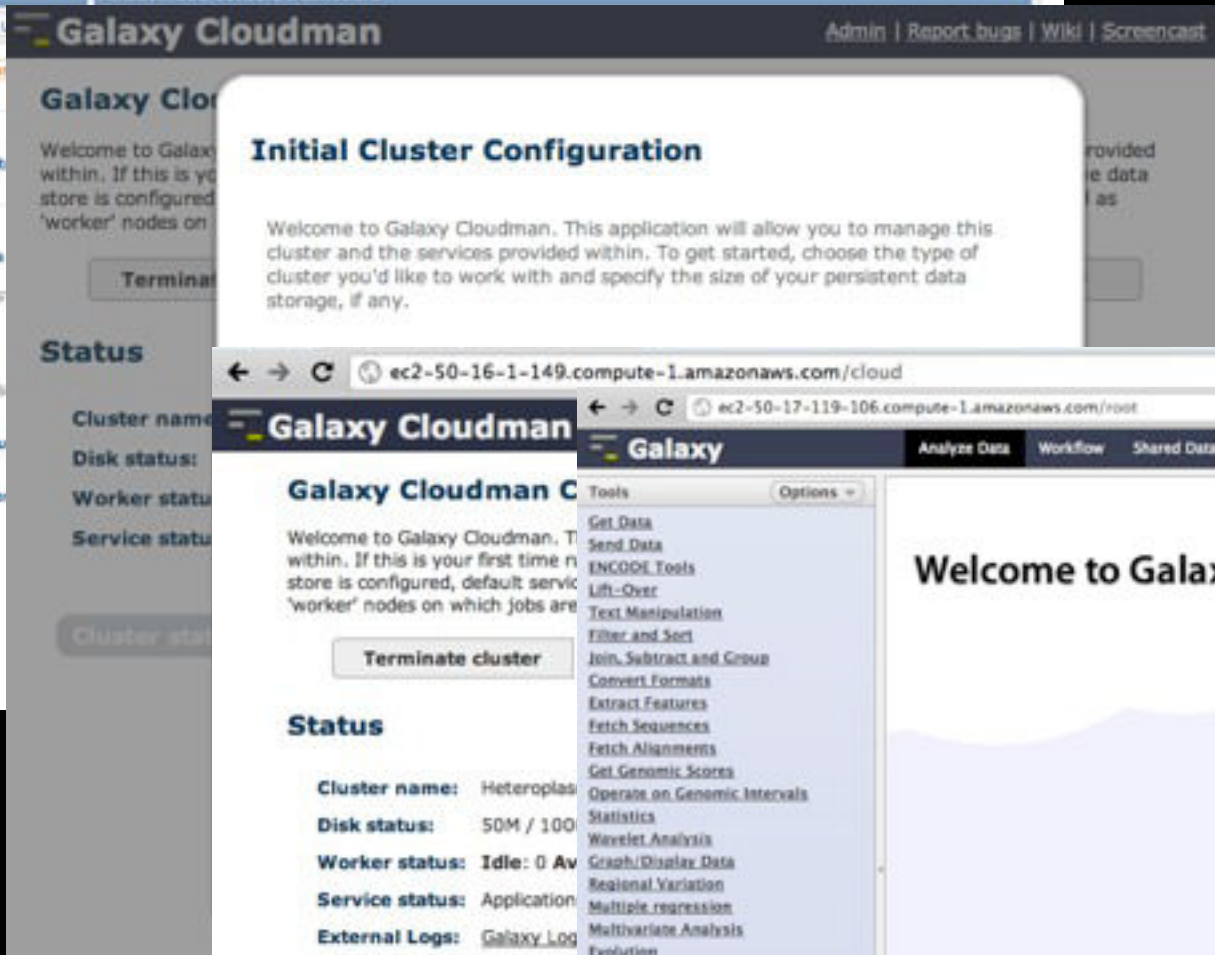


Deploying a cluster on AWS

