Managing a local Galaxy Instance
Agenda

- Who are we
- Why a local installation
- Local infrastructure
- Local installation
- Tips and Tricks
- SlipStream Appliance
WHO ARE WE
Who We are

Over a decade of Life Sciences IT consulting

- Staffed by **scientists** forced to learn IT to get research done
- Served over **400** organizations
  - Academic, Non-profit
  - Government, Military
  - Pharm, AgBio, Biotech
  - Cloud & Datacenter Providers
• Active contributors to open-source projects
Galaxy Project

- Decrease the barrier to entry into data analysis by improving accessibility of the Galaxy platform

BioTeam

- Encapsulate a decade of IT best-practices expertise to eliminate redundant effort spent building IT systems and installing software
• BioTeam offers a all-in-one solution to help run a local instance Galaxy

• BioTeam is the the official appliance provider for Galaxy platform
  • Exclusive partnership with the Galaxy Team
  • Donations back to the Galaxy Project
Why a local installation

Galaxy is available in several different ways.

Which Option to Choose?
Your choices depend upon your needs. Here are the options depending on what you need:

<table>
<thead>
<tr>
<th></th>
<th>Main</th>
<th>Local</th>
<th>Cloud</th>
<th>Appliance</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your data sets are moderately sized</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Your computational requirements are moderate</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>You want to share your Galaxy objects with others</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>All needed Tools are installed on Main.</td>
<td>Yes</td>
<td>?</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Your data sets are very large</td>
<td>No</td>
<td>?</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Your computational requirements are very large</td>
<td>No</td>
<td>?</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>You have absolute data security requirements</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>No network transfer of data</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Current challenges

- Galaxy Main is a fantastic public resource!!
- Limitations due to popularity
  - Wait times
  - Job and storage quotas
  - Data transfer bottlenecks
  - Pre-defined set of tools and datasets
Current challenges

- Cloud is a good option but has its challenges
  - Understand how to price out jobs
  - Data transfer bottlenecks
  - Familiarity with AWS

- Running Galaxy locally solves these issues
• IT/informatics expertise
  • Acquire and set up infrastructure
  • Install Galaxy, tools, necessary dependencies
  • Optimize/customize for your use cases

• Define policies
  • Managing usage
  • Data back-up
  • Software updates/upgrades
• Informatics support
  • Handle user questions/requests
  • Gather user feedback

• Ongoing dedicated resource
  • Manage updates
  • Facilitate user support
  • Maintain infrastructure
## Benefits of a local Galaxy

It’s all about control

<table>
<thead>
<tr>
<th>YOU CONTROL</th>
<th>BENEFIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of storage</td>
<td>Handle large datasets</td>
</tr>
<tr>
<td>Type of hardware</td>
<td>Run compute intensive jobs</td>
</tr>
<tr>
<td>What tools to install</td>
<td>Customize to your research</td>
</tr>
<tr>
<td>Data access</td>
<td>Granular control of security</td>
</tr>
<tr>
<td>Networking architecture</td>
<td>No data transfer bottleneck</td>
</tr>
<tr>
<td>Software behavior</td>
<td>Optimize how jobs are run</td>
</tr>
</tbody>
</table>
Local Instance of Galaxy

Get Galaxy: Galaxy Download and Installation

In addition to using the public Galaxy server (a.k.a. Main), you can also install your own instance of Galaxy (what this page is about), or create a cloud-based instance of Galaxy. Another option is to use one of the ever-increasing number of Public Galaxy Servers hosted by other organizations.

See Big Picture: Choices for help on deciding which of these options may be best for your situation.

Reasons to Install Your Own Galaxy

You only need to download Galaxy if you plan to:

1. Develop it further
2. Add new tools
3. Plug-in new datasources, or
4. Run a local production server for your site because you have
   1. Sensitive data (e.g., clinical)
   2. Large datasets or processing requirements that are too big to be processed on Main

Installation Procedure

The installation procedure is simple and is nearly identical for UNIX/Linux and Mac OS X. We are no longer supporting the Windows platform with...
Benefits of local installation
- Customize galaxy itself
- Install 3rd party/commercial tools
- Develop your own tools
- Add shared genome builds
- Integrate with instruments
- Sensitive or proprietary data
- Caching large datasets
# Base Software

<table>
<thead>
<tr>
<th>Software</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td>Version 2.6 or 2.7</td>
</tr>
<tr>
<td>Galaxy</td>
<td>Most recent version</td>
</tr>
<tr>
<td>Web Server</td>
<td>To enable web hosting</td>
</tr>
<tr>
<td>Database</td>
<td>Recommended for robust performance</td>
</tr>
<tr>
<td>Analysis Tools</td>
<td>Need to follow install directions for each individual tools</td>
</tr>
</tbody>
</table>

• Basic install of Galaxy

• `% hg clone https://bitbucket.org/galaxy/galaxy-dist`

• `% sh run.sh`

• `http://localhost:8080`
• Install tools from toolshed

• Global config file: universe_wsgi.ini
  • Add admin users
  • Configure all galaxy settings

• Tool behavior: tool_conf.xml
Galaxy for NGS requires additional tools

http://wiki.galaxyproject.org/Admin/Tools/Tool%20Dependencies

Set up reference genomes or fetch indexes

Fabric scripts provided with CloudMan
Production Instance of Galaxy

Get Galaxy: Galaxy Download and Installation

In addition to using the public Galaxy server (a.k.a. Main), you can also install your own instance of Galaxy (what this page is about), or create a cloud-based instance of Galaxy. Another option is to use one of the ever-increasing number of Public Galaxy Servers hosted by other organizations.

