PIETRO MANDREOLI

Consortium GARR | THE ITALIAN EDUCATION & RESEARCH NETWORK

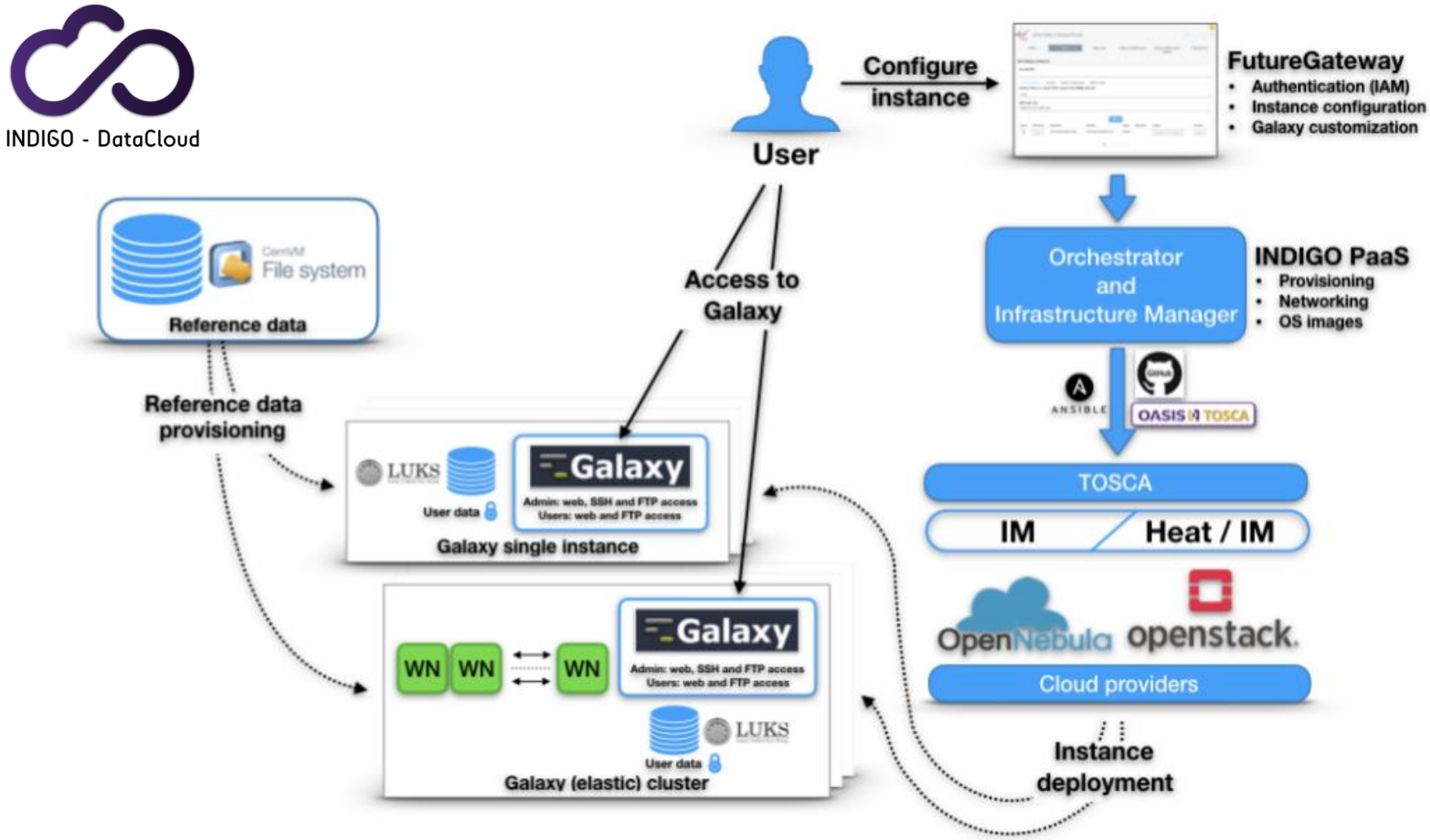
Portabilità su GARR Cloud di Laniakea : un servizio Galaxy on-demand basato su tecnologia INDIGO-Datacloud

