Setting up a Galaxy Instance as a Service

Nikolay Vazov
Jochen Bick
Hans-Rudolf Hotz

You are given the task to set up a Galaxy instance for others (i.e. as a core service in your institute) and you are not really familiar with Galaxy.
outline

Introduction

first steps in setting up a server

Apache and PostgreSQL

configure Galaxy for a production environment

add a tool

extras
Introduction

first steps in setting up a server

Apache and PostgreSQL

configure Galaxy for a production environment

add a tool

extras

will be demonstrated using the provided VM
What this training session is not about:

- how to use galaxy

- writing tool wrappers
who we are

Hans-Rudolf Hotz
Jochen Bick
Nikolay Vazov
'10 rules' for Setting up a Galaxy Instance as a Service
10 rules for setting up a galaxy instance as a service

There is no such thing as 'Free Lunch'
Check: what are you actually asked for
Check: what resources do you have / need
Start small, but right
Set up only what you have been asked to
Know the tools you offer
Set up 'reports'
Prevent data duplication
Offer training
Keep the Galaxy software (and you) up to date
1) There is no such thing as 'Free Lunch'
1) There is no such thing as 'Free Lunch'

Time invested in setting up and running Galaxy
1) There is no such thing as 'Free Lunch'

Number of Bioinformatics 'Helpdesk' jobs you can delegate to Galaxy

Time invested in setting up and running Galaxy
1) There is no such thing as 'Free Lunch'
2) Check: what are you actually asked for
2) Check: what are you actually asked for

- talk to the person(s) who contacted you
  - why Galaxy?
2) Check: what are you actually asked for

- talk to the people who will use your service

  - using Galaxy for what?

  - do they know use.galaxy.org?

  - alternatives?
2) Check: what are you actually asked for

- talk to the people who will use your service

  - using Galaxy for what?

  - do they know use.galaxy.org?

  - alternatives?

  define the tools
2) Check: what are you actually asked for

- talk to the people who will use your service
  - using Galaxy for what?
  - do they know use.galaxy.org?
  - alternatives?

define the tools

use the toolshed

develop your own tools
2) Check: what are you actually asked for

- talk to the people who will use your service

  - using Galaxy for what?

  - do they know use.galaxy.org?

  - alternatives?

define the tools

use the toolshed

develop your own tools
2) Check: what are you actually asked for

Visibility

- internal web site
- public web site

Access

- everybody can create an account
- accounts are created for the users
- external authentication
3) Check: what resources do you need?
3) Check: what resources do you need?

...to run the tools you want to offer to the expected number of users:
3) Check: what resources do you need?

...to run the tools you want to offer to the expected number of users:

- hardware
  - cpu / memory
- storage
  - fast (local) storage
  - slow (network) storage
3) Check: what resources do you need?

...to run the tools you want to offer to the expected number of users:

- people / knowledge
  - system administration for the Galaxy server
- Bioinformatics background
3) Check: what resources do you need?

make sure you have a 'Galaxy Admin' person

https://wiki.galaxyproject.org/Admin/Interface

'Galaxy Admin':
- does not need to be a 'Sys Admin'
- write access to Galaxy code
- can restart Galaxy server
- write access to Galaxy database
4) Start small, but right
4) Start small, but right

follow the suggestions on the wiki

https://wiki.galaxyproject.org/Admin

https://wiki.galaxyproject.org/Admin/Config/Performance/ProductionServer
4) Start small, but right

follow the suggestions on the wiki

https://wiki.galaxyproject.org/Admin

https://wiki.galaxyproject.org/Admin/Config/Performance/ProductionServer

switching to a database server

SQLite  ➡️ PostgreSQL
4) Start small, but right

check other installations:
https://wiki.galaxyproject.org/Community/Deployments

Galaxy Community Log Board:
(a place to share how you addressed a particular task)

https://wiki.galaxyproject.org/Community/Logs
4) Start small, but right

you don't need a cluster to set up different queues
4) Start small, but right

you don't need a cluster to set up different queues

job.conf.xml

job_conf.xml.sample_advanced

https://wiki.galaxyproject.org/Admin/Config/Jobs
4) Start small, but right

you can change the hardware
4) Start small, but right

you can change the hardware, as long as you keep the ‘database/’ directory and the SQL DB in sync.
5) Set up only what you have been asked to
5) Set up only what you have been asked to

.....at least in the beginning:
don't confuse your clients with too many tools
5) Set up only what you have been asked to

.....at least in the beginning: don't confuse your clients with too many tools

offer group/user specific tools

https://wiki.galaxyproject.org/UserDefinedToolboxFilters
https://wiki.galaxyproject.org/Admin/Config/Access%20Control
5) Set up only what you have been asked to

