nextflow

CZI EOSS

December 2020

seqeralabs

Project Mission

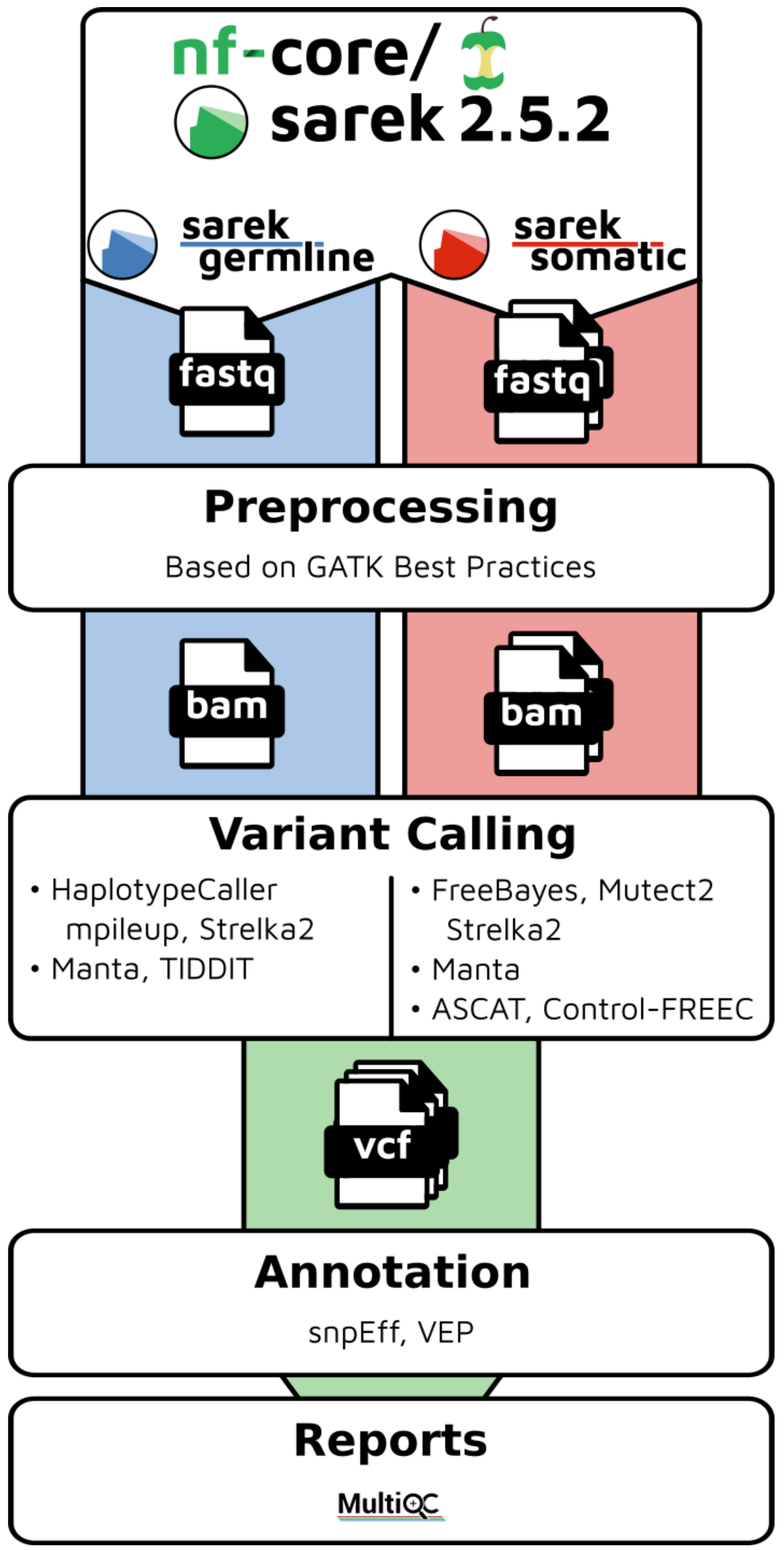
- Deploy anywhere, securely
- Enable best-practise software engineering
- Operate from GUI to API

seqeralabs

Seamless data pipelines at scale

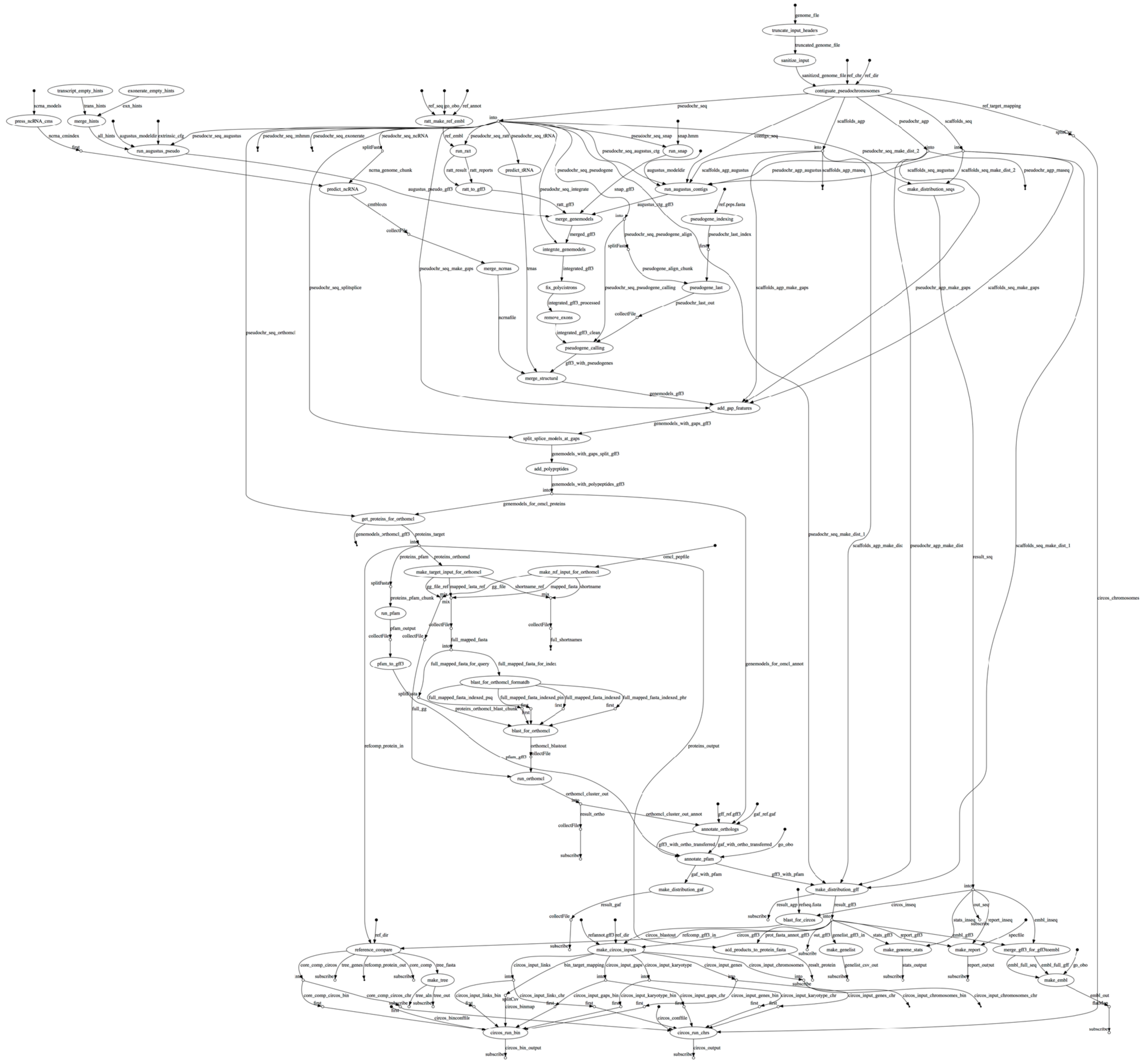
NGS Variant Calling

https://github.com/nf-core/sarek



Parasite Genome Annotation

https://github.com/sanger-pathogens/companion



seqeralabs

What is Nextflow?

nextflow script

Write code
in any language.

R BASH THE BOURNE-AGAIN SHELL

Define orchestration with
dataflow programming.

Define software
dependencies
with containers.

CONDA docker

Git GitHub GitLab Bitbucket

Version
control.

nextflow runtime

Orchestration of tasks to
deploy anywhere with ease.

aws

HTCondor High Throughput Computing

GRID ENGINE

slurm workload manager

Platform Computing an IBM Company

kubernetes

PBS Works™

Task example

```
bwa mem reference.fa sample.fq \
      | samtools sort -o sample.bam
```

Task example

```
process align_sample {

   input:
   path ‘reference.fa’ from genome_ch
   path 'sample.fq' from reads_ch

   output:
   path 'sample.bam' into bam_ch


   script:
   """
   bwa mem reference.fa sample.fq \
         | samtools sort -o sample.bam
   """

}
```

Task composition

```
process align_sample {

   input:
   file 'reference.fa' from genome_ch
   file 'sample.fq' from reads_ch

   output:
   file 'sample.bam' into bam_ch


   script:
   """
   bwa mem reference.fa sample.fq \
         | samtools sort -o sample.bam
   """

}
```

```
process index_sample {

   input:
   file 'sample.bam' from bam_ch

   output:
   file 'sample.bai' into bai_ch


   script:
   """
   samtools index sample.bam
   """
}
```

How does it work?

