aws genomics

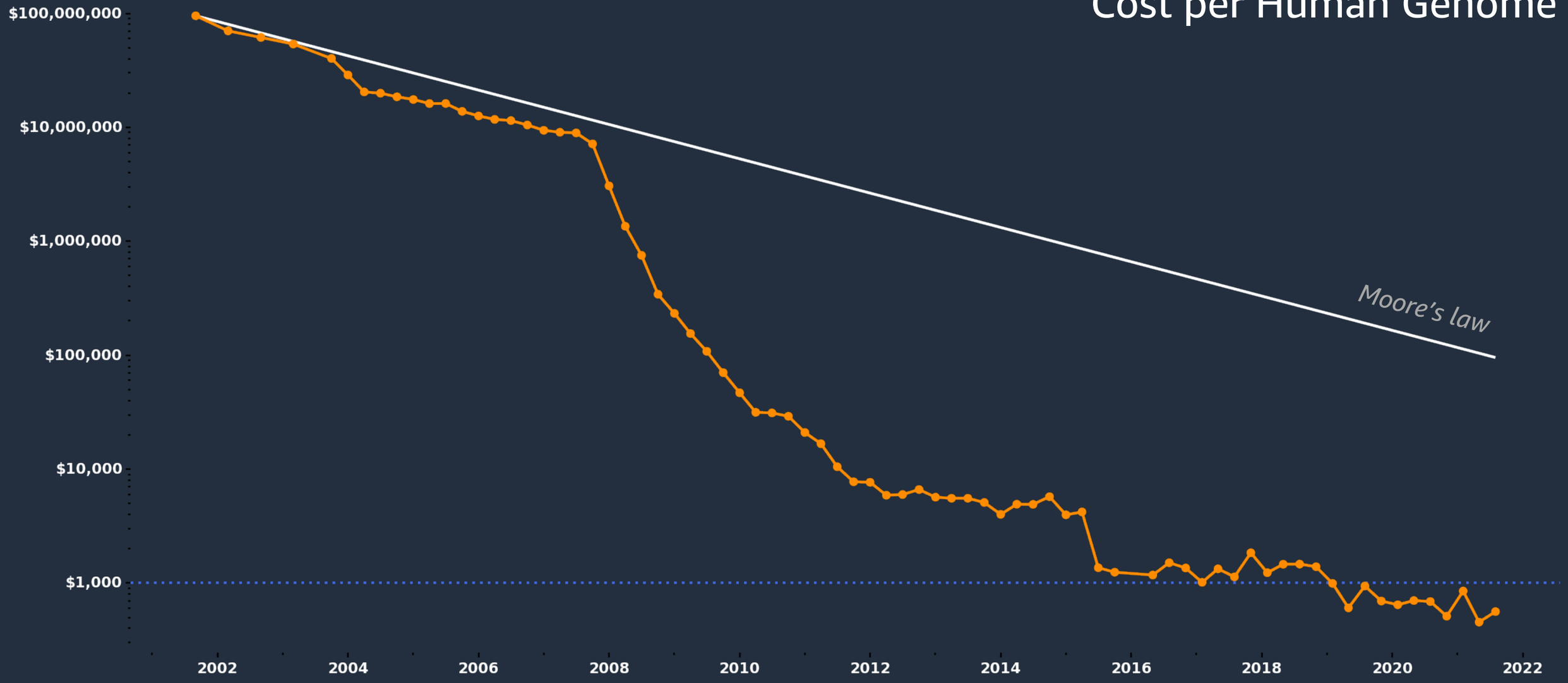
Scalable and reproducible genomics data analysis on AWS

W. Lee Pang, PhD
Pr. Developer Advocate – AWS HealthAI

The genomics challenge

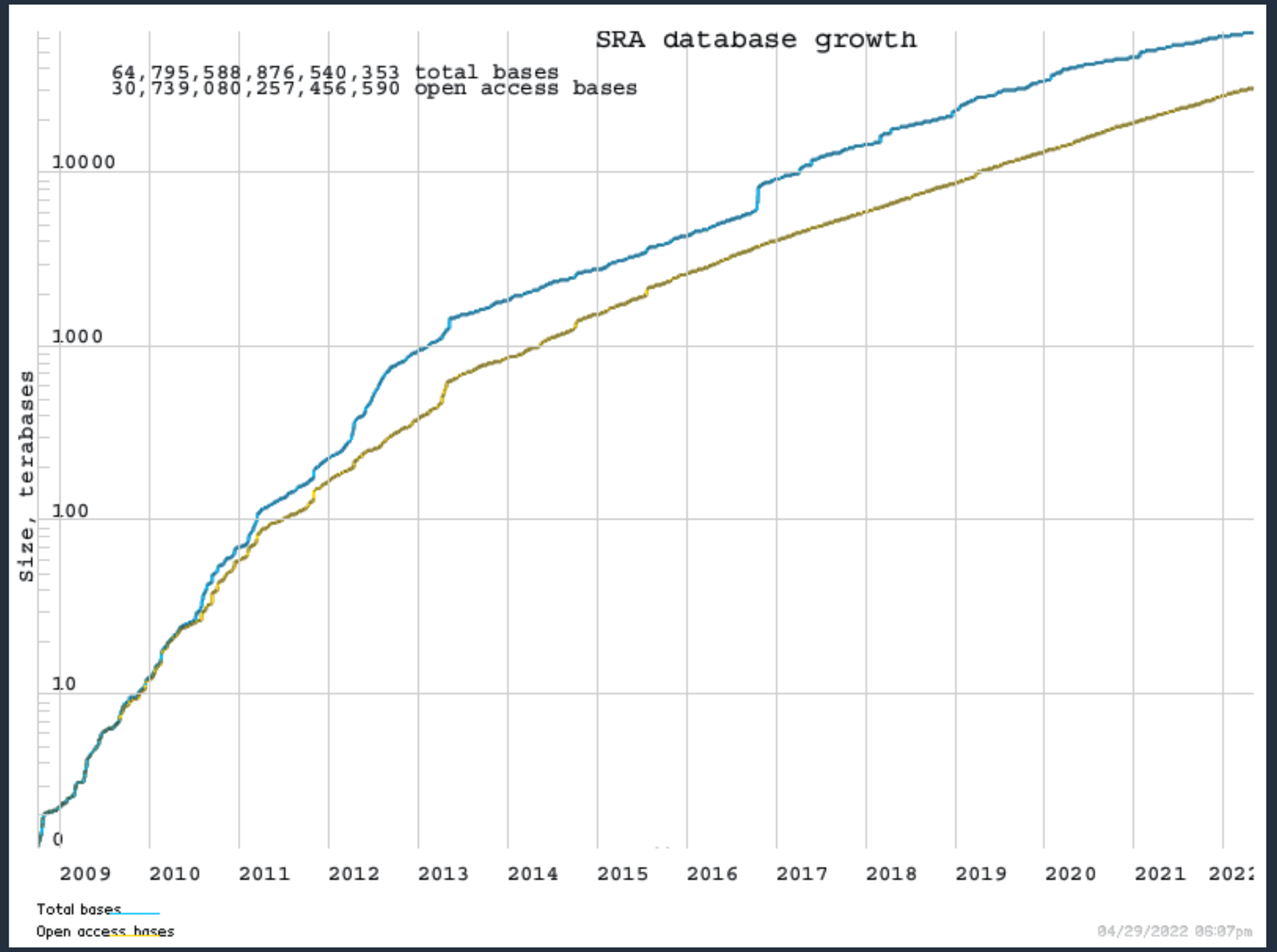
aws genomics

Cost per Human Genome



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

aws genomics



https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

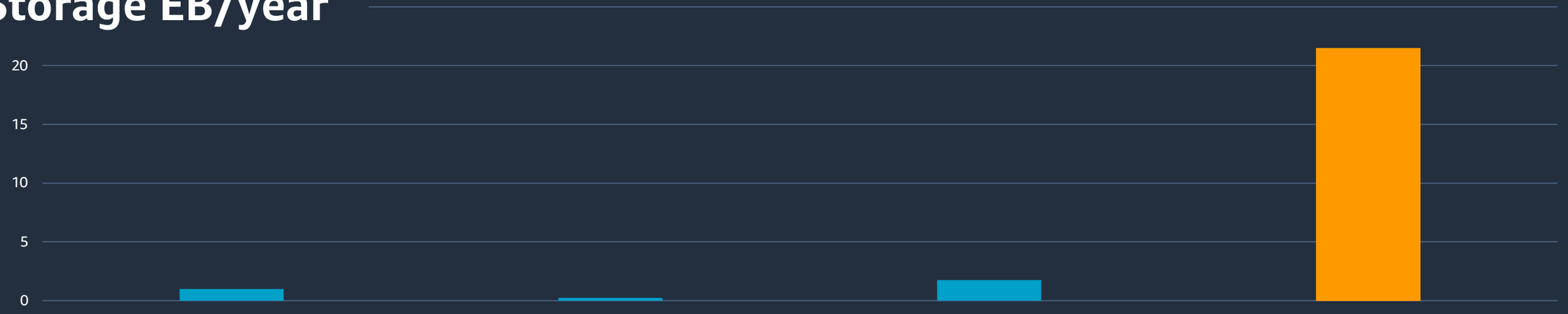
The amount of genomics data is growing **exponentially**

logarithmic y-axis!

aws genomics

Genomic footprints require scalable storage and compute

Storage EB/year



Data phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta/bytes/year	0.5/15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction Real-time processing Massive volumes	Topic and sentiment mining Metadata analysis	Limited requirements	Heterogeneous data and analysis Variant calling, ~2 trillion central processing unit (CPU) hours All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Source: Stephens, et al., Big Data: Astronomical or Genomical? (2015)

aws genomics

A common scenario

Newly published genomics analysis workflow

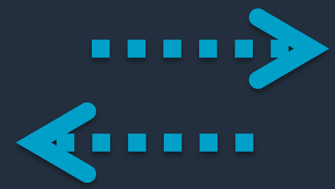
written in workflow language **X**

requires **accelerated** tools

examples use **1000s** of genomes

?