GIORNATA DI INCONTRO
BORSE DI STUDIO GARR
"ORIO CARLINI"
ROMA

Roma 27 giugno 2019

Borsisti day

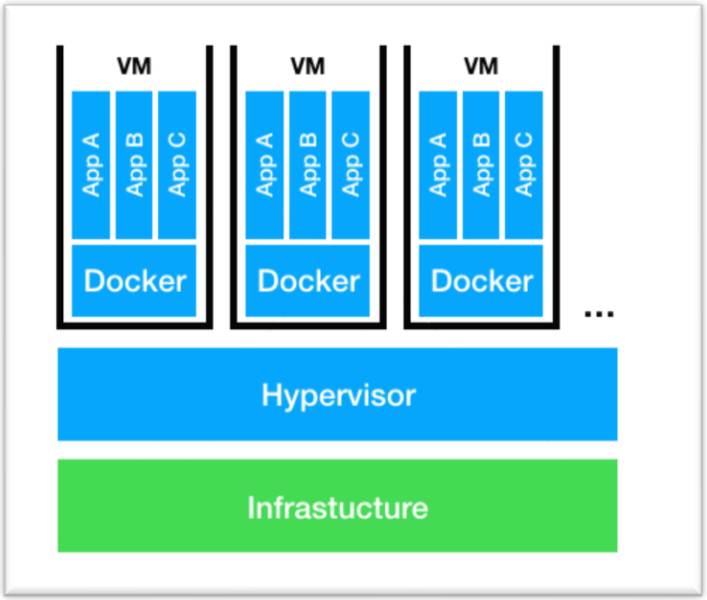
ARCHITETTURA DEL SERVIZIO



Componenti di Laniakea installati

installazione su GARR Cloud dei seguenti componenti INDIGO :

- INDIGO Infrastucture Manager (IM)
- INDIGO Change management database (CMDB)
- Cloud Provider Ranker CPR
- Service Level Agreement Manager (SLAM)
- FutureGateway Portal



I componenti INDIGO di Laniakea sono dockerizzati e installati su virtual machine

questo sistema crea un livello addizionale di virtualizzazione unendo la caratteristica di isolamento del container con la sicurezza della VM.

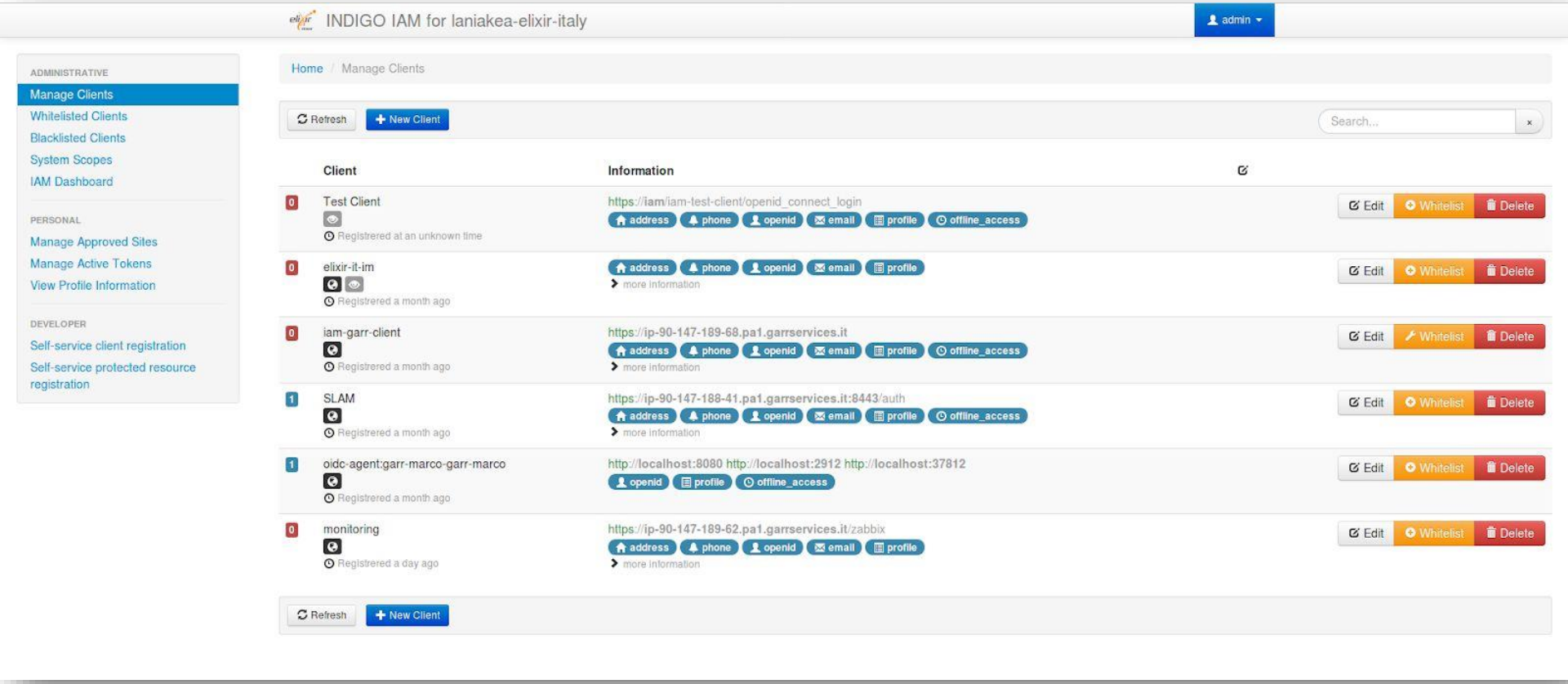
Procedura di installazione dei componenti: IM, CMDB, CPR, SLAM e proxy server