- **Fast prototyping** ⇒ custom DSL that enables tasks composition, simplifies most use cases + general purpose programming language for corner cases
- **Easy parallelisation** ⇒ declarative reactive programming model based on dataflow paradigm, implicit portable parallelism
- **Self-contained** ⇒ functional approach, a task execution is idempotent ie. cannot modify the state of other tasks + isolate dependencies with containers
- **Portable deployments** ⇒ executor abstraction layer + deployment configuration from implementation logic

DSL2

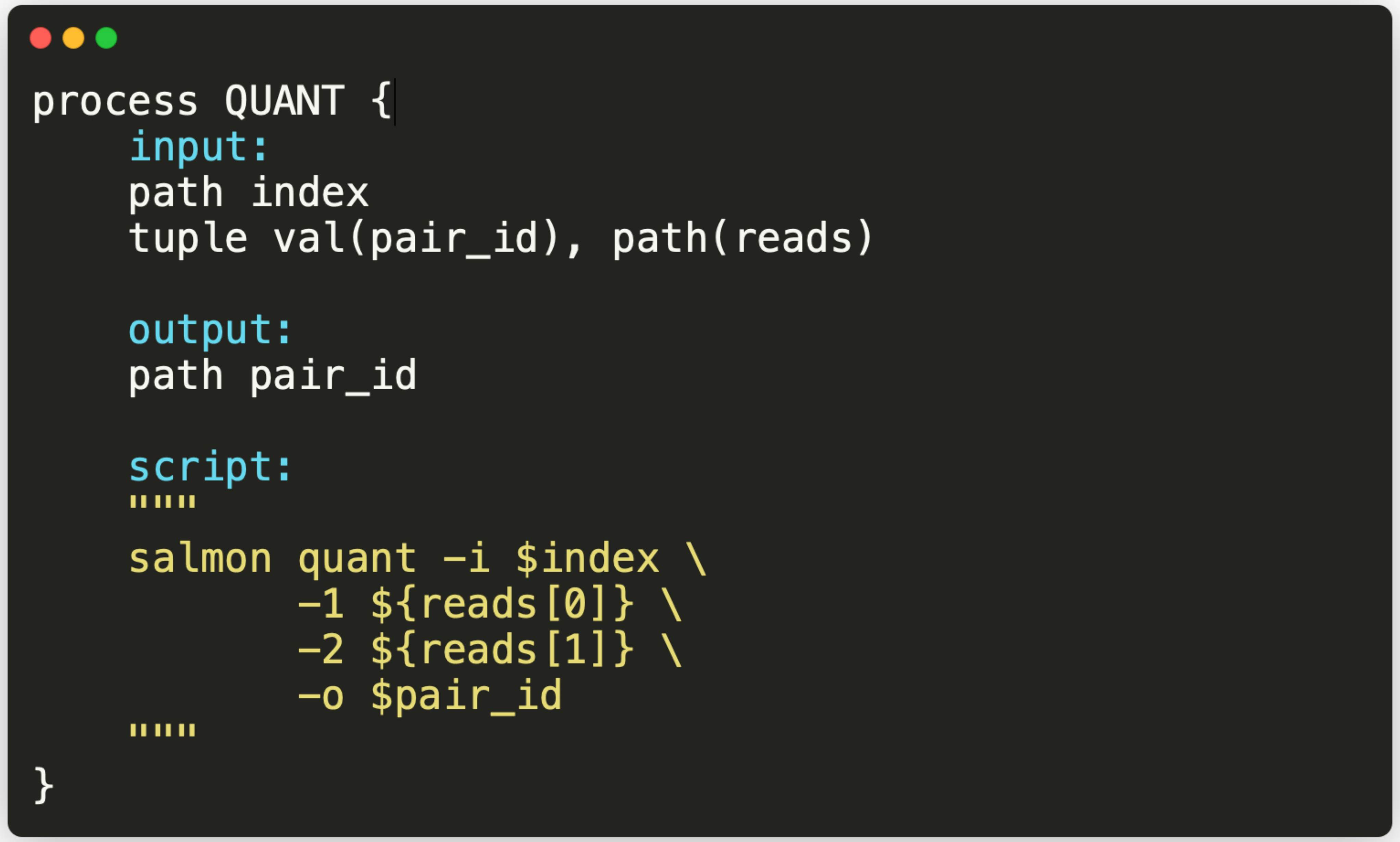
A major revision of the Nextflow DSL

- Pipeline modularisation
- Component reuse
- Fluent definition of recurrent implementation patterns

seqeralabs

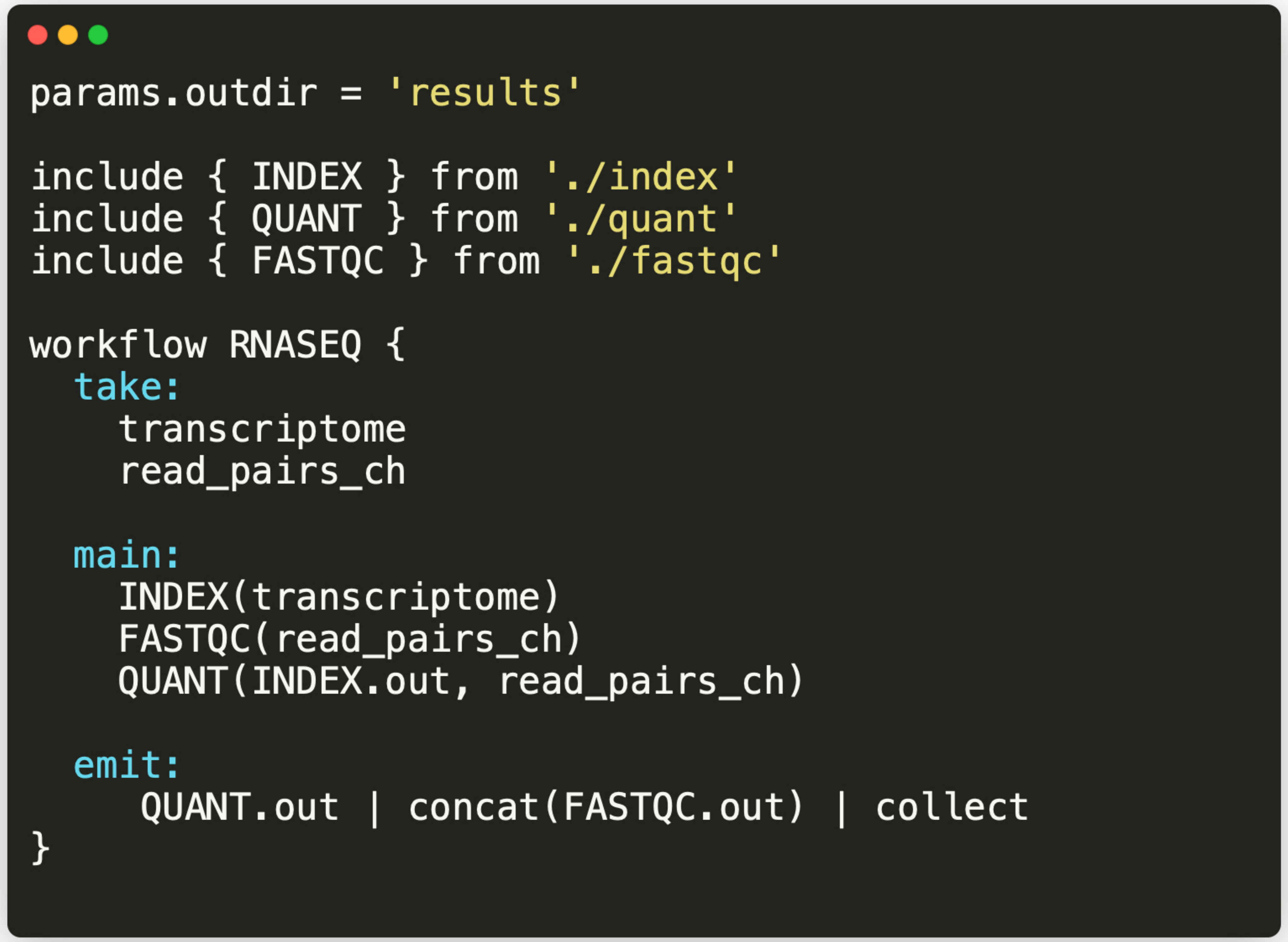
Nextflow syntax - DSL 2

task



```
process QUANT {
    input:
    path index
    tuple val(pair_id), path(reads)

    output:
    path pair_id

    script:
    """
    salmon quant -i $index \
           -1 ${reads[0]} \
           -2 ${reads[1]} \
           -o $pair_id
    """
}
```

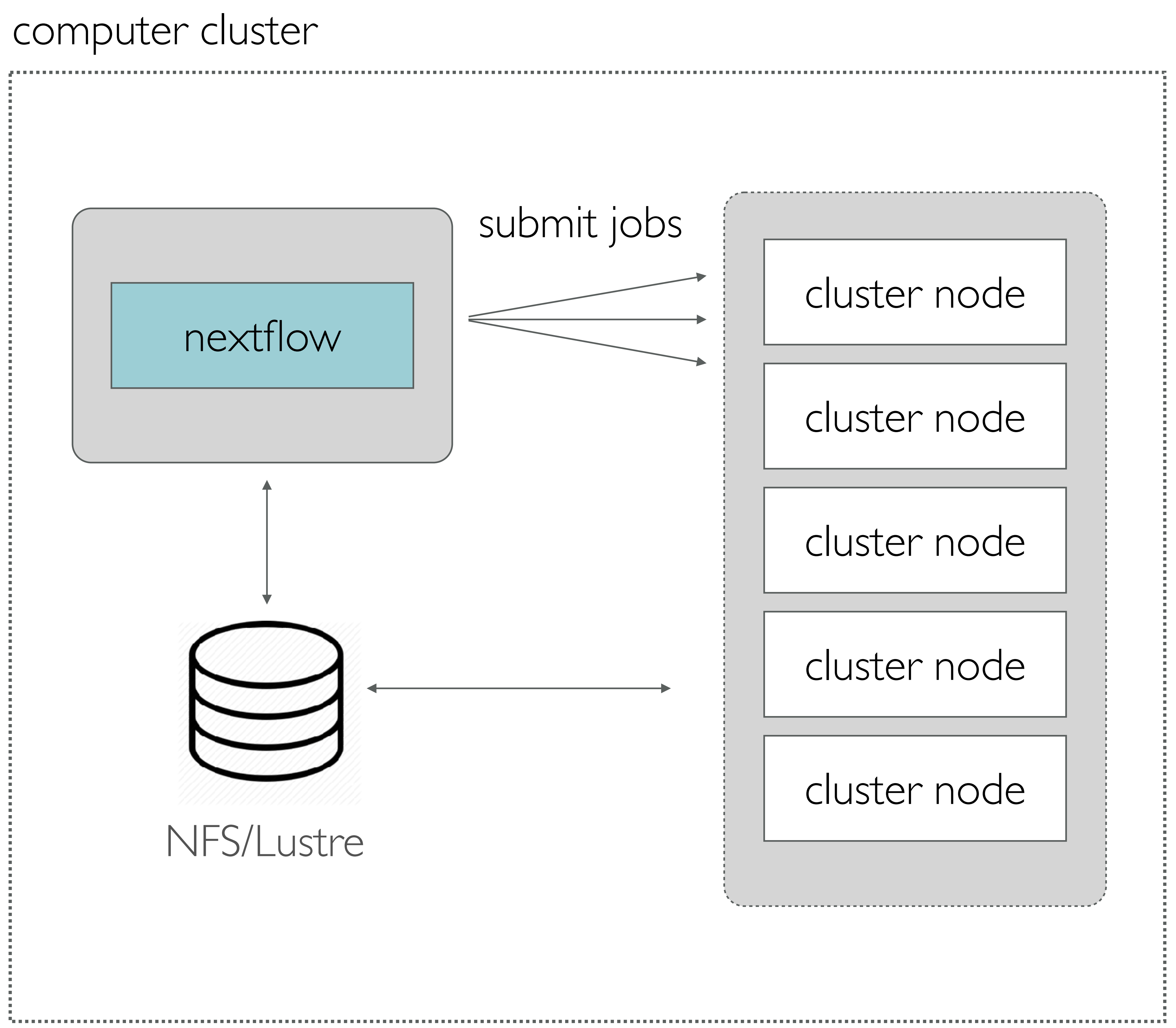
workflow



```
params.outdir = 'results'

include { INDEX } from './index'
include { QUANT } from './quant'
include { FASTQC } from './fastqc'

workflow RNASEQ {
  take:
    transcriptome
    read_pairs_ch

  main:
    INDEX(transcriptome)
    FASTQC(read_pairs_ch)
    QUANT(INDEX.out, read_pairs_ch)

  emit:
     QUANT.out | concat(FASTQC.out) | collect
}
```

Scientist and engineers can now write complex, distributed and parallel data pipelines without requiring a degree in computer science.

