Managing a local Galaxy Instance

Anushka Brownley / Adam Kraut
BioTeam Inc.

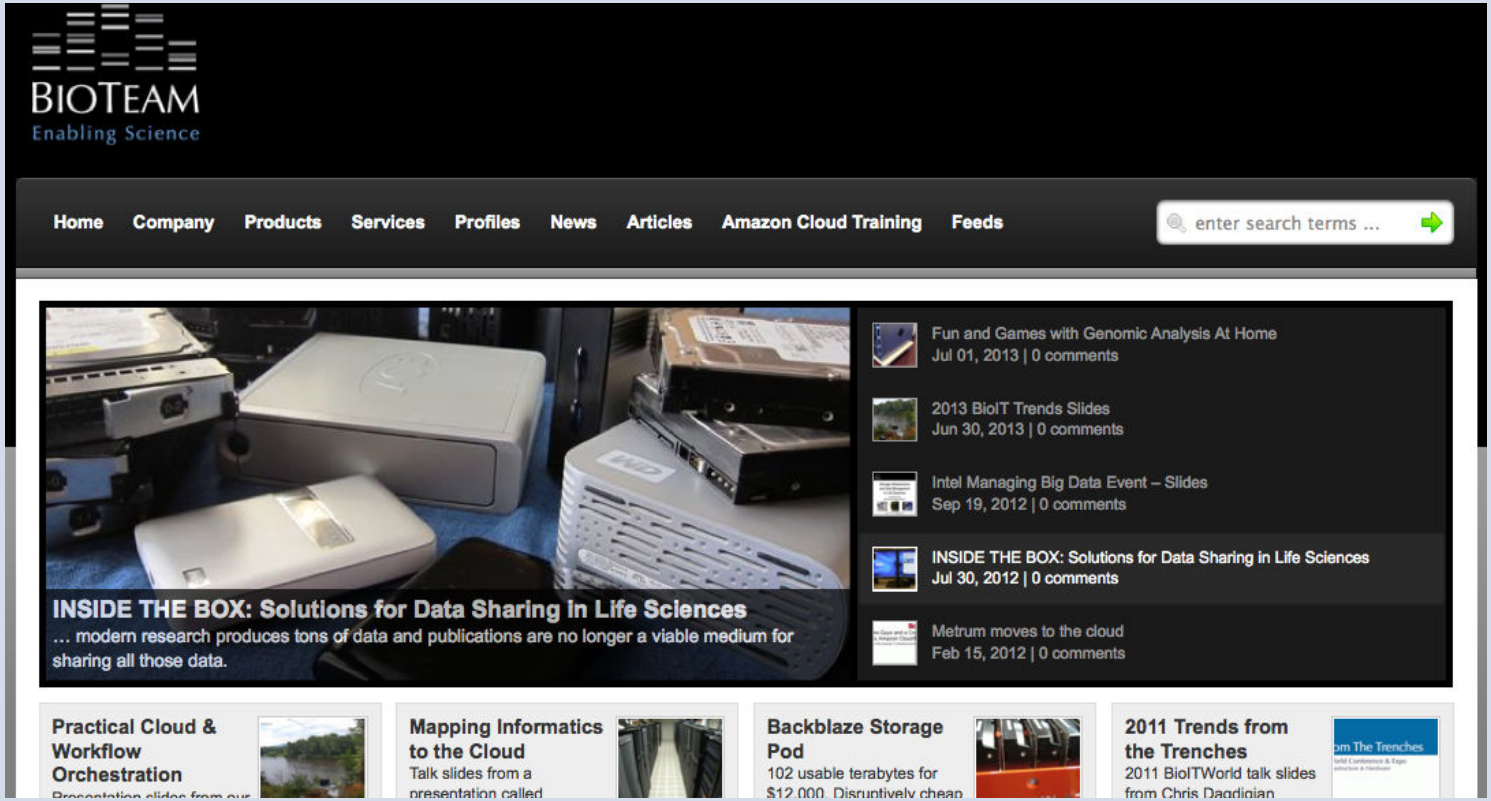
BIOTEAM
Enabling Science

Agenda

Galaxy

- Who are we
- Why a local installation
- Local infrastructure
- Local installation
- Tips and Tricks
- SlipStream Appliance

