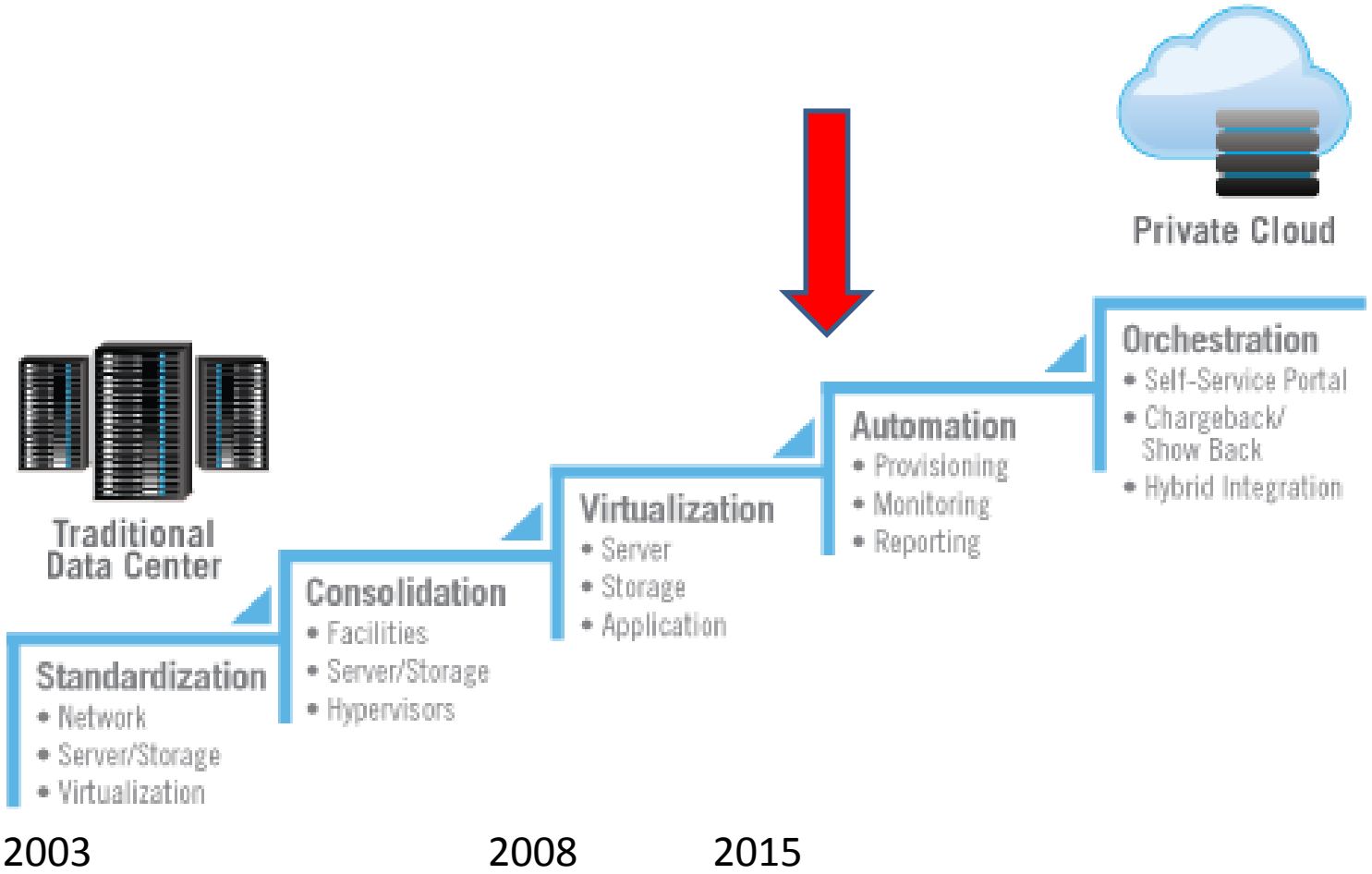
Basic Course on Bioinformatics tools for Next Generation Sequencing data mining