.....at least in the beginning:
don't confuse your clients with too many tools

offer group/user specific tools

https://wiki.galaxyproject.org/UserDefinedToolboxFilters
https://wiki.galaxyproject.org/Admin/Config/Access%20Control

good for testing
5) Set up only what you have been asked to

.....but make sure you have:

production server / development server
5) Set up only what you have been asked to

.....but make sure you have:

production server / development server

back up of production server (including the PostgreSQL DB)
6) Know the tools you offer
6) Know the tools you offer

First, make sure you know how to use galaxy

https://wiki.galaxyproject.org/Learn
6) Know the tools you offer

First, make sure you know how to use galaxy

https://wiki.galaxyproject.org/Learn

Second, understand the tools you offer

- can you execute them on the command line
7) Set up 'reports'
7) Set up 'reports'

https://wiki.galaxyproject.org/Admin/UsageReports
7) Set up 'reports'

it is a second web site
7) Set up 'reports'

Today's jobs

All unfinished jobs

Jobs per tool

Jobs per user

User disk usage
7) Set up 'reports'

....but nevertheless, get familiar with the database
7) Set up 'reports' 

....but nevertheless, get familiar with the database 

- execute queries which are not covered by 'reports'
7) Set up 'reports'

....but nevertheless, get familiar with the database

- execute queries which are not covered by 'reports'

- fix the database
7) Set up 'reports' 

....but nevertheless, get familiar with the database 

- execute queries which are not covered by 'reports'

- fix the database
8) Prevent data duplication
8) Prevent data duplication - user data

~/.tools/fasta_tools
  fmi_tools
  ...

Galaxy
8) Prevent data duplication - user data

~/.tools/fasta_tools
  fmi_tools
  ...
~/.database/files/001/dataset_1000.dat
  dataset_1001.dat
  ...
  002/dataset_2000.dat
  ...

Galaxy
8) Prevent data duplication - user data

```
~/.tools/fasta_tools
    fmi_tools
    ...
~/.database/files/001/dataset_1000.dat
    dataset_1001.dat
    ...
002/dataset_2000.dat
    ...
```

each History item is a 'dataset file'
8) Prevent data duplication - user data

`~/tools/fasta_tools`
`  fmi_tools`
`  ...
`~/database/files/001/dataset_1000.dat`
`  dataset_1001.dat`
`  ...
`  002/dataset_2000.dat`
`  ...

*each History item is a 'dataset file'*

consider user quotas
8) Prevent data duplication - user data

use 'Data Libraries'

https://wiki.galaxyproject.org/Admin/DataLibraries
8) Prevent data duplication - user data

use 'Data Libraries'

https://wiki.galaxyproject.org/Admin/DataLibraries

- 'Link to files without copying into Galaxy, 
- enable 'Upload files from filesystem paths'

https://wiki.galaxyproject.org/Admin/DataLibraries/UploadingLibraryFiles
8) Prevent data duplication - user data

promote history sharing

promote Galaxy 'pages'
8) Prevent data duplication - user data

promote history sharing

promote Galaxy 'pages'

allow user to see the full path of datasets
(expose_dataset_path = True)
8) Prevent data duplication - reference data
8) Prevent data duplication - reference data

tools (like bowtie, BLAST, etc) rely on '.loc' files
8) Prevent data duplication - reference data

tools (like bowtie, BLAST, etc) rely on '.loc' files

'Data Managers'
available from the toolshed
will download/index the data
will adjust the '.loc' files
8) Prevent data duplication - reference data

tools (like bowtie, BLAST, etc) rely on '.loc' files

'Data Managers'

available from the toolshed
will download/index the data
will adjust the '.loc' files

existing data/indices can be used
9) Offer training
9) Offer training

individual training

run training courses

https://wiki.galaxyproject.org/Teach
9) Offer training

individual training

run training courses

https://wiki.galaxyproject.org/Teach

stress testing for the server
10) Keep your server (and you) up to date
10) Keep your server up to date

unless you have a very good reason, make sure you are running the latest (or at least a recent) code version

- it is easier for others to help you
- the reported issue might already be fixed in the current release
10) Keep your server up to date

unless you have a very good reason, make sure your are running the latest (or at least a recent) code version

- it is easier for others to help you
- the reported issue might already be fixed in the current release

find a balance between updates (with new or different features) and continuity

we do 3 update per year
10) Keep your server up to date

doing an update is easy
10) Keep your server up to date

doing an update is easy ....most of the time
10) Keep your server up to date

doing an update is easy ....most of the time

- announce the down time one week in advance
- install a new server
- update the 'new server' from last time
- update the development server
- update the production server
10) Keep your server up to date

doing an update is easy ....most of the time

- announce the down time one week in advance
- install a new server
- update the 'new server' from last time
- update the development server
- update the production server

*goal: minimize the down time of the production server*
10) Keep yourself up to date
10) Keep yourself up to date

