Running Galaxy on the Cloud

Enis Afgan & Dannon Baker
Ruđer Bošković Institute & Emory University

@ GCC 2013, Oslo

Need an analysis? There's a tool for that.

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- Wormbase test server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- EpiGRAPH test server

NGS: Mapping

- Lastz map short reads against reference sequence
- Lastz paired reads map short paired reads against reference sequence
- Map with Bowtie for Illumina
- Map with Bowtie for SOLiD
- Map with BWA for Illumina
- Map with BWA for SOLiD
- Map with BFAST
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output
- Map with PerM for SOLiD and Illumina
- Re-align with SRMA
- Map with Mosaik

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for SOLiD Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- Depth of Coverage on BAM files

REALIGNMENT

- Realigner Target Creator for use in local realignment
- Indel Realigner - perform local realignment

BASE RECALIBRATION

- Count Covariates on BAM files
- Table Recalibration on BAM files
- Analyze Covariates - draw plots

GENOTYPING

- Unified Genotyper SNP and indel caller

ANNOTATION

- Variant Annotator

FILTRATION

- Variant Filtration on VCF files

VARIANT QUALITY SCORE RECALIBRATION

- Variant Recalibrator
- Apply Variant Recalibration

VARIANT UTILITIES

- Validate Variants
- Eval Variants
- Combine Variants

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 1%

Tools

search tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- Genome Diversity
- Phenotype Association
- EMBOSS

NGS TOOLBOX BETA

- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- Depth of Coverage on BAM files

Depth of Coverage (version 0.0.2)

Choose the source for the reference list:
Locally cached

BAM files
-I,--input_file <input_file>

BAM file 1

BAM file:
9: https://galaxy-vic.genome.edu.au/datasets/a6aaf8a0aa187613/display?to

Add new BAM file

Using reference genome:
Human (Homo sapiens) (b37): hg_g1k_v37
-R,--reference_sequence <reference_sequence>

RefSeq Rod:
Selection is Optional
-geneList,--calculateCoverageOverGenes <calculateCoverageOverGenes>

Partition type for depth of coverage:
Select All | Unselect All

- [x] sample
- [ ] readgroup
- [ ] library

-pt,--partitionType <partitionType>

Summary coverage thresholds
-ct,--summaryCoverageThreshold <summaryCoverageThreshold>

Add new Summary coverage threshold

Output format:
rtable
--outputFormat <outputFormat>

Basic or Advanced GATK options:
Basic

Basic or Advanced Analysis options:
Basic

Execute

History

Random samples
1.1 GB

- 14: GenomeSpace Exporter on data 8
- 13: GenomeSpace import on Galaxy History Item: Generate pileup
- 10: Generate pileup on data 9: converted pileup
- 9: https://galaxy-vic.genome.edu.au/datasets/a6aaf8a0aa187613/display?to_ext=bam
- 8: Count on data 7
- 7: https://galaxy-vic.genome.edu.au/datasets/fdff468d39537a58/display?to_ext=tabular
- 6: FastQC http://static.vlsci.unimelb.edu.au/nathanh/NA12878.chr22_exome.BWA_mapped.chr22_filtered.bam.html
- 5: http://galaxy-vic.genome.edu.au/datasets/870093c60d62e4ae/display?to_ext=gtf
- 4: Generate pileup on data 3: converted pileup
- 3: http://static.vlsci.unimelb.edu.au/nathanh/NA12878.chr22_exome.BWA_mapped.chr22_filtered.bam
- 2: Map with Bowtie for Illumina on data 1
- 1: p1-c-b-1.fastq

usegalaxy.org

Getting your own Galaxy

Local install

http://wiki.galaxyproject.org/Admin/Get%20Galaxy

Galaxy on the Cloud

http://wiki.galaxyproject.org/CloudMan/


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0%
Tools
search tools
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
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+
NGS: QC and manipulation
NGS: Picard (beta)
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: GATK Tools (beta)
Welcome to Galaxy on the Cloud
managed by CloudMan
History
0 bytes
Your history is empty. Click 'Get Data' on the left pane to start
Researchers
CloudMan
100+
100sGB
Cloud resources

usecloudman.org

When to use the cloud?

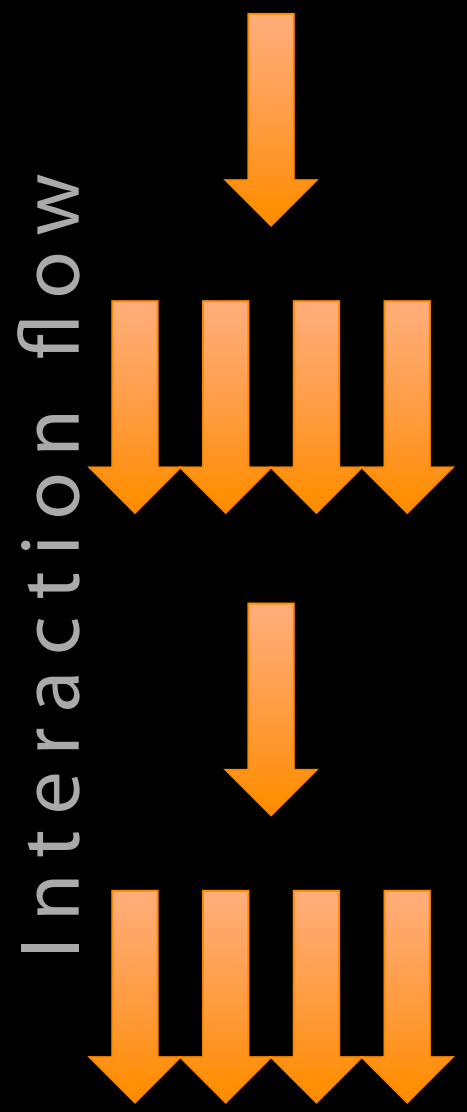
- Don't have informatics expertise or the infrastructure to run and maintain
- Have variable or particular resource needs
- Cannot upload data to a shared resource
- Need for customization
- Have oscillating data volume
- Want to test or share a tool, quickly & safely
- Want to make your analysis readily available to others
- Want fast access to AWS public datasets

How to use the cloud?