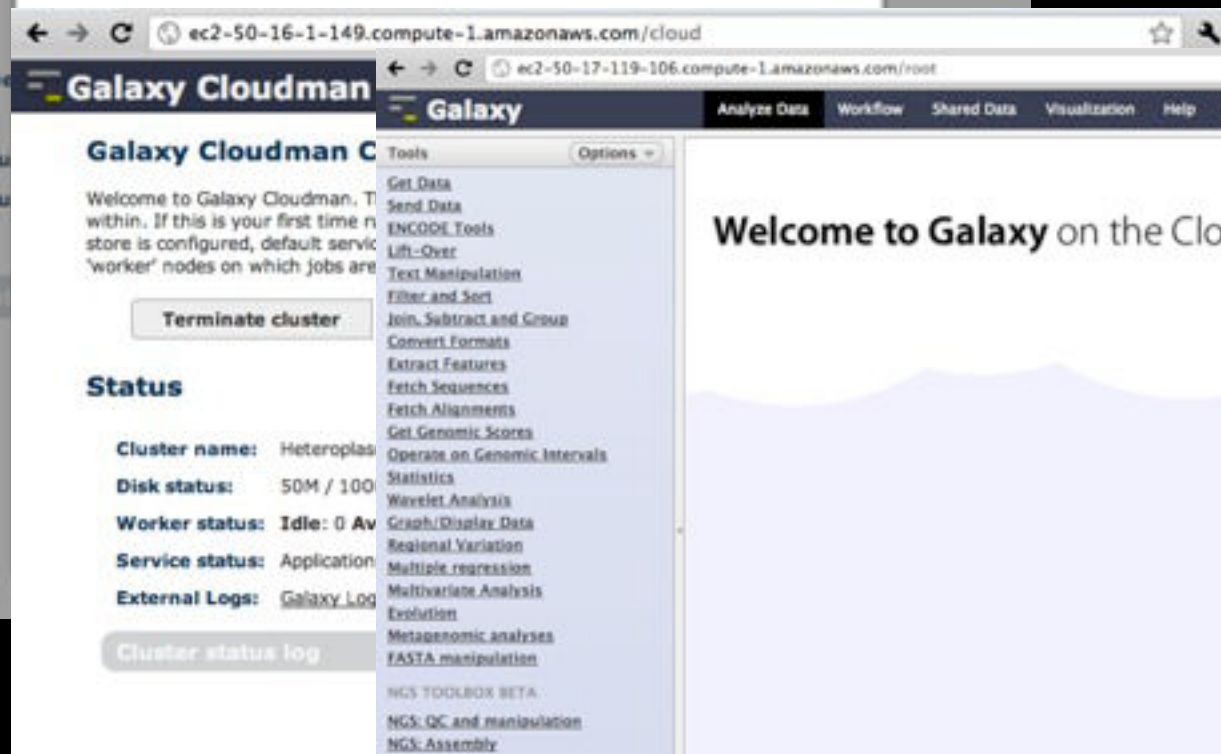
1.



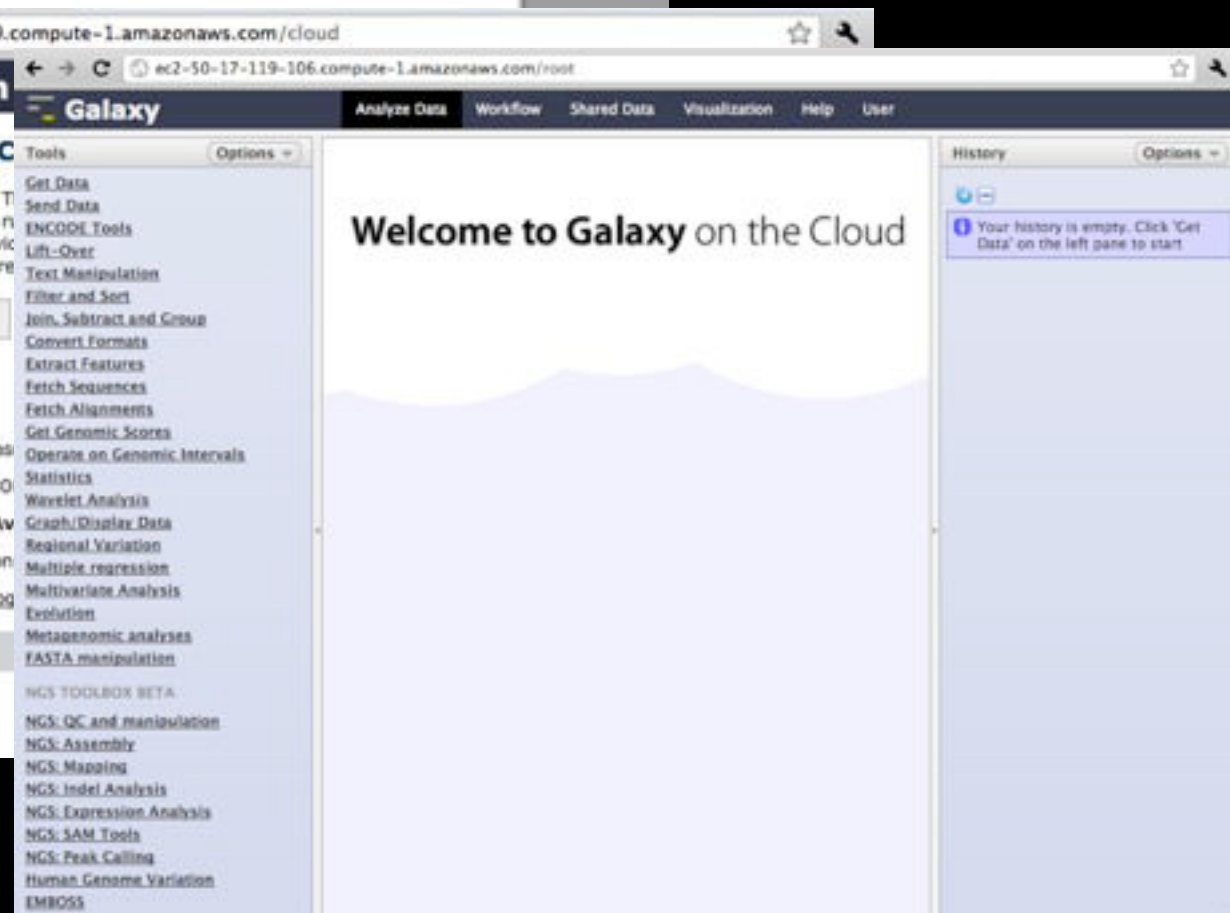
2.