Centralised cluster orchestration

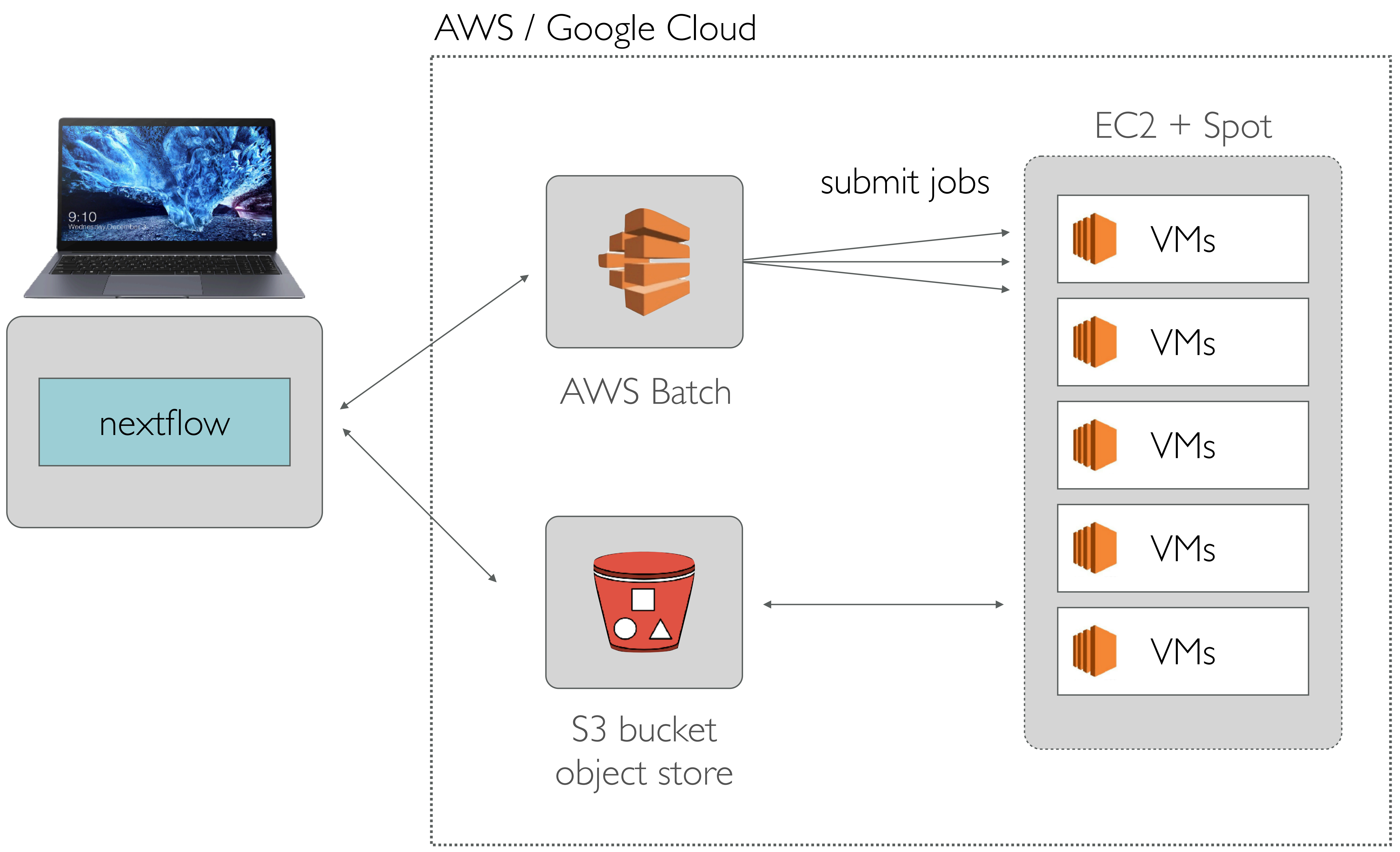


- Nextflow orchestrates workflow execution submitting jobs to a compute schedular
- Can run in the head node or a compute node
- Requires a shared storage to exchange data between tasks
- Ideal for corse-grained parallelism

seqeralabs

slurm workload manager PBS Works GRID ENGINE HTCondor Platform Computing an IBM Company

Cloud orchestration

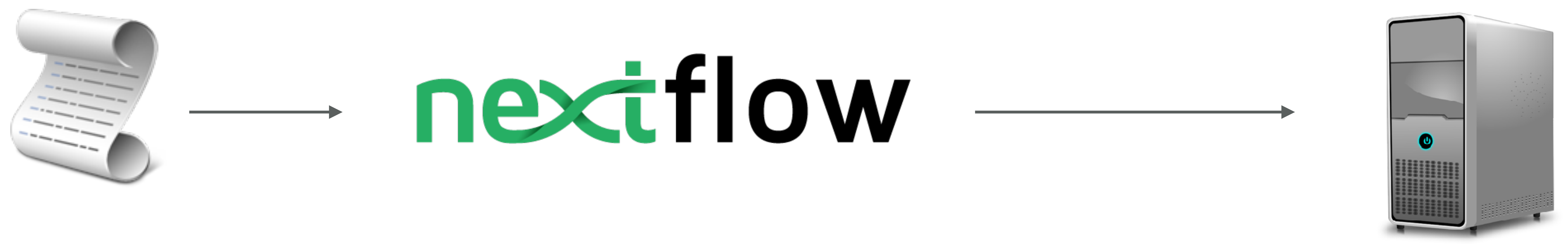


- Nextflow orchestrates workflow execution via AWS Batch
- Launched workflow from anywhere into the cloud
- Transfer of data between local environment and cloud storage
- Requires a shared object storage to exchange data between VMs.

