

Snakemake et docker dans le cloud IFB

Journée Snakemake

Institut Pasteur le 12/12/2016

Sandrine Perrin



ifb
INSTITUT FRANÇAIS
DE BIOINFORMATIQUE

Institut Français de Bioinformatique - IFB
French Institute of Bioinformatics - ELIXIR-FR
CNRS UMS360I - Gif-sur-Yvette - FRANCE

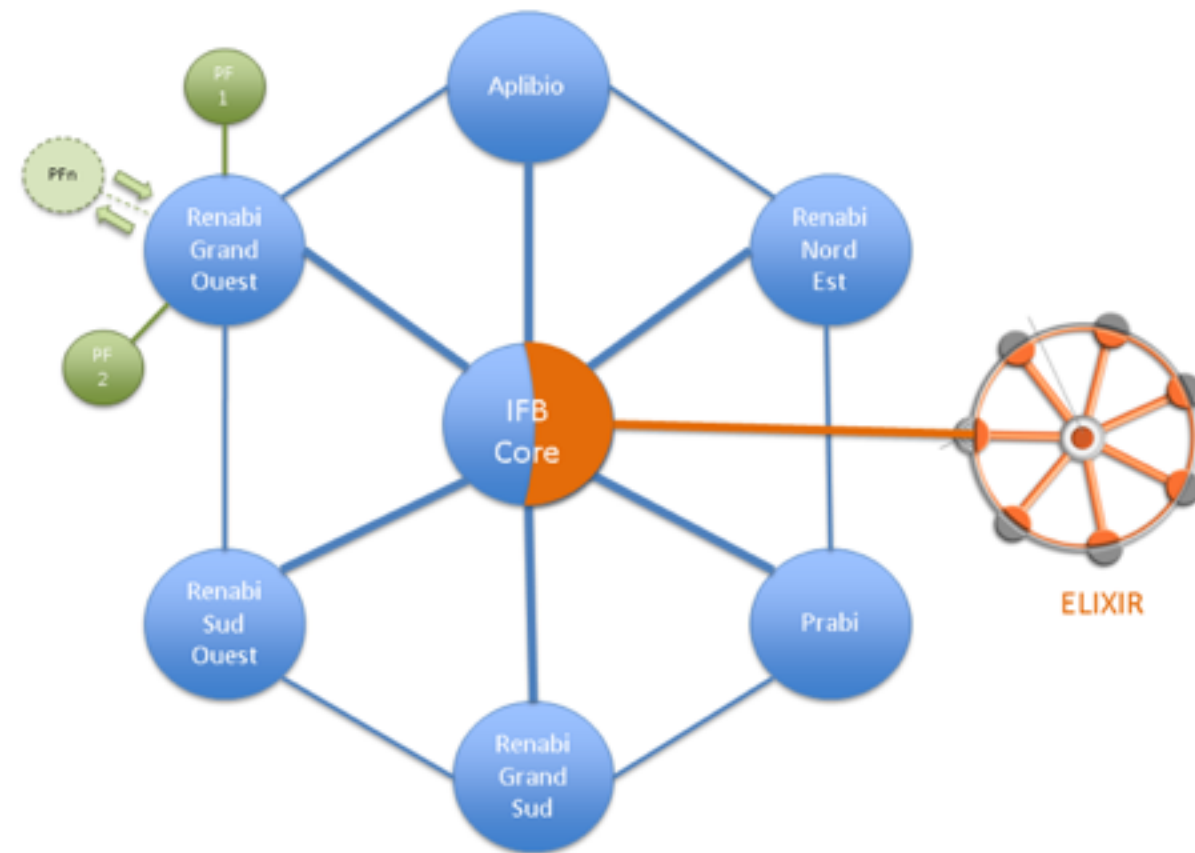
Plan

- I. Introduction IFB Cloud**
- II. Snakemake & Docker**
- III. Demo : workflow procNCRNAdisc**






