An innovative PaaS solution to support Big Data Analytics and Workflow management via Galaxy

G. Donvito, M. Antonacci, V. Spinoso, S. Nicotri, F. Zambelli, <u>M.A. Tangaro</u>

> Conferenza Annuale di Lifewatch Italia Roma 25-27 Giugno 2018

1



Service architecture

Galaxy production environment

Reference data availability

Storage encryption

Automatic elasticity

Conclusions and outlook

-Galaxy

Galaxy is a workflow manager adopted in many life science research environments in order to facilitate the interaction with bioinformatics tools and the handling of large quantities of biological data.

Through a coherent work environment and an user-friendly web interface it organizes data, tools and workflows providing reproducibility, transparency and data sharing functionalities to users.

Galaxy instances can be deployed in three ways, each one with pros and cons:

- public servers;
- local servers;
- commercial cloud solutions.

👼 🗋 Galaxy 🛛 🗙			
🚍 Galaxy	Analyze Data Workflow Shared Data • Visualization • Admin Help • User •	=	Using 5.2 GB
Tools 🗘	Map with Bowtie for Illumina (Galaxy Tool Version 1.1.3) Options	History	COI
bowtie	Will you select a reference genome from your history or use a built-in index?	search datasets	0
NGS: Mapping Map with Bowtie for Illumina	Use a built-in index	Unnamed history 17 shown, 5 <u>deleted</u> , 4 <u>hidd</u>	
Bowtie2 - map reads against reference genome	Select a reference genome	4.01 GB	8 🔌 🖻
NGS: RNA Analysis TopHat for Illumina Find spice junctions using RNA-seq data TopHat Gapped-read mapper for RNA-seq data RSEM calculate reference BSEM calculate regression RNA-Seq by Expectation-Maximization	Homo supports (Pg15) • Tryour genome interred is not listed - contact Galaxy team • Is this library mate-paired? • Single-rel • FASTQ file • Must have ASCH encoded apply scores • Bowle settings to use •	26: Map with BWA for II umina on data 22, data 3, and data 25: mapped 25: https://renodo.org/r ecord/61377/files/hg19 chr/s.fa 24: https://renodo.org/r ecord/61377/files/dbsnp _138.hg19.chr8.vcf 23: https://renodo.org/r 23: https://renodo.org/r	
NGS: Peak Calling homer_makeTagDirectory Simple wrapper for makeTagDirectory. Used	Commonly used For most mapping needs use Commonly used settings. If you want full control use Full parameter list	ecord/61377/files/proba nd_R1.fg	
by findPeaks Workflows	Suppress the header in the output SAM file (sam-nohead)	22: https://zenodo.org/r ecord/61377/files/proba nd R2.fq	● / ×
<u>All workflows</u>	Bowke produces SAM with several lines of header information by default	21: galaxy portal test	@ / X
	What It does	20: Sc IP.fastq 14: FastQC on collectio Data	● / × n 8: Raw ×
▲vascript:void(0) III	Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole	• II	

	Ready to use	Quota	Galaxy Custom.	Maintenance	Costs	Data Privacy	
Public Servers		Strongly Limited	\bigotimes	Up to service provider	No costs (usually)	\bigotimes	
Local Install	\bigotimes			Required	Costly		
Cloud* (e.g. Amazon)		Costs Dependent		Only Galaxy Maintenance	Costly	⊗ <	

(*) Over 2400 Galaxy cloud servers launched in 2015 (Nucleic Acids Research (2016) doi: 10.1093/nar/gkw343)

ELIXIR-Italy in the framework of the INDIGO-DataCloud project has developed a cloud Galaxy instance provider, allowing to fully customize each virtual instance through a user-friendly web interface, overcoming the limitations of others galaxy deployment solutions.

In particular, our goal was to develop a PaaS architecture to automate the creation of Galaxy-based virtualized environments exploiting the software catalogue provided by the INDIGO-DataCloud community.

The **INDIGO-DataCloud** project (H2020-EINFRA-2014-2) aimed to to develop an open source computing and data platform, targeted at multi-disciplinary scientific communities, provisioned over public and private e-infrastructures.

https://www.indigo-datacloud.eu/

www.indigo-datacloud.eu/service-component

INDIGO-DataCloud started in April 2015 and ended in September 2017.



Service architecture

Integrating different INDIGO-DataCloud technologies to automatically deploy a ready-to-use Galaxy production environment.

All Galaxy required components automatically deployed (Orchestrator and IM).

User friendly access, allowing to easily configure and launch a Galaxy instance (INDIGO FutureGateway portal).