1. Get an account on the supported cloud
2. Start a master instance via the cloud web console or CloudLaunch
3. Use CloudMan's web interface on the master instance to manage the platform
4. Use or customize Galaxy

Workshop plan

- Launch an instance
- Demonstrate the following CloudMan features and prepare for the data analysis part:
  - Auto-scaling
  - Using an S3 bucket as a data source
  - Accessing an instance over ssh
  - Customizing an instance
  - Controlling Galaxy
  - Sharing-an-instance



YOUR TURN

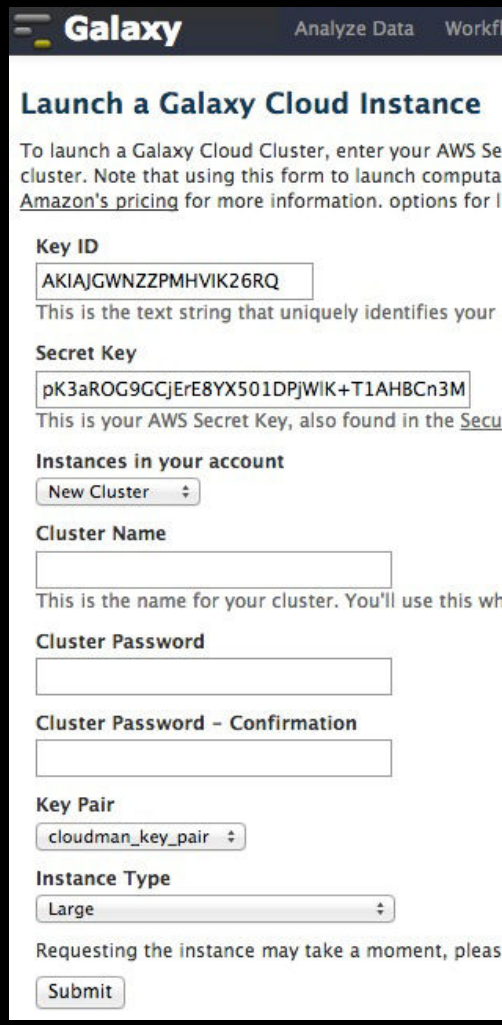
Internet connection

SSID: conferences

Password: uio202aar

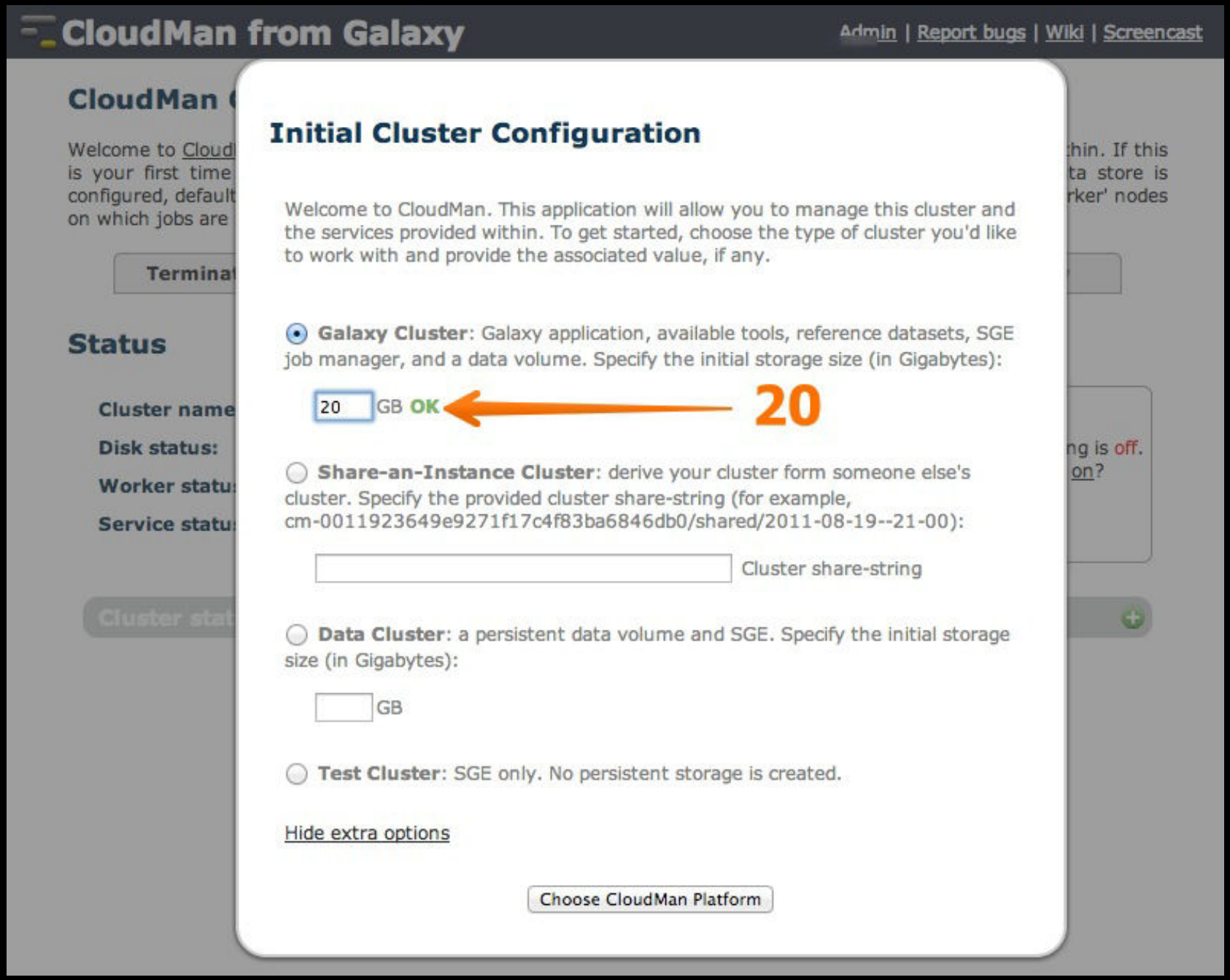
Launch an instance

1. Visit usegalaxy.org/cloudlaunch
2. Enter the provided access key and secret key
   Available from http://bit.ly/gcw2013
3. Choose New cluster
4. Set any name as the cluster name
5. Set any password
6. Choose CloudManKP1
7. Choose Large instance type
8. Launch your instance
   Wait for the instance to start (~2-3 minutes)

