Running a Local Galaxy

Anushka Brownley
BioTeam Inc.

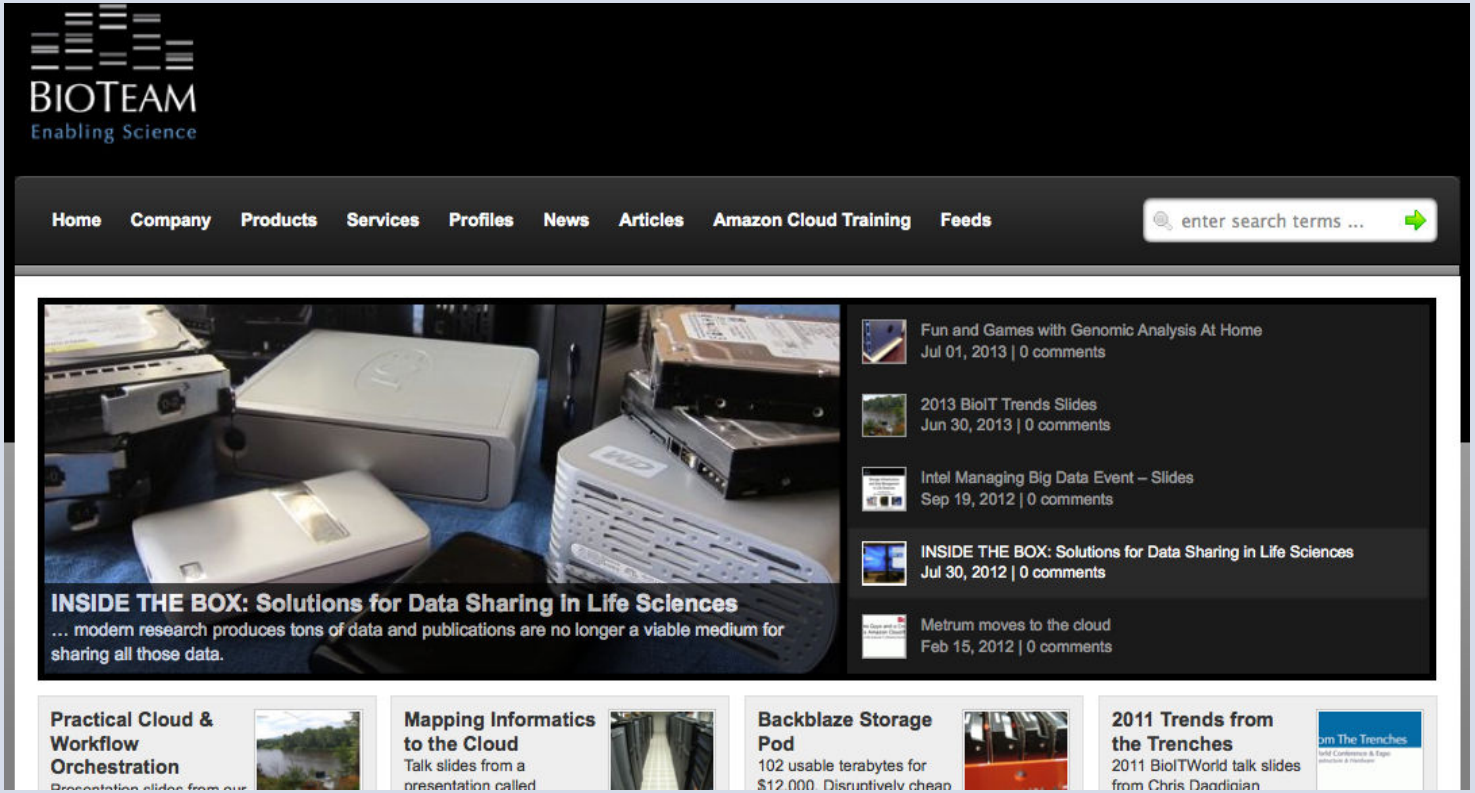
BIOTEAM
Enabling Science

Agenda

Galaxy

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

Who is BioTeam



BIOTEAM
Enabling Science

Who We Are

Galaxy

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

Who We Are

- Active contributors to open-source projects

BioPerl

O|B|F

BIOTEAM
Enabling Science

Bridging the IT Gap

Galaxy

- **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
Enabling Science

Bridging the IT Gap

Galaxy

- BioTeam offers a all-in-one solution to help run a local instance Galaxy

SLIPSTREAM GALAXY EDITION
APPLIANCE
Galaxy made easy.