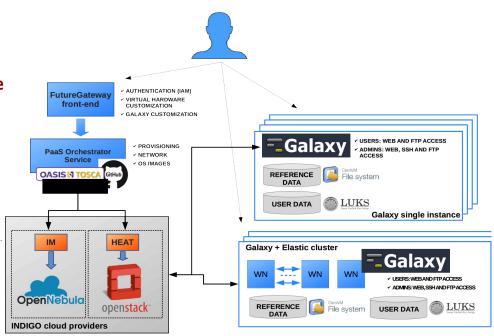
Authentication (IAM and FutureGateway).

Tools (Orchestrator and IM).

Reference data availability (CernVM File System).



Cluster support with automatic elasticity (INDIGO CLUES).



Running Instances

Galaxy production environment

Galaxy is deployed for a multi-user production environment, i.e. there are some additional auxiliary application needed for the best performance (the basic Galaxy installation is suitable for development by a single user):

- PostgreSQL as database
- NGINX as web server (+ upload module)
- uWSGI link between the service and the web server
- Proftpd as FTP server



Each Galaxy instance is customizable, through the web front-end, with different sets of pre installed tools (e.g. SAMtools, BamTools, Bowtie, RSEM, etc...), exploiting CONDA as default dependency resolver and YAML recipe.

Current available tools presets:

- galaxy-no-tools
- galaxy-rna-workbench
- galaxy-epigen
- galaxy-testing (for test purpose)

History History Unsame Instany Literature Instany L	0 ≠ 9 0 ≠ X 0 ≠ X	Administration Server Contractors reporter Contractors reporter Contractors reporter Contractors reporter Contractors Contractors Administration Administration Tools and Tool Shed Contractors Tool Shed		0 0 0 0
Unsamed history 32 shows 5 genesis 4 last 605 9 Hill 21: galaxy partial test 20: So. Pitang 14: PastQC on potentio	× • ≠ ×	Contr. Sport. registry Costs. station.registry Costs. station.registry Chooley.opplications Manager.ichts Tools and Tool Shed		•
14: PastQC on collectio	and the second s			
		Manage installed tools Reset metadata		•••
in the of channels 13: FastQC on collection Weigenge in the of channels	. x	Heselinetadata Deventeed local tool Tool lineage Persail a tools configuration		•••
A let of calculate	*	Wine Noti Entrillogs	-	•
6: FastQC on data 1: Writemat		Security • Liters		•••
data 4	*/×	Crospo Ecolea Additional		. •
	A NET OF CREATER T. FastingConcentral Li BaneData Mc FastingConcentral Li Writeman Sc BAM 30 BigHilg on data 4	LifatiQCandiaL → / x Breban Schultz → / x Writesu Schultzshaft	Land Constanting Tandoo and Constanting Tandoo and Constanting Refault Constanting Refau	Each e encoderation Database Image: An and an

AND AND AREAD BOARD	-		NFN Decas 🖌 OperProject 🗳 Ope	ale weekstoore dit talance Protect	No. Porta	in Plancha	a decision a	-	de P	COTTAN				
Galaxy			-			lelp= Usar+			00 L.	- Contine			Using	1.8 68
ministration	-							-						_
Ner Date types registry		•	barn_tr_bipelg +	Generate BigWig coverage files from files. Allows gapped reads to be split for RSuA-Geq.	Ered 1	brad-chapman	\$20cdb5ee3d6	insta	led		100	cred g	be pound	× .
Data tebles registry Display applications		•	bars, to_sam i v	Converta BAM format to SAM format.		desteam	#fc50182808	insta	led		100	ind g	2.bx.psu.ed	
Manage jobs		•	basecoverage -	Rose Coverage of all intensits		devteam	6a/w2133b69d	insta	lied		190	ered g	2 be proved	ł.,
Search Tool Shod		•	bed_overlap_significance +	permutation test of two bed file over	p	xuebing	#de3diffebal	irsta	led.		190	stel g	techened	,
Manage installed tools Reset metadata		•	bectools 🖉	bedtools: a powerful toolset for gena arithmetic	10	100	206703216258	Irola	Net		100	594.5	.bxpsu.ed	,
International In		•	best_regranics_subsets +	Portern Best-subsets Regression		devican	4536736445	Feli	ned.		.00	5101.5	Ltxpsu.ed	,
Relaad a tool's configuration Review tool intendion allages		•	bowle2	Gostie2: Fast and sensitive road alig	rment	deuteam	019023815478	Insta	led		100	ret g	2.bc.psu.ed	ε.
óme Tool Error Logs		•	bowie_wappers	Galaxy wrappers for the Bowle shar mapping tools.	read	devteam	#B0CSbedee	Insta	led		190	stred g	2.bx.pou.ed	ł.
Annage Disality WhiteIst arity Jams		•	twa_wappers +	Galaxy arappens for the BMA short r ofigner.	ad	devican	8442708060ed	Fela	Red.		300	stel g	techoned	,
83.63		•	canonical_consisten_analysis [+	Canonical Correlation Analysis		devlean	\$90004840271	ir sta	694		300	shed g	t techou ea	,
izina Etimpi		•	ckater 💌	Cluster the intervals of a dataset		devleam	785008056362	Folia	ned.		80	514.5	t to pound	,
	i m.	0	column join	Join tabular files		ishred.	6056c8a28c87	Insta	led.		100	the disc	becould	u l