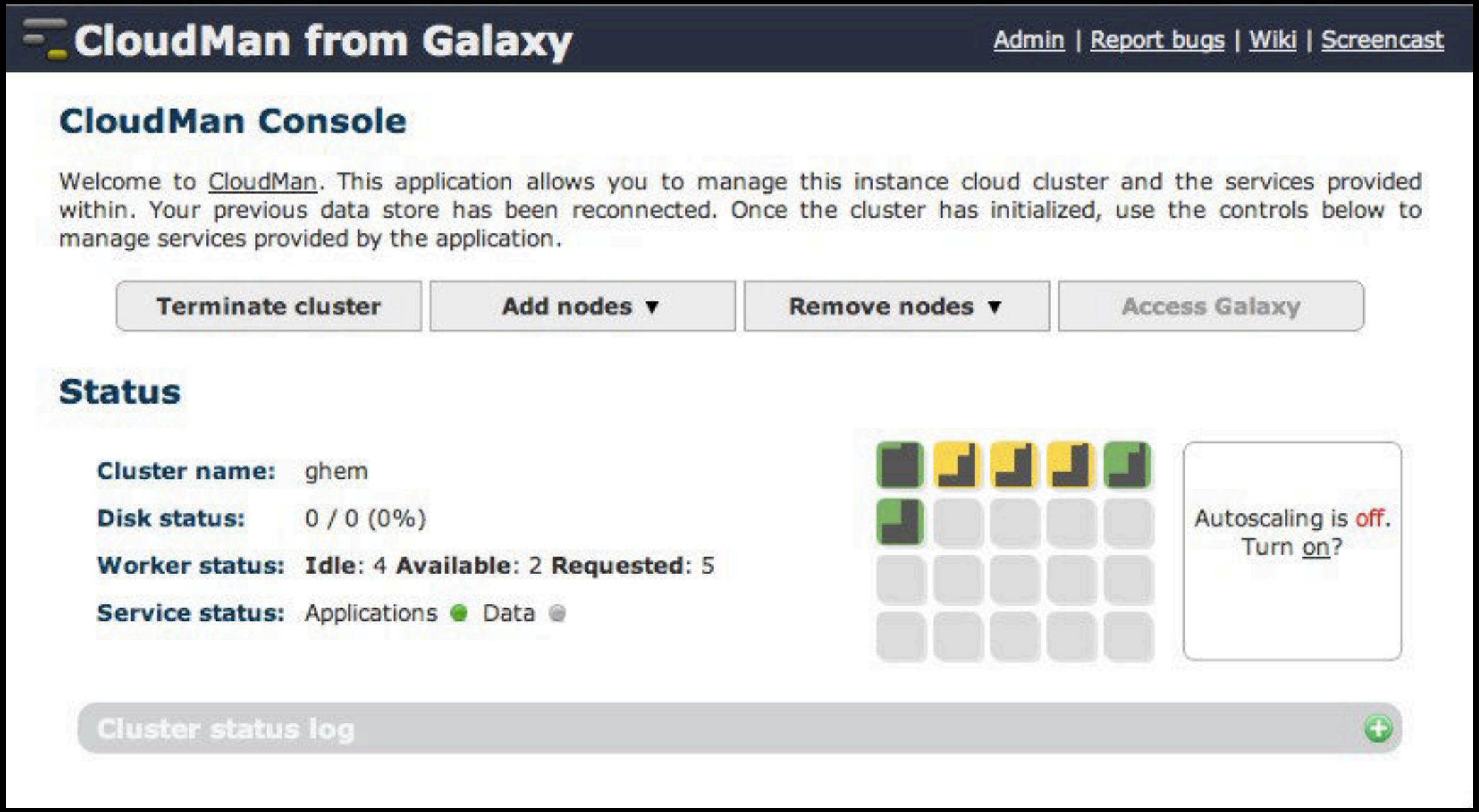


For more details, see
wiki. galaxyproject.org/CloudMan

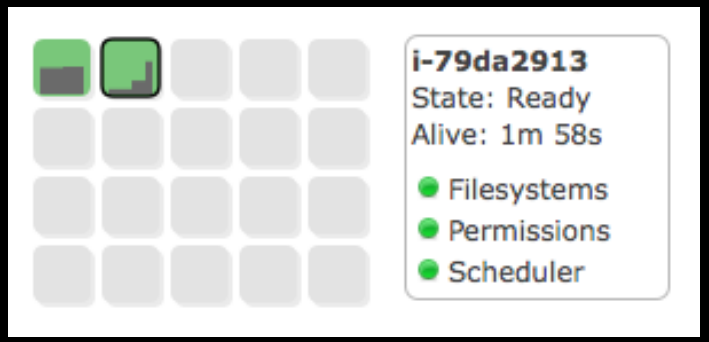
Configure Your Cluster

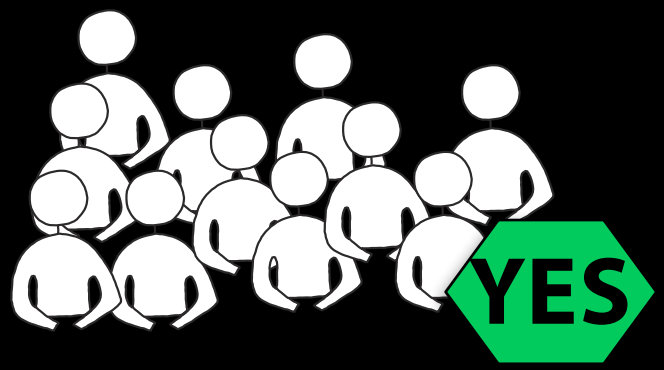


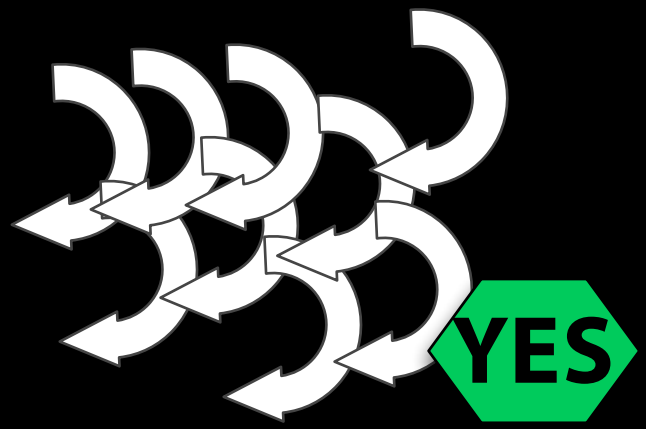
Manage Your Cluster

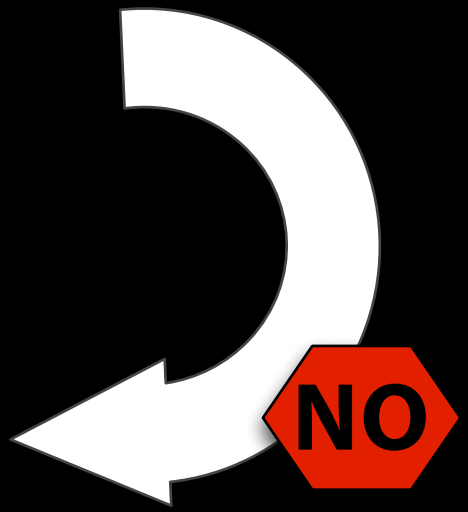


Scaling computation









Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 0%

Tools

search tools

- 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