n/a



aws genomics

Key considerations for genomics workloads

Data gravity and access

- Raw human genomics data is 100s of GBs per sample
- Population sequencing initiatives are growing genomics data for 100K to 1M individuals
- Human genomes are sensitive information that require controlled access

Scalable compute

- Varied computational needs depending upon genomic data type and tool
- Access to compute capacity and specialized resources like acceleration

Evolving tools and methods

- Multiple tooling options to implement generalized ”best practices”
- Several key high-level analysis concepts – e.g. variant calling and joint genotyping

aws genomics

Why AWS for genomics

aws genomics

Computing as a utility

Focus on applications
and not infrastructure

Pay as you go, and only
for what you use

On Demand and fit for
purpose

aws genomics

AWS core cloud capabilities facilitating Genomics

Compute

Amazon EC2

Storage

Amazon Elastic Block Store (EBS)

Amazon S3

FSX for Lustre

Identity

AWS Identity and Access Management (IAM)

Managed services

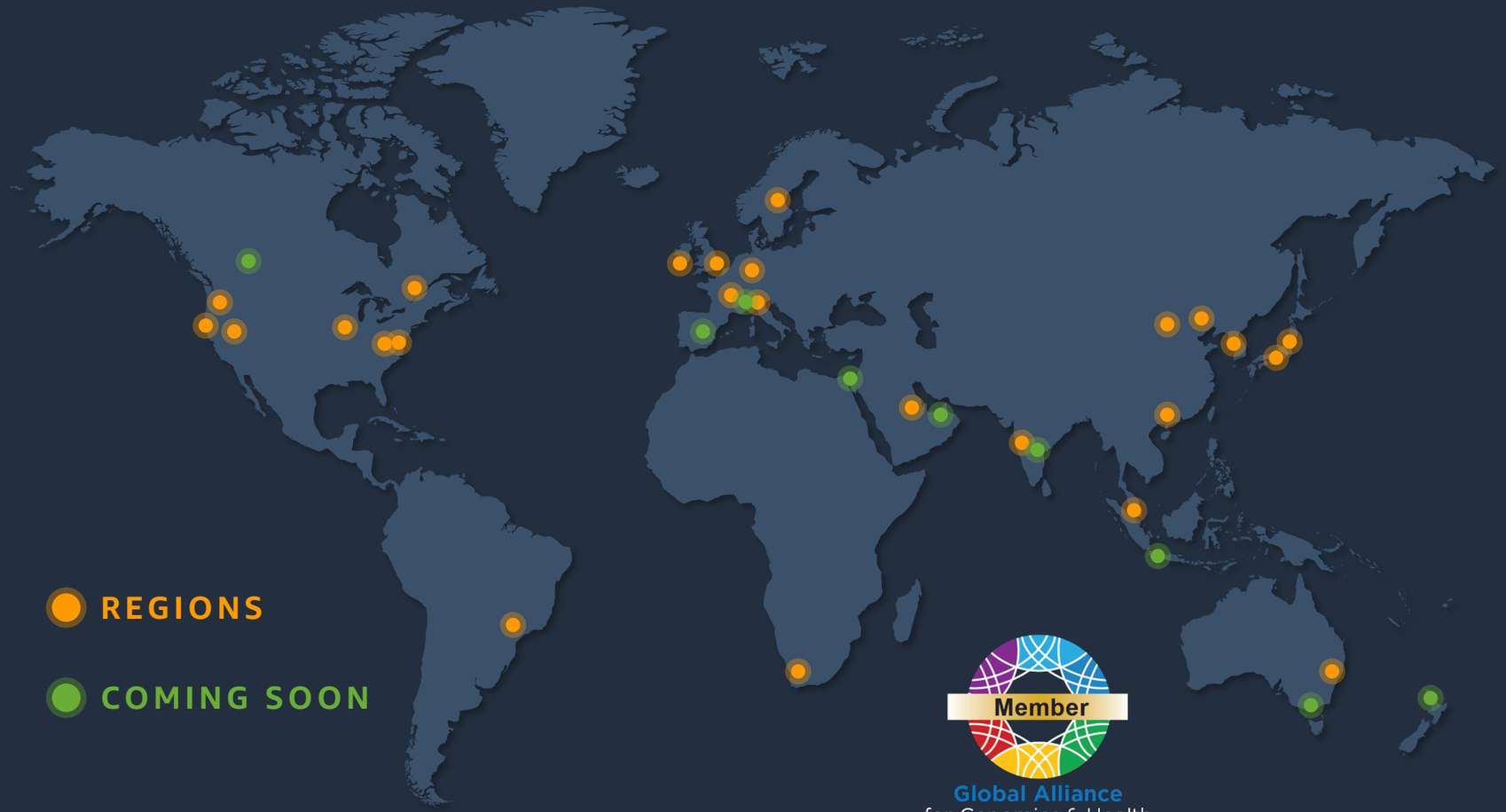
AWS Batch

Amazon Elastic Container Service

AWS Step Functions

aws genomics

AWS is global



Member

Global Alliance
for Genomics & Health

- Over 1 million active customers across 190 countries
- 2,000+ government agencies
- 5,000+ educational institutions
- 26 regions (+8 Planned)
- 84 availability zones with 3+ data centers per zone
- 310+ POPs
- 245 countries and territories served

aws genomics

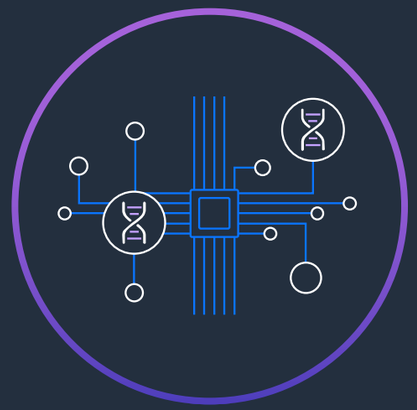
Customer benefits of the AWS Global Infrastructure

aws genomics

AWS for Genomics solution areas

AWS provides solutions and tools across the Genomics workflow

Data transfer & storage



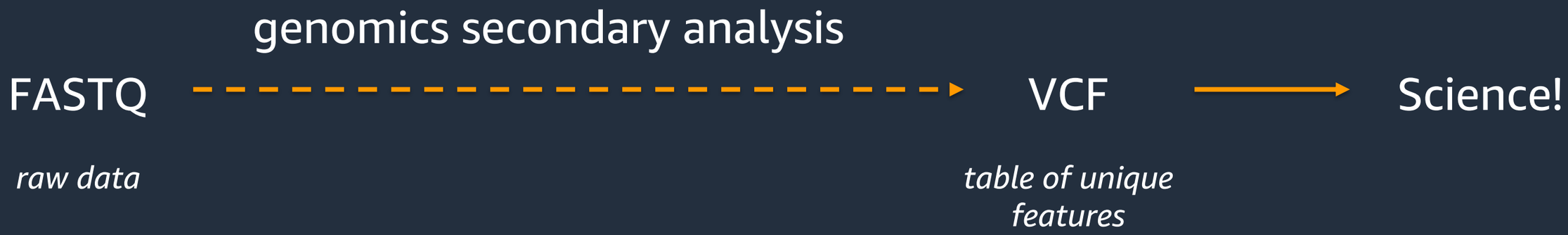
Workflow automation and secondary analysis

Data aggregation & governance

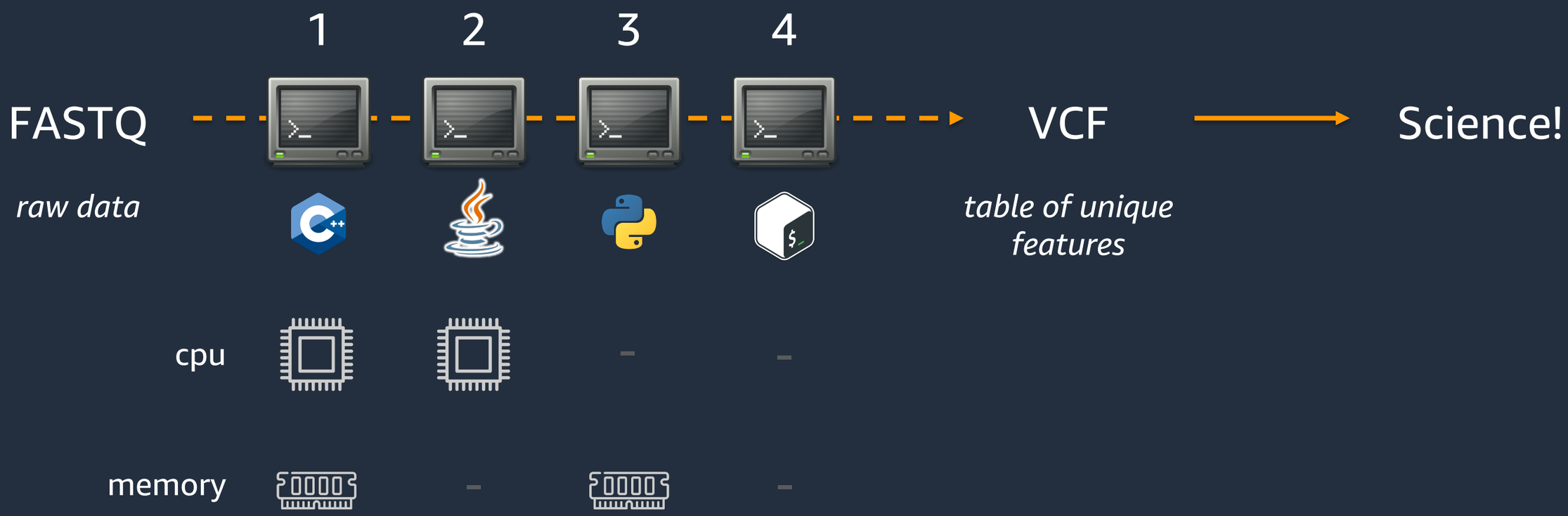
Interpretation & ML for tertiary analysis

