Agenda

- Who is BioTeam
- Why Local Galaxy
- Running Galaxy Locally
- Local Galaxy Best Practices
Who is BioTeam
Who We Are

Over a decade of Life Sciences IT consulting

• We are 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
• BioTeam
Encapsulate IT best-practices expertise to eliminate redundant effort spent designing and building infrastructure

• Galaxy Project
Decrease the barrier to entry into data analysis by improving accessibility of the Galaxy platform
• BioTeam offers a all-in-one solution to help run a local instance Galaxy

• BioTeam is the official appliance provider for Galaxy
  • Exclusive partnership with the Galaxy Team
  • Donations back to the Galaxy Project
## FEATURES

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Powerful Server</td>
<td>Complete analysis tasks quickly</td>
</tr>
<tr>
<td>Optimized Galaxy</td>
<td>Configured for stable production use</td>
</tr>
<tr>
<td>Analysis Tools</td>
<td>Provides comprehensive set of open-source tools</td>
</tr>
<tr>
<td>Automated Updates</td>
<td>Software can be updated automatically</td>
</tr>
<tr>
<td>Preinstalled Datasets</td>
<td>Includes 5 model organisms (additional upon request)</td>
</tr>
<tr>
<td>Open Platform</td>
<td>Install other software for your own use</td>
</tr>
</tbody>
</table>

**Price:** $19,995 (USD)
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>Condition</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>
• 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 scalable option but has its challenges
  - Understand how to manage pricing
  - Data transfer bottlenecks
  - Familiarity with cloud (e.g. Amazon Web Services)
  - Dealing with sensitive data

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...
• Customizing a local installation
  • Customize galaxy itself
  • Install 3rd party/commercial tools
  • Develop your own tools
  • Add shared genome builds
  • Integrate with instruments
  • Manage sensitive or proprietary data
<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>To store histories, users, track datasets, etc.</td>
</tr>
<tr>
<td>Analysis Tools</td>
<td>Need to follow install directions for each individual tools</td>
</tr>
</tbody>
</table>

http://wiki.galaxyproject.org/Admin/GetGalaxy
• Basic install of Galaxy

% hg clone https://bitbucket.org/galaxy/galaxy-dist
% sh run.sh

• In your browser type: http://localhost:8080
• Install tools from toolshed

• Global config file: universe_wsgi.ini
  • Over 100 configurable parameters
  • Add admin users
  • Configure all galaxy settings

• Tool behavior
  • tool_conf.xml
  • Individual tool wrappers (also xml files)
Galaxy for NGS requires additional tools


- Set up reference genomes or fetch indexes
• Scalability
  • Handle more users (>5)
  • Run more jobs (>8 concurrent)
  • Support large datasets (>3TB)

• Efficiency
  • Schedule jobs
  • Optimize runs
  • Manage data
<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>RAM</td>
<td>256-512 GB</td>
</tr>
<tr>
<td>Storage Amount</td>
<td>&gt;10TB</td>
</tr>
</tbody>
</table>
Best Practices

Galaxy Software Development Best Practices

Lists software development best practices for the Galaxy Project. These are works in progress and practices vary in how broadly they have been applied.

Contents

1. Metastandards
2. Python Standards
3. JavaScript Standards
   1. Backbone
4. Email Threads
Galaxy Best Practices

- Turn off developer settings
  - Speed up the galaxy server

- Switch databases (e.g. PostgreSQL)
  - handle multiple database requests at once

- Switch web servers (e.g. Apache, nginx)
  - Handle file transfers, external authentication
• Minimize redundant storage
  • Galaxy stores everything!
  • Filesystem compression
  • Deleting datasets and histories

• Improve Data Transfer
  • FTP, HTTP, SCP are slow
  • Consider data transfer products (e.g. Aspera, Globus)
• Dev, Test, Production environment

• Automated testing

• Leverage toolshed

• DEDICATED RESOURCE!!
  • ½ FTE to support a large production installation
• Galaxy is a fantastic public resource

• Local Galaxy offers a high level of control over your analysis environment

• There is tremendous value in being aware of the underlying components of Galaxy

• The flexibility makes it complex to manage so plan accordingly
Thank You

Contact BioTeam to discuss your Galaxy needs

• Website: www.bioteam.net
• Email: anushka@bioteam.net
EXTRA SLIDES
SlipStream Galaxy Appliance
Lowering the barriers

Galaxy made easy.

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>2x Intel® Xeon® Processor E5-2690, 8-core (16 cores total)</td>
</tr>
<tr>
<td>Memory</td>
<td>24x 16 GB RDIMM (384 GB)</td>
</tr>
<tr>
<td>Storage</td>
<td>7x 3TB SAS 6 Gbps HDD (16 TB usable)</td>
</tr>
<tr>
<td></td>
<td>1x 100GB SSD</td>
</tr>
<tr>
<td>Network</td>
<td>Dual Gigabit network adaptor</td>
</tr>
<tr>
<td>Power</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>Comprehensive set of open-source tools</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>Hardware Maintenance</td>
<td>Warranty includes maintenance</td>
</tr>
<tr>
<td>Open Platform</td>
<td>Install other software for your own use</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
• Become an Early Access Partner Today!!

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