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard (beta)
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools (beta)

Welcome to Galaxy on the Cloud

managed by CloudMan

History

0 bytes

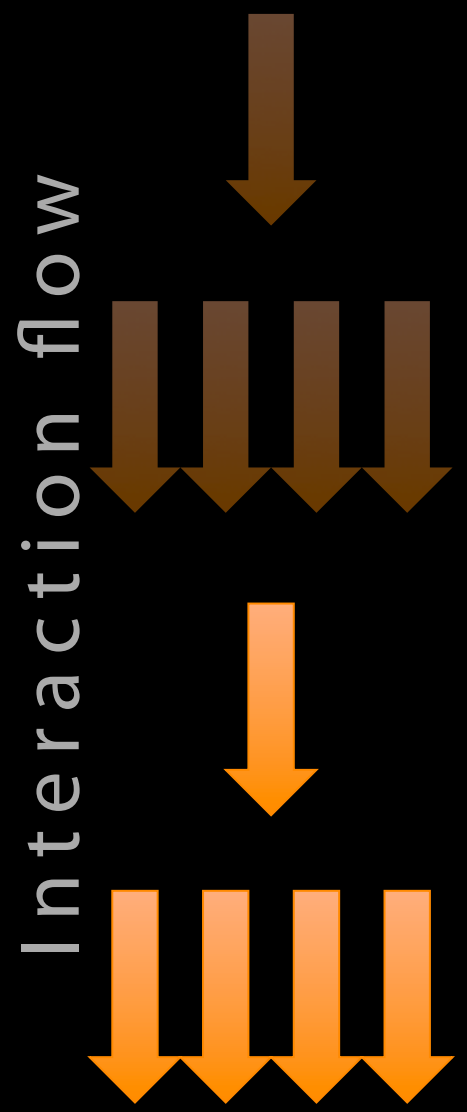
Your history is empty. Click 'Get Data' on the left pane to start

Use your own Galaxy instance

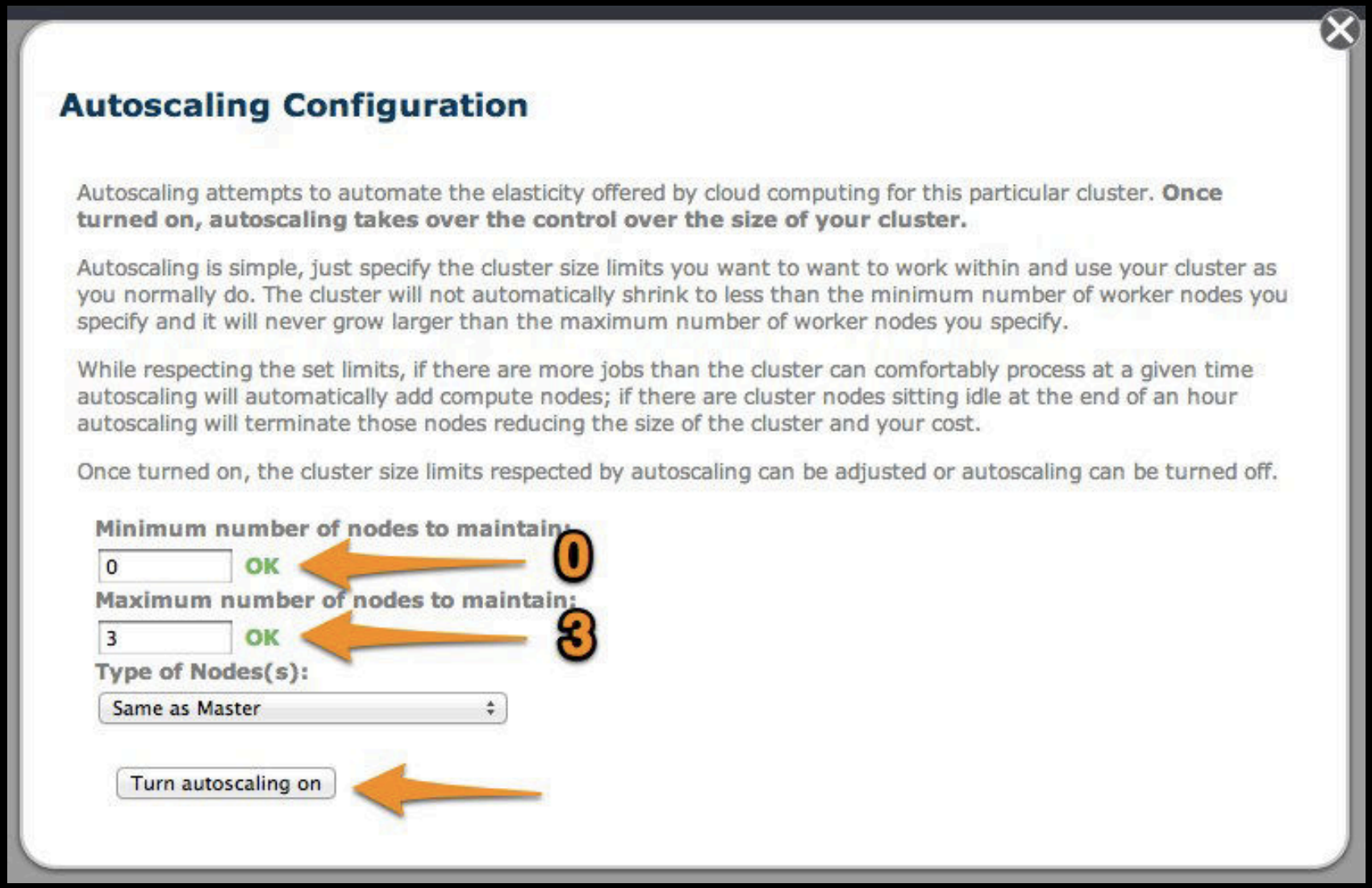
- Register as a new user
- Become Admin user
- Use or customize