Clinical translation

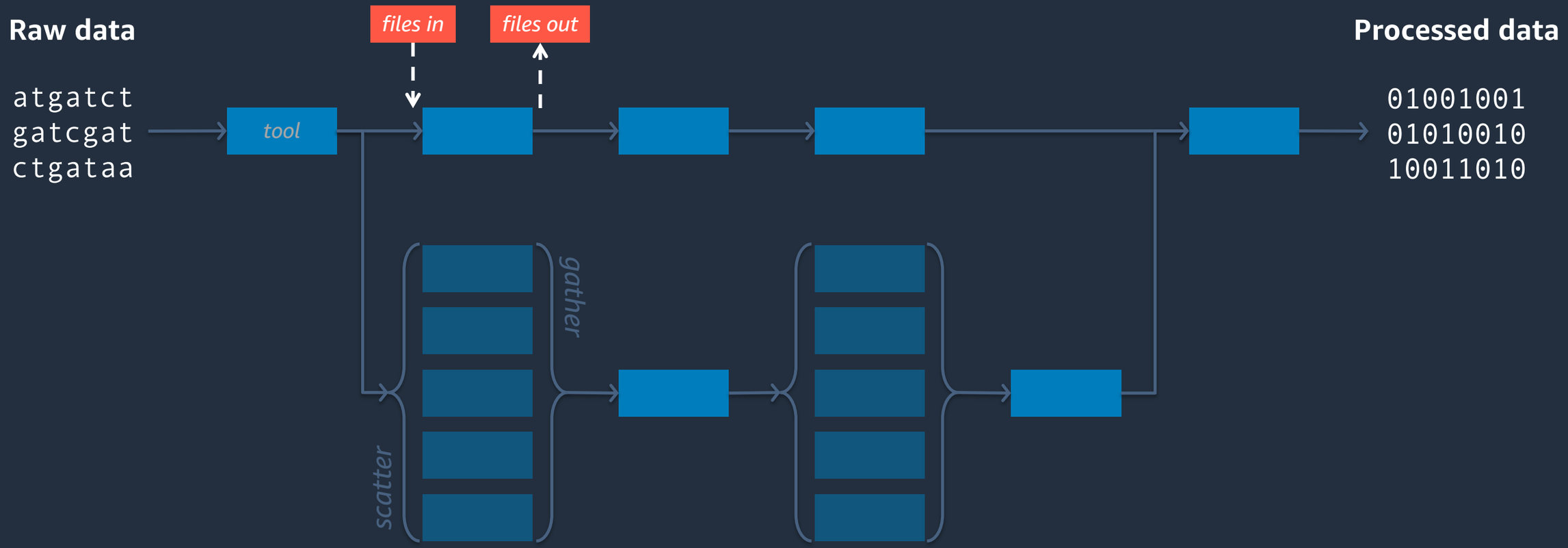
Genomics data analysis



aws genomics



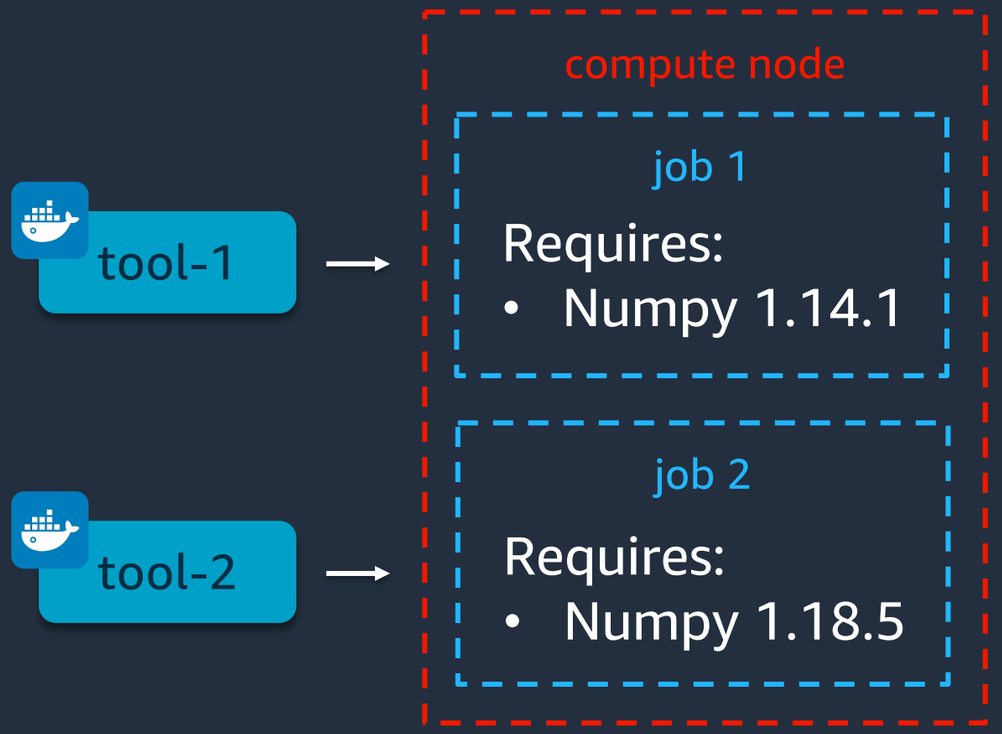
aws genomics



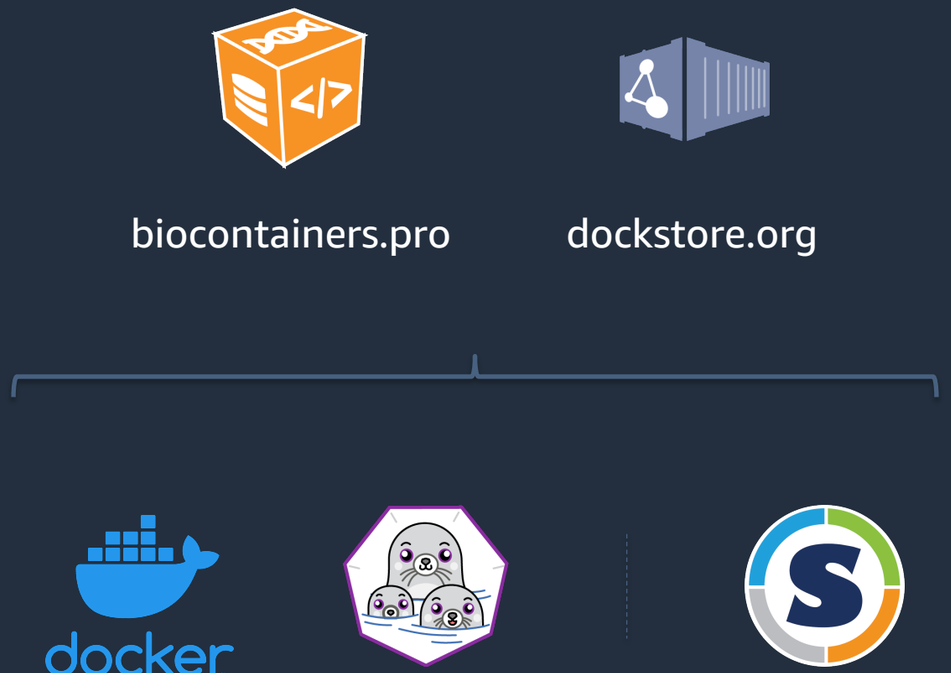
aws genomics

Containerized tools are portable

Isolate dependencies and concerns



many containerized bioinformatics tools



aws genomics

Workflow languages abstract complexity

Workflow Definition Language

Common Workflow Language

Nextflow (DSL 2)