WHO ARE WE


BIOTEAM
Enabling Science
Home
Company
Products
Services
Profiles
News
Articles
Amazon Cloud Training
Feeds
enter search terms ...
INSIDE THE BOX: Solutions for Data Sharing in Life Sciences
... modern research produces tons of data and publications are no longer a viable medium for sharing all those data.
Fun and Games with Genomic Analysis At Home
Jul 01, 2013 | 0 comments
2013 BioIT Trends Slides
Jun 30, 2013 | 0 comments
Intel Managing Big Data Event – Slides
Sep 19, 2012 | 0 comments
INSIDE THE BOX: Solutions for Data Sharing in Life Sciences
Jul 30, 2012 | 0 comments
Metrum moves to the cloud
Feb 15, 2012 | 0 comments
Practical Cloud & Workflow Orchestration
Mapping Informatics to the Cloud
Talk slides from a
Backblaze Storage Pod
102 usable terabytes for
2011 Trends from the Trenches
2011 BioITWorld talk slides

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

Who We are

- Active contributors to open-source projects

emboss

GMOD

DIYA

BioPerl

Homebrew
The missing package manager for OS X

OBF

GRID ENGINE

BIOTEAM
Enabling Science

Bridging the IT gap

Galaxy

- **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
Enabling Science

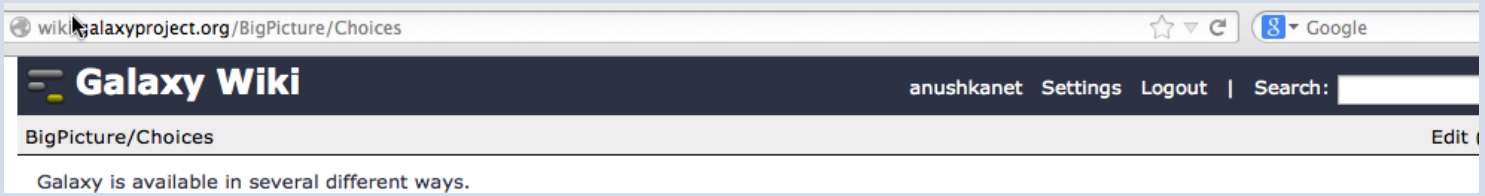
