CloudMan: Galaxy on the Cloud

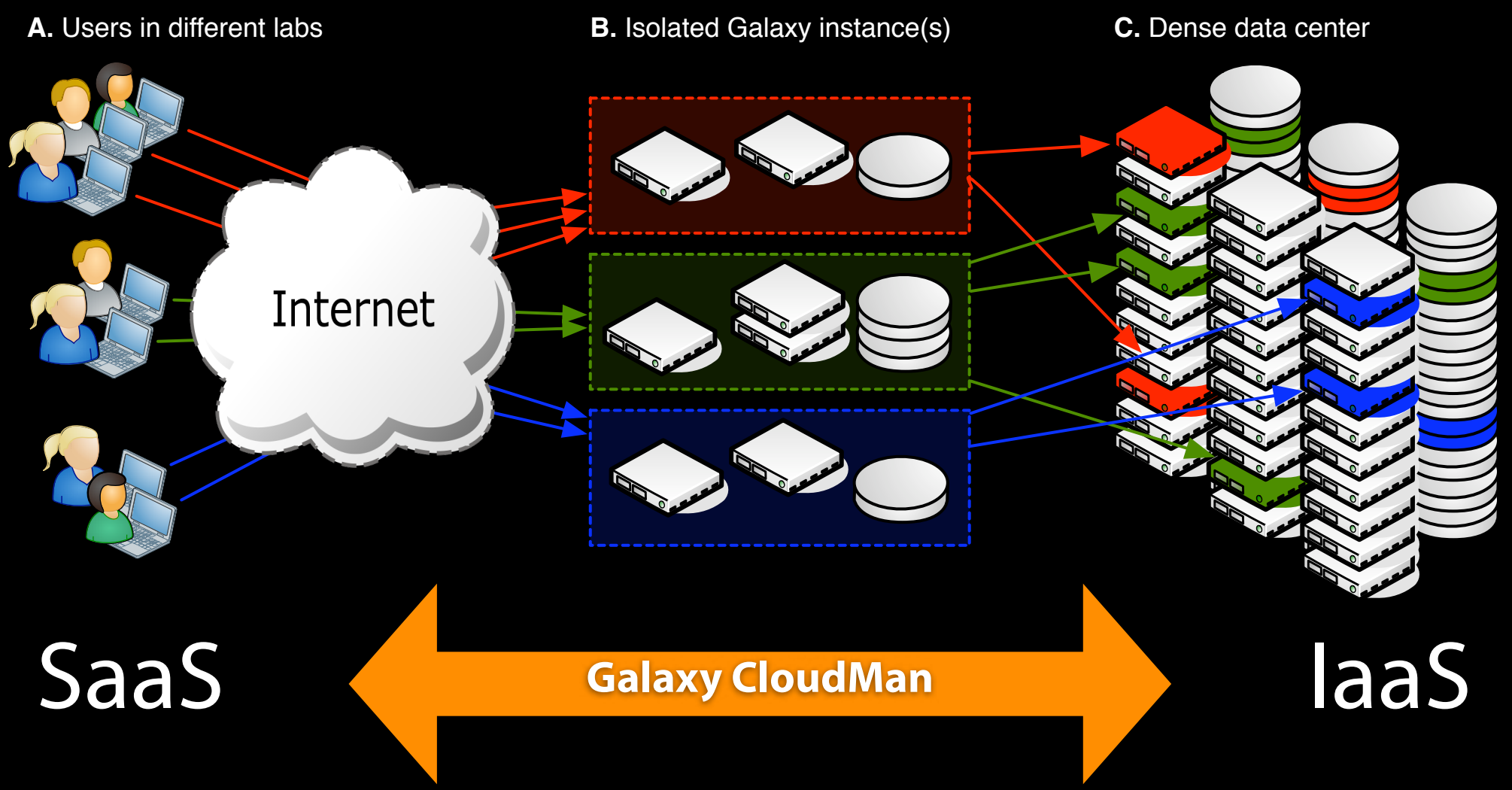
Enis Afgan

Galaxy Community Conference
May 26, 2011 - Lunteren

When to use the cloud?

- Don’t have informatics expertise or infrastructure to run/maintain
- Have extended or particular resource needs
- Cannot upload data to a shared resource
- Need for customization
- Have oscillating data volume