- Studio Ansible roles messi a disposizione da INDIGO
- Creazione VM necessarie e security group

name	service	RAM	vCPU	storage	Porte aperte
vm_master	Proxy server	2Gb	1	20Gb	443, 8080
IAM	Identity and Access Manager (IAM)	4Gb	2	40Gb	80, 443
Elixir-it-IM	Infrastructure Manager	8Gb	4	80Gb	8800
Cmdb	INDIGO Change management database (CMDB) and CPR	4Gb	2	40Gb	443, 5984, 8080, 8081
SLAM	Service Level Agreement Manager (SLAM)	8Gb	4	80Gb	8443, 443
FGW	FutureGateway Portal	8Gb	4	80Gb	80,443

Installazione componenti e integrazione con IAM

- Creazione dei client necessari per l'integrazione dei componenti INDIGO in IAM attraverso MitreID dashboard
- Ottenuti i client ID e client Secret questi sono stati opportunamente inseriti negli Ansible role modificati per poter essere eseguiti su GARR Cloud
- I componenti sono stati quindi installati nelle rispettive VM
- È stato poi installato il server proxy e sono stati definiti tutti gli endpoint dei componenti
- Per FutureGateway portal è stato seguito il processo di configurazione descritto nella documentazione di Laniakea


INDIGO IAM for laniakea-elixir-italy
admin
ADMINISTRATIVE
Manage Clients
Whitelisted Clients
Blacklisted Clients
System Scopes
IAM Dashboard
PERSONAL
Manage Approved Sites
Manage Active Tokens
View Profile Information
DEVELOPER
Self-service client registration
Self-service protected resource registration
Home / Manage Clients
Refresh
New Client
Search...
Client
Information
Test Client
Registered at an unknown time
https://iam/iam-test-client/openid_connect_login
address phone openid email profile offline_access
Edit Whitelist Delete
elixir-it-im
Registered a month ago
address phone openid email profile
more information
iam-garr-client
Registered a month ago
https://ip-90-147-189-68.pa1.garrservices.it
SLAM
Registered a month ago
https://ip-90-147-188-41.pa1.garrservices.it:8443/auth
oidc-agent:garr-marco-garr-marco
Registered a month ago
http://localhost:8080 http://localhost:2912 http://localhost:37812
openid profile offline_access
monitoring
Registered a day ago
https://ip-90-147-189-62.pa1.garrservices.it/zabbix

IM tests

- Si è presa dimestichezza con le REST API di IM
- Sono state definite le chiamate *curl* per l'esecuzione e il management dei deployment

```
curl -k -H 'Content-type: text/yaml' -H 'AUTHORIZATION: type = InfrastructureManager; username = pietro; token = <token_from_IAM>\nid = ost; type = OpenStack; host = https://keystone.cloud.garrservices.it:5000; domain = cloudusers; auth_version = 3.x_password; username = ******; password = *****; tenant = elixir; service_region = garr-pa1' -i –X POST https://ip-90-147-189-62.pa1.garrservices.it/im/infrastructures --data-binary"@TOSCA_TEMPLATE.yaml"
```

(Chiamata curl per l'esecuzione del deployment)

- Scrittura e modifica dei TOSCA template
- Test di IM per differenti infrastrutture contenenti Galaxy