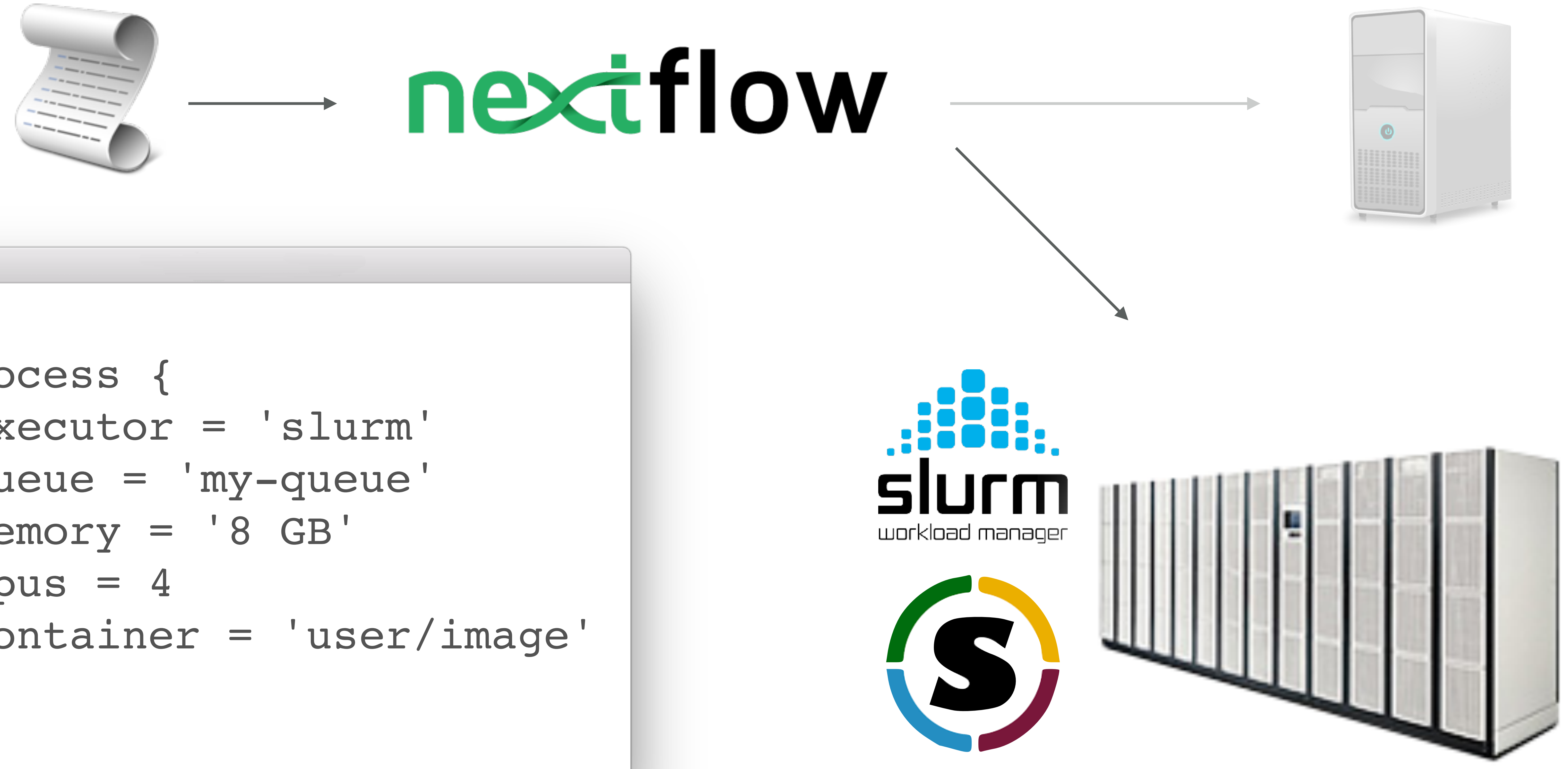
kubernetes Azure aws

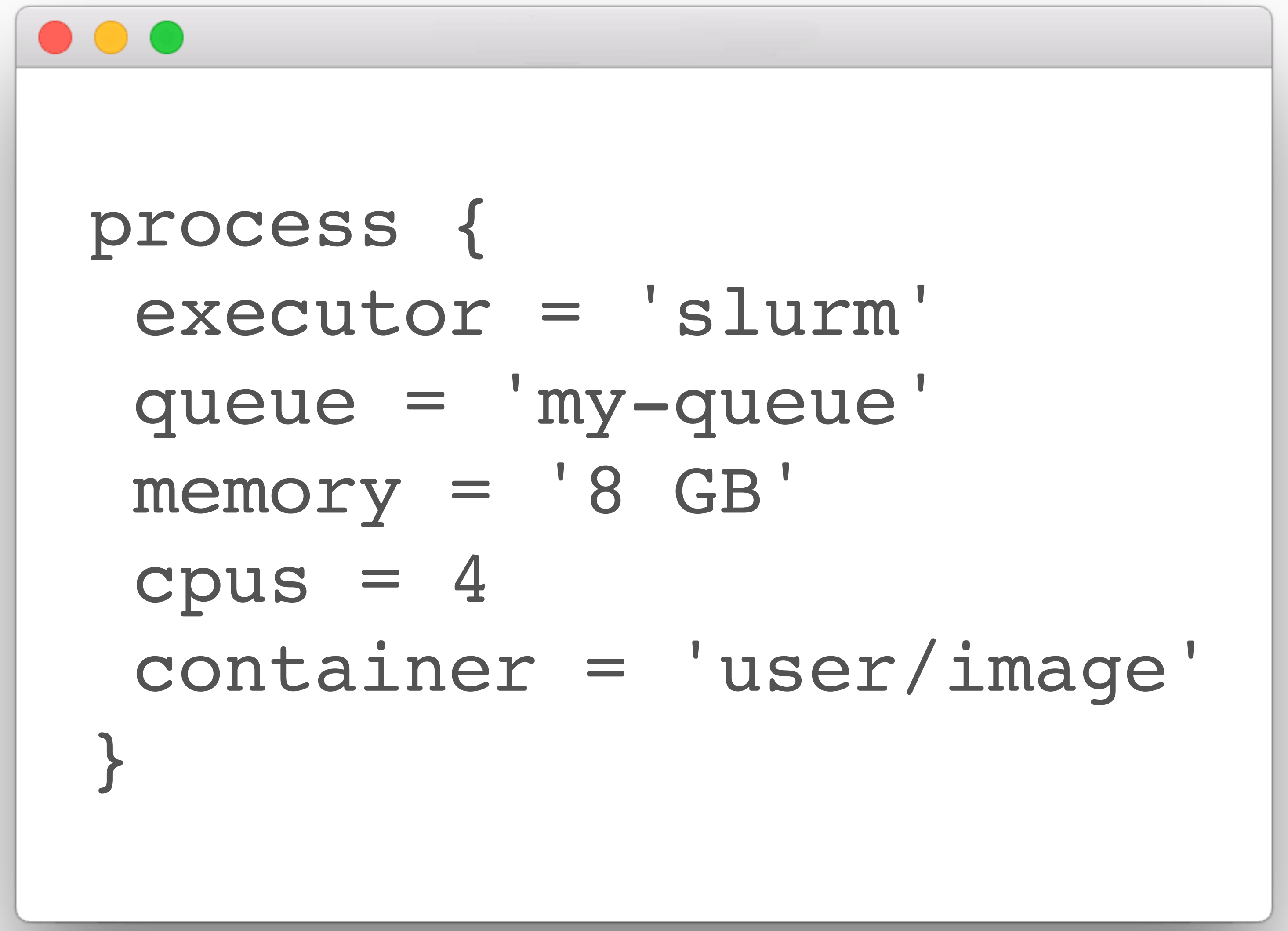
seqeralabs

PORTABILITY



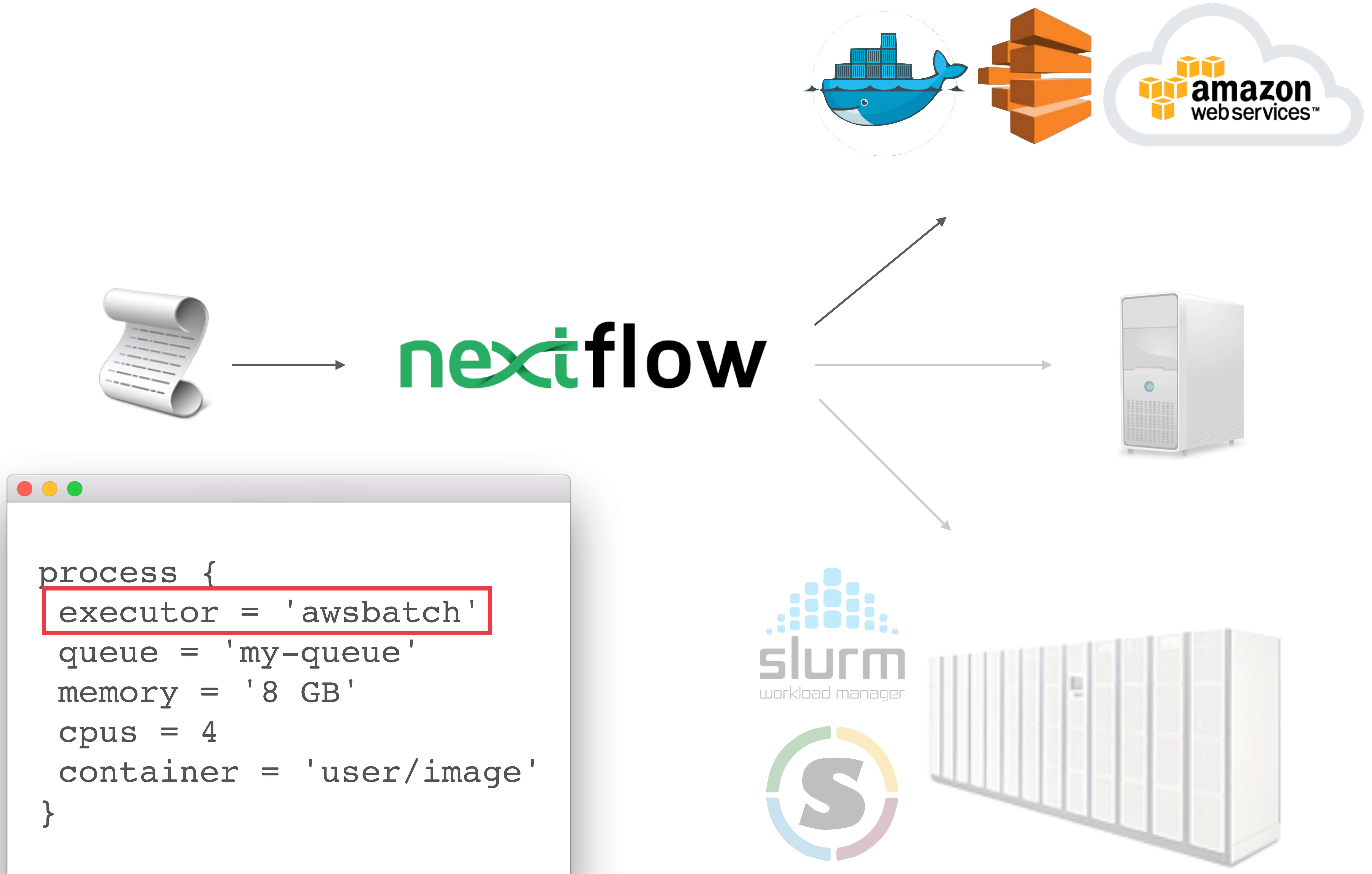
PORTABILITY

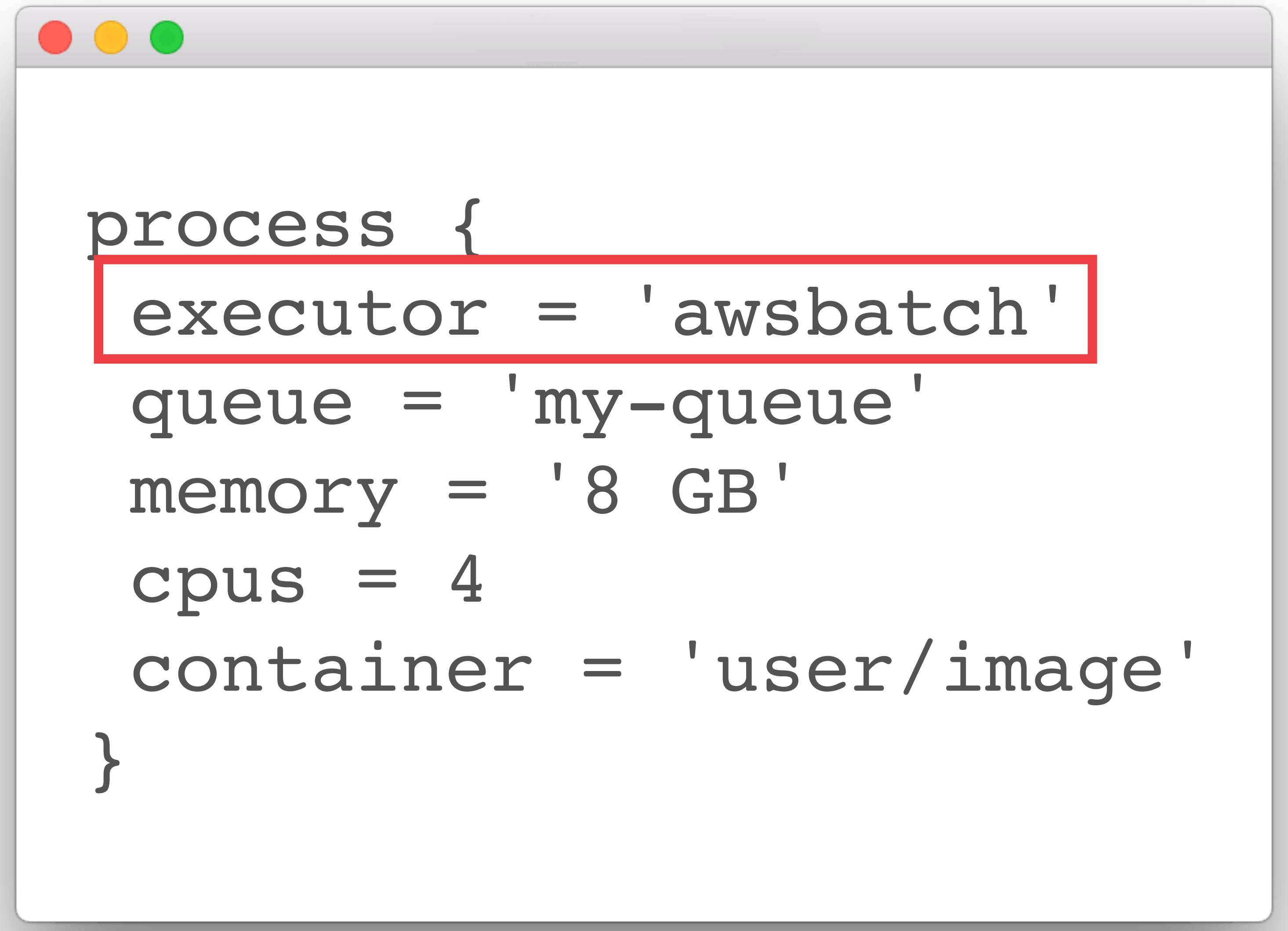




```
process {
 executor = 'slurm'
 queue = 'my-queue'
 memory = '8 GB'
 cpus = 4
 container = 'user/image'
}
```