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

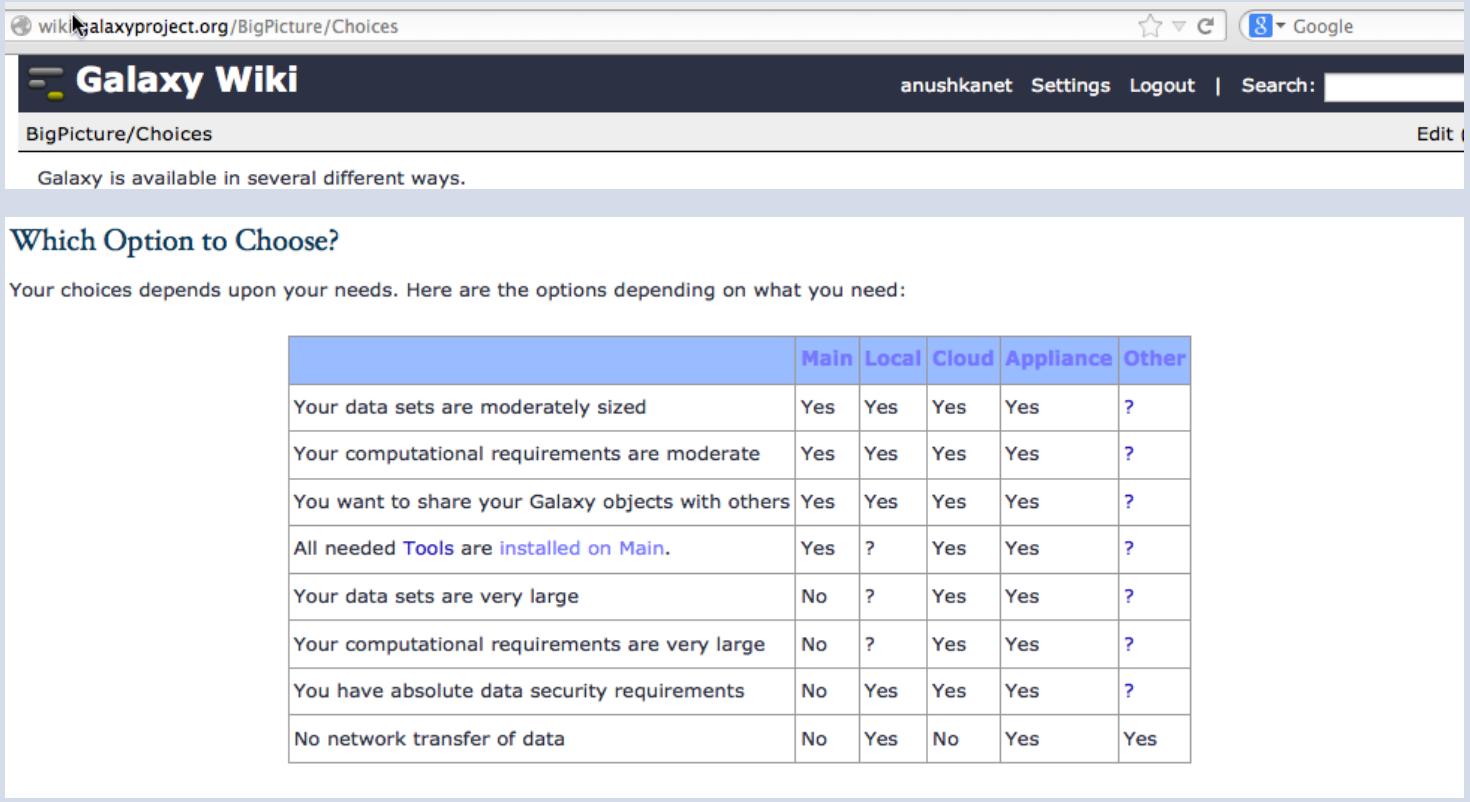
BIOTEAM
Enabling Science

SlipStream Galaxy

Galaxy

FEATURES	
Powerful Server	Complete analysis tasks quickly
Optimized Galaxy	Configured for stable production use
Analysis Tools	Provides comprehensive set of open-source tools
Automated Updates	Software can be updated automatically
Preinstalled Datasets	Includes 5 model organisms (additional upon request)
Open Platform	Install other software for your own use
Price: $19,995 (USD)	

Why a Local Installation



wiki.galaxyproject.org/BigPicture/Choices

Google

Galaxy Wiki

anushkanet Settings Logout | Search:

BigPicture/Choices Edit

Galaxy is available in several different ways.