Snakemake

Define tasks

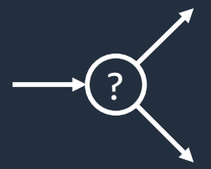
container image

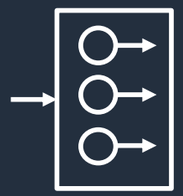
compute resources

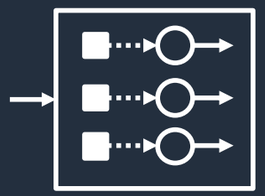
command

data inputs / outputs

Define task relationships

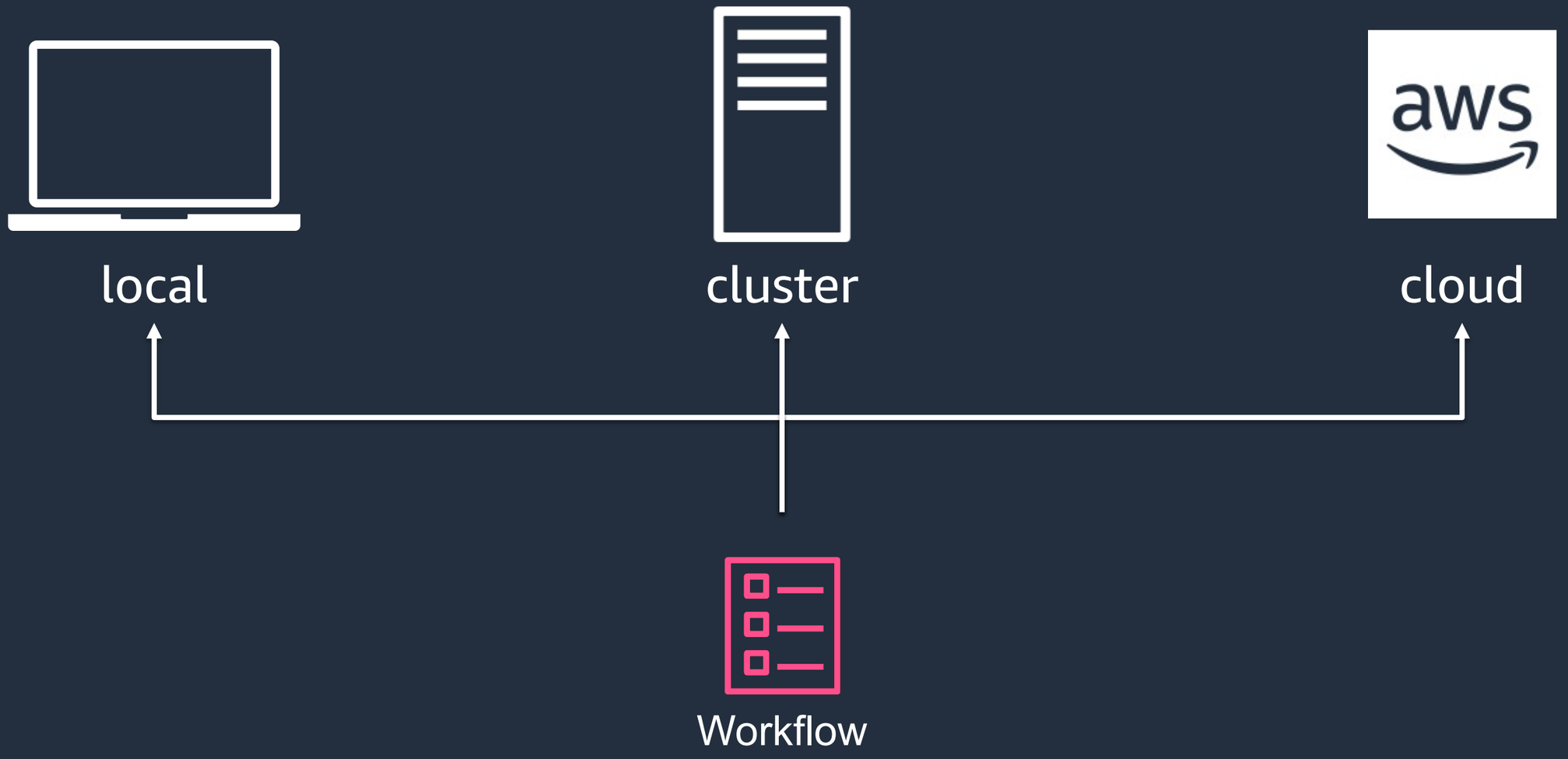






aws genomics

Workflows are portable



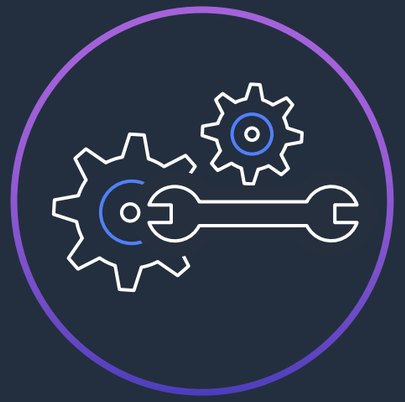
aws genomics

Major architectural components

Workflow orchestration

AWS Step Functions

Job execution

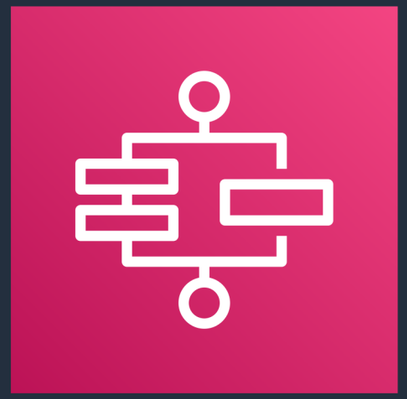


AWS Batch

Data storage

Amazon S3

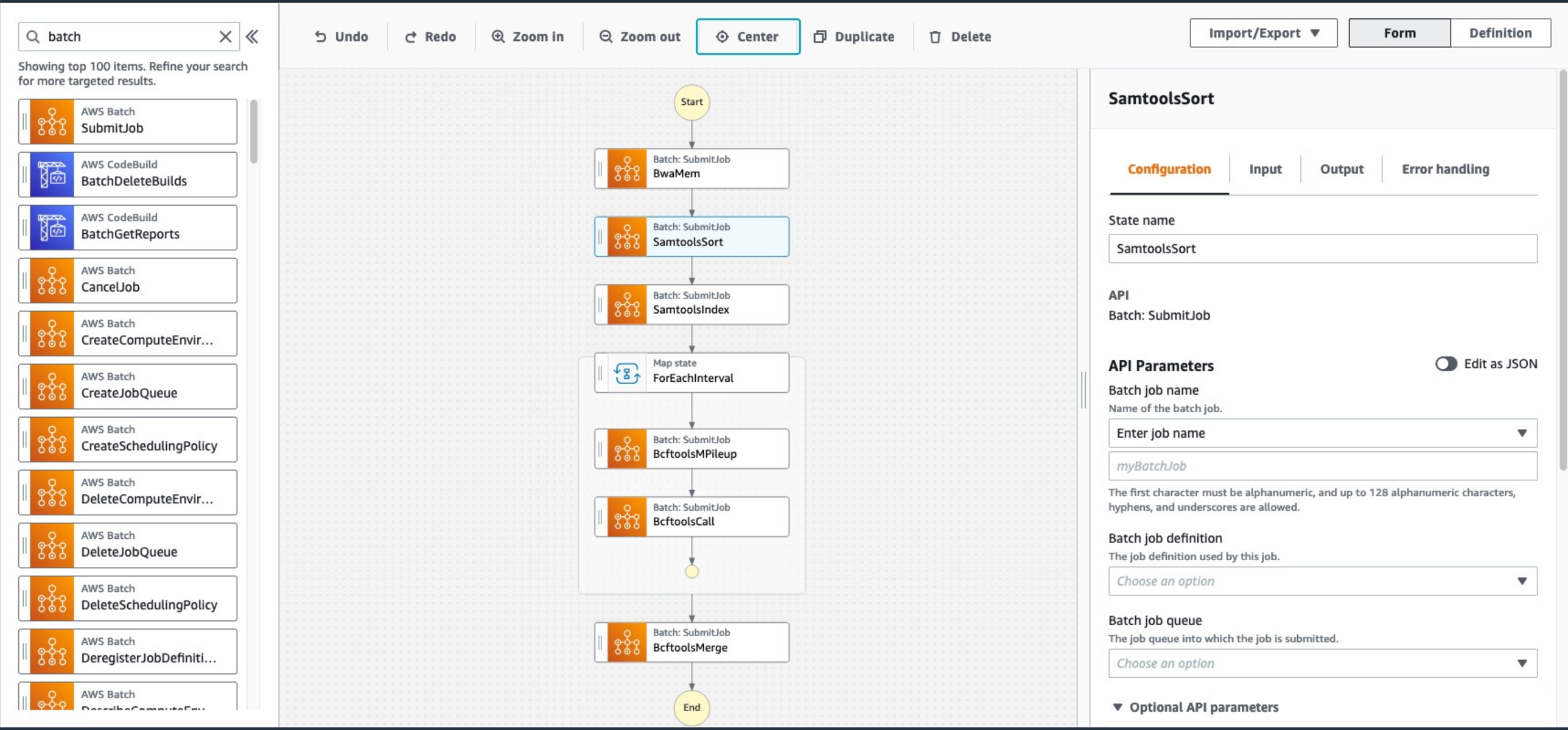
Workflow orchestration with AWS Step Functions



Design and run workflows that stitch together services such as AWS Lambda, AWS Batch, and Amazon ECS into feature-rich applications

Easily build workflows as a series of steps, with the output of one step acting as input into the next

aws genomics



aws genomics

Job execution with AWS Batch

Fully managed, optimized resources

No software to install or servers to manage. AWS Batch provisions and scales the optimal quantity and type of compute resources

Integrated with AWS

AWS Batch jobs can easily and securely interact with services such as Amazon S3, DynamoDB, and more

Cost efficient

AWS Batch launches compute resources tailored to your jobs and can provision Amazon EC2 and EC2 Spot instances

AWS Batch

compute environment(s)

On-Demand instances

SPOT instances

inputs

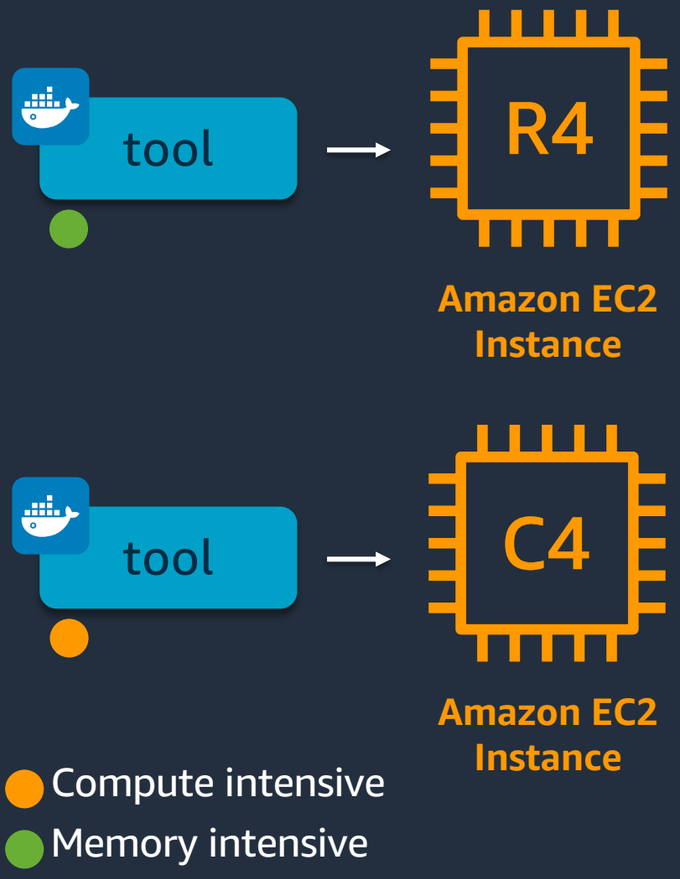
outputs

Amazon Simple Storage Service (S3)