Galaxy: creazione differenti configurazioni

Effettuato il deployment di diverse istanze Galaxy su GARR Cloud e sono state create, definite e testate tre differenti configurazioni di deployment

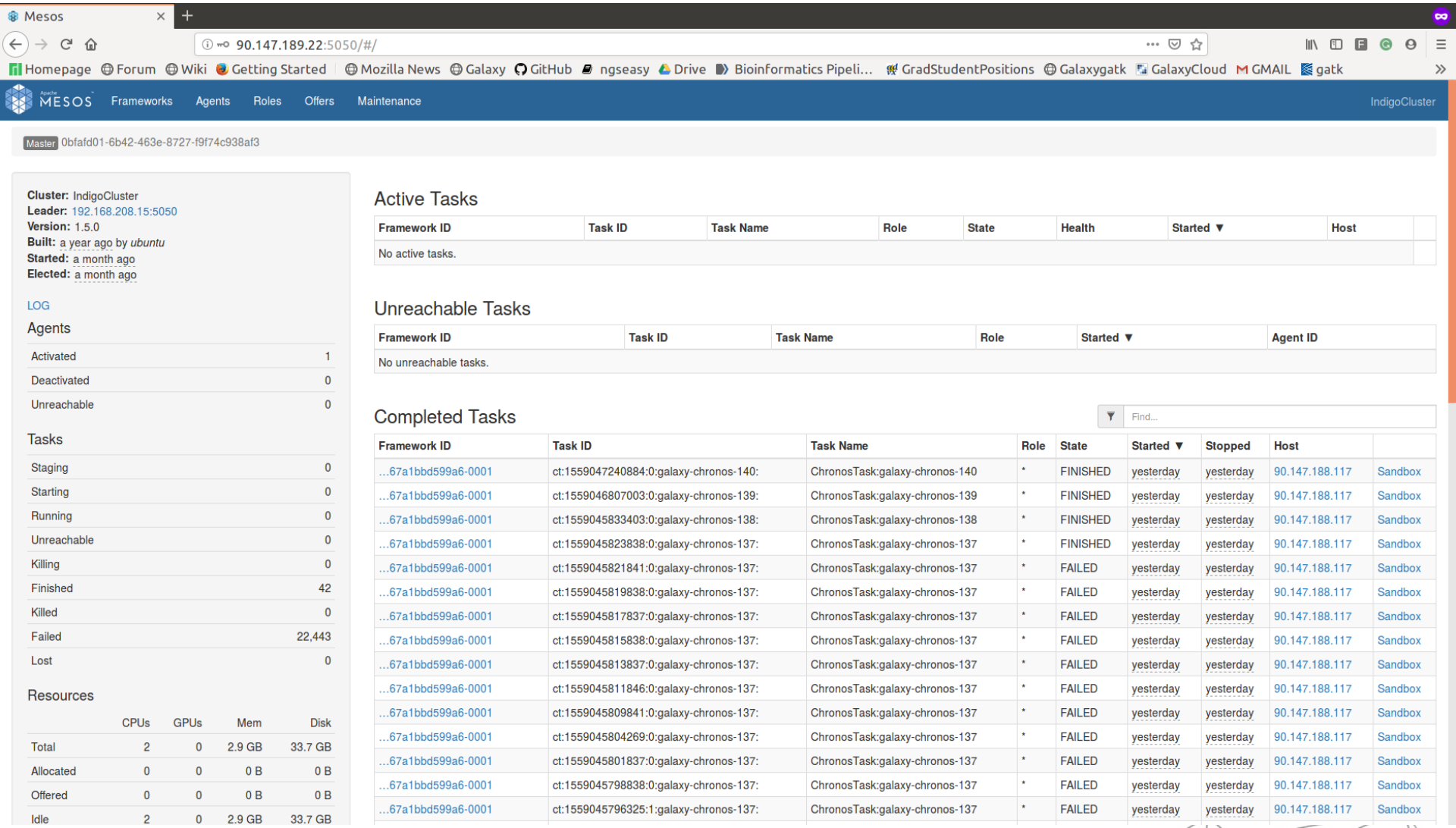
	Full Docker	Hibrid Docker	Full Singularity
Caratteristiche	esecuzione di tutti i tools presenti in Galaxy in Docker container	esecuzione di solo determinati tool in Docker container, il resto utilizza Conda	Esecuzione di tutti i tools in container Singularity
Configurazione	▪ Installazione dell'utility Mulled ▪ Cambiamento file di configurazione	▪ Cambiamento file di configurazione	▪ Cambiamento file di configurazione ▪ Montaggio CVMFS contenente le immagini