- read the DevNewsBriefs
  https://wiki.galaxyproject.org/DevNewsBriefs

- follow the mailing lists
  https://wiki.galaxyproject.org/MailingLists

- join the "Galaxy Admins"
  https://wiki.galaxyproject.org/Community/GalaxyAdmins
  BoF: GalaxyAdmins, Tuesday lunchtime

- follow Twitter
  @galaxyproject  #usegalaxy
10) Keep yourself up to date

- read the DevNewsBriefs
  https://wiki.galaxyproject.org/DevNewsBriefs

- follow the mailing lists
  https://wiki.galaxyproject.org/MailingLists

- galaxyproject IRC Channel
  https://wiki.galaxyproject.org/Support/IRC

- join the "Galaxy Admins"
  https://wiki.galaxyproject.org/Community/GalaxyAdmins
  BoF: GalaxyAdmins, Tuesday lunchtime

- follow Twitter
  @galaxyproject  #usegalaxy

*take the time to look at new features*
10) Keep yourself up to date

- go to "GCC2017"
10) Keep yourself up to date

- go to the next "Swiss/German Galaxy Day"

Swiss German Galaxy Days

Freiburg (Germany) 20/21 October 2016
Galaxy-Cluster integration
Main issues in galaxy – cluster configuration

1. File structure
2. Software configuration
1. File structure

- Galaxy & cluster partitions
- Common files (tool-data & libraries)
### Galaxy & cluster partitions

![Diagram showing Galaxy and cluster partitions]

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>1K-blocks</th>
<th>Used</th>
<th>Available</th>
<th>Use%</th>
<th>Mounted on</th>
</tr>
</thead>
<tbody>
<tr>
<td>/dev/mapper/internvg-root</td>
<td>12319920</td>
<td>5645068</td>
<td>6045836</td>
<td>49%</td>
<td>/</td>
</tr>
<tr>
<td>/dev/sda1</td>
<td>245679</td>
<td>85344</td>
<td>147228</td>
<td>37%</td>
<td>/boot</td>
</tr>
<tr>
<td>/dev/mapper/internvg-opt</td>
<td>999320</td>
<td>409484</td>
<td>537408</td>
<td>44%</td>
<td>/opt</td>
</tr>
<tr>
<td>/dev/mapper/internvg-tmp</td>
<td>999320</td>
<td>4456</td>
<td>942436</td>
<td>1%</td>
<td>/tmp</td>
</tr>
<tr>
<td>/dev/mapper/internvg-usr</td>
<td>3997376</td>
<td>2040116</td>
<td>1747548</td>
<td>54%</td>
<td>/usr</td>
</tr>
<tr>
<td>/dev/mapper/internvg-var</td>
<td>3997376</td>
<td>1219144</td>
<td>2568520</td>
<td>33%</td>
<td>/var</td>
</tr>
<tr>
<td>fghfs_nodev</td>
<td>298765170688</td>
<td>157997927936</td>
<td>140767242752</td>
<td>53%</td>
<td>/cluster</td>
</tr>
<tr>
<td>fghfs_nodev</td>
<td>363204909056</td>
<td>222982917120</td>
<td>140221991936</td>
<td>62%</td>
<td>/work</td>
</tr>
</tbody>
</table>

[Check df command output for more details]
Galaxy files on the cluster

As the Galaxy jobs run on the cluster, some setup on the cluster is required.

The files live in

/cluster/software/galaxy

/work/projects/galaxy/database
Galaxy files on the cluster

All Galaxy data must be on a cluster partition:

```
lrwxrwxrwx 1 galaxy galaxy 49 Jan 6 11:05 ./database -> /work/projects/galaxy/data/database_galaxy_prod01
```
Galaxy files on the cluster

Galaxy python libraries

Place the whole

.venv/lib64
.venv/lib

tree in

/cluster/software/galaxy/.venv directory.
Galaxy files on the cluster

Scripts called by tools XML wrappers

Galaxy tools developed by the Galaxy team are distributed in the same directory as the tool xml files.

For cluster execution, these scripts have to live on the cluster.

The /cluster/software/galaxy/tools directory holds the tools scripts.
2. Software configuration

- Galaxy code modifications
- Third-party libraries (SLURM-DRMAA)
- Slurm binaries / munge
Information flow in cluster job submission

Submit host

Controller

Cluster SLURM scheduler

Towards compute nodes
Define plugin

```xml
<plugins>
  <plugin id="slurm" type="runner"
    load="galaxy.jobs.runners.slurm:SlurmJobRunner"/>
  <plugin id="local" type="runner"
    load="galaxy.jobs.runners.local:LocalJobRunner"/>
</plugins>
```

Define handler

```xml
<handlers>
  <handler id="handler0"/>
  <handler id="handler1"/>
</handlers>
```