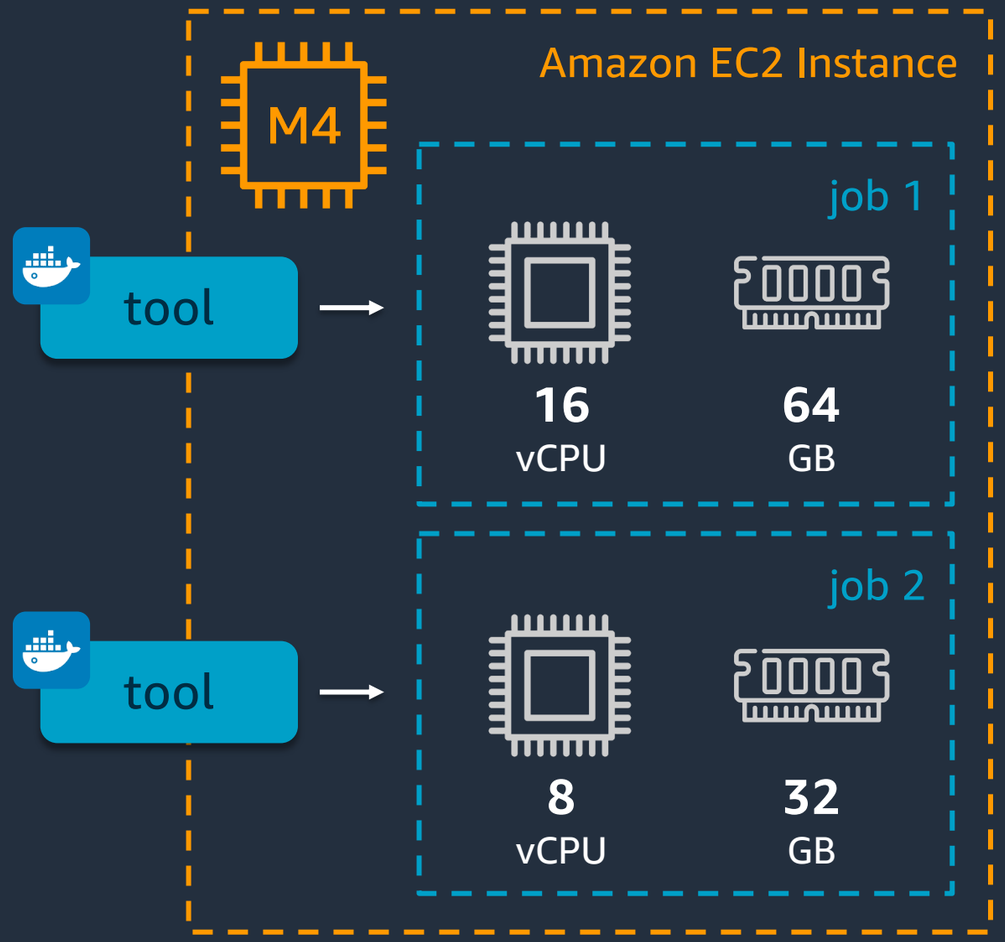
aws genomics

Cost efficiency by right sizing compute for jobs

Match jobs to instance types



Bin pack similar sized jobs



aws genomics

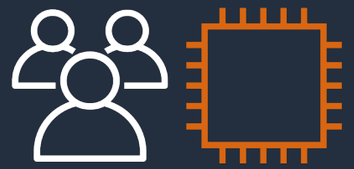
Data storage with Amazon S3

Amazon S3

Cloud object storage, like Amazon S3, provides an easy way to store and securely share data. It can provide a cloud-native single source of truth for your data heavy applications.

Durable

robust storage for active and archive data



Available

to you, your collaborators, and your compute

Secure

protect sensitive information

aws genomics

MODIS NOAA GHE NCBI SRA NWM GEFS

TCGA OpenStreetMap GEOS-Chem ECMWF ERA5 CIViC

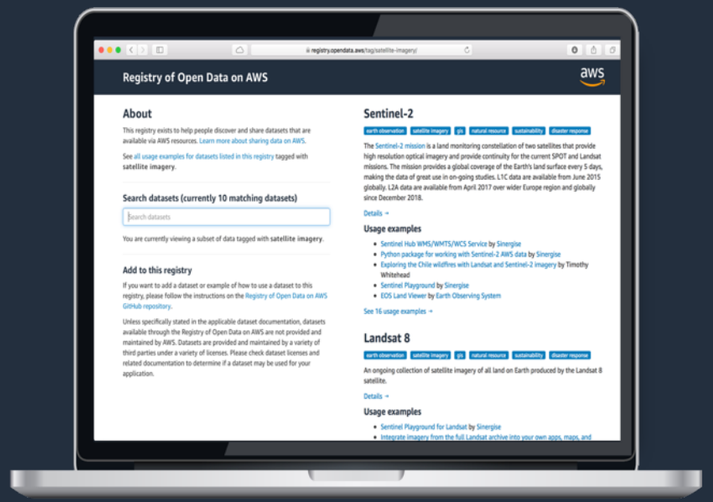
Sentinel-1 HRRR Sentinel-2 OFS eBird Terrain Tiles