The big picture



Galaxy CloudMan

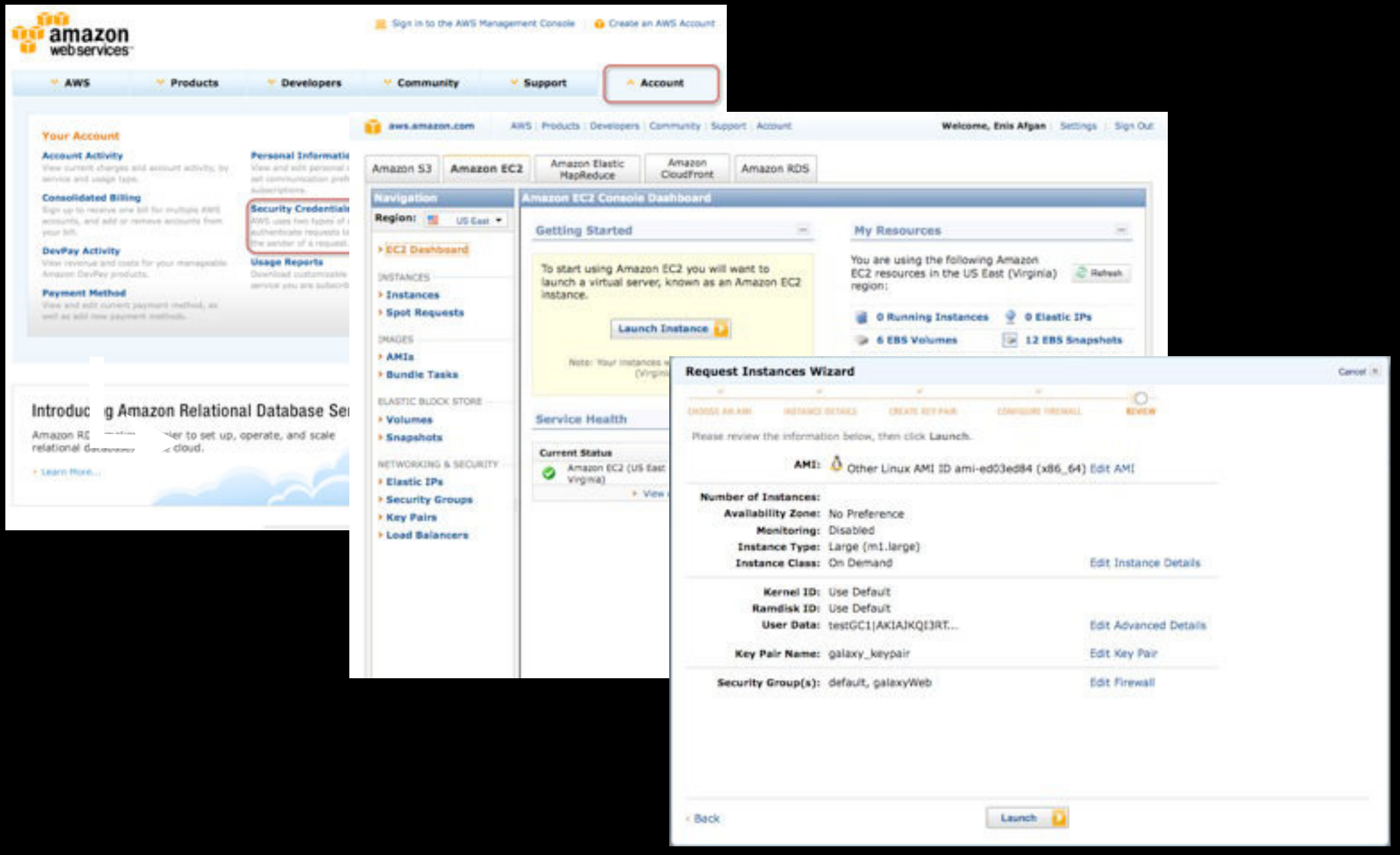
- Complete solution for instantiating, running and scaling cloud resources with automatically configured Galaxy
  - Scope of tools and reference datasets exceed Galaxy Main
- Deployment on Amazon Web Services Cloud
  - Wizard-guided setup: requires no computational expertise, no infrastructure, no software
- Automated configuration for machine image, tools, and data
- Self-contained deployment
- Dynamic persistent storage
- Sharing of derived cluster instances
- Elastic resource scaling: manual or automatic