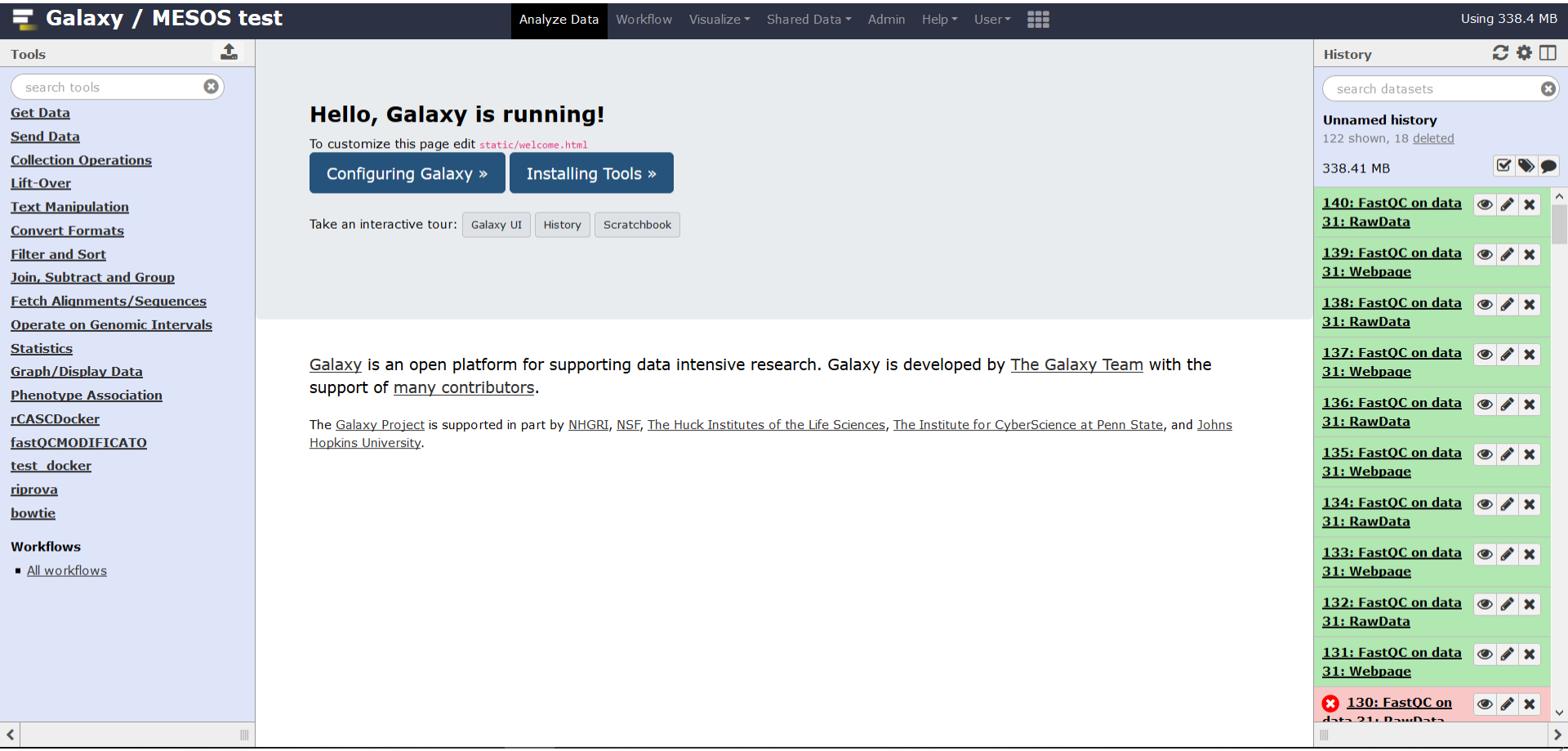
Integrazione Galaxy-ApacheMESOS

Apache MESOS è un Resource Manager per HPC cluster che permette una migliore condivisione delle risorse hardware da parte delle applicazioni sfruttando le tecnologie di containerizzazione.