Workshop plan

- Launch an instance ✓
- Demonstrate the following CloudMan features and prepare for the data analysis part:
  - Auto-scaling
  - Using an S3 bucket as a data source
  - Accessing an instance over ssh
  - Customizing an instance
  - Controlling Galaxy
  - Sharing-an-instance



Auto-scaling



Auto-scaling in action

Fixed cluster size

5 nodes

Computation time: 9 hrs

Computation cost: $20

20 nodes

Computation time: 6 hrs

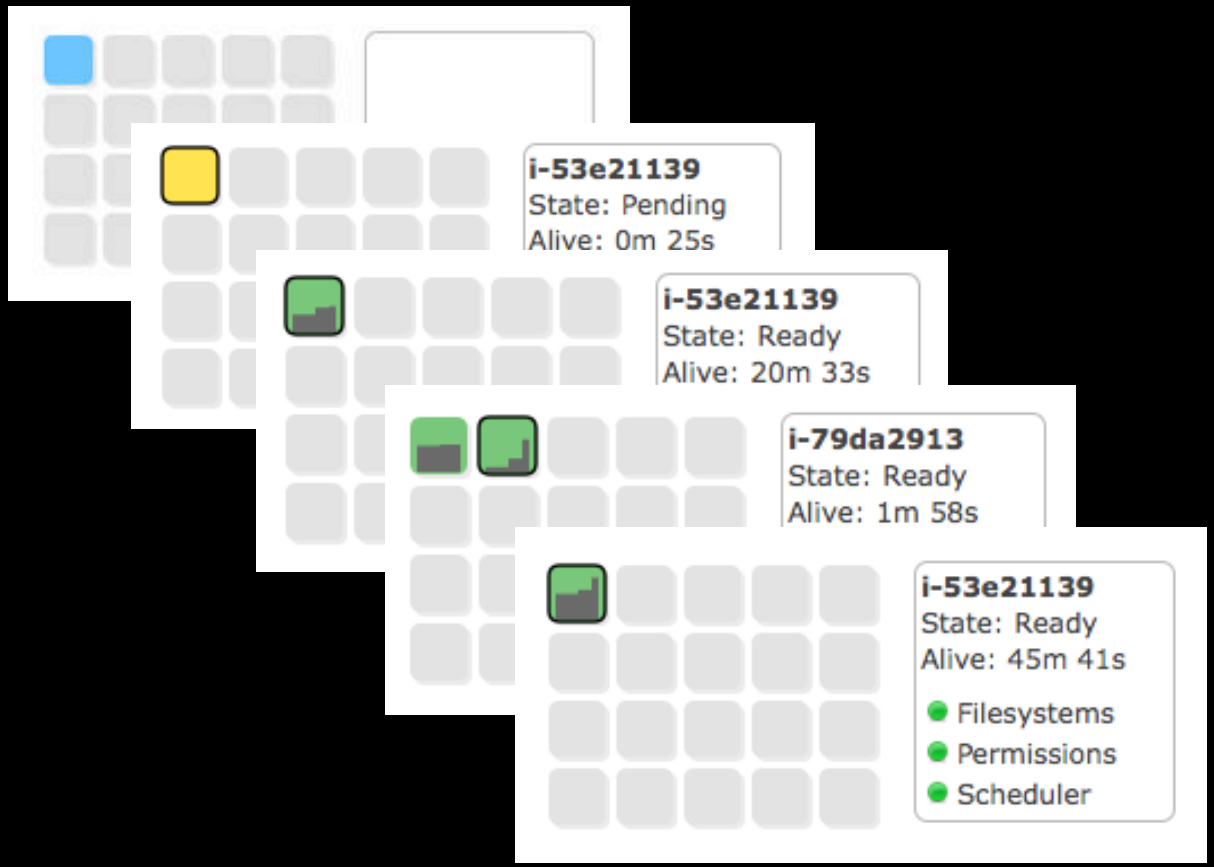
Computation cost: $50

Dynamic cluster size

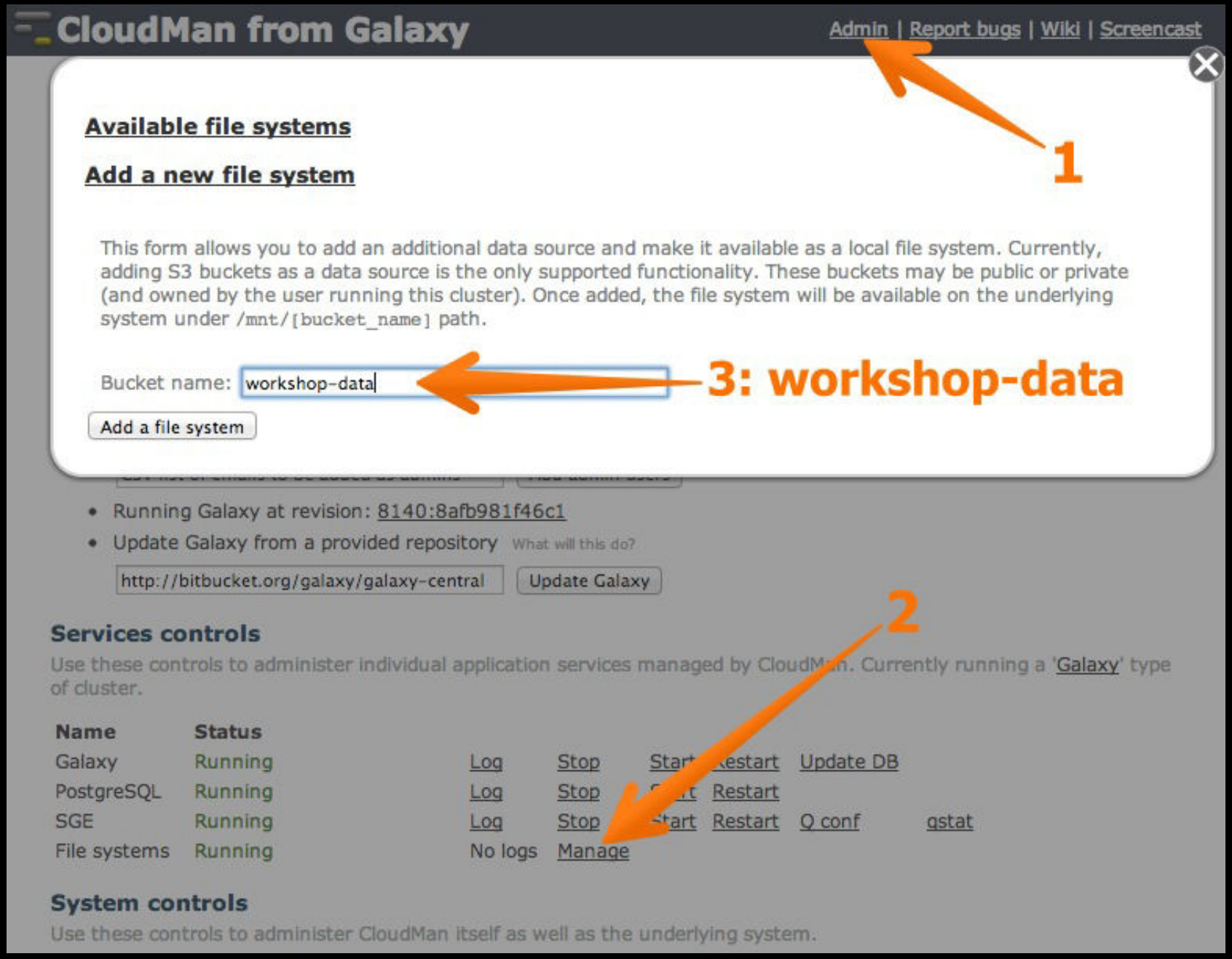
1 to 16 nodes

Computation time: 6 hrs

Computation cost: $20



Using an S3 bucket as a data source



Accessing an instance over ssh