PORTABILITY





```
process {
  executor = 'awsbatch'
  queue = 'my-queue'
  memory = '8 GB'
  cpus = 4
  container = 'user/image'
}
```

OPEN SOURCE COMMUNITY DRIVEN

45K+

monthly downloads

6,800+

active developers /month

18

international Workshops

150 k+

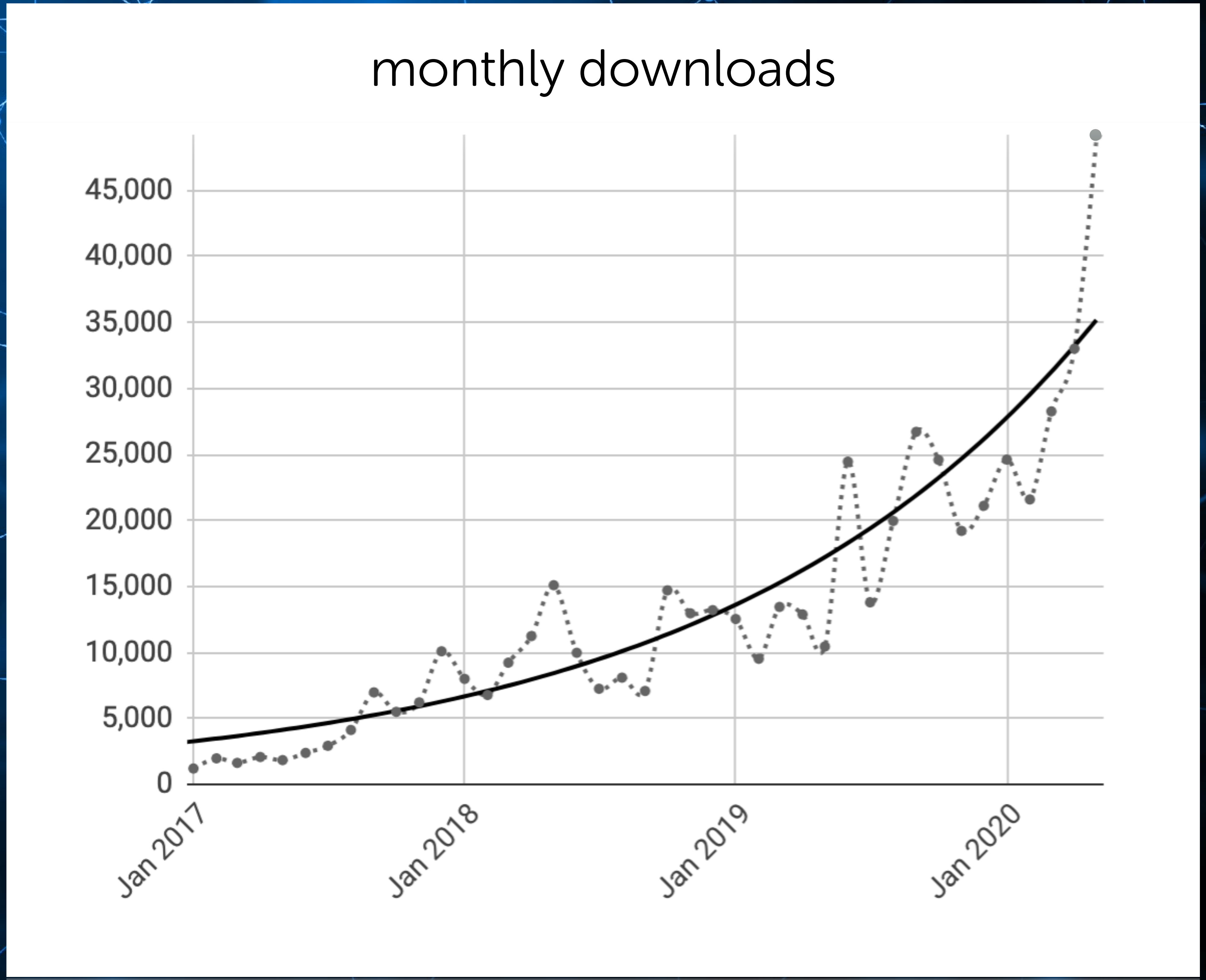
lines of code

1.3k+

stars on github

80

contributors



Enterprise adoption

AstraZeneca
Roche
FRED HUTCH CURES START HERE™
MERCK INVENTING FOR LIFE
Boehringer Ingelheim
THE FRANCIS CRICK INSTITUTE

EMBL-EBI
wellcome sanger institute
WuXiNextCODE Leading Genomics
Institut Pasteur
IQVIA™
UCSF University of California San Francisco

FABRIC GENOMICS™
Genentech A Member of the Roche Group
SciLifeLab
CHAN ZUCKERBERG BIOHUB
enterome bioscience

BASECLEAR
KAROLINSKA INSTITUTET ANNO 1810 Karolinska Institutet
CANCER RESEARCH UK
International Agency Research on Cancer World Health Organization
IMP Research Institute of Molecular Pathology
SINGAPORE Agency for Science, Technology and Research

VE RI TAS HARVARD UNIVERSITY
gmi genomics medicine ireland
SOPHiA™
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY
BENAROYA RESEARCH INSTITUTE Virginia Mason
Weill Cornell Medicine

ACHILLES THERAPEUTICS
PAWSEY supercomputing centre
VIR
SYNTHETIC GENOMICS
institutCurie
QBiC Quantitative Biology Center

Berkeley UNIVERSITY OF CALIFORNIA
Kintai THERAPEUTICS
gritstone ONCOLOGY
UTSouthwestern Medical Center®
University of Colorado Boulder
READCOOR

nextflow

data pipelines at scale

massively scalable pipelines across cluster & cloud.

The world's leading workflow software for genomics, biopharma and life sciences.

features.

Powerful HPC execution engines

Deploy across cloud & clusters effortlessly

Out-of-the-box support for AWS and GCP plus schedulers including SLURM, LSF & Grid Engine.

Portable & reproducible

Containers without the hassle

Docker and Singularity integrations encapsulate all pipeline dependencies across environments.

Language agnostic

Write pipelines in your language

The flexibility to develop, port and reuse code allows you to do things your way.

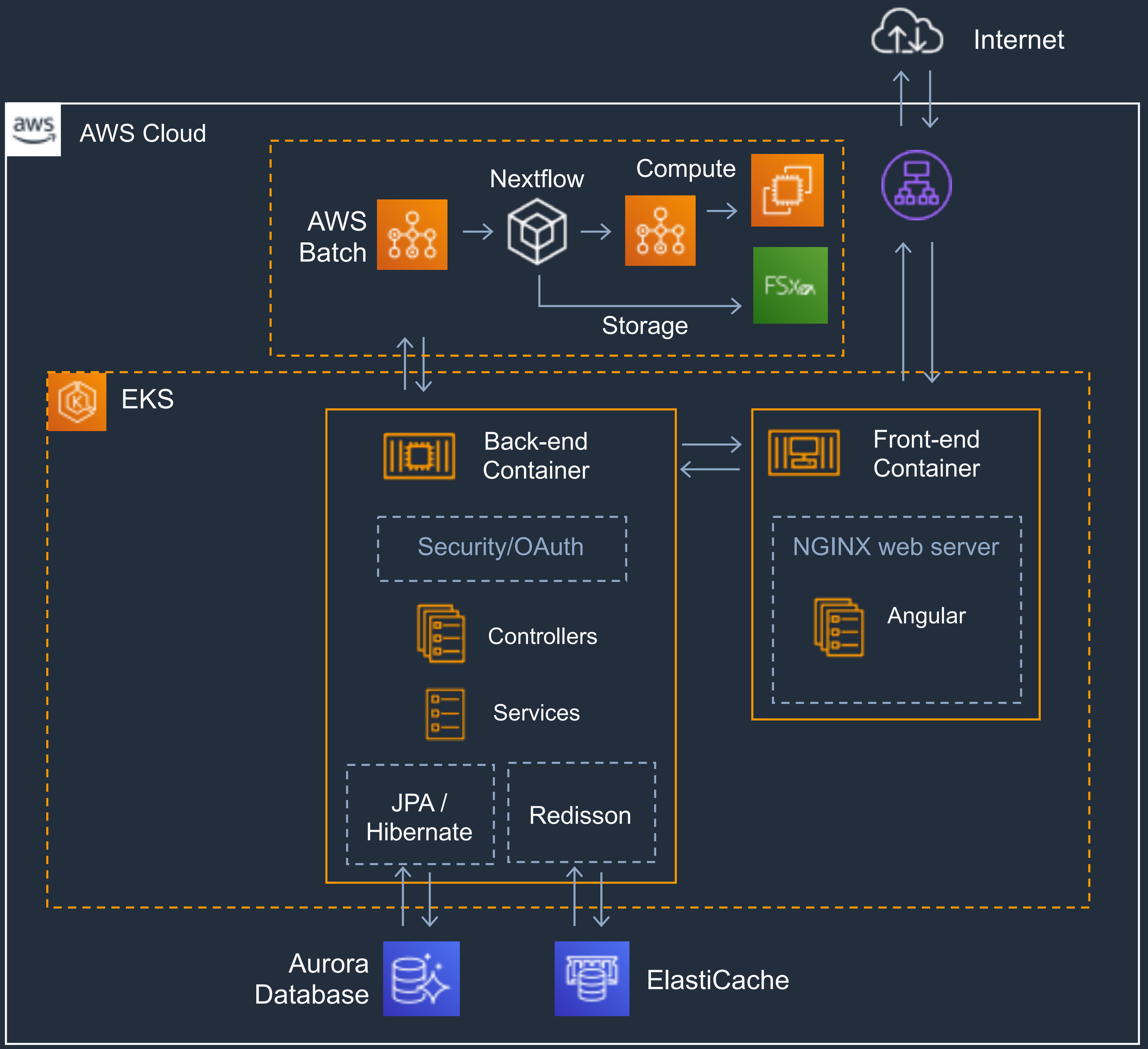
nextflow *tower*

delivering discovery.