Présentation de l'IFB

- ✓ Objectif : structurer l'ensemble de la communauté française des PF de service en bioinformatique et y associer la communauté recherche en bioinformatique
- ✓ Organisation :
 - ❖ 37 nœuds/plates-formes organisées en 6 pôles régionaux
 - ❖ Un nœud national, IFB-core, chargé d'impulser et de coordonner la mise en place de l'infrastructure.
- ✓ Mission générale : fournir des ressources de base en bioinformatique à la communauté des sciences de la vie
- ✓ Infrastructure nationale de **service** en bioinformatique
 - ❖ **Données & Outils** ;
 - ❖ **Appui & Formations**
 - ❖ **Infrastructure** : Mettre à disposition une infrastructure informatique dédiée à l'analyse des données des sciences du vivant (matériel, données, outils)
- ✓ Noeud français pour **ELIXIR**



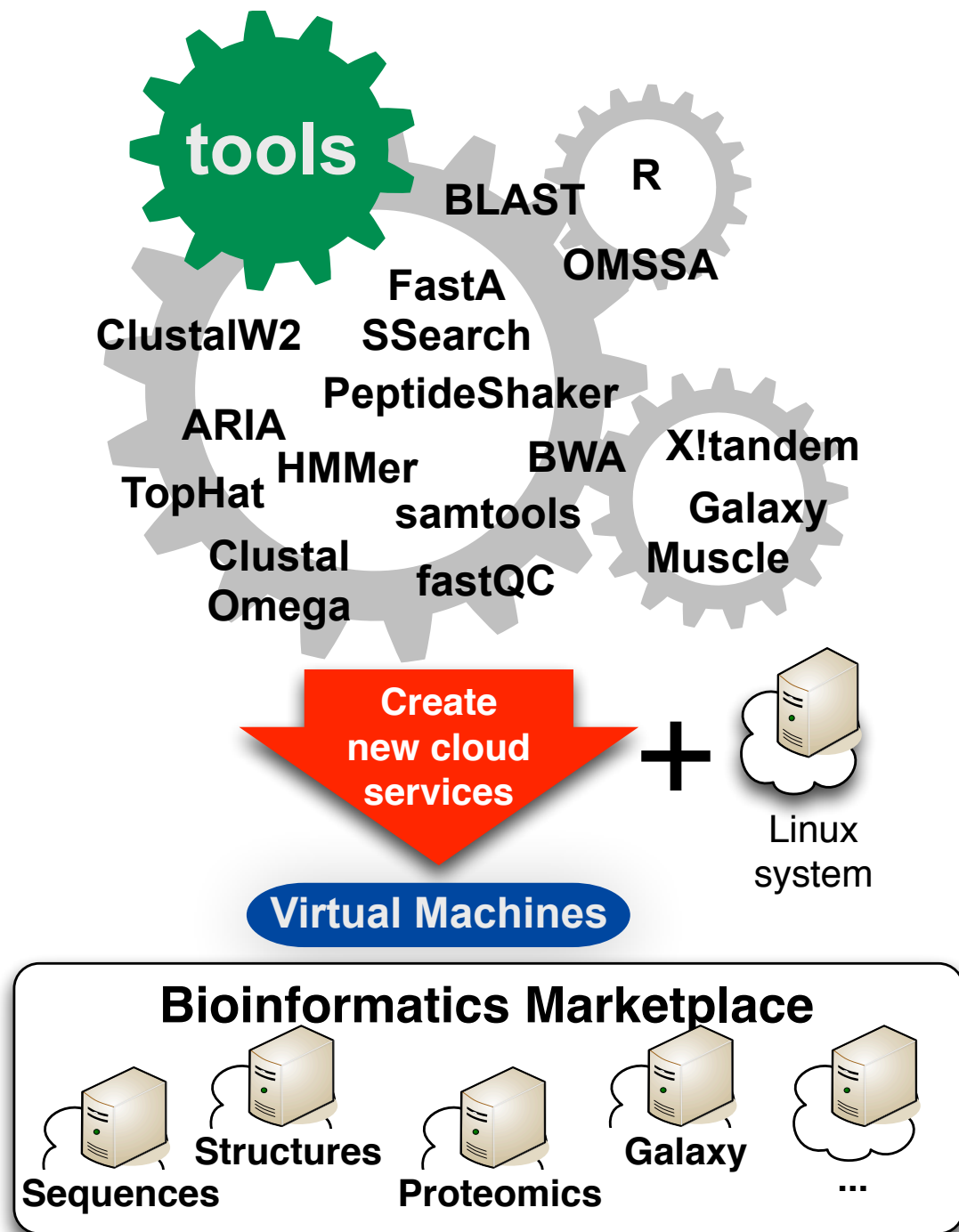
<https://www.elixir-europe.org/>

Hardware Characteristics

SITE	Compute #cores	Storage #TB	RAM #GB	Largest VM	Technology	Location
IFB-core <i>2014-16</i>	200 (+160)	50 (+96)	2,000 (+1)	20c 256GB		CNRS- IDRIS, Paris
début 2017	5,000	1,000	40,800	128c 3TB		CNRS-IDRIS, Paris
fin 2017	10,000	2,000	-	-		CNRS-IDRIS, Paris
GenOuest <i>2014-16</i>	220 (+96)	8 (+20)	685	8c 32GB	OpenNebula	IRISA, Rennes



Appliances sur le cloud



Appliance : machine préconfigurée disponible sur le cloud, dédiée aux traitements de données biologiques.

Elles sont :

- rapides à lancer;
- prête à l'emploi;
- personnalisable en fonction des besoins (manuellement, avec des scripts) ;
- petites tailles.

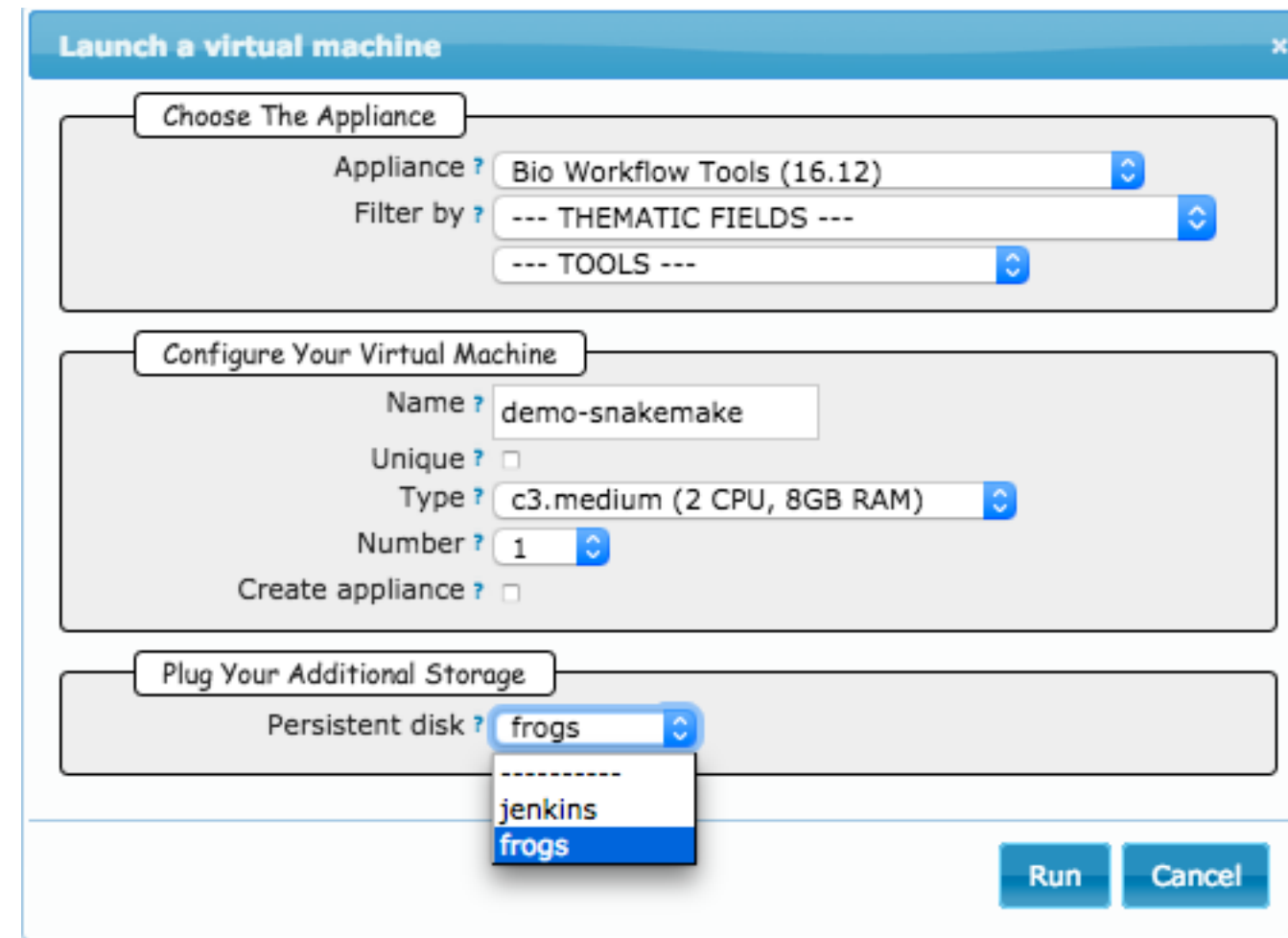
Elles peuvent être utilisées :