3.



4.



CloudBioLinux + Galaxy + CloudMan =

- A lot of (NGS) tools immediately available and easily accessible
- 700GB of reference genome data

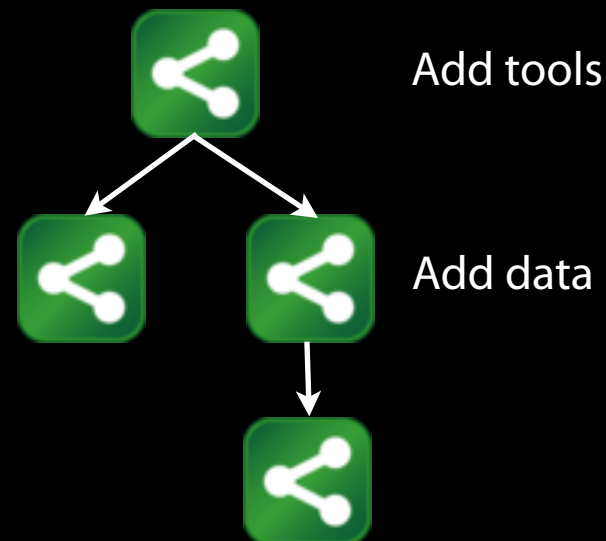
Bowtie, BWA, Samtools, MAQ, BFAST, ABySS, Velvet, MACS, Tophat, Cufflinks, MegaBLAST, BLAST, Sputnik, Taxonomy, HyPhy, Lastz, Perm, GATK, Srma, Beam, Pass, LPS, Plink, Haploview, Freebayes, Mosaik, Picard, ...

But what if **your tool** (or data) is missing?

1. Add it! (via automation)
 - CloudMan instances are self-contained

```
def _install_picard():  
    version = '1.45'  
    mirror_info = "Use mirror voxel"  
    url = 'http://downloads.sourceforge.net/project/picard/picard-tools/%s/picard-tools-%s.zip' %  
        (version, version)  
    pkg_name = 'picard'  
    install_dir = os.path.join(env.install_dir, pkg_name, version)  
    install_cmd = sudo if env.use_sudo else run  
    if not exists(install_dir):  
        install_cmd("mkdir -p %s" % install_dir)  
    with _make_tmp_dir() as work_dir:  
        with cd(work_dir):  
            run("wget %s -O %s" % (url, mirror_info, os.path.split(url)[-1]))  
            run("unzip %s" % (os.path.split(url)[-1]))  
            install_cmd("mv picard-tools-%s/*.* %s" % (version, install_dir))  
        sudo("touch %s/env.sh" % install_dir)  
        sudo("chmod +x %s/env.sh" % install_dir)  
    install_dir_root = os.path.join(env.install_dir, pkg_name)  
    sudo("if [ ! -d %s/default ]; then ln -s %s %s/default; fi" % (install_dir_root, install_dir, install_dir_root))  
    print(green("----- Picard %s installed to %s -----" % (version, install_dir)))
```

2. Save & share
 - With individual users or make it public



Deployment sharing

Galaxy Cloudman

Info: [report bugs](#) | [wiki](#) | [screencast](#)

Currently shared instances

Share-an-instance

This form allows you to share this cluster instance, at its current state, with others. You can make the instance public or share it with specific users by providing their account information below. You may also share the instance with yourself by specifying your own credentials, which will have the effect of saving the instance at its current state.