Procedura di Integrazione:

- Attraverso l'Infrastructure Manager e un TOSCA file specifico è stato eseguito il deployment di un cluster MESOS e di una Galaxy Virtual Machine
- È stato attivato il protocollo HTTPS creando i certificati Let's Encrypt
- È stato configurato il Network File System (NFS) tra Galaxy e i nodi del cluster e modificato Galaxy per poter lanciare job su MESOS
- Configurazione di Galaxy per permettere l'interazione con Chronos il job scheduler del sistema
- Test del sistema


Mesos
90.147.189.22:5050/#/
Homepage Forum Wiki Getting Started Mozilla News Galaxy GitHub ngseasy Drive Bioinformatics Pipeli... GradStudentPositions Galaxygatk GalaxyCloud GMAIL gatk
MESOS Frameworks Agents Roles Offers Maintenance IndigoCluster
Master 0bfafd01-6b42-463e-8727-f9f74c938af3
Cluster: IndigoCluster
Leader: 192.168.208.15:5050
Version: 1.5.0
Built: a year ago by ubuntu
Started: a month ago
Elected: a month ago
LOG
Agents
Activated 1
Deactivated 0
Unreachable 0
Tasks
Staging 0
Starting 0
Running 0
Unreachable 0
Killing 0
Finished 42
Killed 0
Failed 22,443
Lost 0
Resources
CPUs GPUs Mem Disk
Total 2 0 2.9 GB 33.7 GB
Allocated 0 0 0 B 0 B
Offered 0 0 0 B 0 B
Idle 2 0 2.9 GB 33.7 GB
Active Tasks
Framework ID Task ID Task Name Role State Health Started ▼ Host
No active tasks.
Unreachable Tasks
Framework ID Task ID Task Name Role Started ▼ Agent ID
No unreachable tasks.
Completed Tasks
Find...
Framework ID Task ID Task Name Role State Started ▼ Stopped Host
...67a1bbd599a6-0001 ct:1559047240884:0:galaxy-chronos-140: ChronosTask:galaxy-chronos-140 * FINISHED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559046807003:0:galaxy-chronos-139: ChronosTask:galaxy-chronos-139 * FINISHED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045833403:0:galaxy-chronos-138: ChronosTask:galaxy-chronos-138 * FINISHED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045823838:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FINISHED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045821841:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045819838:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045817837:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045815838:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045813837:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045811846:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045809841:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045804269:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045801837:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045798838:0:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox
...67a1bbd599a6-0001 ct:1559045796325:1:galaxy-chronos-137: ChronosTask:galaxy-chronos-137 * FAILED yesterday yesterday 90.147.188.117 Sandbox


Galaxy / MESOS test
Analyze Data
Workflow
Visualize
Shared Data
Admin
Help
User
Using 338.4 MB
Tools
search tools
Get Data
Send Data
Collection Operations
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
Operate on Genomic Intervals
Statistics
Graph/Display Data
Phenotype Association
rCASCDocker
fastQCMODIFICATO
test_docker
riprova
bowtie
Workflows
All workflows
Hello, Galaxy is running!
To customize this page edit static/welcome.html
Configuring Galaxy »
Installing Tools »
Take an interactive tour:
Galaxy UI
History
Scratchbook
Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.
The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.
History
search datasets
Unnamed history
122 shown, 18 deleted
338.41 MB
140: FastQC on data 31: RawData
139: FastQC on data 31: Webpage
138: FastQC on data 31: RawData
137: FastQC on data 31: Webpage
136: FastQC on data 31: RawData
135: FastQC on data 31: Webpage
134: FastQC on data 31: RawData
133: FastQC on data 31: Webpage
132: FastQC on data 31: RawData
131: FastQC on data 31: Webpage
130: FastQC on

Integrazione Galaxy-ApacheMESOS: fase di debug

- Test effettuati con due differenti tool: uno semplice creato ad hoc per la fase di debug e FastQC (tool molto utilizzato e quindi stabile)
- Galaxy è stato integrato correttamente i tools sono eseguiti su MESOS ma gli outputs non sono linkati nel pannello «history» dell'applicazione
- Testate differenti configurazioni cambiando il file job_conf.xml file che gestisce le destinazioni e le dipendenze dei tool lanciati
- Controllati i i vari log sia di Galaxy che di Chronos e i comandi Docker dei tools lanciati su MESOS

➢ Il problema è stato individuato nel runner che si occupa dell'invio dei job da Galaxy a Chronos, questo infatti non setta in modo giusto la working directory del Docker causando l'unlinking tra i risultati e il pannello «history» inoltre non permette il montaggio di volumi ulteriori per l'accesso ai reference data

➢ Sono stati messi al corrente del bug i developer del progetto Galaxy inviando un report dettagliato della fase di debug contenente tutti i test effettuati, la configurazione utilizzata e i log

Progetto rCASC

Durante questi mesi ho inoltre partecipato ad un progetto parallelo per rendere disponibile la pipeline rCASC collaborando con il per l'analisi di dati provenienti da single cell RNA-seq, su Galaxy collaborando con Dipartimento di Biotecnologie Molecolari e Scienze per la salute dell'università di Torino.