Install *Secure Shell from Chrome*

SSH using user `ubuntu` and the password you chose when launching an instance

Once logged in

- You have full system access to your instance, including `sudo`; use it as any other system
- *galaxy* user exists on the system and should be used when manipulating Galaxy (`sudo su galaxy`)
- Can submit any jobs via the standard `qsub` command

Customizing an instance

- Edit Galaxy's configuration

```
$ sudo su galaxy
$ cd /mnt/galaxy/galaxy-app
$ vi universe_wsgi.ini
```

allow_library_path_paste = True

Controlling Galaxy

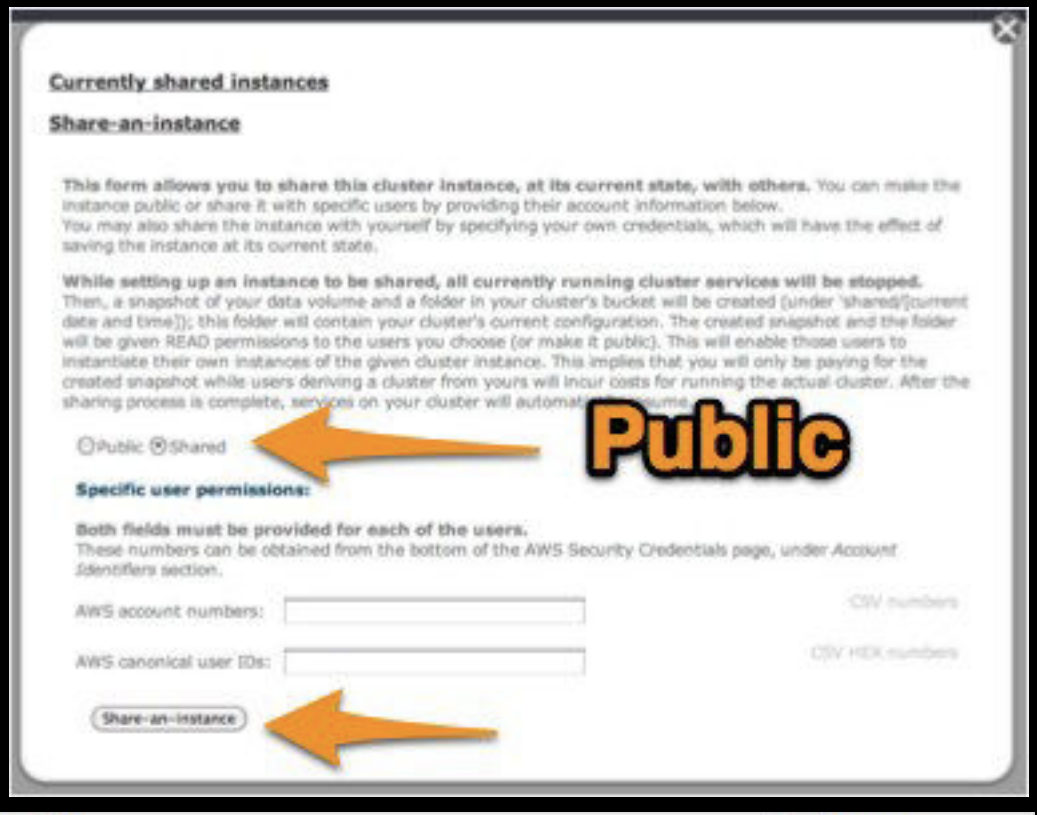
- Start/stop Galaxy application
- Add an admin user
  - Use the email you registered with