Which Option to Choose?

Your choices depends upon your needs. Here are the options depending on what you need:

	Main	Local	Cloud	Appliance	Other
Your data sets are moderately sized	Yes	Yes	Yes	Yes	?
Your computational requirements are moderate	Yes	Yes	Yes	Yes	?
You want to share your Galaxy objects with others	Yes	Yes	Yes	Yes	?
All needed Tools are installed on Main.	Yes	?	Yes	Yes	?
Your data sets are very large	No	?	Yes	Yes	?
Your computational requirements are very large	No	?	Yes	Yes	?
You have absolute data security requirements	No	Yes	Yes	Yes	?
No network transfer of data	No	Yes	No	Yes	Yes

BIOTEAM
Enabling Science

Current Challenges

Galaxy

- 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

BIOTEAM
Enabling Science

Current Challenges

Galaxy

- 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

BIOTEAM
Enabling Science

Be Prepared

Galaxy

- 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

BIOTEAM
Enabling Science

Be Prepared

Galaxy

- 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

YOU CONTROL	BENEFIT
Amount of storage	Handle large datasets
Type of hardware	Run compute intensive jobs
What tools to install	Customize to your research
Data access	Granular control of security
Networking architecture	No data transfer bottleneck
Software behavior	Optimize how jobs are run

Local Instance of Galaxy

Galaxy Wiki

anushkanet Settings Logout | Search:

Admin/Get Galaxy Edit (

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.

Contents

1. Reasons to Install Your Own Galaxy
2. Installation Procedure
    1. Check your Python version
    2. Get the latest copy from the repository
    3. Start it up
    4. Join the Mailing List
    5. Keep your instance backed up
    6. Keep your code up to date
3. Advanced Configuration
4. Other Help

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

BIOTEAM
Enabling Science

Customizing Galaxy

Galaxy

- 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

BIOTEAM
Enabling Science

Base Software

Galaxy

Software	Comment
Python	Version 2.6 or 2.7
Galaxy	Most recent version
Web Server	To enable web hosting
Database	To store histories, users, track datasets, etc.
Analysis Tools	Need to follow install directions for each individual tools

http://wiki.galaxyproject.org/Admin/GetGalaxy

BIOTEAM
Enabling Science

Simple Start-up

Galaxy

- Basic install of Galaxy

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

- In your browser type: http://localhost:8080

BIOTEAM
Enabling Science

Additional Configuration

Galaxy

- 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)

BIOTEAM
Enabling Science

Galaxy for NGS

Galaxy

- Galaxy for NGS requires additional tools
- http://wiki.galaxyproject.org/Admin/Tools/Tool%20Dependencies
- Set up reference genomes or fetch indexes

BIOTEAM
Enabling Science

Production Instance of Galaxy

Galaxy

- Scalability
  - Handle more users (>5)
  - Run more jobs (>8 concurrent)
  - Support large datasets (>3TB)
- Efficiency
  - Schedule jobs
  - Optimize runs
  - Manage data