Integrazione pipeline:

- modifica codice R delle funzioni, non adatte in primis per Galaxy
- Scrittura wrapper relativi ai vari tool che compongono il pacchetto rCASC

Il progetto è ancora nella sua fase iniziale ma siamo già stati in grado di definire le linee guida per l'integrazione di rCASC in Galaxy.

Galaxy-Consensus Variant Calling System (CoVaCS)

Il flavour CoVaCS è stato validato confermando che l'implementazione del CoVaCS workflow in Laniakea funziona come l'implementazione originale di CoVaCS al CINECA.

Attualmente un istanza di Galaxy di Laniakea presente al RECAS di Bari è utilizzata dal reparto di bioinformatica della SSD di Genetica Medica dell'Istituto Ortopedico Rizzoli.

Il workflow Covacs è stato modificato dagli utenti per poter essere più performante su dati provenienti da IonTorrent

«l'introduzione di CoVaCS nel processo di pre-screening ha permesso un'identificazione più veloce delle varianti su cui tutte le piattaforme ThermoFisher sono particolarmente deboli, integrando e complementando uno strumento che è ottimizzato per l'identificazione di variati complesse (SeqPilot).» (Emanuele Bonetti, bioinformatico, SSD di Genetica Medica dell'Istituto Ortopedico Rizzoli)

Conclusioni

- Durante questi 9 mesi di borsa sono stati installati e configurati e testati su GARR Cloud i diversi componenti del sistema Laniakea
- Sono state definite e testate diverse configurazioni del Workflow manager bioinformatico Galaxy che permetteranno in un futuro una diversificazione delle istanze supportate dal sistema Laniakea
- Per ottenere un sistema di produzione completo e accessibile da utenti generici, è necessario che il servizio di Autenticazione INDIGO IAM sia incluso tra quelli supportati dal GARR, per questo motivo si è preso contatto con in personale tecnico di GARR Cloud per discutere la possibilità di ottenere un'istanza OpenStack di pre-produzione che permetterà il supporto di IAM e quindi l'installazione di un sistema in fase di produzione
- Durante i prossimi 2 mesi di borsa procederò con l'ottimizzazione del sistema e continuerò a lavorare all'integrazione di Galaxy con Apache MESOS.

Pubblicazioni:

- Poster **"rCASC implementation in Laniakea: porting containerization-based-reproducibility to a cloud Galaxy on-demand platform",** Authors: Alessandri L , Mandreoli P, Tangaro MA, Beccuti M, Calogero RA, Zambelli F. **Presentato a** : 16th Annual Meeting of the Bioinformatics Italian Society June 26-28, 2019, Palermo, Italy **e al** ELIXIR - EXCELERATE All Hands meeting 2019, Lisbon,17 July 2019
- Talk **"Laniakea: A Galaxy-on-demand Provider Platform Through Cloud Technologies"** ,Authors:Tangaro Marco Antonio 1 , Donvito Giacinto 2 , Antonacci Marica 2 , Chiara Matteo 3 , Mandreoli Pietro 3 , Pesole Graziano 1,4 , Zambelli Federico 1,3,2019 Galaxy Community Conference (GCC2019)Freiburg, Germany, 1-6 July 2019
- Talk **"Laniakea@ReCaS: an ELIXIR-ITALY Galaxy on-demand cloud service"** Authors:Marco Antonio Tangaro, Giacinto Donvito, Marica Antonacci, Pietro Mandreoli, Matteo Chiara, Graziano Pesole and Federico Zambelli. **Presentato al** ELIXIR - EXCELERATE All Hands meeting 2019, Lisbon,17 July 2019 **e al** 16th Annual Meeting of the Bioinformatics Italian Society June 26-28, 2019, Palermo, Italy

Proposta di progetto per proroga borsa: Messa in produzione, verifica e validazione su GARR Cloud di Laniakea, servizio Galaxy on-demand basato su tecnologia INDIGO-DataCloud

Il progetto prevede di portare il servizio di Laniakea alla fase di pre-produzione e la sua validazione con utenti esterni che lo utilizzeranno per il loro lavoro di ricerca.