IT infrastructure and user interface: The Galaxy architecture and ARIES cluster

Arnold Knijn

IT Sector - ISS

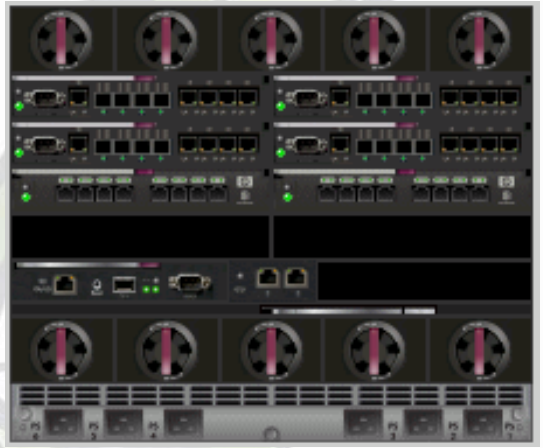
Data Center evolution



ISS IT infrastructure

> 130 virtual servers

50 TB



180 TB

256 CPU/512 thread

32 x 1 GB/s

Storage

1,552 GB RAM

Networking

Galaxy

Data intensive biology *for everyone*.

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

Mailing lists, Tool Shed, wiki

Search all resources

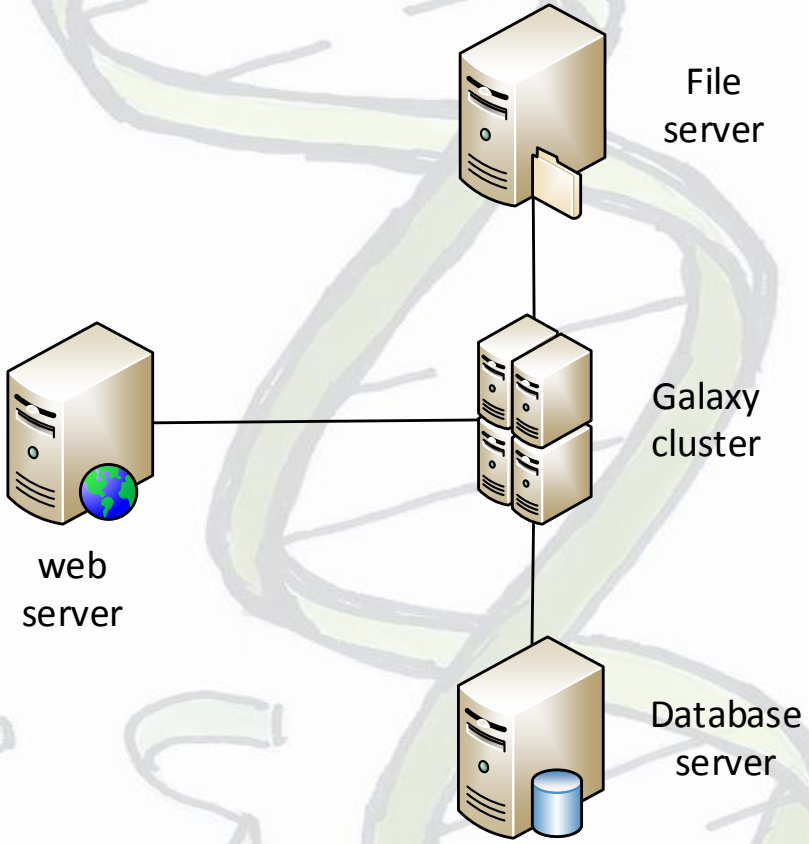
The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State University, and the Department of Biology at Johns Hopkins University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

Which Galaxy?

- Public server (80+)
- Own computer
- Appliance Galaxy Edition
- Data Center cluster
- Cloud standalone: Galaxy on Jetstream
- Cloud virtual cluster: Cloudman