GOES-16 NAIP CBERS NOAA ERI

Registry of Open Data on AWS

HIRLAM ISD gnomAD

PubSeq Kids First Landsat GFS



SILAM Air Quality NREL Solar Radiation

CESM LENS TARGET ENCODE OpenAQ

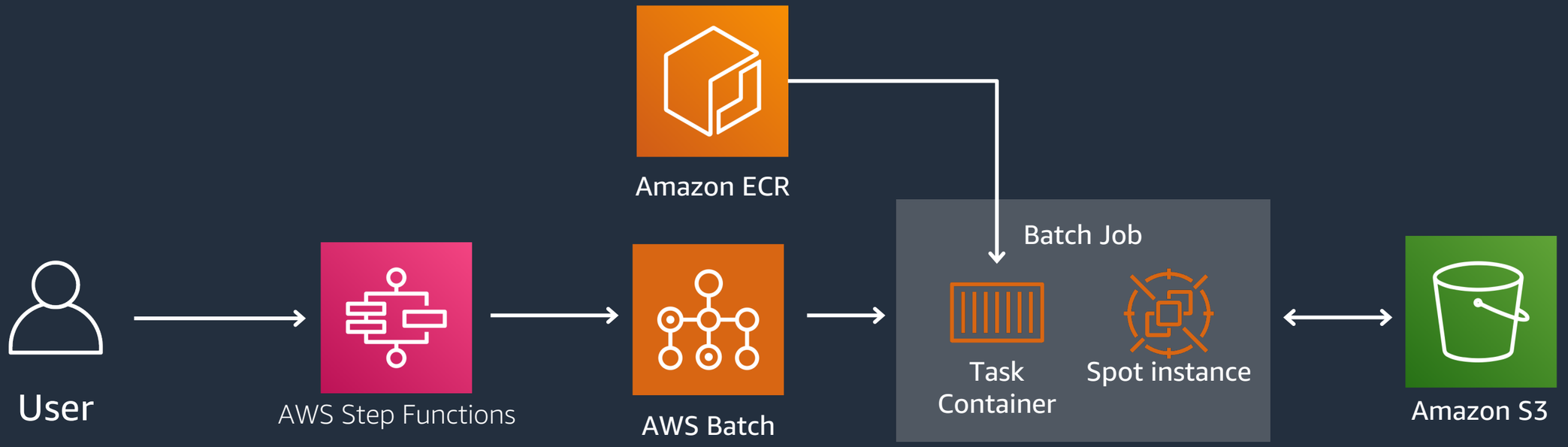
NREL Wind Integration Common Crawl

registry.opendata.aws

aws genomics

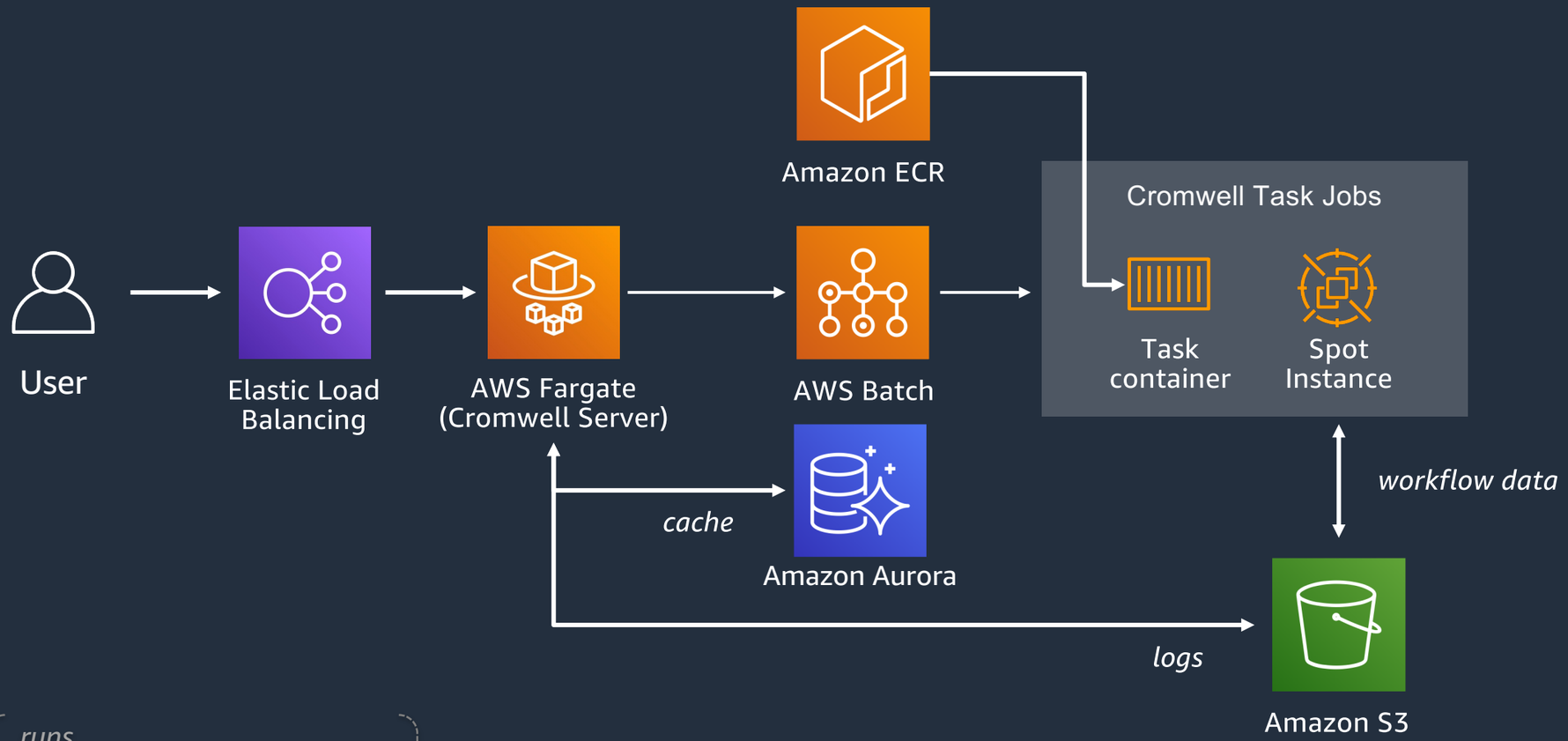
Example architectures

AWS Reference Architecture

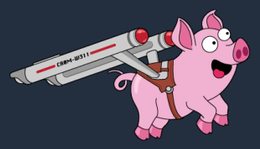


aws genomics

Cromwell on AWS



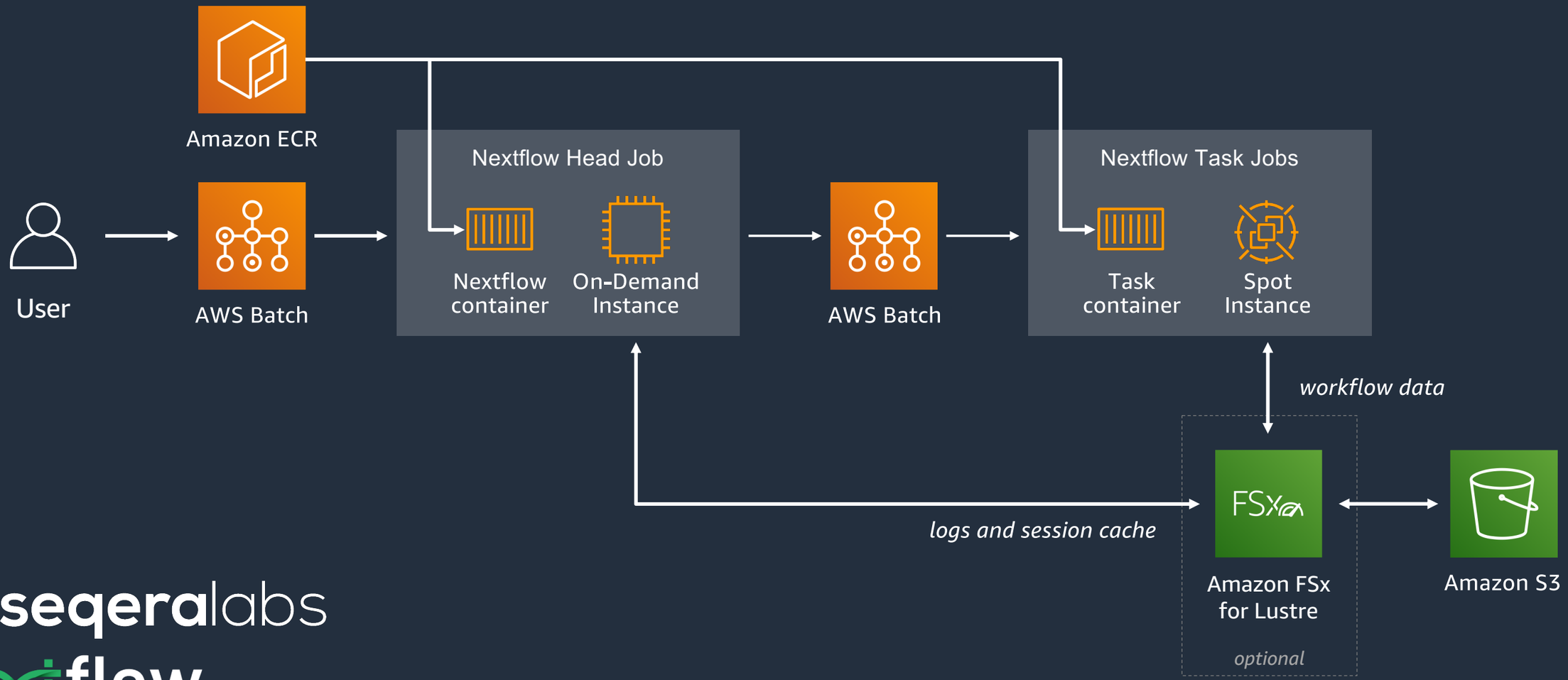
BROAD INSTITUTE



runs {wdl} CWL

aws genomics

Nextflow on AWS

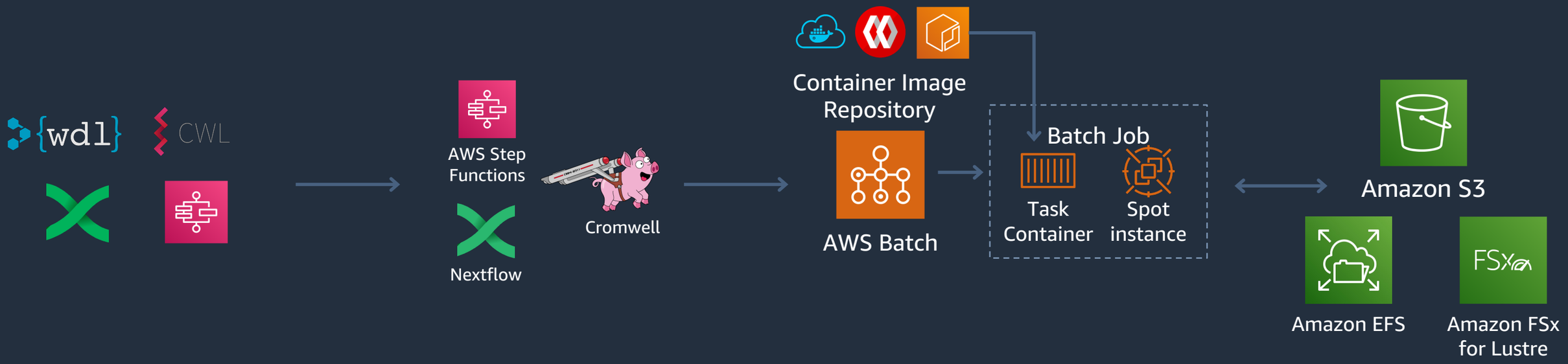


seqeralabs

nextflow

aws genomics

Common architectural pattern



Workflow definition

Develop a workflow using a definition language and containerized tools

Workflow orchestrator

Submit your workflow to a workflow engine

Job execution

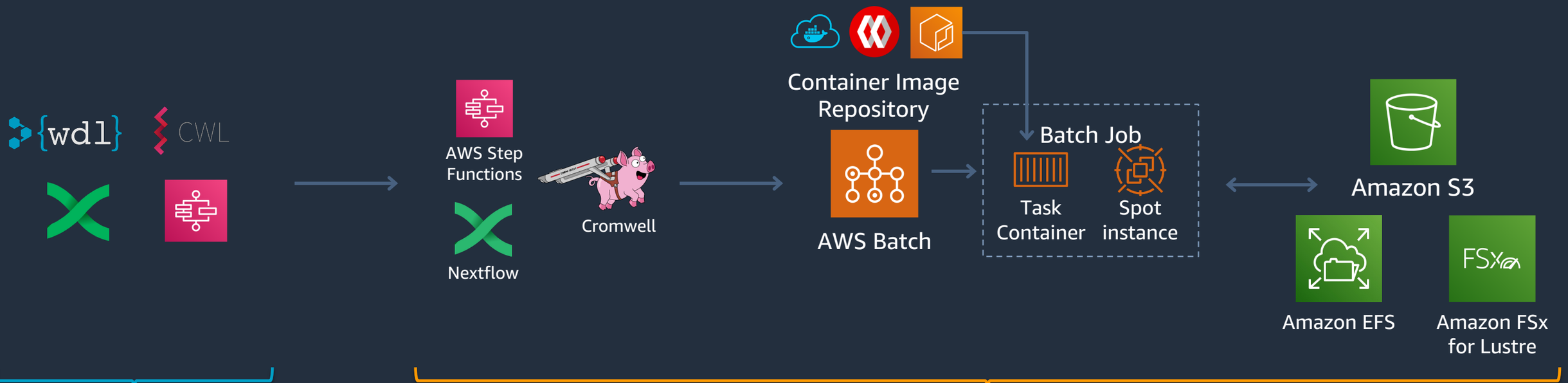
Workflow engine submits tasks to cloud compute resources (e.g., AWS Batch)

Data storage

Tasks retrieve and store data in cloud object storage (e.g., Amazon S3)

aws genomics

Reproducible architectural pattern



Code

Also Code

AWS CloudFormation

AWS CloudFormation gives you an easy way to model a collection of related AWS and third-party resources, provision them quickly and consistently, and manage them throughout their lifecycles, by treating infrastructure as code.

aws genomics

Amazon Genomics CLI

Amazon Genomics CLI is an **open source command line interface (CLI)** that helps customers new to AWS run Genomics workflows in the cloud by **automating deployment of best practices architectures** for workflow engines. Amazon Genomics CLI reduces the time for scientists and developers to start running existing Genomics workflows at scale and speed up iteration cycles as they develop new ones.