- lancer des analyses sur ces données;
- tester des outils;
- environnements de développements.

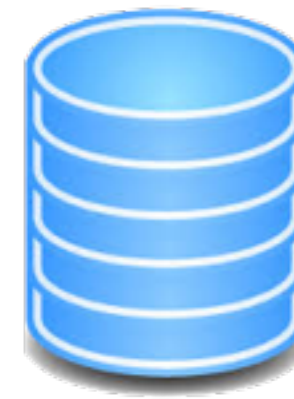
Les VMs reposent sur une Centos (6 ou 7) ou une Ubuntu 14.04.

Appliance Bio workflow tools

- Créée à l'occasion de cette journée, elle comprend :
 - ✓ snakemake ;
 - ✓ docker ;
 - ✓ conda ;
 - ✓ ensemble d'outils bio-informatiques.
- Utilisation classique :
 - ✓ machine virtuelle instanciée à partir de l'image ;
 - ✓ disque virtuel pour le traitement des données.



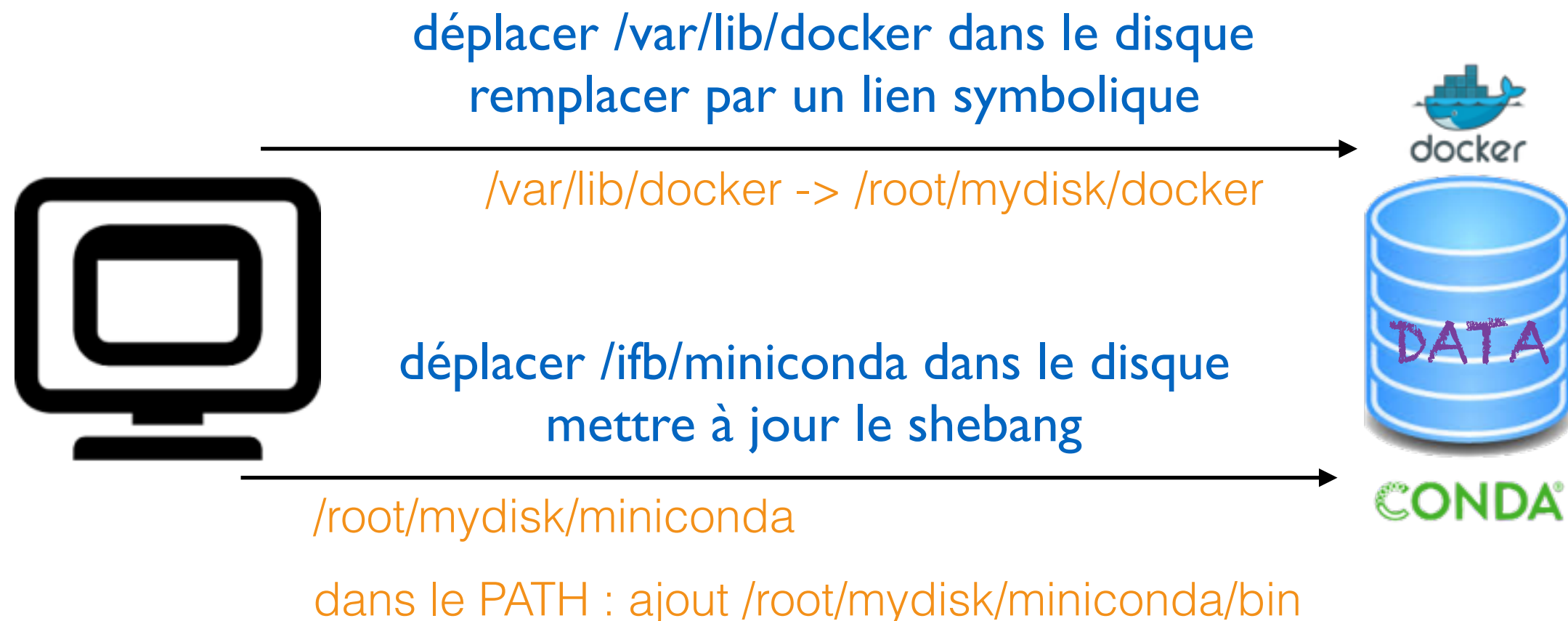
machine
virtuelle



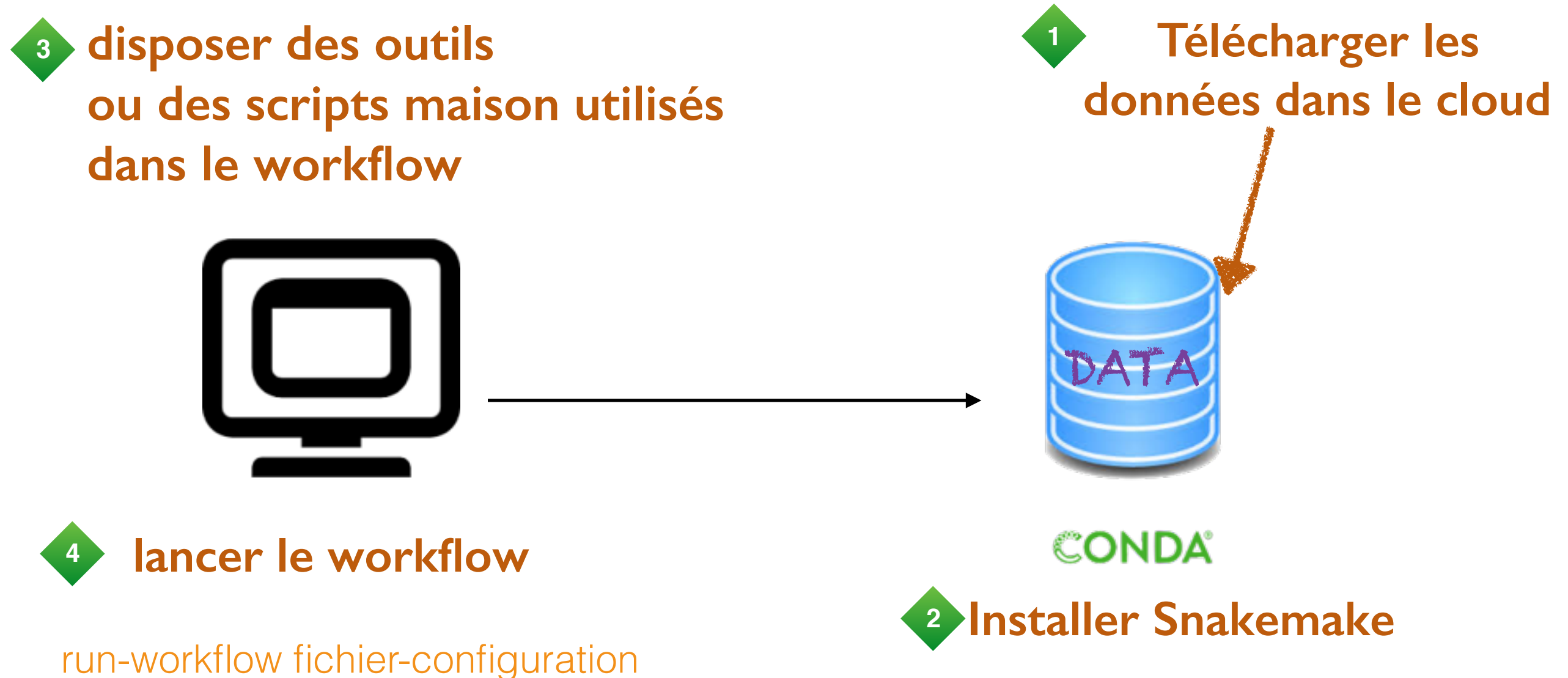
disque
virtuel

Optimisation de l'instance

- Machine virtuelle : dédiée au calcul, peu d'espace disque pour le stockage des données.
- Conda et Docker consomment beaucoup d'espace s'il y a beaucoup d'outils installés.
 - ➔ déplacer les dossiers /var/lib/docker, <path>/conda dans le disque virtuel ;
 - ➔ conserver les données, transférable entre machines virtuelles.



Workflow Snakemake