7

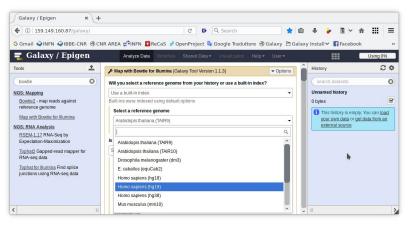
Reference data availability

Many Galaxy platform tools rely on the presence of reference data, such as alignment indexes or reference genome sequences, to efficiently work.

Each instance comes with reference data (e.g. genomic sequences) already available for many species, shared among all the instances through the CERN-VM FileSystem (cernvm.cern.ch) technology, thus avoiding unnecessary and costly data duplication.

Galaxy automatically is configured to properly use them.

Recipes and documentation to automatically setup your own CVMFS server and ship your reference data.



Storage encryption

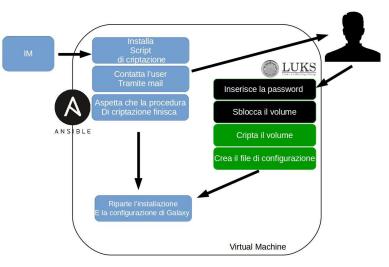
While the adoption of a distributed environment for data analysis makes data difficult to be tracked and identified by a malevolus attacker, full data anonymity and isolation is still not granted.

Data privacy is granted through LUKS storage encryption as a service: Users will be required to insert a password to encrypt/decrypt data directly on the virtual instance during its deployment, avoiding any interaction with the cloud administrator(s).

A notification mail is, sent to users describing how-to log into the VM and encrypt/decrypt the system.

User is only asked to insert their alphanumeric password 3 times:

- 1. Set password
- 2. Confirm password
- 3. Open LUKS volume.



Automatic logout after password injection: the encryption procedure continues in background.

Default encryption algorithm:

- aes-xts-plain64 encryption
- 256 bit key
- sha256 as hash algorithm used for key derivation.

Script to easily manage the LUKS volume is added to each virtual instance:

- check if the volume is correctly mounted,
- Mount and open LUKS volumes.
- Close and umount LUKS volumes.

```
marco — ssh -i cloud.key galaxy@90.147.170.75 — 127×40
Last login: Wed Mar 21 10:28:09 on ttys006
MacBook-Pro-di-Marco:~ marco$ ssh -i cloud.key galaxy@90.147.170.75
The authenticity of host '90.147.170.75 (90.147.170.75)' can't be established.
ECDSA key fingerprint is SHA256:K9G2eDRQUYtrk1UZWk9Iq25ZM7pXOwcj1xDqnVvq9C4.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '90.147.170.75' (ECDSA) to the list of known hosts.
Enter passphrase for key 'cloud.key':
         ELIXIR-Italy
              Filesystem encryption script
A password with at least 8 alphanumeric string is needed
There's no way to recover your password.
Example (automatic random generated passphrase):
                   zykhZyu6
You will be required to insert your password 3 times:
 1. Enter passphrase
 2. Verify passphrase
 3. Unlock your volume
The connection will be automatically closed.
         _____
INFO 2018-03-21 19:47:32 [fast-luks] Check if the required applications are installed...
INFO 2018-03-21 19:47:32 [fast-luks] Using aes-xts-plain64 algorithm to luksformat the volume
Enter passphrase:
Verify passphrase:
Command successful.
INFO 2018-03-21 19:47:32 [fast-luks] Open LUKS volume.
Enter passphrase for /dev/vdb1:
```

10

Test scenario (see backup slides):

- Volume not mounted. Impossible to access to its content.
- Volume opened and mounted, Galaxy running. Impossible to read data, even using cloud controller.

Automatic Elasticity

Virtual clusters through a dedicated section of the web front-end: allows to instantiate Galaxy with SLURM as Resource Manager and to customize the number of virtual nodes, nodes and master virtual hardware.

Automatic elasticity, provided using the Infrastructure Manager and CLUES service components, enables **dynamic cluster resources scaling,** providing an efficient use of the resources, making them available only when really needed.