Galaxy components

- Job runner
- File server
- Database server
- Web server



Default vs production installation

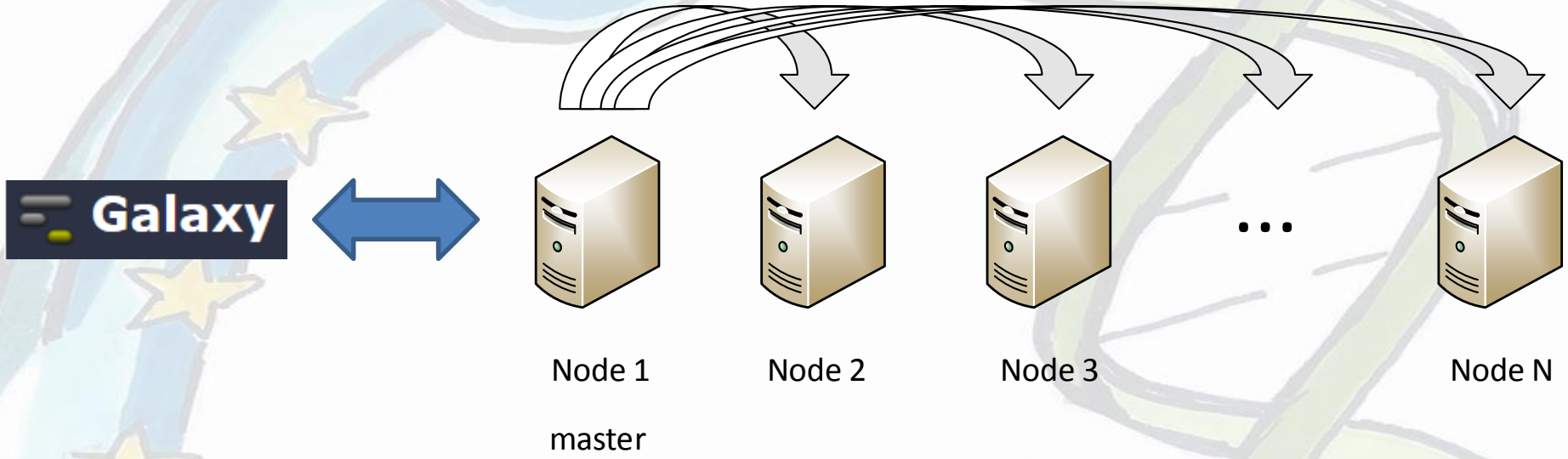
All-in-one (default)

- ✗ Internal file system
- ✗ SQLite
- ✗ Built-in HTTP server for all tasks
- ✗ Local job runner
- ✗ Single process
- ✓ Simplest error-proof configuration

Production (scalable)

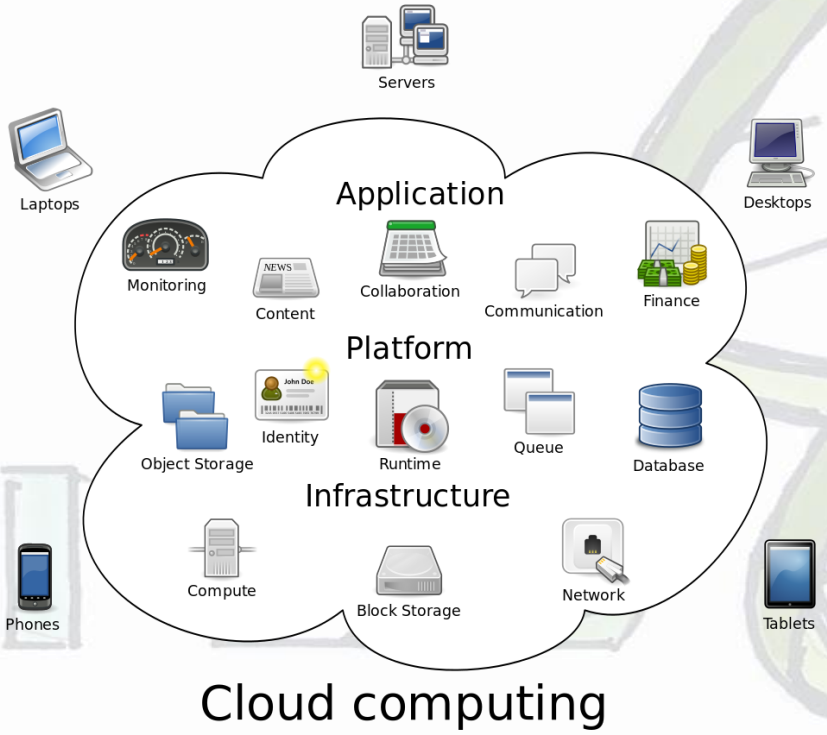
- ✓ External file server
- ✓ Real database
- ✓ Real HTTPS server for many tasks
- ✓ Cluster job runner
- ✓ Multi process
- ✗ More complex configuration

Galaxy cluster



Cloud?