Inoltre il tempo a disposizione permetterebbe di integrare altri servizi di supporto come il repository CVMFS per i reference data ed il supporto a Pulsar

CernVM-File System stratum 1

- File system sviluppato dal CERN
- Miglioramento esperienza utente
- Maggior velocità dell'accesso ai Reference Data

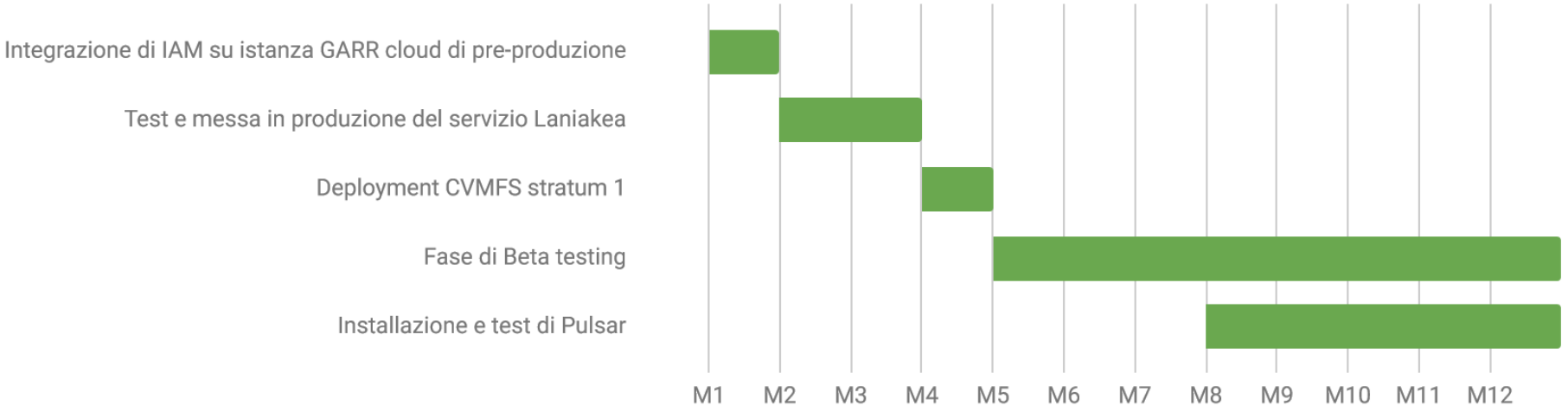
PULSAR

- Strumento che permette a risorse computazionali remote di supportare il carico di lavoro dell'istanza Galaxy ufficiale usegalaxy.eu o, potenzialmente, di qualsiasi altra istanza Galaxy, senza necessità di condividere un file system

Step del progetto e timeline

Il progetto sarà suddiviso in quattro parti:

- Porting del sistema Laniakea sull'istanza di pre-produzione
  - Integrazione di IAM tra i provider OpenID connect dell'istanza di pre-produzione
  - Riconfigurazione e validazione dei Componenti INDIGO sull'istanza di preproduzione
  - Sostituzione di FutureGateway con un nuovo front-end più leggero e flessibile
- Installazione di CernVM-File System stratum 1
- fase di beta-testing
  - Selezione e formazione dei Beta tester
  - Risoluzione possibili problematiche del sistema
- Installazione e test di Pulsar
  - Installazione, configurazione e test di una Virtual Machine Pulsar su GARR Cloud


Integrazione di IAM su istanza GARR cloud di pre-produzione
Test e messa in produzione del servizio Laniakea
Deployment CVMFS stratum 1
Fase di Beta testing
Installazione e test di Pulsar
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12

Consortium GARR | THE ITALIAN EDUCATION & RESEARCH NETWORK

Grazie

TOSCA

Topology and Orchestration Specification for Cloud Applications (TOSCA)

- Linguaggio open-source usato per descrivere le dipendenze e le relazioni tra i servizi e le applicazioni presenti su un'infrastruttura cloud
- Sintassi YAML
- Possibilità di creare i "customized TOSCA" types per la descrizione di applicazioni specifiche
- Integrazione con Ansible