Manage, optimize and launch data analysis pipelines from a secure command-post.

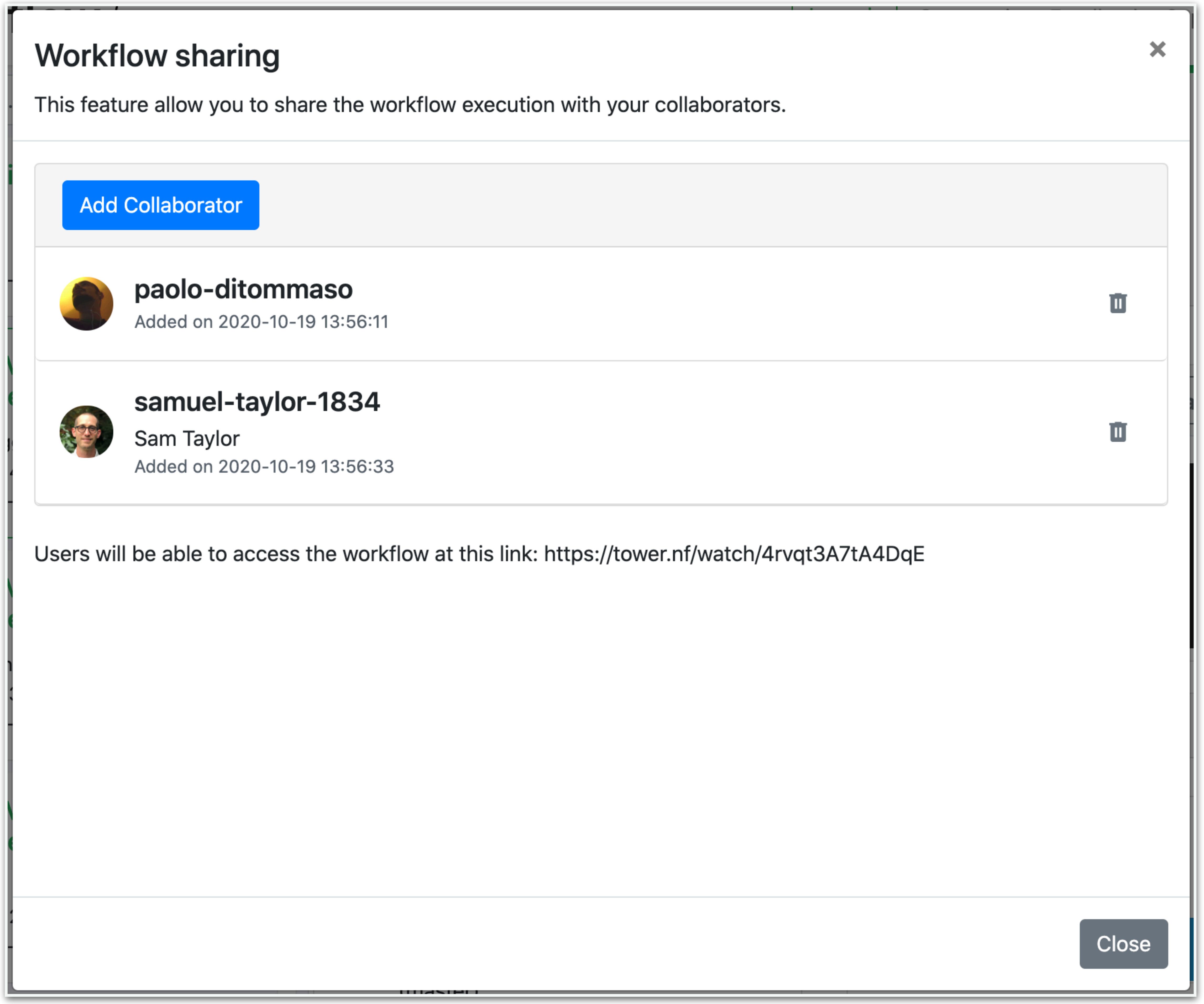
Tower for Nextflow

AWS Reference Deployment



* *Support available for all cloud or on-premise infrastructure.*

Sharing and collaboration



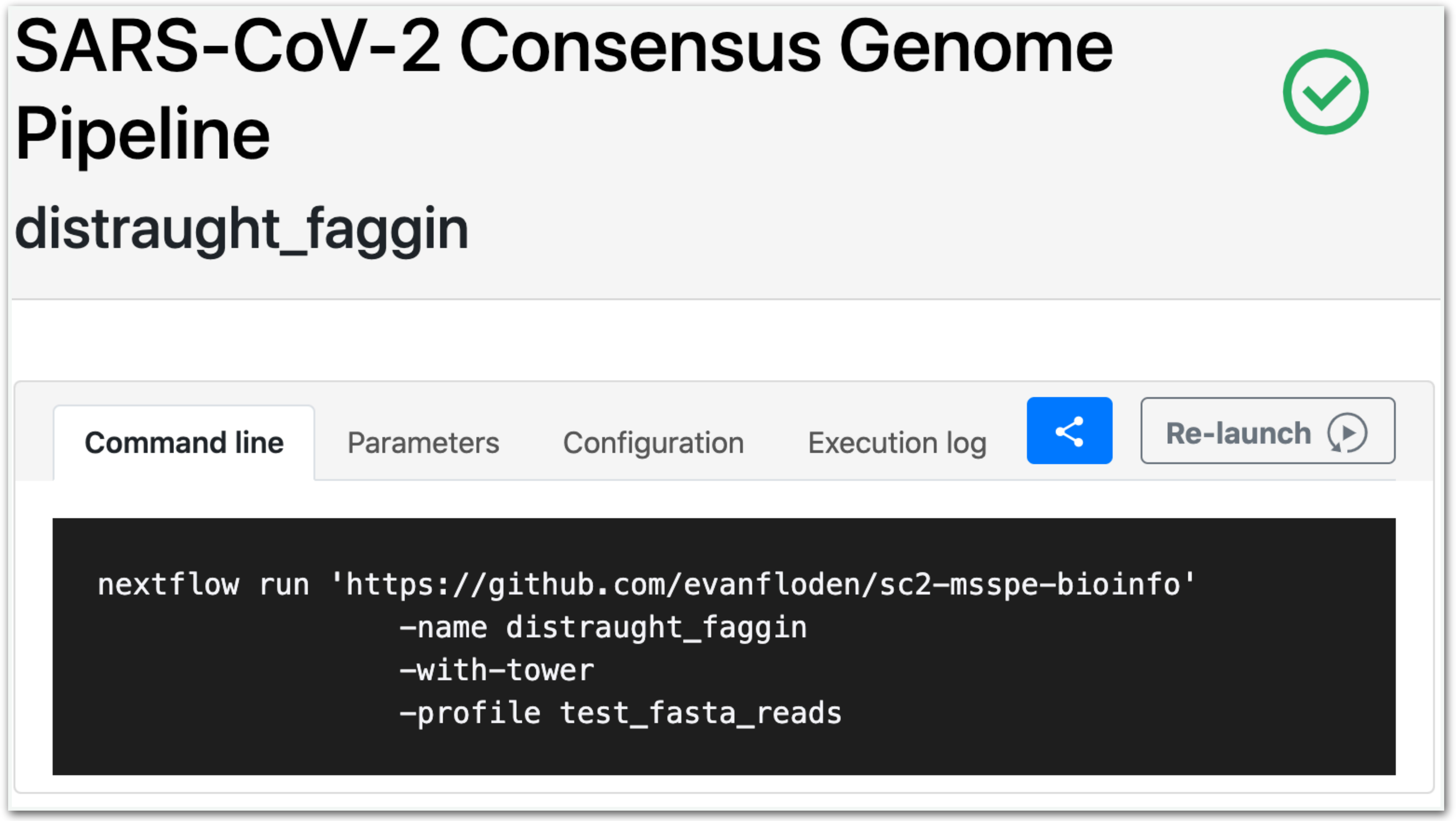
Workflow sharing

This feature allow you to share the workflow execution with your collaborators.

paolo-ditommaso
Added on 2020-10-19 13:56:11

samuel-taylor-1834
Sam Taylor
Added on 2020-10-19 13:56:33

Users will be able to access the workflow at this link: https://tower.nf/watch/4rvqt3A7tA4DqE



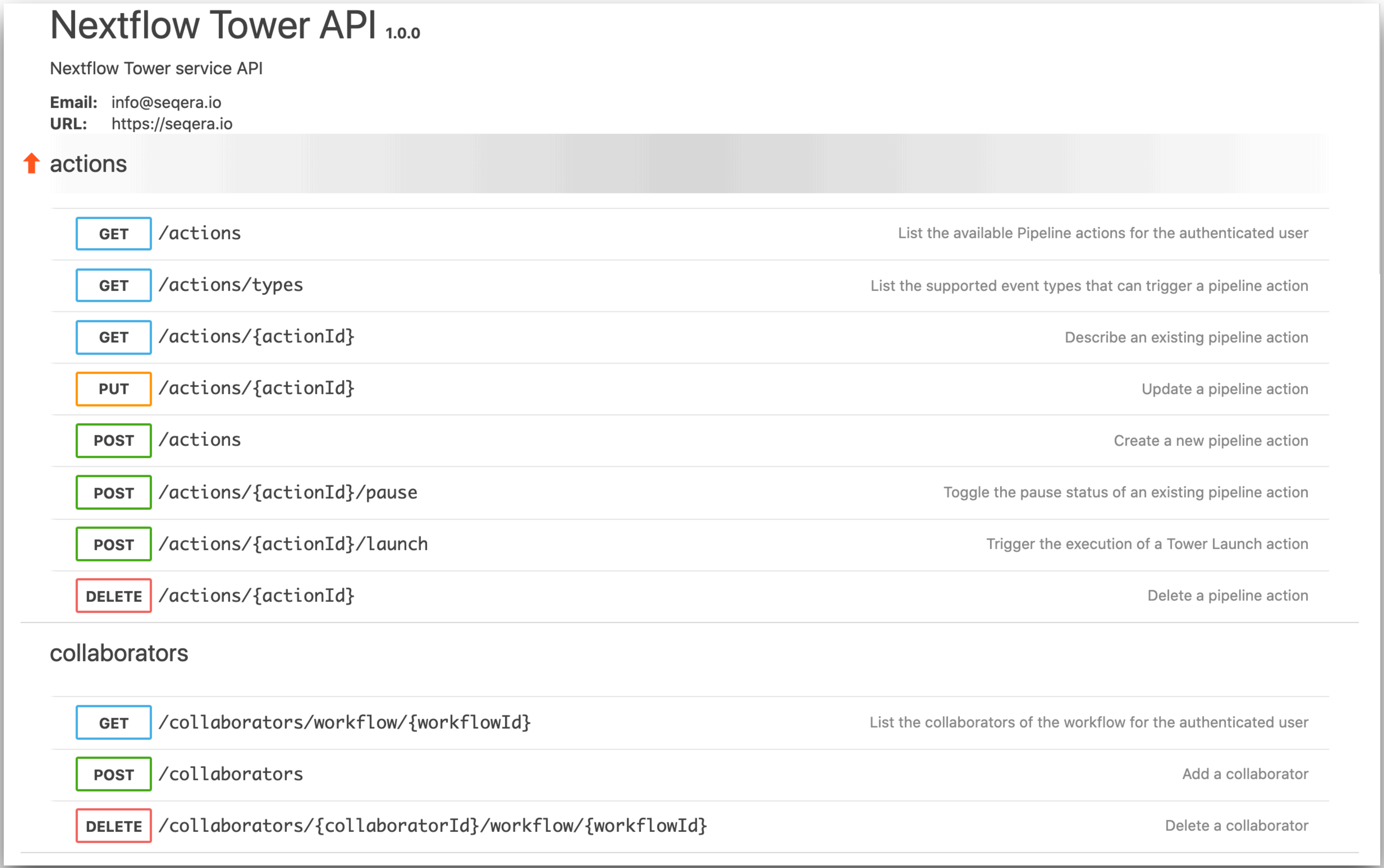
SARS-CoV-2 Consensus Genome Pipeline
distraught_faggin

```
nextflow run 'https://github.com/evanfloden/sc2-msspe-bioinfo'
            -name distraught_faggin
            -with-tower
            -profile test_fasta_reads
```