See Big Picture: Choices for help on deciding which of these options may be best for your situation.

Reasons to Install Your Own Galaxy

You only need to download Galaxy if you plan to:

1. Develop it further
2. Add new tools
3. Plug-in new datasources, or
4. Run a local production server for your site because you have
   1. Sensitive data (e.g., clinical)
   2. Large datasets or processing requirements that are too big to be processed on Main

Installation Procedure

The installation procedure is simple and is nearly identical for UNIX/Linux and Mac OS X. We are no longer supporting the Windows platform with...
Benefits of a Production Install

• Scalability
  • Handle more users (>5)
  • Run more jobs (>8 concurrent)
  • Large datasets (>3TB)

• Efficiency
  • Schedule jobs
  • Optimize runs
  • Manage data
### Recommended Hardware requirements

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Cores</td>
<td>32-64 processors</td>
</tr>
<tr>
<td>Form Factor</td>
<td>Rack mount server with a capable RAID (6, 6+0)</td>
</tr>
<tr>
<td>RAM</td>
<td>256-512 GB</td>
</tr>
<tr>
<td>Storage Amount</td>
<td>10TB</td>
</tr>
<tr>
<td>Storage Type</td>
<td>DAS SATA/SAS</td>
</tr>
</tbody>
</table>
N-TIER Architecture

- End User
- Apache
- Galaxy (Python, WSGI)
- Postgres
- Host Server
• Turn off developer settings
  • Memory profiling, debug mode

• Switch to a database server
  • SQLite → PostgreSQL

• Use a proxy server
  • Apache/nginx
  • Handles static files, uploads, downloads
  • External authentication
Galaxy Best Practices

- Load balance the Galaxy application
- Single file upload can stall the web app
  - Python GIL
- Split into multiple web server processes and job handler processes
- Use the proxy server as front-end
- Add service control scripts
• Use a compute cluster
  • DRMAA – Distributed Resource Management Application API
  • Works with Grid Engine, PBS, Slurm, LSF
  • Run multiple queues, enforce project quotas
  • Target specific resources

• Cleaning up datasets
  • Galaxy stores everything!
  • Automatically purge histories and datasets
Minimizing redundant storage
  • Filesystem compression
  • Deleting datasets/histories

Data Transfer
  • FTP, HTTP, SCP are slow
  • Consider Aspera, Globus, or UDP-based transfers
• Dev, Test, Production systems
  • Automated testing
    • Continuous integration
  • Leverage toolshed
    • Version controlled
    • Dependency injection
    • Test framework
• DEDICATED RESOURCE!!!
  • ½ FTE to support a large production installation
SlipStream Galaxy Appliance
Lowering the barriers

Powerful dedicated desktop server pre-configured with a fully operational production instance of Galaxy
### Hardware Specifications

<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>2x Intel® Xeon® Processor E5-2690, 8-core (16 cores total)</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>24x 16 GB RDIMM (384 GB)</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>7x 3TB SAS 6 Gbps HDD (16 TB usable) 1x 100GB SSD</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>Dual Gigabit network adaptor</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>Dual redundant power supplies</td>
</tr>
<tr>
<td>FEATURES</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Optimized Galaxy</td>
<td>Production configuration, optimized data transfer</td>
</tr>
<tr>
<td>Analysis Tools</td>
<td>All open-source tools available in Main</td>
</tr>
<tr>
<td>Automated Updates</td>
<td>All software can be updated automatically</td>
</tr>
<tr>
<td>Preinstalled Datasets</td>
<td>5 model organisms (additional upon request)</td>
</tr>
<tr>
<td>Grid Compute</td>
<td>Grid Engine -based job management</td>
</tr>
<tr>
<td>Data Back-up</td>
<td>Copy contents of Appliance locally or to AWS</td>
</tr>
<tr>
<td>Hardware Maintenance</td>
<td>Warranty includes maintenance</td>
</tr>
<tr>
<td>Open Platform</td>
<td>Can ben used as your own powerful server</td>
</tr>
</tbody>
</table>

**Price:** $19,995 (USD)
• **EARLY ACCESS PROGRAM (Limited Availability)**
  • Seamless Adoption
  • Dedicated Support
  • Workflow Generation

• **Early Development Partner Feedback**
  • “A device that centralizes functions with respect to data archives, storage, and analysis is a tremendous aid.” – Ed DeLong, MIT
SlipStream Galaxy

• Become an Early Access Partner Today!!

• Web: www.bioteam.net/slipstream/galaxy-edition