- **Deploy a Galaxy cluster in minutes!**

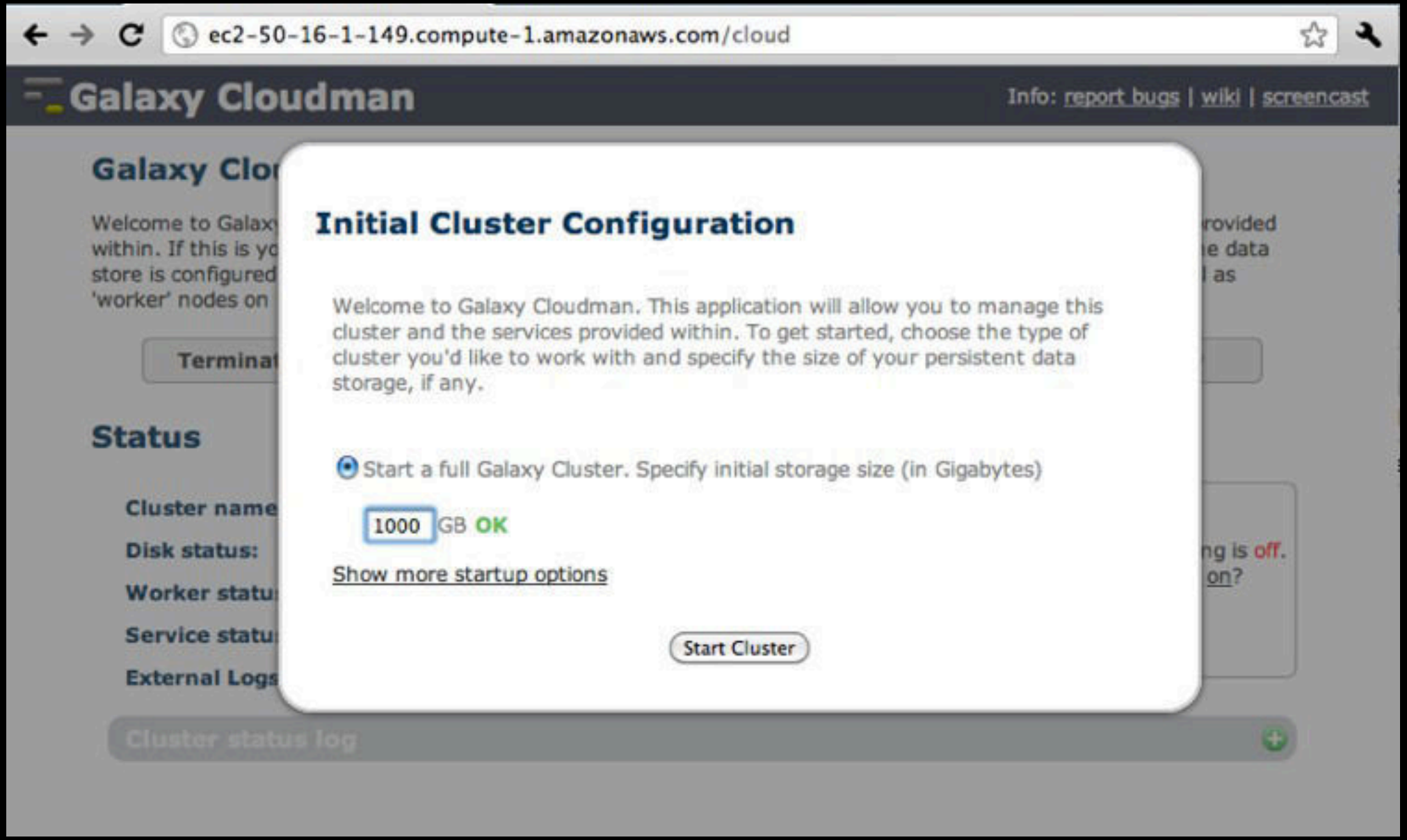
Deploying Galaxy on the AWS Cloud

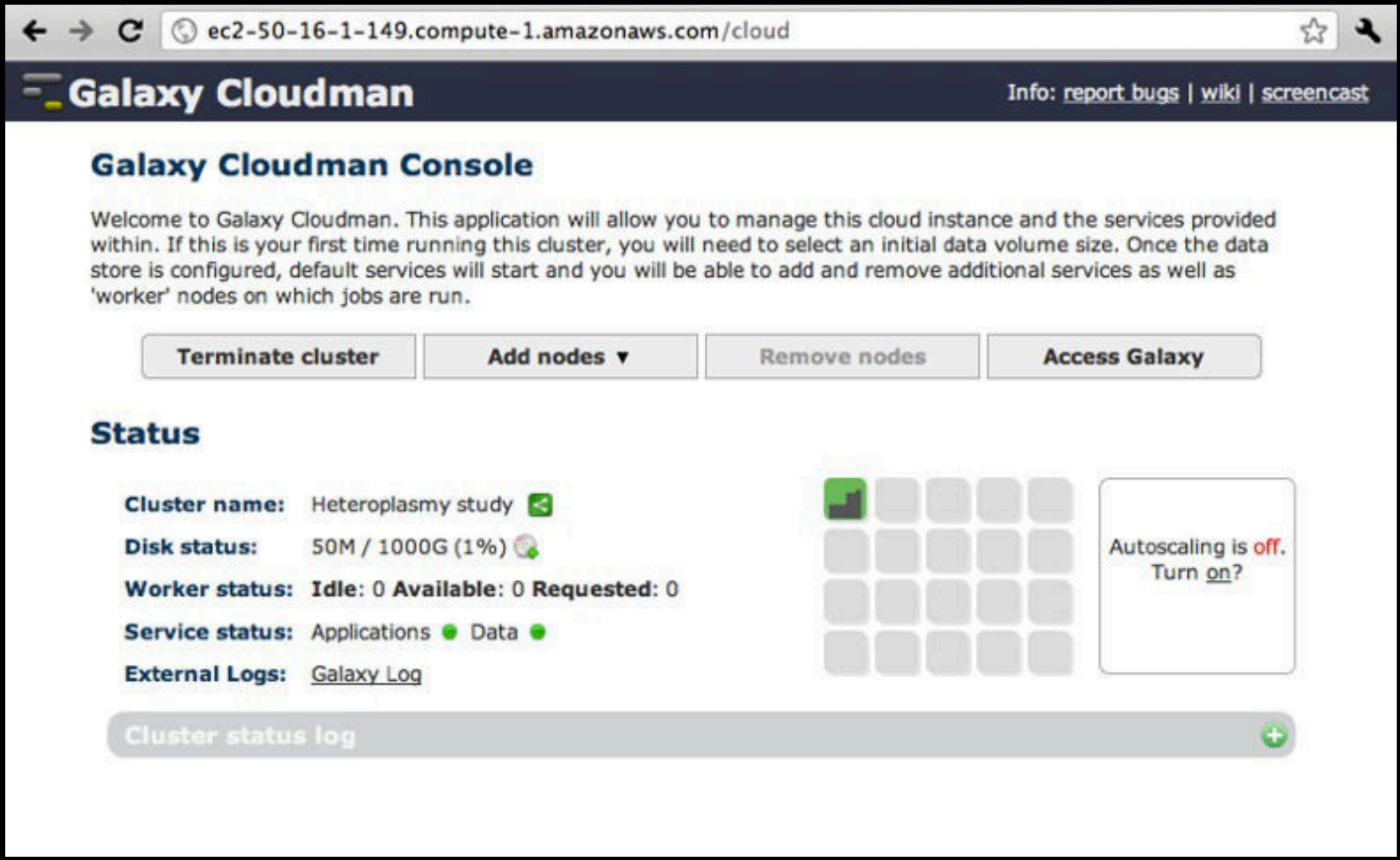
1. Create an AWS account and sign up for EC2 and S3 services
2. Use the AWS Management Console to start a master EC2 instance
3. Use the Galaxy CloudMan web interface on the master instance to manage the cluster