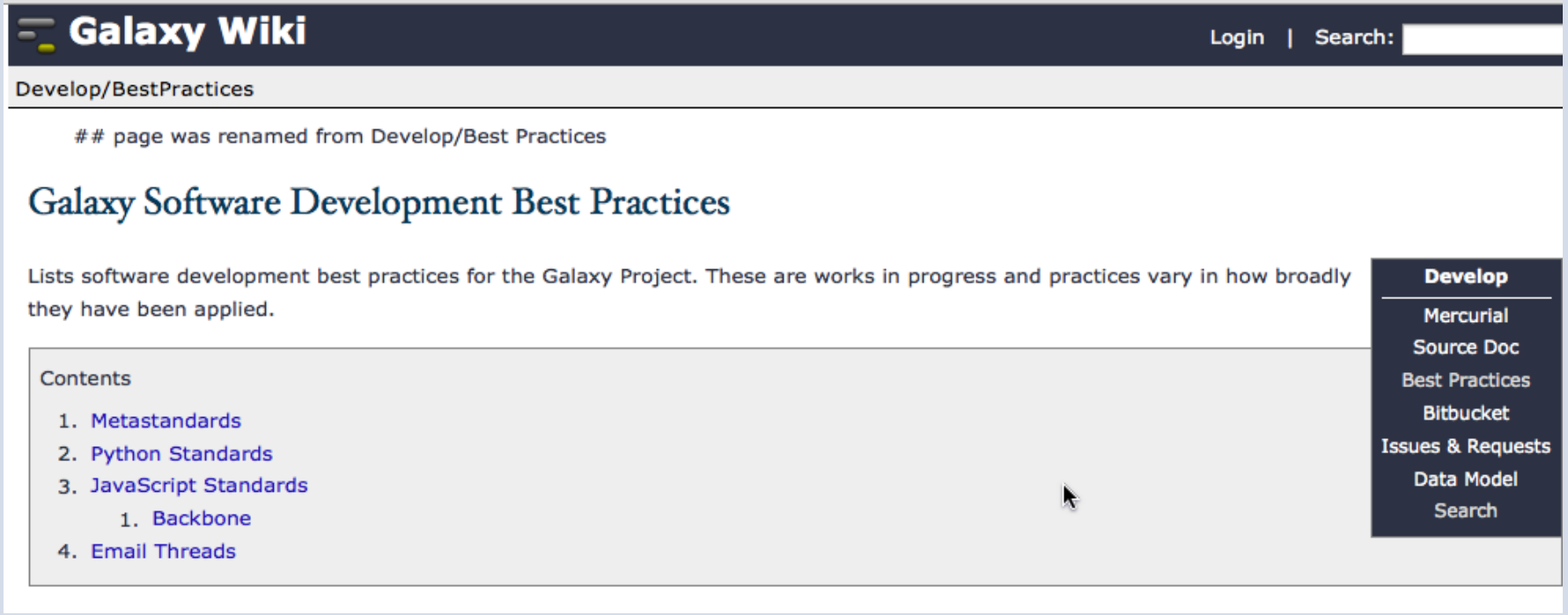
BIOTEAM
Enabling Science

Production Instance of Galaxy

Galaxy

Hardware	Recommendation
# of Cores	32-64 processors
RAM	256-512 GB
Storage Amount	>10TB

Best Practices



BIOTEAM
Enabling Science

Galaxy Best Practices

Galaxy

- 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

BIOTEAM
Enabling Science

Galaxy Best Practices

Galaxy

- 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)

BIOTEAM
Enabling Science

Galaxy Best Practices

Galaxy

- Dev, Test, Production environment
- Automated testing
- Leverage toolshed
- DEDICATED RESOURCE!!!
  - ½ FTE to support a large production installation

BIOTEAM
Enabling Science

Take Away

Galaxy

- 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

SLIPSTREAM GALAXY EDITION
APPLIANCE
Galaxy made easy.

EXTRA SLIDES

SlipStream Galaxy Appliance

BIOTEAM
Enabling Science

Lowering the barriers

Galaxy

SLIPSTREAM GALAXY EDITION
APPLIANCE
Galaxy made easy.

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

BIOTEAM
Enabling Science

SlipStream Galaxy

Galaxy

Hardware Specifications	
CPU	2x Intel® Xeon® Processor E5-2690, 8-core (16 cores total)
Memory	24x 16 GB RDIMM (384 GB)
Storage	7x 3TB SAS 6 Gbps HDD (16 TB usable) 1x 100GB SSD
Network	Dual Gigabit network adaptor
Power	Dual redundant power supplies

BIOTEAM
Enabling Science

SlipStream Galaxy

Galaxy

FEATURES	
Optimized Galaxy	Production configuration, optimized data transfer
Analysis Tools	Comprehensive set of open-source tools
Automated Updates	All software can be updated automatically
Preinstalled Datasets	5 model organisms (additional upon request)
Grid Compute	Grid Engine -based job management
Hardware Maintenance	Warranty includes maintenance
Open Platform	Install other software for your own use
Price: $19,995 (USD)	

BIOTEAM
Enabling Science

Partnering with Thought Leaders

Galaxy

- **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