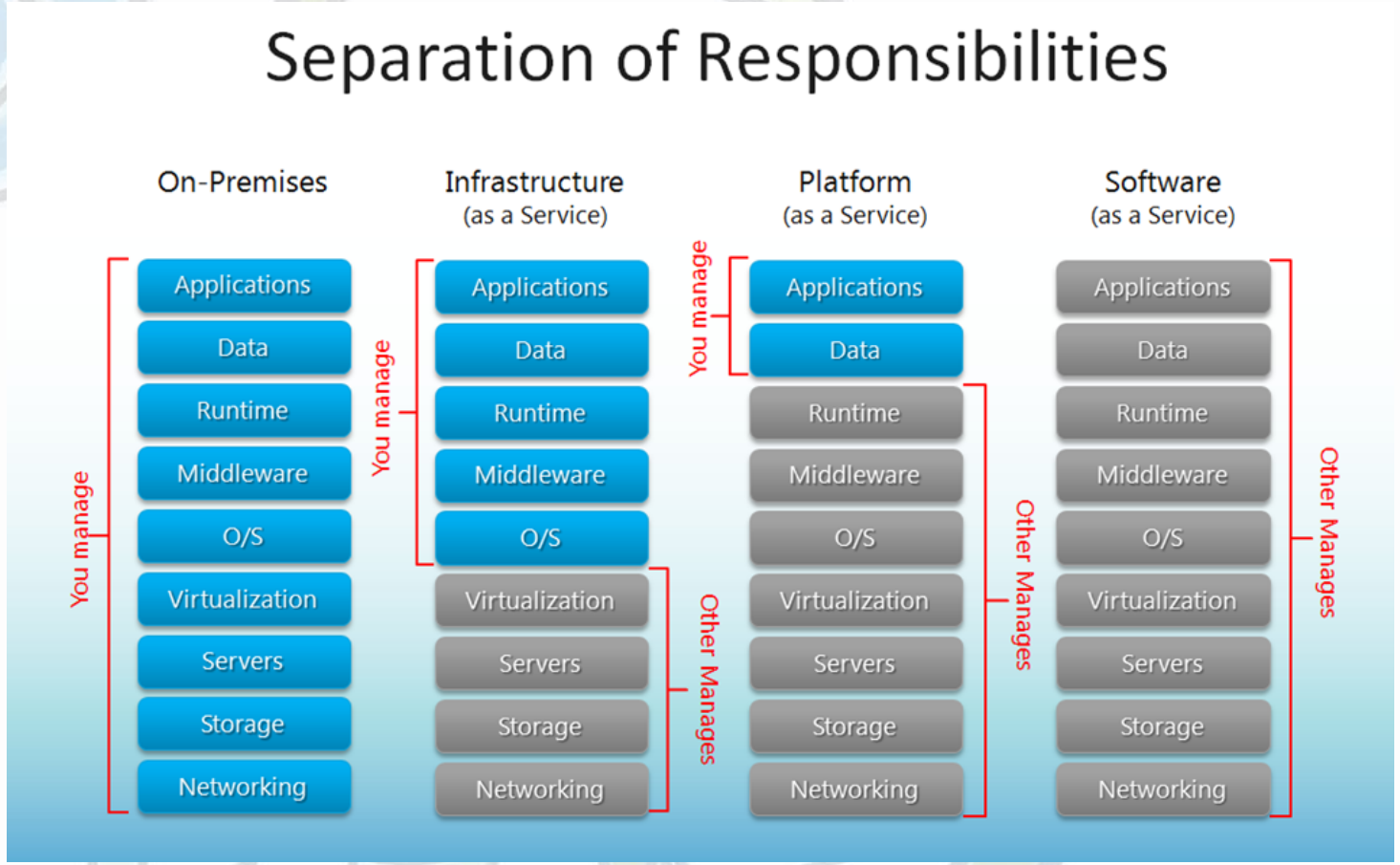
- Cloud Storage (Google drive, iCloud, OneDrive, DropBox, ecc.)
- Cloud Computing: internet-based on-demand access to a shared pool of configurable computing resources