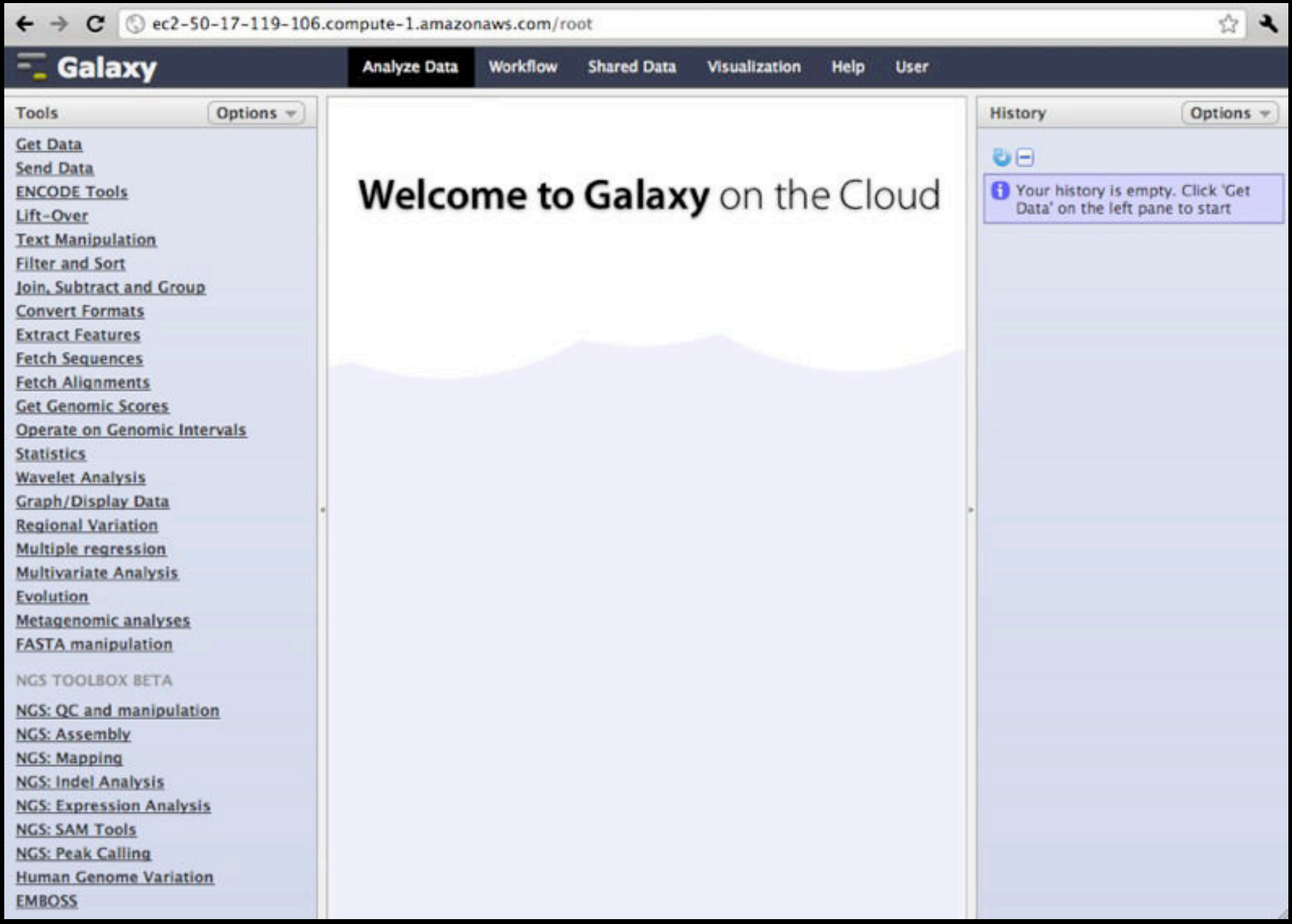
2. Start an EC2 Instance



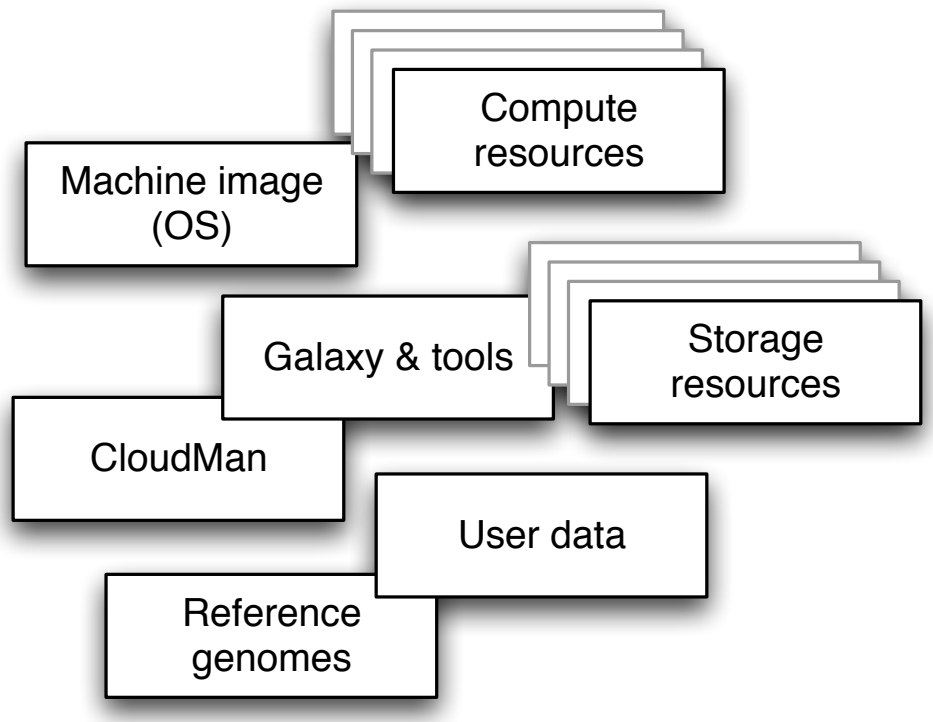
3. Configure Your Cluster

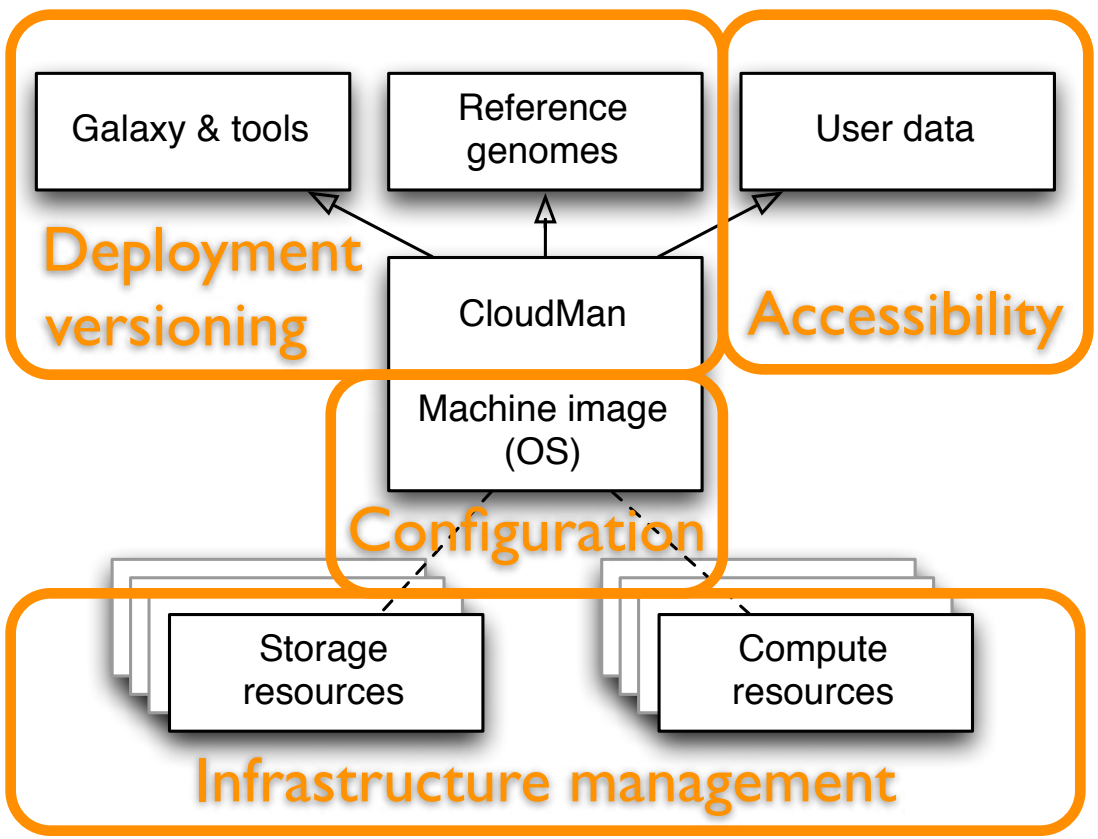



ec2-50-16-1-149.compute-1.amazonaws.com/cloud
Galaxy Cloudman
Info: report bugs | wiki | screencast
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.
Terminate cluster
Add nodes ▾
Remove nodes
Access Galaxy
Status
Cluster name: Heteroplasmy study
Disk status: 50M / 1000G (1%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data
External Logs: Galaxy Log
Autoscaling is off.
Turn on?
Cluster status log


ec2-50-17-119-106.compute-1.amazonaws.com/root
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
FASTA manipulation
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Indel Analysis
NGS: Expression Analysis
NGS: SAM Tools
NGS: Peak Calling
Human Genome Variation
EMBOSS
Welcome to Galaxy on the Cloud
History
Options
Your history is empty. Click 'Get Data' on the left pane to start