Define destination

```xml
<destinations default="slurm">
  <destination id="local" runner="local"/>
  <destination id="slurm" runner="slurm">
    <param id="nativeSpecification">-A nn9108k -t 05:00 --mem-per-cpu=4G -D dummy -J dummy -n 1</param>
  </destination>
</destinations>
```

Define tool configuration

```xml
<tools>
  <tool id="hello_cluster" destination="slurm" resources="hello_cluster"/>
  <tool id="upload1" destination="local" resources="default"/>
</tools>
```

Define resources

```xml
<resources default="default">
  <group id="default">processors, memory, time, project</group>
  <group id="hello_cluster">processors_threads, memory, time, project</group>
</resources>
```
<parameters>

  <param label="Processors" name="processors" type="integer" size="2" min="1" max="16" />

  <param label="Memory" name="memory" type="integer" size="3" min="1" max="900" />

  <param label="Time" name="time" type="integer" size="3" min="1" max="672" />

  <param label="Project" name="project" type="select" >
    <option value="gx_default">Default Geoportal 200hrs project</option>
    <option value="lp21">lp21</option>
    ...
    <option value="staff">staff</option>
  </param>

  <param label="Processors (threads)" name="processors_threads" type="integer" size="2" min="1" max="16" />

  <param label="Processors (processes)" name="processors_processes" type="integer" size="2" min="1" max="16" />

</parameters>
Setting job parameters at job runtime

(screenshot)

Abel handshake (Galaxy Tool Version 1.0.0)

Remaining CPU hrs in your default project gx_default are: undefined

Job Resource Parameters

Specify job resource parameters

Processors (threads)

Number of processing cores, 'ppn' value (1–16). Leave blank to use default value.

Memory

Memory size in gigabytes, 'pmem' value (1–900). Leave blank to use default value.

Time

Maximum job time in hours, 'walltime' value (1–672) or 28 days. Leave blank to use default value.

Project

Default Geoportal 200hrs project

Project to assign resource allocation to. Leave blank to use default value.

☑ Execute
Galaxy code – drmee.py

Works out of the box, but can be used for:

Editing the job parameters (native specs)

- account (redirect all galaxy accounts towards one cluster account)
- memory (hugemem vs. normal node)
- partition (long vs. normal)

Checking allocation settings for errors (inconsistencies)

- memory
- walltime
- memory + ntasks

Debugging/printing out the sbatch script (executable by the queueing system)
Configuration files for cluster communication

runners/utils/jobscript/DEFAULT_JOB_FILE_TEMPLATE.sh

template for the slurm scripts

#!/bin/sh

$headers
$slots_statement

$working_directory $command

echo $? > $exit_code_path

runners/utils/jobscript/CLUSTER_SLOTS_STATEMENT.sh

template for the slurm scripts (creates the script itself, jobsetup)

source /etc/profile
source /cluster/bin/jobsetup
Third-party slurm-drmaa (C library)

Edited to accommodate slurm switches which are not shipped with the source code:

- SLURM_NATIVE_TMP,
- SLURM_NATIVE_WORKDIR
- SLURM_NATIVE_CLUSTERS (for multiple cluster support)

Must be always compiled again after the slurm client update

Establishes the communication between DRMAA egg (session) and Slurm client (sbatch)
SLURM client binaries

[root@galaxy-test submit_host]# ls -l
total 29504
-rw-r--r-- 1 root root  135224 Jan 16 12:43 munge-0.5.10-1.el6.x86_64.rpm
-rw-r--r-- 1 root root  34492 Jan 16 12:43 munge-devel-0.5.10-1.el6.x86_64.rpm
-rw-r--r-- 1 root root  42164 Jan 16 12:43 munge-libs-0.5.10-1.el6.x86_64.rpm
-rw-r--r-- 1 root root 28533636 Jan 16 12:43 slurm-2.4.3-1.el6.x86_64.rpm
-rw-r--r-- 1 root root  122400 Jan 16 12:43 slurm-devel-2.4.3-1.el6.x86_64.rpm
-rw-r--r-- 1 root root  16732 Jan 16 12:44 slurm-munge-2.4.3-1.el6.x86_64.rpm
-rw-r--r-- 1 root root  415304 Jan 16 12:43 slurm-perlapi-2.4.3-1.el6.x86_64.rpm
-rw-r--r-- 1 root root  890428 Jan 16 12:44 slurm-plugins-2.4.3-1.el6.x86_64.rpm

[root@galaxy-test submit_host]# rpm -Uvh munge-*0.5.10*.rpm
[root@galaxy-test submit_host]# rpm -Uvh slurm-*2.4.3*.rpm

bash-4.1$ cat /etc/slurm_drmaa.conf

job_categories: {default: "--account=3572 --time=00:01:00 --mem-per-cpu=1000"},