Share workflow executions with colleagues who can follow along live on the progress of the pipeline or catch up on the results at any time in the future.

seqeralabs

API Release



Nextflow Tower API 1.0.0

Nextflow Tower service API

Email: info@seqera.io
URL: https://seqera.io

actions

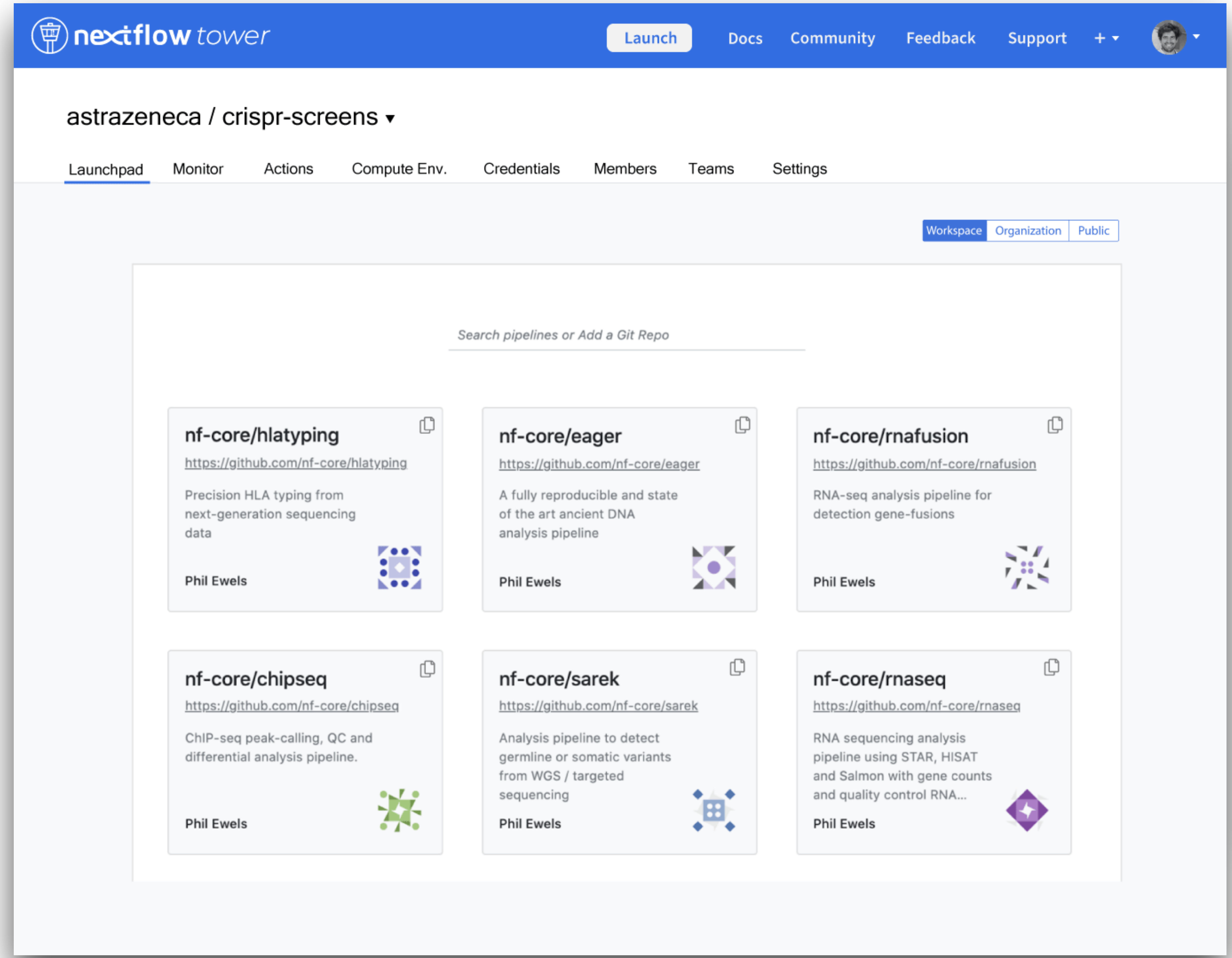
GET	/actions	List the available Pipeline actions for the authenticated user
GET	/actions/types	List the supported event types that can trigger a pipeline action
GET	/actions/{actionId}	Describe an existing pipeline action
PUT	/actions/{actionId}	Update a pipeline action
POST	/actions	Create a new pipeline action
POST	/actions/{actionId}/pause	Toggle the pause status of an existing pipeline action
POST	/actions/{actionId}/launch	Trigger the execution of a Tower Launch action
DELETE	/actions/{actionId}	Delete a pipeline action

collaborators

GET	/collaborators/workflow/{workflowId}	List the collaborators of the workflow for the authenticated user
POST	/collaborators	Add a collaborator
DELETE	/collaborators/{collaboratorId}/workflow/{workflowId}	Delete a collaborator

seqeralabs

Pipelines Feature



astrazeneca / crispr-screens ▾

Launchpad Monitor Actions Compute Env. Credentials Members Teams Settings

Workspace Organization Public

Search pipelines or Add a Git Repo

nf-core/hlatyping
https://github.com/nf-core/hlatyping
Precision HLA typing from next-generation sequencing data
Phil Ewels

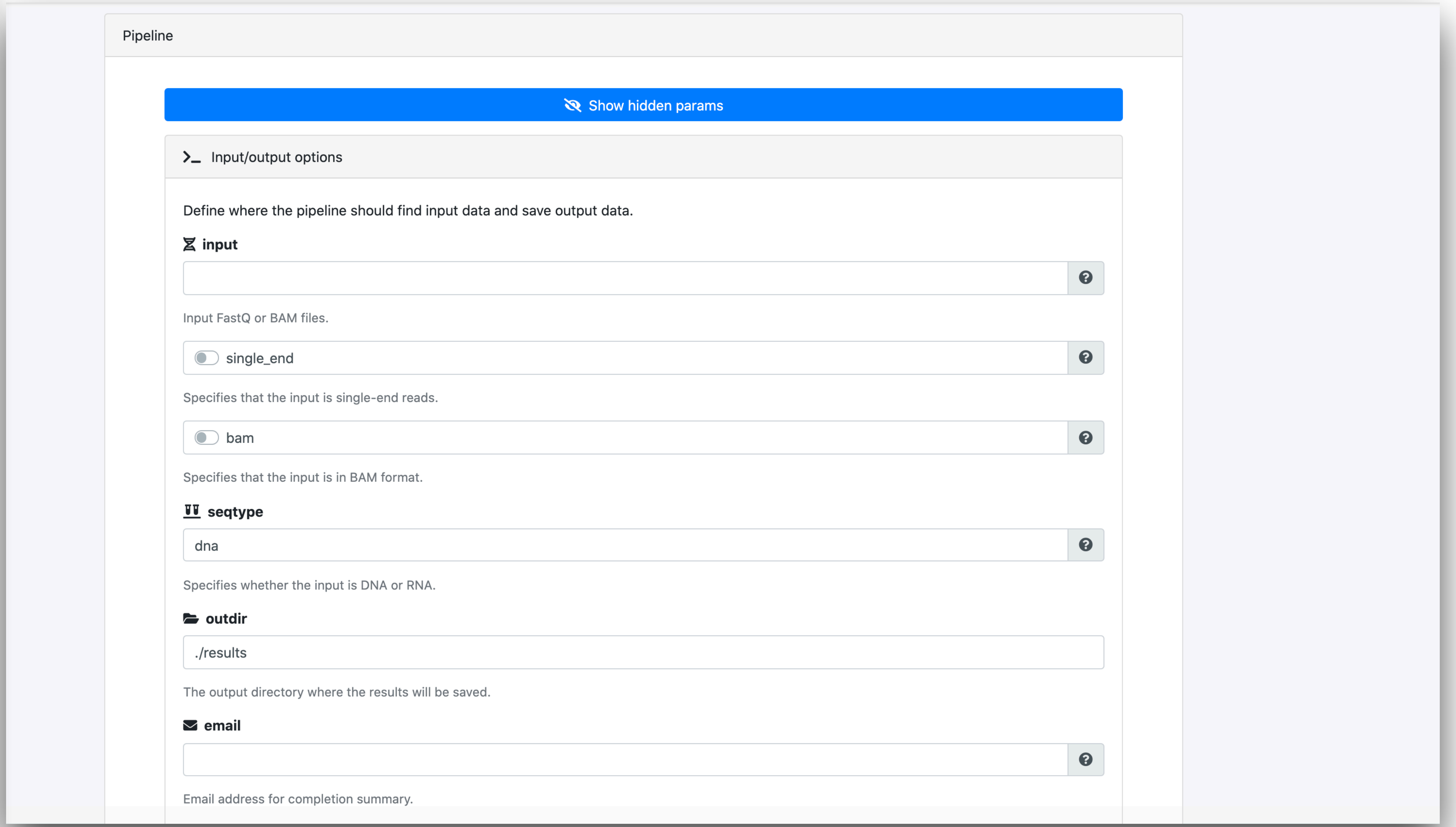
nf-core/eager
https://github.com/nf-core/eager
A fully reproducible and state of the art ancient DNA analysis pipeline
Phil Ewels

nf-core/rnafusion
https://github.com/nf-core/rnafusion
RNA-seq analysis pipeline for detection gene-fusions
Phil Ewels

nf-core/chipseq
https://github.com/nf-core/chipseq
ChIP-seq peak-calling, QC and differential analysis pipeline.
Phil Ewels

nf-core/sarek
https://github.com/nf-core/sarek
Analysis pipeline to detect germline or somatic variants from WGS / targeted sequencing
Phil Ewels

nf-core/rnaseq
https://github.com/nf-core/rnaseq
RNA sequencing analysis pipeline using STAR, HISAT and Salmon with gene counts and quality control RNA...
Phil Ewels