Setup a new project and run a Genomics secondary analysis workflow in the cloud with a few CLI commands

Open source and built on community open standards

Amazon Genomics CLI

Start running genomics workflows on AWS with a few easy steps and familiar tooling

Configure

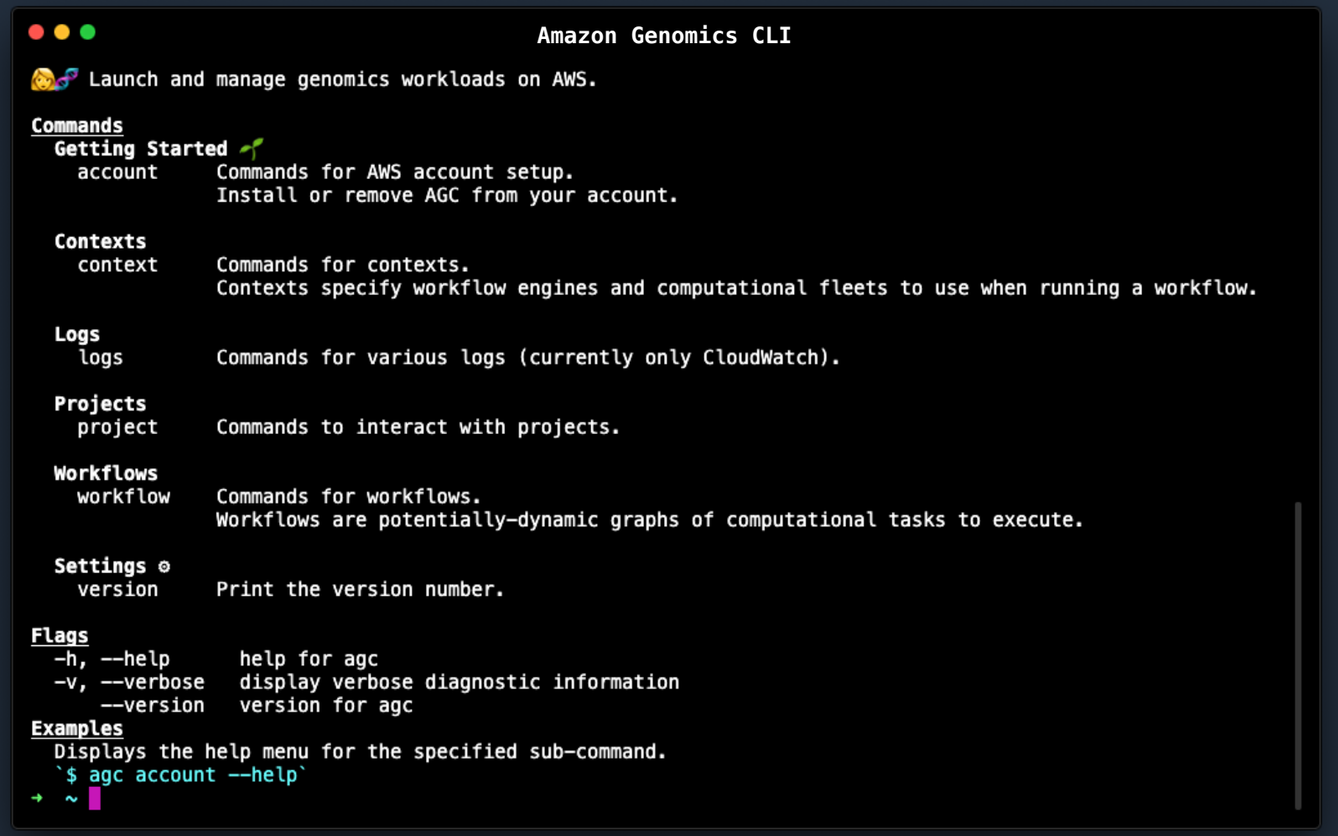
Create a project, define compute resources and workflows

Deploy

Deploy compute resources and container clusters to execute workflow engines

Run

Process genomic data and derive research insights



Amazon Genomics CLI

Start running genomics workflows on AWS with a few easy steps and familiar tooling

Configure

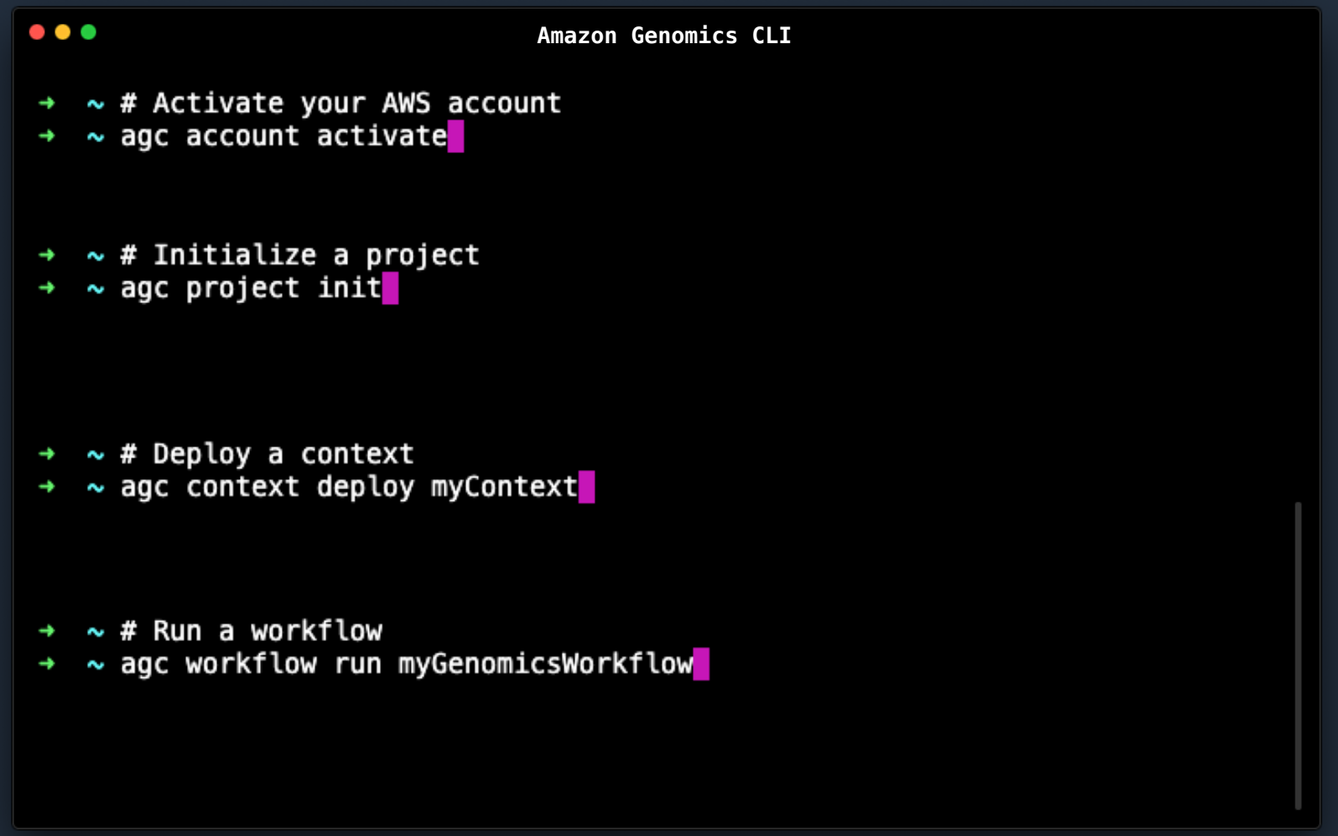
Create a project, define compute resources and workflows

Deploy

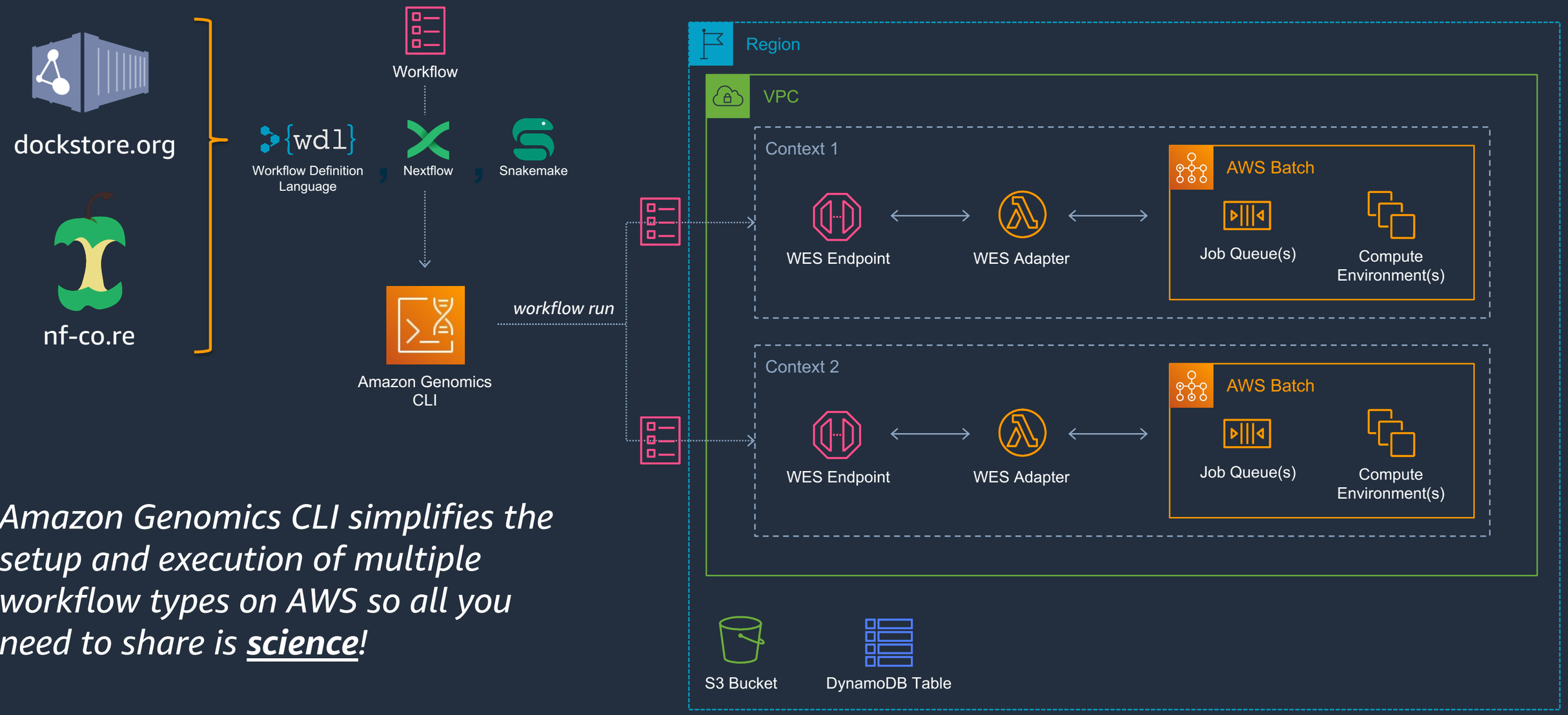
Deploy compute resources and container clusters to execute workflow engines