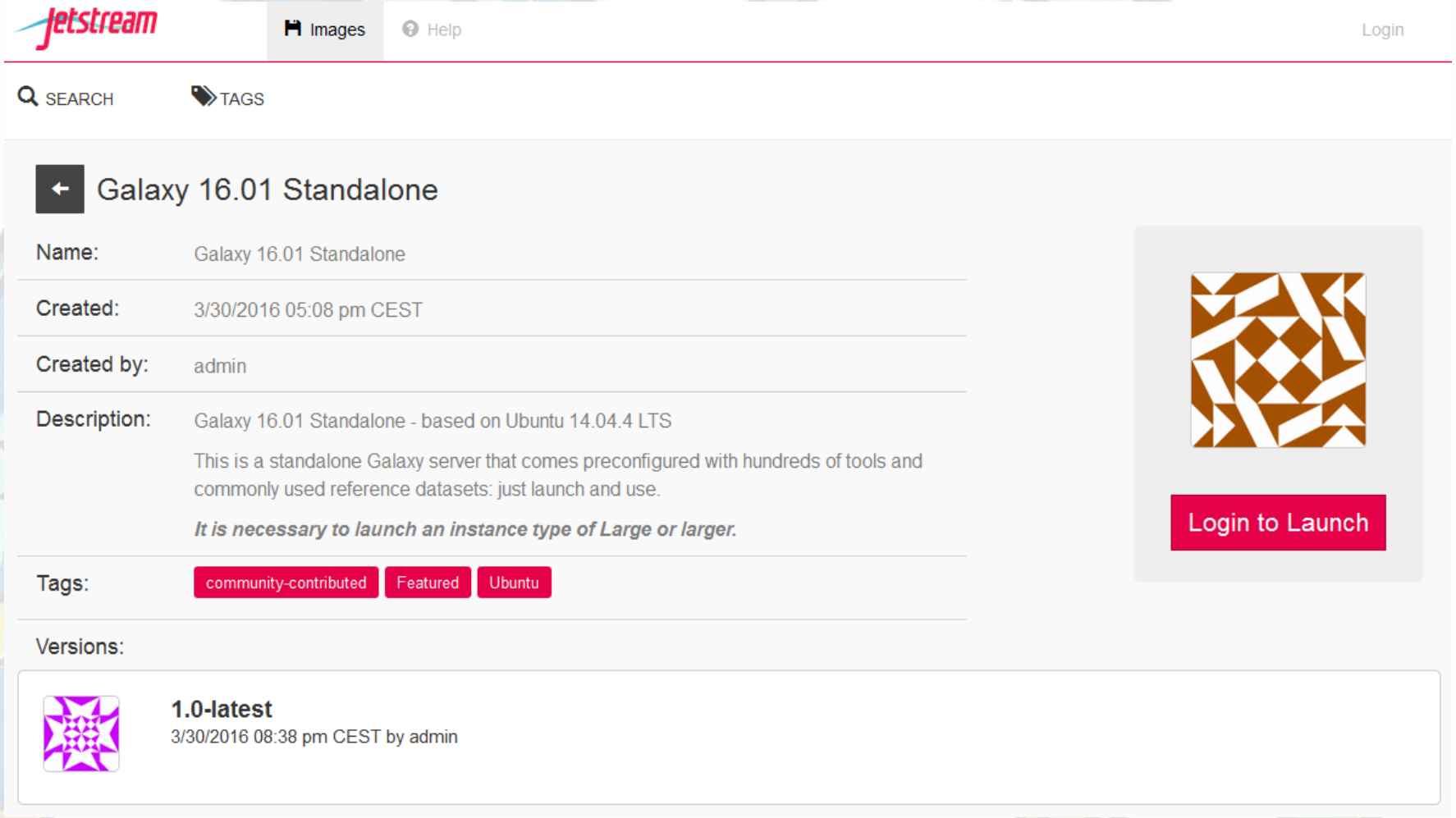
Galaxy on the cloud

- Infrastructure as a Service (system administrators)
- Platform as a Service (software engineers)
- Software as a Service (users)

- Self-service
- Your wallet is the limit



Galaxy on Jetstream (USA)



Galaxy on Amazon

Galaxy Cloud Launch

Easily launch your own cloud servers for use with Galaxy and CloudMan. See this page for detailed instructions on how to get started.

Release Candidate 1 for Galaxy on the Cloud with Galaxy 16.04 is available. If you would like to use/test it, choose *Testing: Galaxy 16.04* flavor under the Advanced options dropdown.