S3 bucket as a data library

- Within Galaxy, create a Data Library, using S3 bucket path as the data source (`/mnt/workshop-data`)
  - This will import all the datasets into the Data Library
- Import that dataset into a history

Sharing-an-Instance

- Share the entire Galaxy CloudMan platform
  - Includes all of user data and even the customizations
- Publish a self-contained analysis
- Make a note of the *share-string* and send it to your neighbor

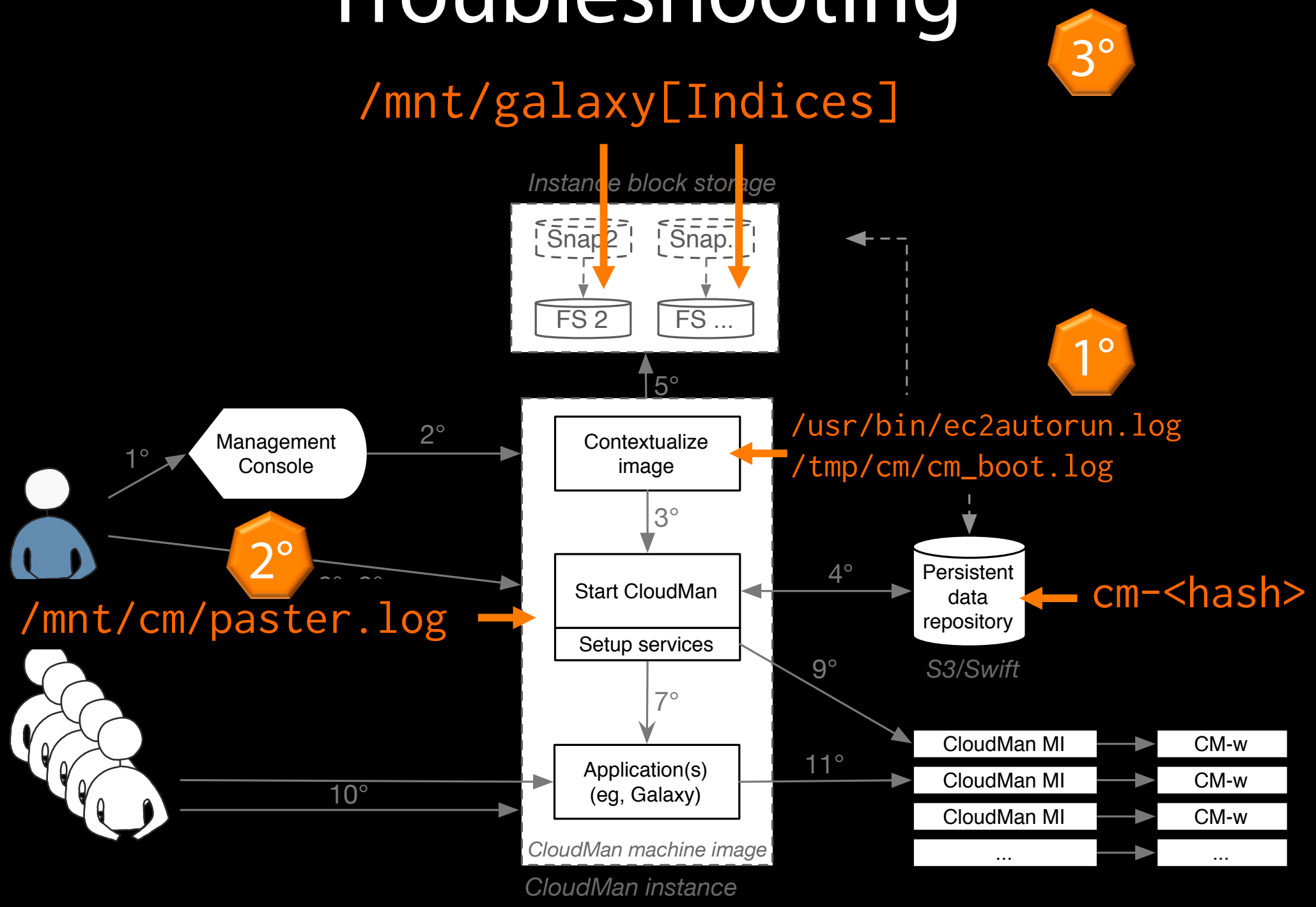


Name	Instance ID	References
Exome sequencing pipeline	cm-b53c6f1223f966914df347687f6fc818/shared/2011-10-07--14-00	Pipeline description

CloudMan Features

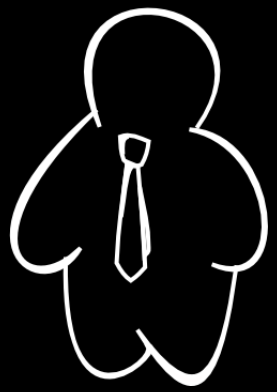
- Start/launch/control through a web browser
- API (via BioBlend library)
- Choose between multiple cluster types
- Terminate and recreate or restart
- Scale the size of the compute cluster
  - Auto-scale
- Support for AWS Spot instances
- Expand the file system
- Customize (via CLI or the Tool Shed): tools, data, references
- Share-an-instance (customized one too)
- Mount an S3 bucket -> data library
- Access via ssh
- Control the Galaxy process
- Deploy on your cloud (automated via CloudBioLinux)