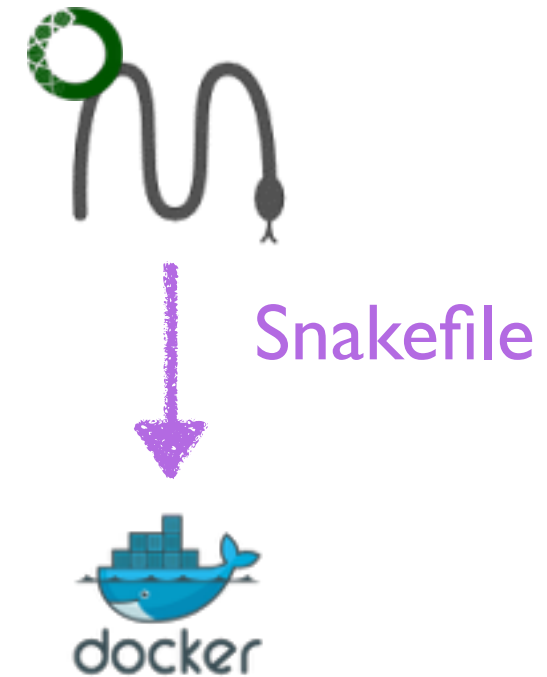


Etape 3: optimiser en utilisant des gestionnaires de packages type conda ou autres solutions telles que docker.

Snakemake & docker : cas 1

Cas 1: aucun outil à installer, pré-requis uniquement docker et snakemake dans la machine virtuelle.

Les outils du workflow sont lancés avec des images docker, il est possible de packager les scripts maisons dans une image docker.



rule docker_h:

input:

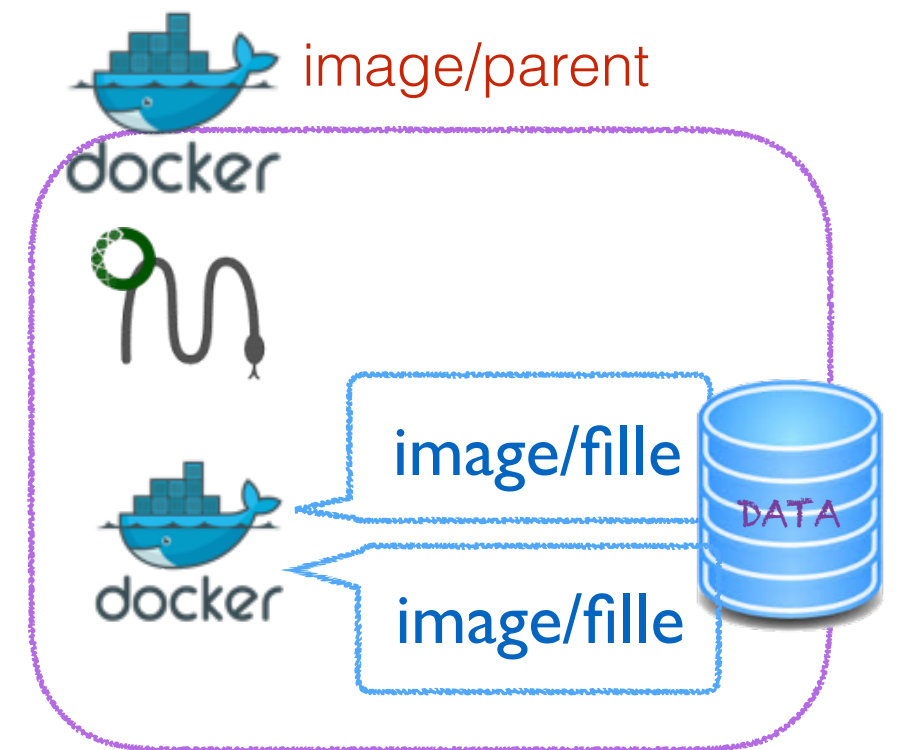
"reads_1.fq"