Cloud: Amazon - Ireland (AWS EC2)
Choose from the available clouds. The credentials you provide below must match (ie, exist on) the chosen cloud.

Access key:
Your cloud account API access key. For the Amazon cloud, available from the security credentials page.

Secret key:
Your cloud account API secret key. For the Amazon cloud, also available from the security credentials page.

Cluster name: Specify a new name or Choose a saved cluster
Name of your cluster used for identification and restarting. If creating a new cluster, type any name you like.

Password:
Your choice of password, for the CloudMan web interface and accessing the server via ssh.

Instance type: Compute Optimized large (c3.large) (2 vCPU / 3.
Type (ie, virtual hardware configuration) of the server to start.

Cluster type:
- (●) Cluster with Galaxy
- ( ) Cluster only
- ( ) Do not set cluster type now

The cluster type determines the initial startup template used by CloudMan. See this page for details on cluster types.

Storage type:
- (●) Persistent volume storage
- ( ) Transient instance storage

The type of storage to use for the main file system. See this page for more details on storage types.

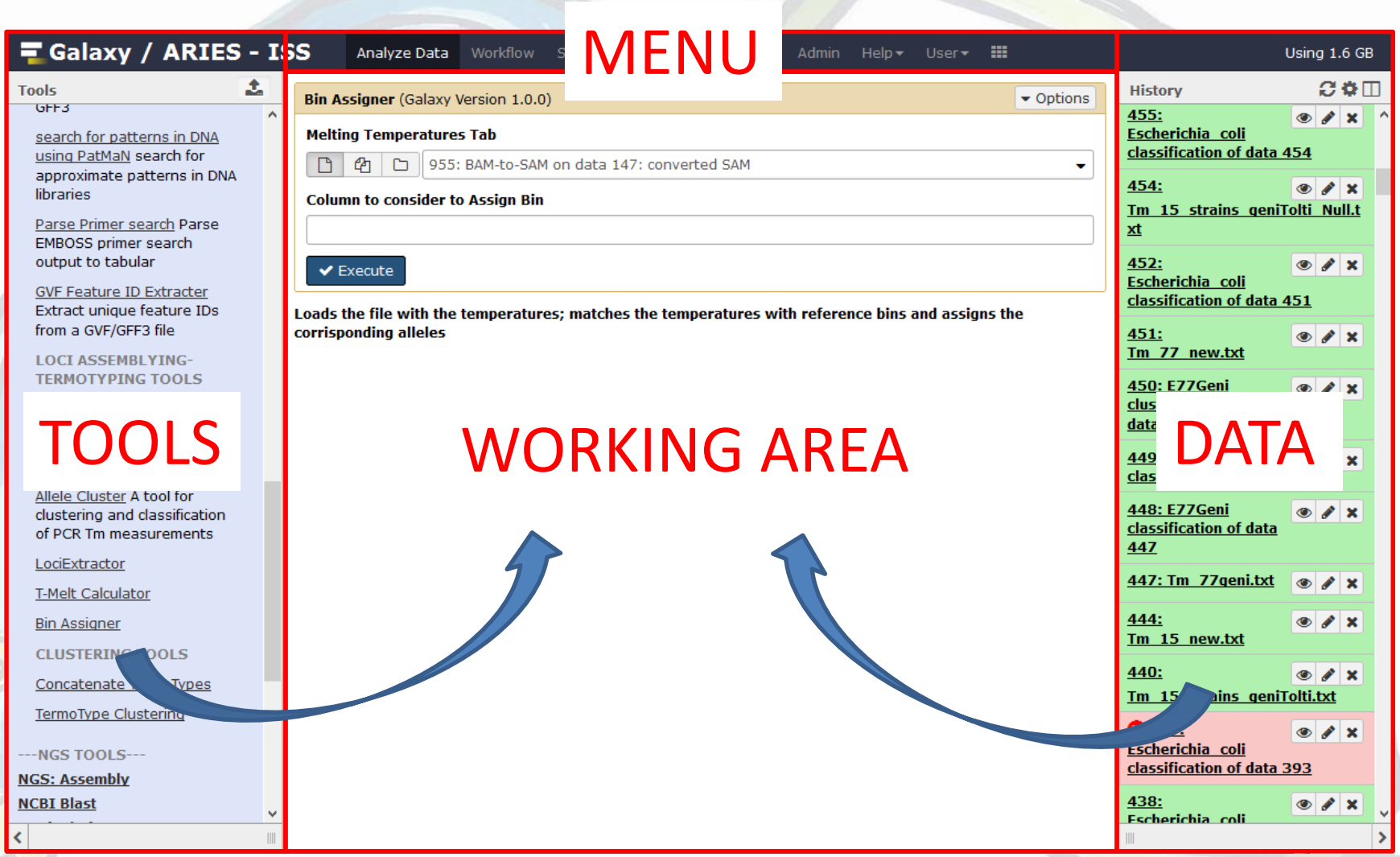
Storage size: 10
The size of the storage (in GB; number only). The default is 10.

