A curated Domain centric shared Docker registry linked to the Galaxy toolshed
François Moreews, Olivier Sallou, Yvan Le Bras, Grosjean Marie, Cyril Monjeaud, Thomas A Darde, Olivier Collin, Christophe Blanchet

To cite this version:
François Moreews, Olivier Sallou, Yvan Le Bras, Grosjean Marie, Cyril Monjeaud, et al.. A curated Domain centric shared Docker registry linked to the Galaxy toolshed. Galaxy Community Conference 2015, Jul 2015, Norwich, United Kingdom. hal-01160514

HAL Id: hal-01160514
https://inria.hal.science/hal-01160514
Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A curated Domain centric shared Docker registry linked to the Galaxy toolshed

François Moreews¹, Olivier Sallou², Yvan le Bras², Marie Grosjean³, Cyril Monjeaud², Thomas Darde⁴, Olivier Collin², Christophe Blanchet³

¹ Genscale team -IRISA -Rennes, France
² Genouest Bioinformatics facility – INRIA/IRISA – Rennes, France
³ French Institute of Bioinformatics – CNRS IFB-Core UMS3601 – Gif-sur-Yvette, France
⁴ INSERM U625 – Rennes France
“Docker is an open-source engine to easily create lightweight, portable, self-sufficient containers from any application. The same container that a developer builds and test on a laptop can run at scale, in production, on Vms, [...], public clouds and more.”
Why using Docker containers to build, deploy and execute applications?

- **Efficient** (no virtualization)
- **Isolation**
- **Build one time, execute “anywhere”, independently of the execution platform (laptops, clusters and clouds with Linux kernels)**
FROM ubuntu:12.0.4

ADD . /script

ENTRYPOINT ["perl /script/analysis.pl"]
docker -run
    containerUniqueID
    -k=31 -i input.fastq -o output.bam

-The docker run command acts as a wrapper of the tool command line.

-Host directories (input, output, work...) can be mounted inside the container.
Google Container Engine

- A container based cloud architecture
- “With container-based computing, application developers can focus on their application code, instead of on deployments and integration into hosting environments”.

Docker on the commercial Cloud
Docker on academic HPC clusters

- **Google Kubernetes**: an open source technology for containers life cycle management.
- **Docker Swarm**: allows to create and access to a pool of Docker hosts.
- **Genouest GO-Docker**: a batch scheduler like SGE, submitting jobs in Docker containers on top of Swarm.
Bioinformatics tools benchmarks with Docker

- cami-challenge.org: Critical Assessment of Metagenomic Interpretation
- http://nucleotid.es: continuous, objective and reproducible evaluation of genome assemblers using docker containers
- bioboxes.org: interchangable bioinformatics software containers
Galaxy Docker integration

- Docker can be used in Galaxy to:
  - manage tools dependencies: one tool, on Docker
  - Distribute populated Galaxy Distribution related to one topic
Shared registries: Docker Hub

- Not structured
- Not curated
- Not domain centric
- Not community driven
BIOSHADOCK

An initiative of the French Bioinformatics Institute & the Genouest Bioinformatics Facility

Goals:

- Federate bioinformatics tools deployment procedures for the IFB cloud infrastructure
- Generate customized Galaxy cloud instances on the fly.
- Docker image indexation (service registry & searches)
BioShaDock registry

A Bioinformatics Shared Docker registry

BioShaDock is a Bioinformatics Docker registry. Here are hosted Docker images dedicated to a broad spectrum of Biological communities as represented by the Biogenouest Western France network. In particular, you will find here:

- Command line tools;
- Complex web server frameworks.
- Galaxy Docker images that you can use with specific Docker Galaxy tools thanks to recent developments through the e-Biogenouest project (https://www.e-biogenouest.org/), (see GUGGO and our Toolshed);

The number of images available will grow following the community needs.

The Docker GenQuest core facility team:
François Moreews, Cyril Monjeaud, Yvan Le Bras, Olivier Sallou
**Shared registries: BIOSHADOCK**

- BIOSHADOCK
  - Focuses on the model “on tool, one docker image”
  - Allows Dockerfile build
  - Manages permissions (private/public images)
  - May integrate meta data to facilitate query and service registry searches
  - One unique repository for softwares with or without tool.xml => SAAS + CMD
  - Integrated to Galaxy by redefining tools dependencies in a Toolshed
### Repositories by Category

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>Alignment tools</td>
</tr>
<tr>
<td>Data management</td>
<td>Tools for data management</td>
</tr>
<tr>
<td>Data manager</td>
<td>Biological data management</td>
</tr>
<tr>
<td>Data source</td>
<td>data source</td>
</tr>
<tr>
<td>Fasta / Fastq manipulation</td>
<td>Fasta / Fastq manipulation</td>
</tr>
<tr>
<td>France Genomique Assembly</td>
<td>Tools for NGS assembly as part of France Genomique Project</td>
</tr>
<tr>
<td>Metadata management</td>
<td>Tools for ISA manipulation</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>NGS metagenomics analysis tools</td>
</tr>
<tr>
<td>Microarray</td>
<td>Analyse gene expression data through microarray technology</td>
</tr>
<tr>
<td>Next Generation Sequencing</td>
<td>NGS analysis tools</td>
</tr>
<tr>
<td>Phylogeny</td>
<td>phylogeny tools</td>
</tr>
</tbody>
</table>
BIOSHADOCK cluster /cloud integration using GO-DOCKER
Build it one time, use it as you want

Galaxy tools & workflows

Command lines

BIOSHADOCK

Other SAAS tools
References

- Genouest GO-DOCKER : http://www.genouest.org/?p=246
- Google Kubernetes, Docker container cluster management : kubernetes.io
- BioShaDock, a Bioinformatics Shared Docker registry : http://docker-ui.genouest.org
- GUGGO Galaxy Tooshed : http://toolshed.genouest.org
- Nucleotid.es, continuous, objective and reproducible evaluation of genome assemblers using docker containers : http://nucleotid.es
- ELIXIR Tools and Data Services Registry : https://elixir-registry.cbs.dtu.dk
- Bioboxes, a standard for creating interchangable bioinformatics software containers : http://bioboxes.org
- IFB academic Cloud : http://www.france-bioinformaticque.fr/?q=en/core/e-infrastructure-team/ifb-cloud