shell:

```
"docker run -v $DIR:/data docker-registry.genouest.org/ifb/tophat:2.1.0 tophat -p 20  
test_ref {input}"
```

Snakemake & docker : cas 2

Cas 2: lancer le workflow avec un docker qui va lancer les dockers du workflow.
Un seul pré-requis, docker; le lancement peut se faire en une seule ligne de commande.



FROM jpetazzo/dind

RUN < install dependencies >
COPY Snakefile /usr/lib
COPY script.sh /usr/lib



VOLUME /data/

CMD ["script.sh"]

FROM ubuntu:14.04

RUN < install snakemake >
RUN < install docker, same version host os >

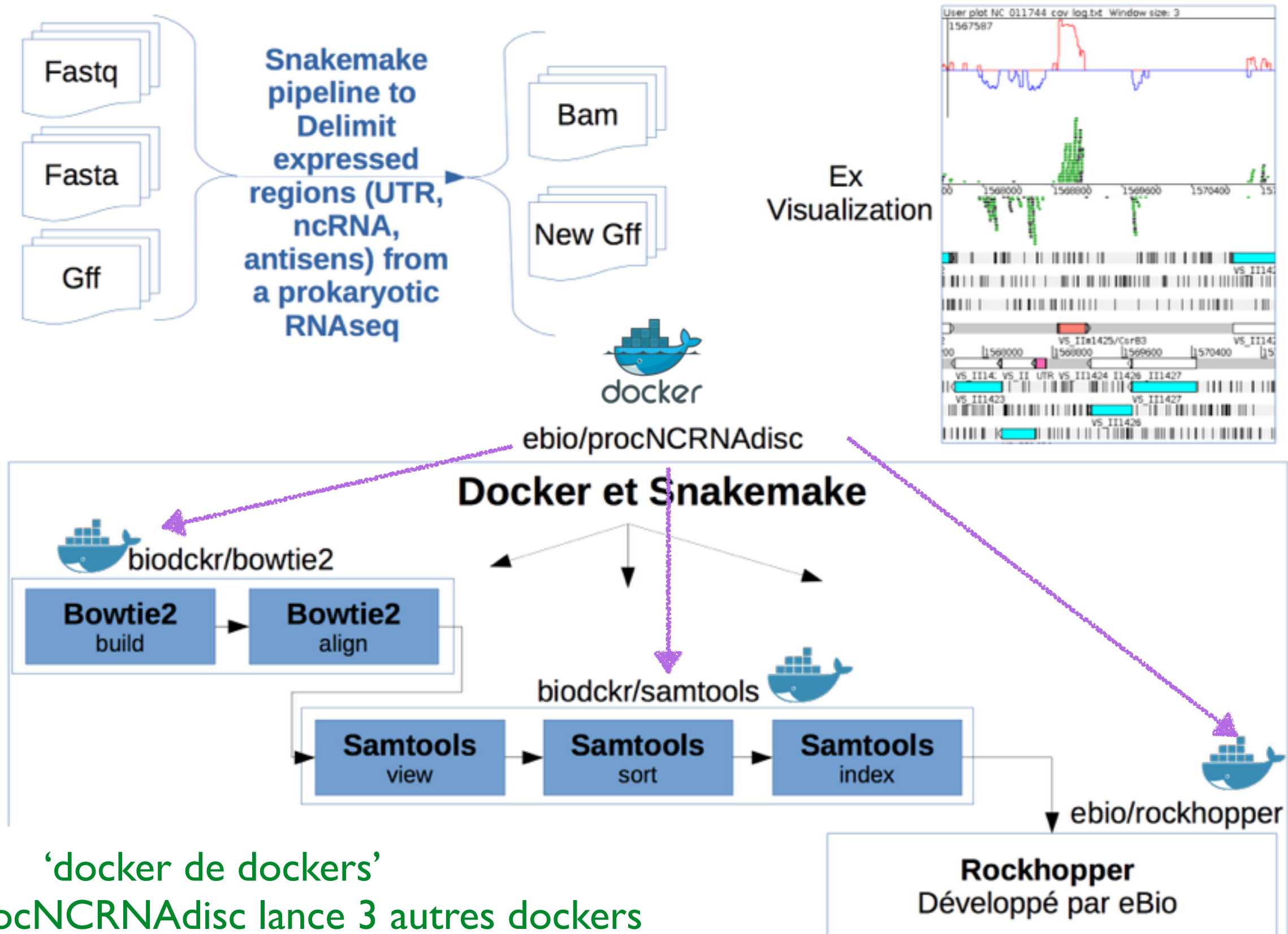


VOLUME /data/

`docker run --privileged -v $DIR:/data image script.sh /data/config.txt`

`docker run -v /var/run/docker.sock:/var/run/docker.sock -v $DIR:/data image /data/script.sh /data/config.txt`

Workflow pour rechercher des régions exprimées non annotées, sRNA avec Rockhopper développé par Emilie Drouineau.



Remarques

- **lancer des dockers dans un docker permet de simplifier l'utilisation d'un workflow en supprimant l'étape d'installation de logiciels, sauf Docker ;**
- **utiliser Docker améliore :**
 - ✦ la reproductibilité des résultats ;
 - ✦ la traçabilité des analyses;
 - ✦ la maintenance du workflow ;
 - ✦ la diffusion du workflow ;
- **utiliser le cloud IFB permet de lancer le workflow dans un cluster adapté aux échantillons à traiter;**
- **travailler dans une machine virtuelle permet de limiter les risques.**

Questions ?



Remerciements à

Emilie Drouineau et Claire Toffano-Nioche
pour le workflow RNASeq bactérien