↓ Show advanced startup options

Create a cluster

Galaxy user interface

Galaxy user interface



Intuitive and self-documenting

Bin Assigner Home-made tool for bin assignment (Galaxy Tool Version 1.0.0) Options

Melting Temperatures Tab

953: Tm_77_new.txt

Column to consider to Assign Bin

7

Execute

Loads the file with the temperatures; matches the temperatures with reference bins and assigns the corrisponding alleles

Citations Show BibTeX

Michelacci, Valeria and Orsini, Massimiliano and Knijn, Arnold and Delannoy, Sabine and Fach, Patrick and Caprioli, Alfredo and Morabito, Stefano (2016). Development of a High Resolution Virulence Allelic Profiling (HReVAP) Approach Based on the Accessory Genome of Escherichia coli to Characterize Shiga-Toxin Producing E. coli (STEC). In *Frontiers in Microbiology, 7.* [doi:10.3389/fmicb.2016.00202][Link]

UI vs Command-Line

Tool: Bin Assigner

Name:	BinAssigner Log File
Created:	Fri Feb 13 07:43:59 2015 (UTC)
Filesize:	877 bytes
Dbkey:	?
Format:	txt
Galaxy Tool ID:	binassigner
Galaxy Tool Version:	1.0.0
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	e9fb797960230e8a
History ID:	f597429621d6eb2b
UUID:	dc1676ef-87b7-48bf-a24e-4359f57cf2fa
Full Path:	/home/galaxy/galaxy-dist/database/files/001/dataset_1439.dat
Job Command-Line	python /home/galaxy/galaxy-dist/tools/Hrevap/BinAssigner.py -t /home/galaxy/galaxy-dist/database/files/001/dataset_1433.dat -o /home/galaxy/galaxy-dist/database/files/001/dataset_1438.dat -c 7 > /home/galaxy/galaxy-dist/database/files/001/dataset_1439.dat
Job Runtime (Wall Clock)	1 seconds
Cores Allocated	1
Job Start Time	2015-02-13 08:44:00
Job End Time	2015-02-13 08:44:01

Input Parameter	Value	Note for rerun
Melting Temperatures Tab	176: TermoTyping Summary File	
Column to consider to Assign Bin	7	