Bridging the IT gap

Galaxy

- 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



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

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

BIOTEAM
Enabling Science

Current challenges

Galaxy

- 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**

Be prepared

- 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

Be prepared

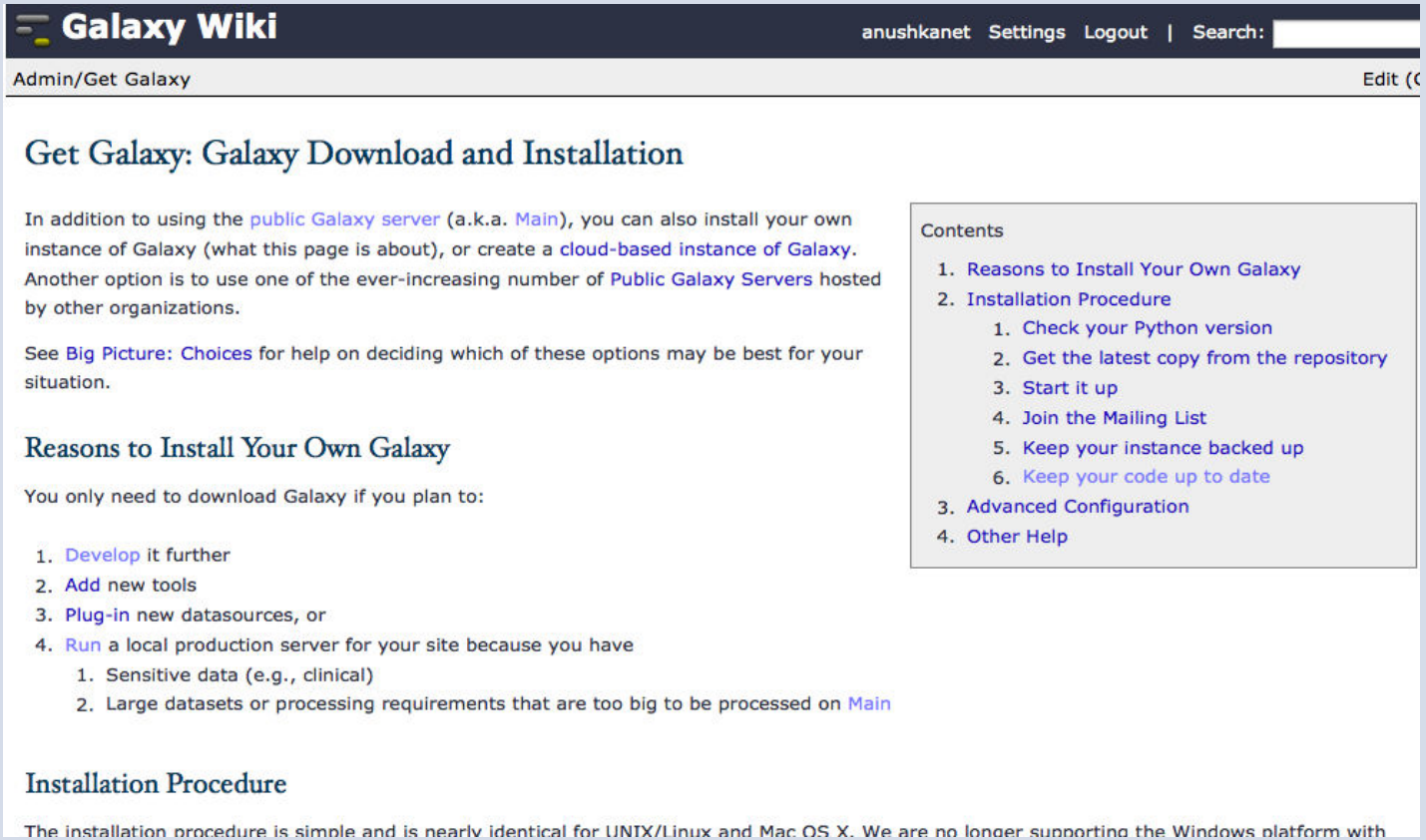
- Informatics support
  - Handle user questions/requests
  - Gather user feedback
- Ongoing dedicated resource
  - Manage updates
  - Facilitate user support
  - Maintain infrastructure

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

Customizing galaxy

- 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

Software	Comment
Python	Version 2.6 or 2.7
Galaxy	Most recent version
Web Server	To enable web hosting
Database	Recommended for robust performance
Analysis Tools	Need to follow install directions for each individual tools

http://wiki.galaxyproject.org/Admin/Get%20Galaxy

BIOTEAM
Enabling Science

Simple Start-up Demo

Galaxy

- Basic install of Galaxy
- % hg clone https://bitbucket.org/galaxy/galaxy-dist
- % sh run.sh
- http://localhost:8080