emilie.drouineau@i2bc.paris-saclay.fr

claire.toffano-nioche@u-psud.fr

- IFB acknowledges funding by
 - the call “**Infrastructures in Biology and Health**” in the framework of the French “Investments for the Future” (ANR-11-INBS-0013) initiative,
 - and EU H2020 projects **CYCLONE** (644925), **EXCELERATE** (676559) and **EGI-Engage** (654142).



Workflow procNCRNAdisc : Snakefile

rule samtools_view:

input:

s = rules.bowtie_map.output

output:

temp("/data/" + "{prefix}_tmp.bam"),

log:

l = ("/data/" + config["log_directory"] +
"log_view_{prefix}.txt")

message:

"run samtools_view: {input} -> {output}"

params:

workDir = checkPath("/data/"),

workDirDock = "/data",

sam = "{prefix}.sam",

bam = "{prefix}_tmp.bam"

shell:

""" /usr/bin/time -v \

docker run --rm -v {params.workDirDock}:{params.workDirDock}:rw

biodckr/samtools samtools view -bSh {params.sam} -o {params.bam} > {log}

"""

Workflow procNCRNAdisc : Dockerfile du docker parent

dépendances

```
FROM jpetazzo/dind
MAINTAINER ebio <plt-ebio@i2bc.paris-saclay.fr>
USER root

RUN apt-get -y update && apt-get install time
RUN mkdir /usr/pipeline

RUN apt-get -y install python3-pip && \
    pip3 install begins && \
    pip3 install snakemake
```

scripts utiles

```
COPY procNCRNAdisc.py /usr/pipeline
COPY Snakefile /usr/pipeline
ENV PATH=$PATH:/usr/pipeline

RUN chmod +x /usr/pipeline/procNCRNAdisc.py
VOLUME /data/

CMD ["python3","/usr/pipeline/procNCRNAdisc.py","-h"]
```

`docker run --privileged -v $(pwd):/data ebio/procncrnadisc procNCRNAdisc.py /data/config_file.txt`