Home–made tools

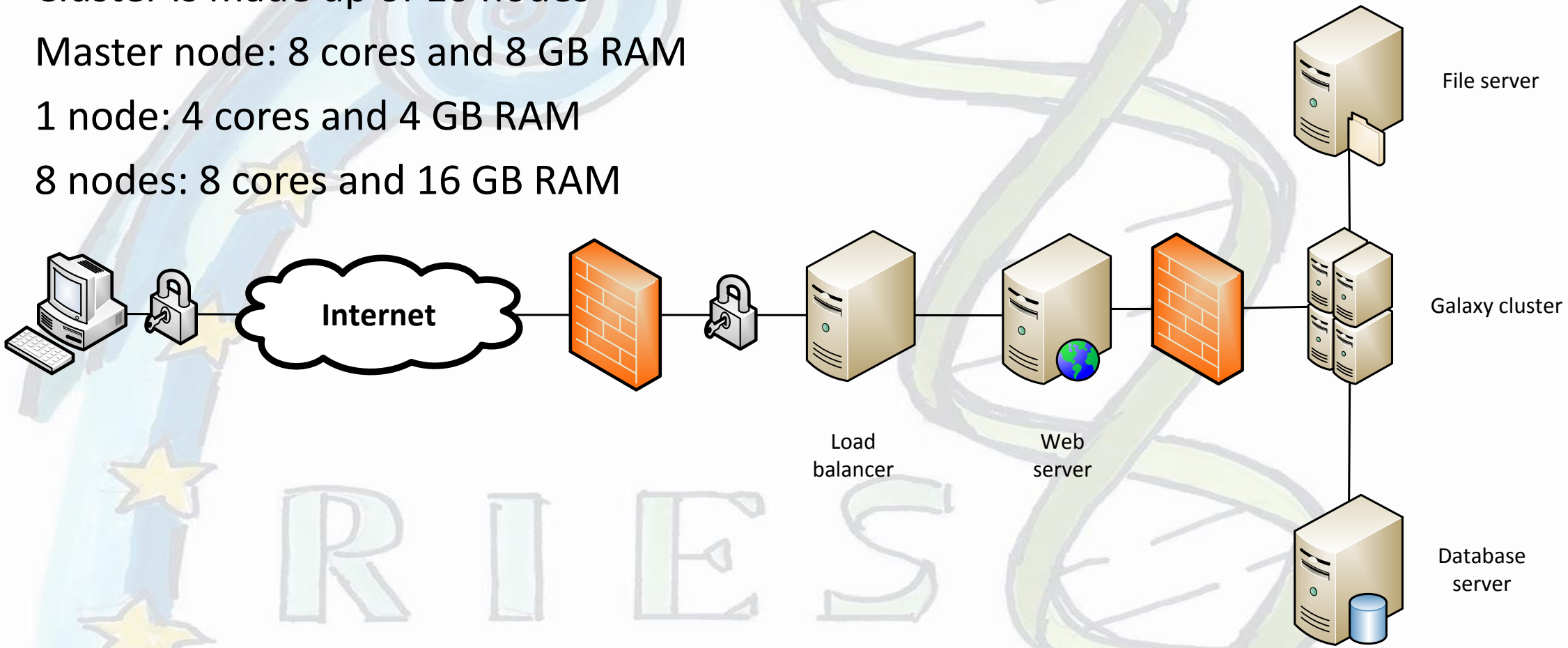
```
<tool id="binassigner" name="Bin Assigner">
  <description>Home-made tool for bin assignment</description>
  <requirements>
    <requirement></requirement>
  </requirements>
  <command interpreter="python">
    BinAssigner.py -t $tmstab -o $output -c $columntab > $logfile
  </command>
  <inputs>
   <param name="tmstab" type="data" format="tabular" label="Melting Temperatures Tab"/>
   <param name="columntab" type="text" format="integer" label="Column to consider to Assign Bin" />
  </inputs>
  <outputs>
   <data format="tabular" name="output" label="Allele Table"/>
   <data format="txt" name="logfile" label="BinAssigner Log File" />
  </outputs>
  <help>
    Loads the file with the temperatures; matches the temperatures with reference bins and assigns
    the corrisponding alleles
  </help>
  <citations>
    <citation type="doi">10.3389/fmicb.2016.00202</citation>
  </citations>
</tool>
```

Galaxy in ISS: ARIES

- First standalone installation in April 2014
- Reinstallation in June 2014
- Installation of the cluster in March 2015
- Alpha test NGS Course last year in June 2015
- Presentation of ARIES in ISS in June 2015
- ARIES opened for ISS users in July 2015
- ARIES went public in September 2015

ARIES cluster

- Cluster is made up of 10 nodes
- Master node: 8 cores and 8 GB RAM
- 1 node: 4 cores and 4 GB RAM
- 8 nodes: 8 cores and 16 GB RAM



ARIES users

Currently 40 users:

- 4 Belgium
- 3 Finland
- 1 France
- 1 Ireland
- 22 Italy (16 ISS)
- 1 Luxembourg
- 5 The Netherlands
- 1 Poland
- 2 Spain

ARIES use

← 1 2 →

Jobs Per Month

Click Month to view details. Graph goes from the 1st to the last of the month.

Max items: 12

Month ↑	User and Monitor Jobs
May 2016	4802
April 2016	4828
March 2016	3285
February 2016	3017
January 2016	2128
December 2015	2712
November 2015	3129
October 2015	1750
September 2015	1480
August 2015	194
July 2015	6266
June 2015	1081

Questions?

- aries@iss.it
- Follow us: @ARIES_GENOMICS