Enabling NGS analysis with(out) the infrastructure

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

Bioinformatics Open Source Conference, July 16, 2011, Vienna
CloudMan-as-a-Bridge

A. Users in different labs

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

A. Users in different labs
B. Isolated Galaxy instance(s)
C. Dense data center

SaaS

IaaS
CloudMan-as-a-Bridge

A. Users in different labs

B. Isolated Galaxy instance(s)

C. Dense data center

Internet

SaaS

CloudMan + Galaxy

IaaS
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. 

2. 

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)
- Share-an-instance
- Data volume and SGE only. Specify initial storage size (in Gigabytes)
- SGE Only. No persistent storage created.

Hide extra options

Start Cluster
Deploying a cluster on AWS

1. Initial Cluster Configuration
   Welcome to Galaxy Cloudman. This application will allow you to manage the 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.

2. 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.

3. Status
   - Cluster name: Heteroplasmy study
   - Disk status: 50M / 10000G (1%)
   - Worker status: Idle: 0 Available: 0 Requested: 0
   - Service status: Applications: Data
   - External Logs: Galaxy Log
   - Cluster status status log
Deploying a cluster on AWS

1. **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.

2. **Galaxy Cloudman**: Welcome to Galaxy Cloudman. This is your first time storing data. Choose default service 'worker' nodes on which jobs are to be run.


4. **Welcome to Galaxy on the Cloud**: Your history is empty. Click 'Get Data' on the left pane to start.
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

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

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 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

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: [Input field]

AWS canonical user IDs: [Input field]

Share-an-instance
Deployment sharing
Deployment sharing
Use CloudMan as SaaS

Use CloudMan as PaaS

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