While setting up an instance to be shared, all currently running cluster services will be stopped. Then, a snapshot of your data volume and a folder in your cluster's bucket will be created (under 'shared/[current date and time]'); this folder will contain your cluster's current configuration. The created snapshot and the folder will be given READ permissions to the users you choose (or make it public). This will enable those users to instantiate their own instances of the given cluster instance. This implies that you will only be paying for the created snapshot while users deriving a cluster from yours will incur costs for running the actual cluster. After the sharing process is complete, services on your cluster will automatically resume.

☐ Public ☒ Shared

Specific user permissions:

Both fields must be provided for each of the users.

These numbers can be obtained from the bottom of the AWS Security Credentials page, under *Account Identifiers* section.

AWS account numbers:

CSV numbers with no dashes

AWS canonical user IDs:

CSV HEX numbers

Share-an-instance

Deployment sharing

The screenshot displays the Galaxy Cloudman web interface. The main header shows the Galaxy Cloudman logo and navigation links: [Info](#), [report bugs](#), [wiki](#), and [screencast](#). The left sidebar contains a navigation menu with items like 'Current', 'Share', 'While', 'Then', 'snapshots', 'enable', 'only be', 'actual', 'Spec', 'Both', 'These', 'Ident', 'AWS', and 'AWS'. The main content area is titled 'Galaxy Cloudman Console' and features a 'Currently shared instances' section. This section explains that users can share bucket names with others to create and instantiate instances of their shared cluster, and provides a list of shared instances with their corresponding snapshot IDs. Below this, there is a 'Share-an-instance' section. The bottom of the console displays worker status (Idle: 0, Available: 0, Requested: 0), service status (Applications: green, Data: green), external logs (Galaxy Log), and a cluster status log with a green checkmark. A 'Min nodes: 0, Max nodes: 15, Adjust limits?' button is also visible.

Galaxy Cloudman Info: [report bugs](#) | [wiki](#) | [screencast](#)

Galaxy Cloudman Console Info: [report bugs](#) | [wiki](#) | [screencast](#)

Currently shared instances

These are the bucket names you can share with others so they can create and instantiate their instances of your shared cluster. Also, for reference, corresponding snapshot ID's are provided and you have an option to delete a given shared instance.

- Shared: `cm-9af6ce6176224e8e18e8b911a05c3ea4/shared/2011-03-31--19-55/ (snap-5d171331)` [✕](#)
- Shared: `cm-9af6ce6176224e8e18e8b911a05c3ea4/shared/2011-04-01--20-51/ (snap-1ec6c572)` [✕](#)

Share-an-instance

Worker status: Idle: 0 Available: 0 Requested: 0

Service status: Applications ● Data ●

External Logs: [Galaxy Log](#)

Min nodes: 0
Max nodes: 15
[Adjust limits?](#)

Cluster status log ●

Deployment sharing

The image shows a screenshot of the Galaxy Cloudman web interface. In the foreground, an "Initial Cluster Configuration" dialog box is open. The dialog has a title bar with the Galaxy Cloudman logo and navigation links. The main content area of the dialog contains a welcome message and three radio button options for cluster configuration. The "Share-an-instance" option is selected. Below this option, a text input field contains a long alphanumeric string, which is highlighted with a blue selection box. To the right of this field, the text "Shared instance" is visible. Below the input field, the text "bucket path" is displayed. At the bottom of the dialog, there is a "Start Cluster" button. The background shows the Galaxy Cloudman console with various sections like "Currently", "Share-an-instance", "Status", and "Cluster status".

Initial Cluster Configuration

Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

☐ Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)

GB

☒ Share-an-instance

Shared instance

bucket path

☐ Data volume and SGE only. Specify initial storage size (in Gigabytes)

GB

☐ SGE Only. No persistent storage created.

[Hide extra options](#)

Start Cluster

Use CloudMan as SaaS

Use CloudMan as PaaS

It's automated, reproducible, extensible, and
transparent.



EMORY

PENNSTATE.



Enis Afgan



Dannon Baker



Dan Blankenberg



Nate Coraor



Dave Clements



Jeremy Goecks



Jennifer Jackson



Greg von Kuster



Kanwei Li



James Taylor



Kelly Vincent



Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health