Components ⇨ deployment





Building CloudMan machine image

- Dependencies and tools
  - Automated the configuration process
  - Go from bare bones OS to Galaxy-ready machine
- Applicable outside the cloud
  - Build-a-machine (yourself)
- Based on Python Fabric
  - Easily extensible: add your own tool

Galaxy VM

- Built a Galaxy pre-configured VM image
  - Galaxy and a range of tools
  - Reference genomes will need to be downloaded (but it's automated)
- Available from S3 bucket `usegalaxy`

CloudMan instances

- Self-contained configuration: CloudMan, Galaxy & tools, reference genomes, data
- Thus:
  - Reproducible
  - Customizable
  - Extensible

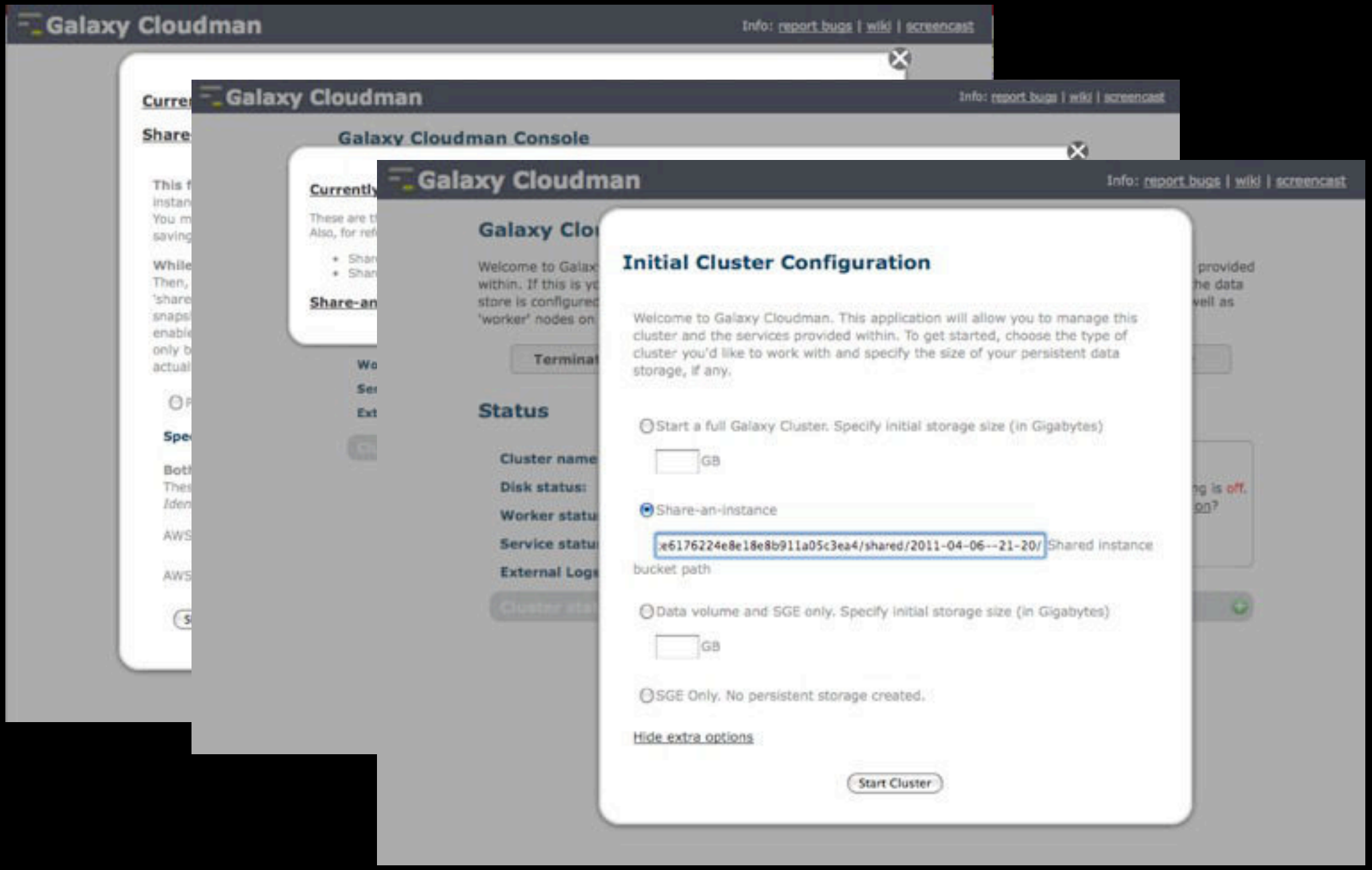
Data

- It is only yours
- Preserved while the cluster is off
- Can grow as the analysis grows
  - Up to 1,000GB currently