New working nodes are powered-on, depending on the cluster workload and powered-off when no longer needed:

- Each Galaxy tools preset installed with Galaxy have been tested to work with SLURM elastic Cluster.
- Each node is configured according to the Galaxy tools installed on the VM as selected by the user during the configuration phase.

Conclusion and outlook

Our service aims to provide the Galaxy workflow manager to end users ranging from small research groups to institutions or SMEs, on suitable computation resources, removing the need to maintain their own hardware and software infrastructure and using resources in a more efficient way, ensuring improved reliability, better performances and the capability to handle larger research activities exploiting the features of the INDIGO-Datacloud components, opening the route for the migration of public Galaxy instances to this service.

INDIGO ended in September 2017, but the development of the services exploited as backend for this service continues in H2020 project Deep-HybridDataCloud and eXtreme-DataCloud:

- Support the transparent access to specialized computing hardware (GPUs, Infiniband, etc.) and HPC resources.
- Improve the workflow for hybrid deployments

Future improvements:

- Deployment of dockerized Galaxy and jobs on Mesos clusters.

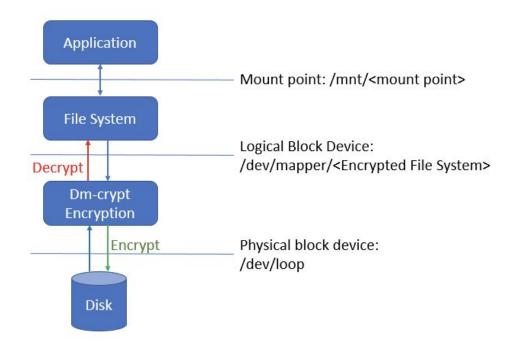


CONTACTS:

- Marco Antonio Tangaro (CNR-IBIOM) ma.tangaro@ibiom.cnr.it
- Federico Zambelli (ELIXIR-ITA technical coordinator) f.zambelli@ibiom.cnr.it
- Giacinto Donvito (INFN) giacinto.donvito@ba.infn.it
- Graziano Pesole (head of ELIXIR-ITALY Node) g.pesole@ibiom.cnr.it







Bash scripting + Ansible + INDIGO PaaS Orchestrator:

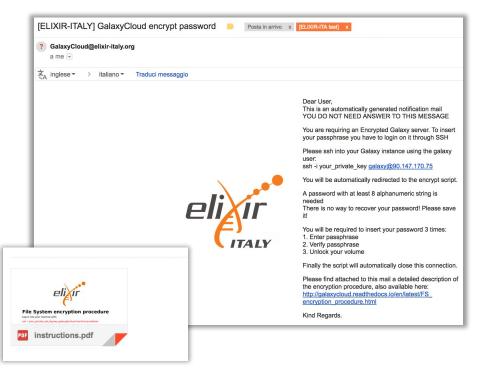
- Storage Encryption as a Service
- Dependency resolution
- Script instance lock, i.e. is not possible to run two instances of the encryption script.
- Configurable (encryption algorithm, key size, hash algorithm, mountpoint, filesystem).
- Automatic configuration file creation to open/close the volume with one command.

Ansible automates the encryption procedure, installing the scripts, informing, by mail, the user once the system is ready to accept the password.

The encryption procedure summary is reported by mail, while a detailed step-by-step how-to is sent attached.

Script to easily manage the LUKS volume is added to each virtual instance:

- check if the volume is correctly mounted,
- Mount and open LUKS volumes.
- Close and umount LUKS volumes.



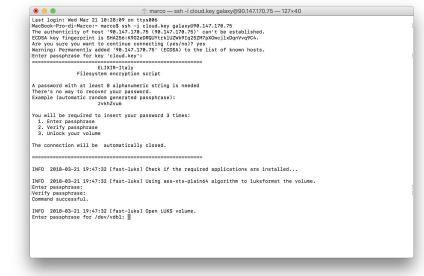
Automatic logout after password injection: the encryption procedure continues in background.

Default encryption algorithm:

- aes-xts-plain64 encryption
- 256 bit key
- sha256 as hash algorithm used for key derivation.

Script to easily manage the LUKS volume is added to each virtual instance:

- check if the volume is correctly mounted,
- Mount and open LUKS volumes.
- Close and umount LUKS volumes.



- Test on unmounted encrypted devices:
 - Create two volumes, one encrypted
 - Put inside the same file
 - Umount volumes
 - Create volume binary images and HexDump the binary image with xdd
 - Grep non-zero bytes and search for the file content

It is possible to see the file content only on the un-encrypted volume.

• Try to open the volume when active (LUKS volume opened and mounted, Galaxy running) in the Virtual Machine.

Test executed on the cloud controller as administrator.

It is not possible to mount the volume without the user password.

Automatic elasticity

ELIXIR-IIB: Galaxy as a Cloud Service