BIOTEAM
Enabling Science

Additional Configuration

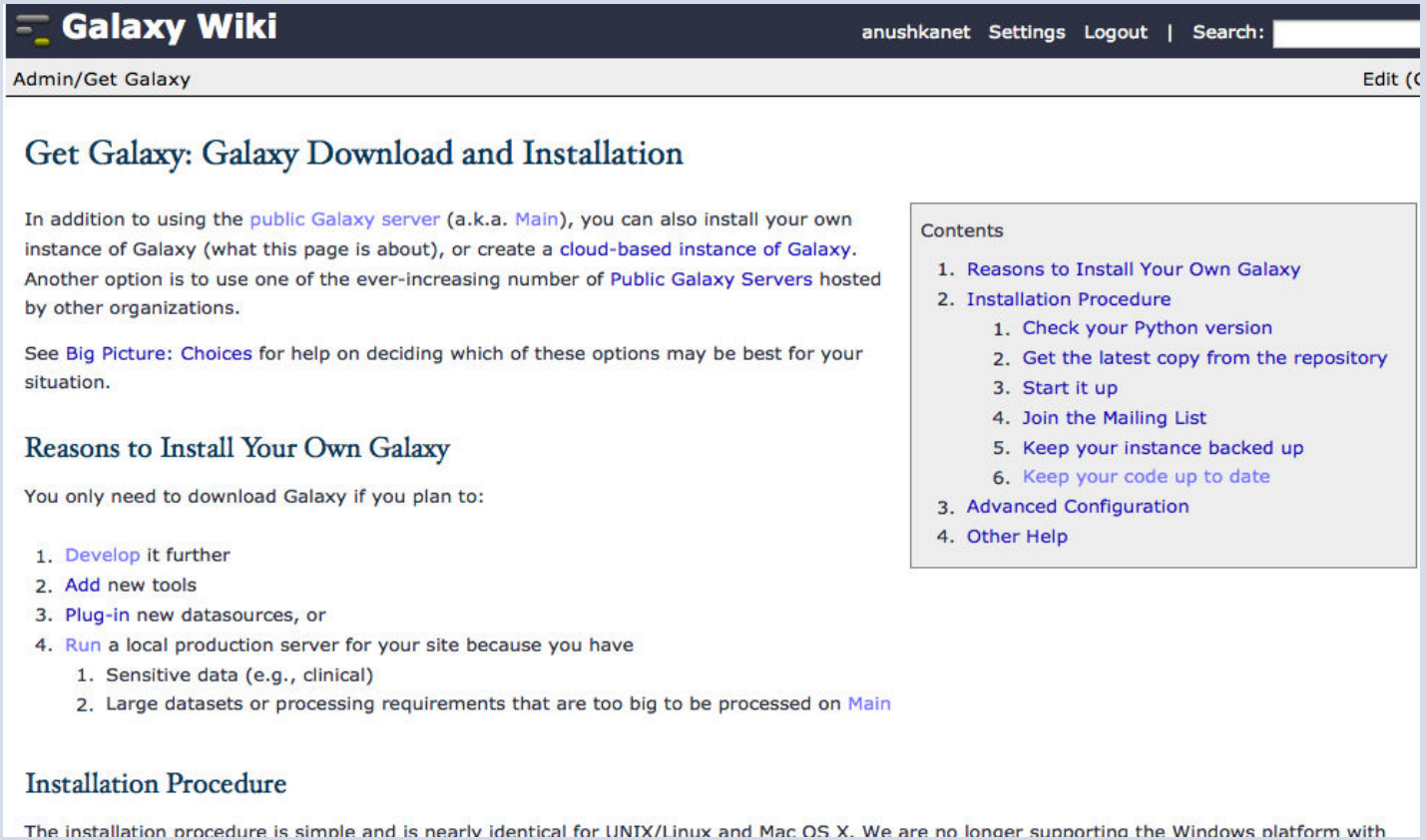
Galaxy

- 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

- 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



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

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

BIOTEAM
Enabling Science

Recommended Hardware requirements

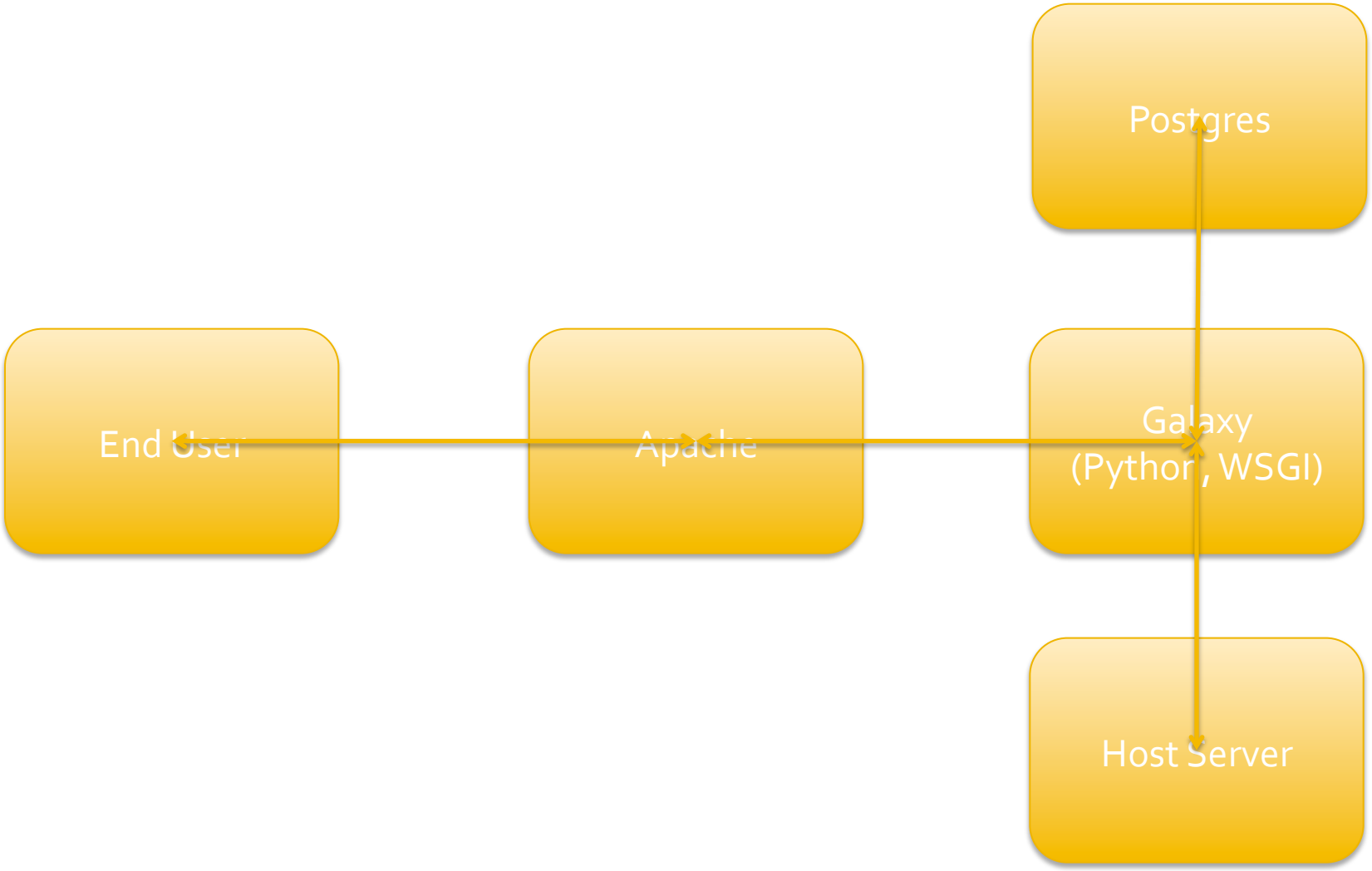
Galaxy

Hardware	Recommendation
# of Cores	32-64 processors
Form Factor	Rack mount server with a capable RAID (6, 6+0)
RAM	256-512 GB
Storage Amount	10TB
Storage Type	DAS SATA/SAS

BIOTEAM
Enabling Science

N-TIER Architecture

Galaxy



Galaxy Best Practices

- 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

BIOTEAM
Enabling Science

Galaxy Best Practices

Galaxy

- 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

Galaxy Best practices

- 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

Efficient Data Management

- Minimizing redundant storage
  - Filesystem compression
  - Deleting datasets/histories
- Data Transfer
  - FTP, HTTP, SCP are slow
  - Consider Aspera, Globus, or UDP-based transfers

Environments

- 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

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

SlipStream 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	All open-source tools available in Main
Automated Updates	All software can be updated automatically
Preinstalled Datasets	5 model organisms (additional upon request)
Grid Compute	Grid Engine -based job management
Data Back-up	Copy contents of Appliance locally or to AWS
Hardware Maintenance	Warranty includes maintenance
Open Platform	Can ben used as your own powerful server
Price: $19,995 (USD)	

Partnering with Thought Leaders

- **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