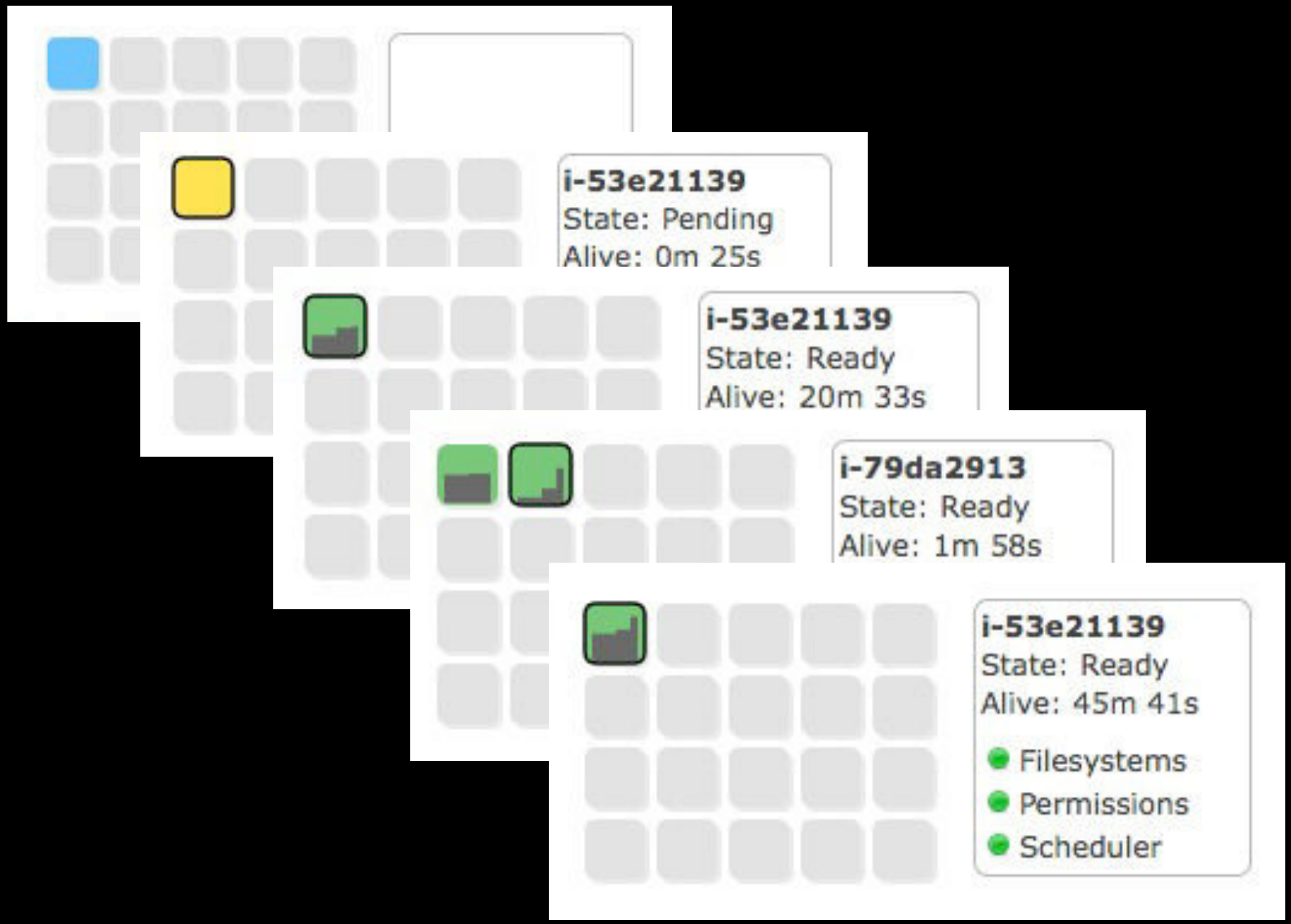
The importance of sharing

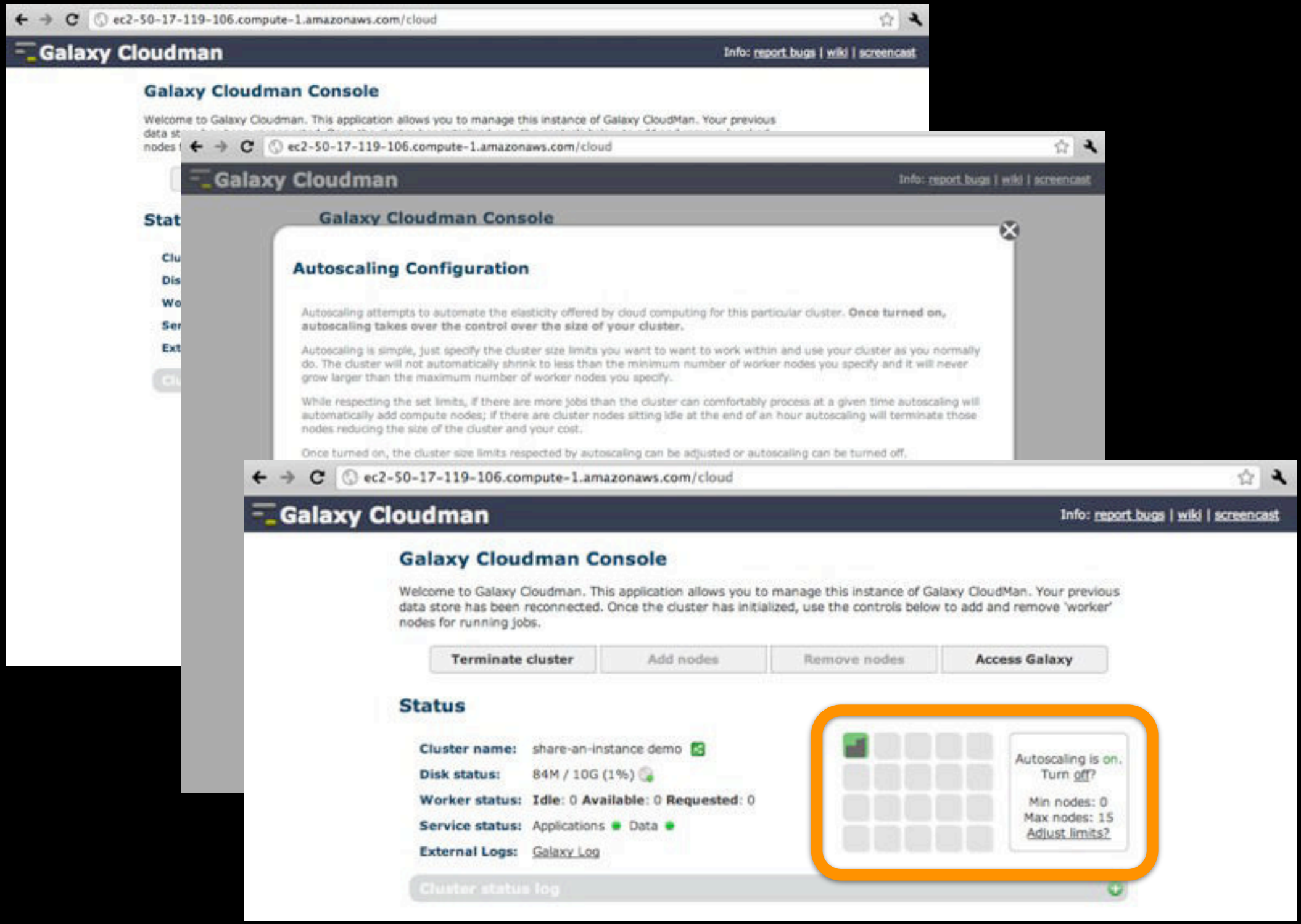
- Share entire Galaxy CloudMan cluster instances
- Publish an analysis
  - In progress or otherwise
- Use CloudMan as PaaS
  - Deploy your own tool and make it available
- Snapshot your instance
  - Data
  - Configuration

Deployment sharing

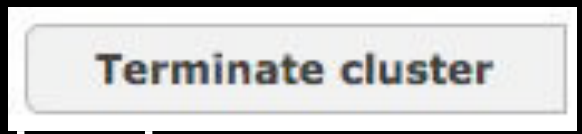


Grow and Shrink




ec2-50-17-119-106.compute-1.amazonaws.com/cloud
Galaxy Cloudman
Info: report bugs | wiki | screencast
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous
Autoscaling Configuration
Autoscaling attempts to automate the elasticity offered by cloud computing for this particular cluster. Once turned on, autoscaling takes over the control over the size of your cluster.
Autoscaling is simple, just specify the cluster size limits you want to want to work within and use your cluster as you normally do. The cluster will not automatically shrink to less than the minimum number of worker nodes you specify and it will never grow larger than the maximum number of worker nodes you specify.
While respecting the set limits, if there are more jobs than the cluster can comfortably process at a given time autoscaling will automatically add compute nodes; if there are cluster nodes sitting idle at the end of an hour autoscaling will terminate those nodes reducing the size of the cluster and your cost.
Once turned on, the cluster size limits respected by autoscaling can be adjusted or autoscaling can be turned off.
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.
Terminate cluster
Add nodes
Remove nodes
Access Galaxy
Status
Cluster name: share-an-instance demo
Disk status: 84M / 10G (1%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data
External Logs: Galaxy Log
Autoscaling is on. Turn off?
Min nodes: 0
Max nodes: 15
Adjust limits?
Cluster status log

Don’t waste the resources

- Once the need for a given cluster subsides, 

 - you can always start it back up
  - Data is preserved while a cluster is down

Summary of the CloudMan architecture

- Minimum setup time and cost
  - No need for an external broker
- Automated configuration
- Data persistence
- Built-in support for managing the oscillating data volume
- Self-contained deployment
  - Customizable instances: CloudMan as PaaS
  - Versioning of tools, data, and configurations

Questions
&
Comments

http://usegalaxy.org/cloud