Pipeline

Show hidden params

Input/output options

Define where the pipeline should find input data and save output data.

input

Input FastQ or BAM files.

single_end

Specifies that the input is single-end reads.

bam

Specifies that the input is in BAM format.

seqtype

dna

Specifies whether the input is DNA or RNA.

outdir

./results

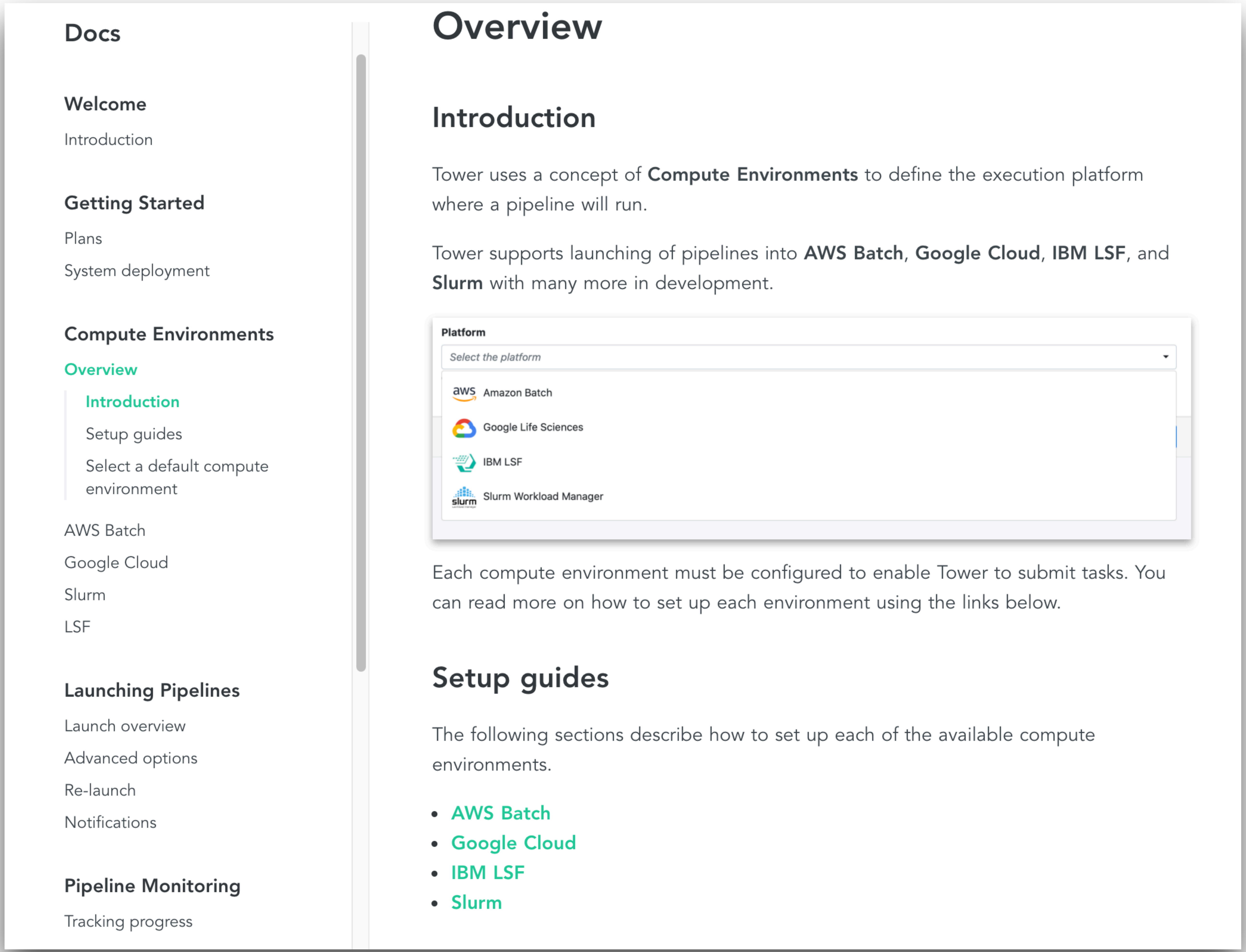
The output directory where the results will be saved.

email

Email address for completion summary.

Preconfigured workflows with input fields based on JSON schema.

Extended documentation



Docs

Welcome

Introduction

Getting Started

Plans

System deployment

Compute Environments

Overview

Introduction

Setup guides

Select a default compute environment

AWS Batch

Google Cloud

Slurm

LSF

Launching Pipelines

Launch overview

Advanced options

Re-launch

Notifications

Pipeline Monitoring

Tracking progress

Overview

Introduction

Tower uses a concept of **Compute Environments** to define the execution platform where a pipeline will run.

Tower supports launching of pipelines into **AWS Batch**, **Google Cloud**, **IBM LSF**, and **Slurm** with many more in development.

Each compute environment must be configured to enable Tower to submit tasks. You can read more on how to set up each environment using the links below.

Setup guides

The following sections describe how to set up each of the available compute environments.

- AWS Batch
- Google Cloud
- IBM LSF
- Slurm

CZI EOSS Goals

WP1

Goal 1.1	Reach 8,000 monthly Nextflow users
Goal 1.2	Reach 150 active nf-core Slack members
Goal 1.3	Sustain community support and advocacy activities
Goal 1.4	Appoint project positions

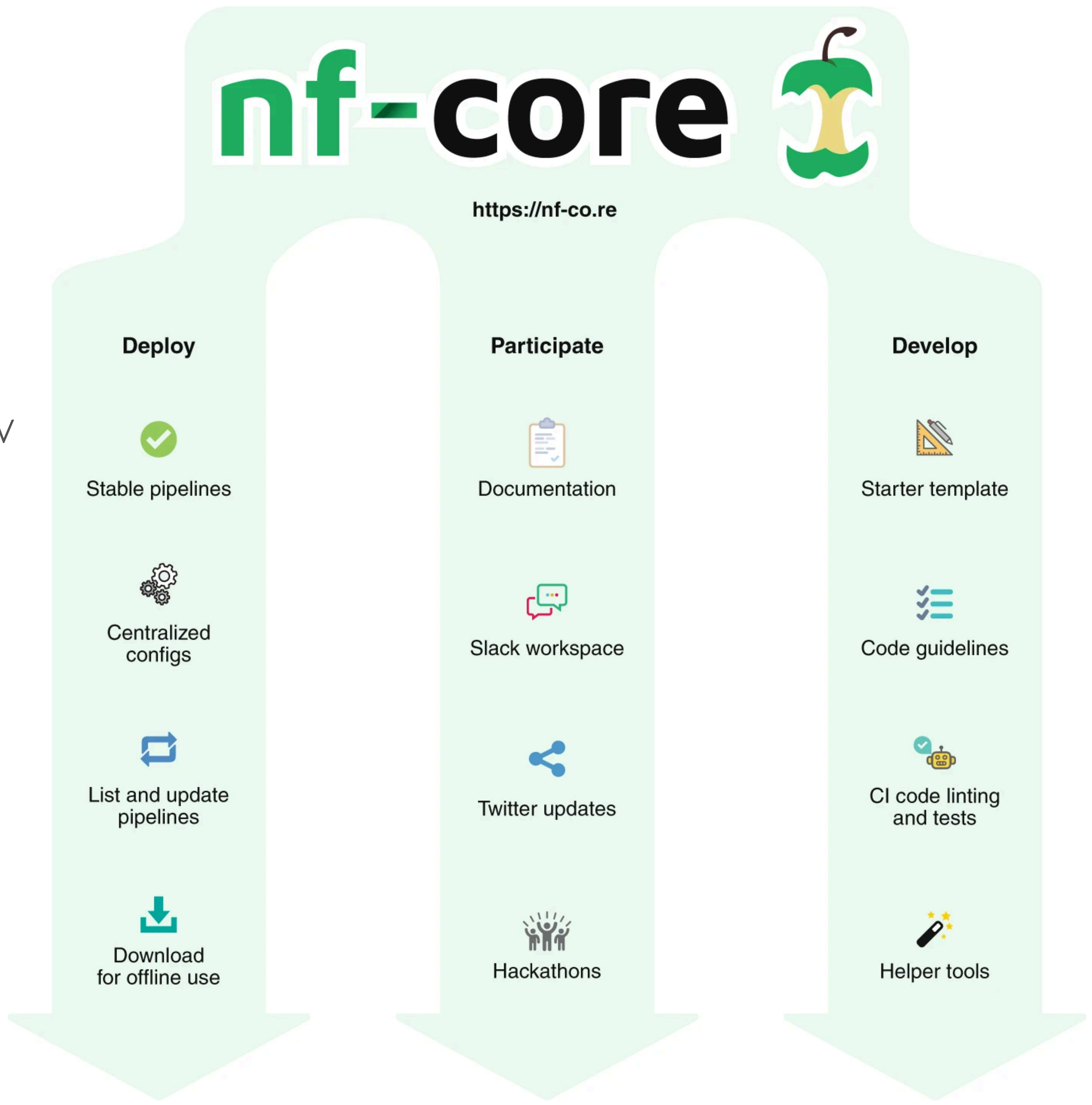
WP2

Goal 2.1	Improve Nextflow scalability and support for public clouds
Goal 2.2	Expand the support for web-based usage of nf-core pipelines
Goal 2.3	Nextflow kernel for Jupyter notebooks
Goal 2.4	GA4GH API compliant TES & WES executors
Goal 2.5	Migrate existing nf-core pipelines to Nextflow DSL2
Goal 2.6	Introduce module-level testing for nf-core
Goal 2.7	nf-core template for DSL2 pipelines

WP3

Goal 3.1	Nextflow website refresh: learning, community and support
Goal 3.2	Four new community Nextflow training events
Goal 3.3	Ten expanded bursaries for 2020 Nextflow Community Conference
Goal 3.4	Establish 2 nf-core hackathons and 2 user workshops
Goal 3.5	Tutorial videos on nf-core website

nextflow +



Community efforts to collect production ready analysis pipelines built with Nextflow

https://nf-co.re

Chan Zuckerberg Initiative