Troubleshooting



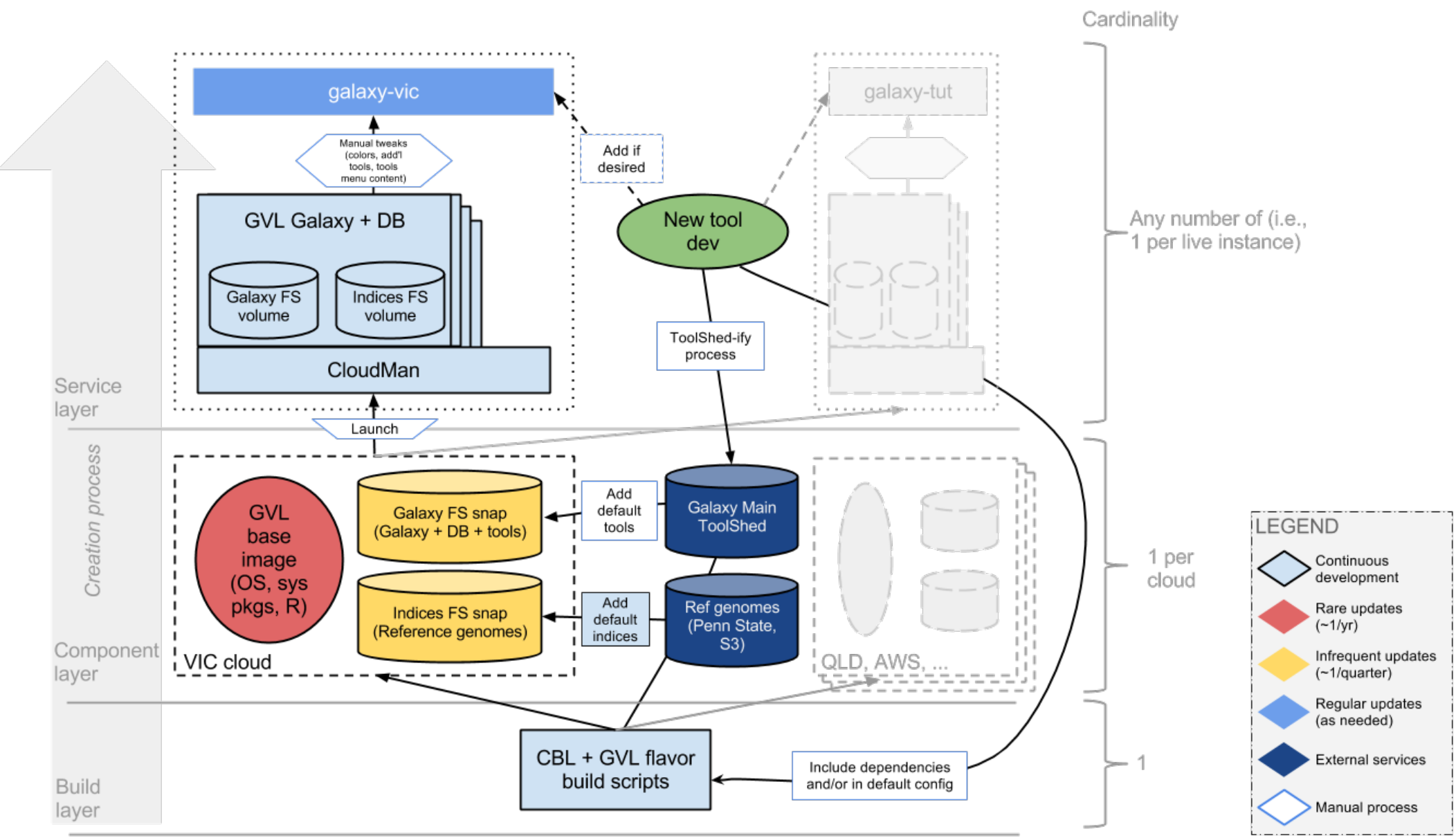
Running Galaxy CloudMan on your own cloud

A brief overview



Supported clouds

- Amazon Web Services
- OpenStack
- Eucalyptus
- OpenNebula


Cardinality
galaxy-vic
Manual tweaks (colors, add'l tools, tools menu content)
GVL Galaxy + DB
Galaxy FS volume
Indices FS volume
CloudMan
Add if desired
New tool dev
galaxy-tut
ToolShed-ify process
Any number of (i.e., 1 per live instance)
Service layer
Launch
Creation process
GVL base image (OS, sys pkgs, R)
Galaxy FS snap (Galaxy + DB + tools)
Indices FS snap (Reference genomes)
VIC cloud
Add default tools
Add default indices
Galaxy Main ToolShed
Ref genomes (Penn State, S3)
QLD, AWS, ...
1 per cloud
Component layer
Build layer
CBL + GVL flavor build scripts
Include dependencies and/or in default config
1
LEGEND
Continuous development
Rare updates (~1/yr)
Infrequent updates (~1/quarter)
Regular updates (as needed)
External services
Manual process

Building

- Leverage CloudBioLinux build framework
- Developed GVL flavor
  - Base Galaxy image
  - Full CloudBioLinux image
- There are also more specific flavors available
  - cloudman

github.com/afgane/gvl_flavor

Deploying

- Integrated with BioCloudCentral.org
  - Use the public one, deploy your own or run locally
- Supports multiple clouds
  - NeCTAR cloud
  - AWS Sydney

Contribute

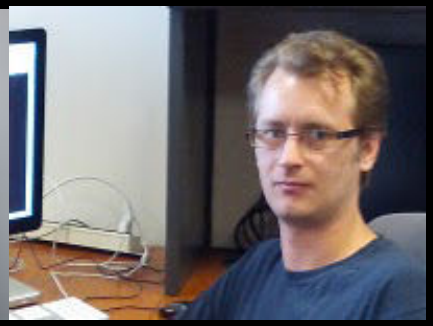
(start with documentation)


Enis Afgan


Guru Ananda


Dannon Baker


Dan Blankenberg


Dave Bouvier


Dave Clements


Nate Coraor


Carl Eberhard


Jeremy Goecks


Nuwan Goonasekera


Jen Jackson


Greg von Kuster


Ross Lazarus


Rémi Marenco


Scott McManus


Anton Nekrutenko

James Taylor

The Galaxy Team

http://galaxyproject.org/wiki/GalaxyTeam