Run

Process genomic data and derive research insights



Amazon Genomics CLI enables collaboration



*Amazon Genomics CLI simplifies the setup and execution of multiple workflow types on AWS so all you need to share is **science**!*

aws genomics

Common tips

keep your containers images small

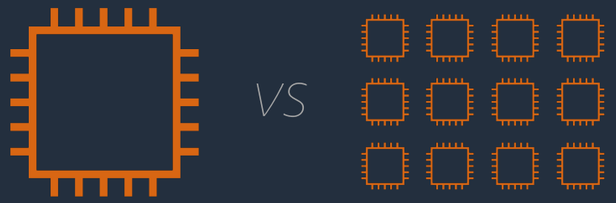
- 1 tool per container
- if they are unavoidably large use cached copies in ECR
- be mindful of “minimal” image containers

write infrastructure agnostic workflows

- move infra specific elements to external config parameters
- change workflow execution ”profiles” quickly without rewriting the workflow definition

leverage “embarrassing parallelism”

1 step = 1 small container



use cloud native integration where possible

e.g. use tools that can read and write directly from/to Amazon S3

aws genomics

Use cases

Tübingen's Quantitative Biology Center (QBiC) Innovates on AWS to Accelerate Genomics Research

Challenges

QBiC enables genomics research for public and private institutions by abstracting the heavy lifting associated with HPC and Data Analysis. As medical imaging and genomic sequencing research data grew rapidly, QBiC foresees difficulties to scale quickly and cost-effectively. Processing data is a bottleneck with on-premises resources routinely at 95% of capacity.

Solution

Through its choice to standardize on NextFlow and nf-core as the workflow solution, QBiC was able to easily integrate with AWS Batch and is currently evaluating its potential. First results hint to a reduced queue time by 50% for all jobs using Amazon EC2 Spot Instances, while driving down the costs of analysis.

Benefits

- Scale processing from 30 samples to 100,000 samples in one research project is possible
- Potentially reduced queue times by 50 percent for genomic data analysis
- Standardized and automated Genomic workflows with 'NextFlow' and nf-core pipelines on AWS and on-premise infrastructure, alike

"Our automated genomic analysis run times on AWS are comparable with on-premises resources but with no wait. It was astonishing how easily Nextflow and nf-core integrate with AWS Batch. With AWS Spot, runtimes for small, medium, and large projects are all similar and at low costs."

Alexander Peltzer,
Research & Development Data Science at Quantitative Biology Center

Quantitative Biology Center, University of Tübingen

Website:

https://qbic.uni-tuebingen.de/en/

About QBiC:

Quantitative Biology Center (QBiC) Tübingen is a research center located at University Tübingen in Tübingen, Germany. QBiC runs a data management platform and high-performance computing (HPC) environment to support genomics research within the university and at private medical research organizations across Germany.

Fred Hutch microbiome researchers use AWS to perform seven years of compute time in seven days

Challenge

Fred Hutch is engaged in analysis of the microbiome. Translating gigabytes of raw microbiome Genomic data into insights about which specific microbes are present in a person is a computationally intensive task requiring highly scalable technology.

Solution

To accelerate its research, the team uses the Nextflow framework to orchestrate AWS Batch processes and scale the high-performance computing platform to accelerate processing time—reducing 7 years of compute time to 7 days.

Benefits

- Processes data from more than 15,000 biological samples
- Reduced 7 years of compute time to 7 days
- Increases resolution on microbiome samples to find links to improve health outcomes

> "AWS Batch integrates well with Nextflow, so it was easy for us to get Nextflow up and running without having to reinvent the wheel."

Sam Minot, PhD and staff scientist at Fred Hutch MRI

FRED HUTCH
CURES START HERE™

Company: Fred Hutch

Industry: Life Sciences

Country: United States

Employees: 3,500

Website: https://research.fredhutch.org/

About Fred Hutch

The Fred Hutch Microbiome Research Initiative, funded by Seattle's Fred Hutchinson Cancer Research Center, includes microbiome investigators with expertise in study design, laboratory methods, animal models, human intervention studies, data analysis, and visualization. These researchers are working to predict health outcomes, understand the pathogenesis of disease, and manipulate the microbiota to promote health.

aws genomics

Melbourne Alliance Uses AWS to Support Genetic Testing

Challenge

To build the GenoVic shared genomics testing system as a modular, flexible, and scalable system, the Melbourne Genomics Health Alliance recognized that using cloud technology would be essential.

Solution

The Alliance runs GenoVic on AWS, using AWS Lambda for serverless compute, AWS Batch to schedule batch processing, Amazon S3 to store genomic data, and AWS CloudFormation to automate cloud resource provisioning.

Benefits

- Sets up a highly secure, scalable system for genomics
- Enables genomic testing for cancer and rare diseases
- Improves testing workflow efficiency
- Onboards new labs in months

Learn more

> "In our first year of rollout, we've seen how GenoVic can support the testing of patients across a wide range of rare diseases, providing greater efficiencies for five laboratories."

Dr. Natalie Thorne, Lead of Innovation & Technology

Melbourne Genomics Health Alliance

Global knowledge. Individual care.

Company: Melbourne Genomics Health Alliance

Country: AU

Employees: 10,000

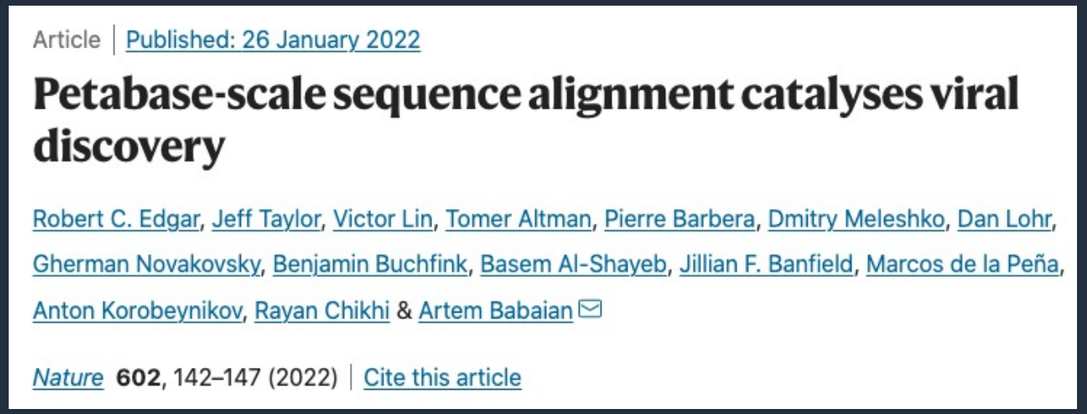
Website: MelbourneGenomics.com.au

About Melbourne Genomics Health Alliance

Working collectively to transform healthcare, the Melbourne Genomics Health Alliance is a group of 10 leading health, research, and teaching institutions in Victoria, Australia, that seeks to show how genomics can deliver more effective medical care that is less wasteful of resources and more likely to succeed the first time.

aws genomics

University of British Columbia identifies 130,000 new viruses in 11 days



Article | Published: 26 January 2022

Petabase-scale sequence alignment catalyses viral discovery

Robert C. Edgar, Jeff Taylor, Victor Lin, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Dan Lohr, Gherman Novakovsky, Benjamin Buchfink, Basem Al-Shayeb, Jillian F. Banfield, Marcos de la Peña, Anton Korobeynikov, Rayan Chikhi & Artem Babaian

Nature **602**, 142–147 (2022) | Cite this article

NIH NCBI SRA
on
aws open data

> Using AWS, Serratus can process over one million libraries of next-generation sequencing data per day for an overall cost of less than half a cent per library.

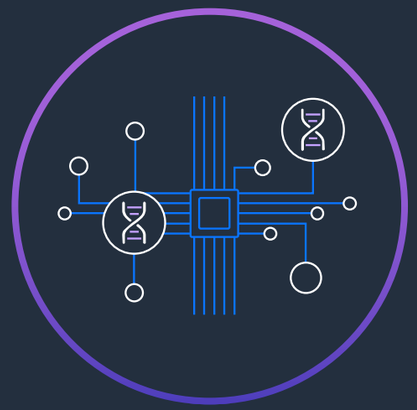
Artem Babaian, Ph.D and Serratus Project Lead at UBC

aws genomics

AWS for Genomics solution areas

AWS provides solutions and tools across the Genomics workflow

Data transfer
& storage



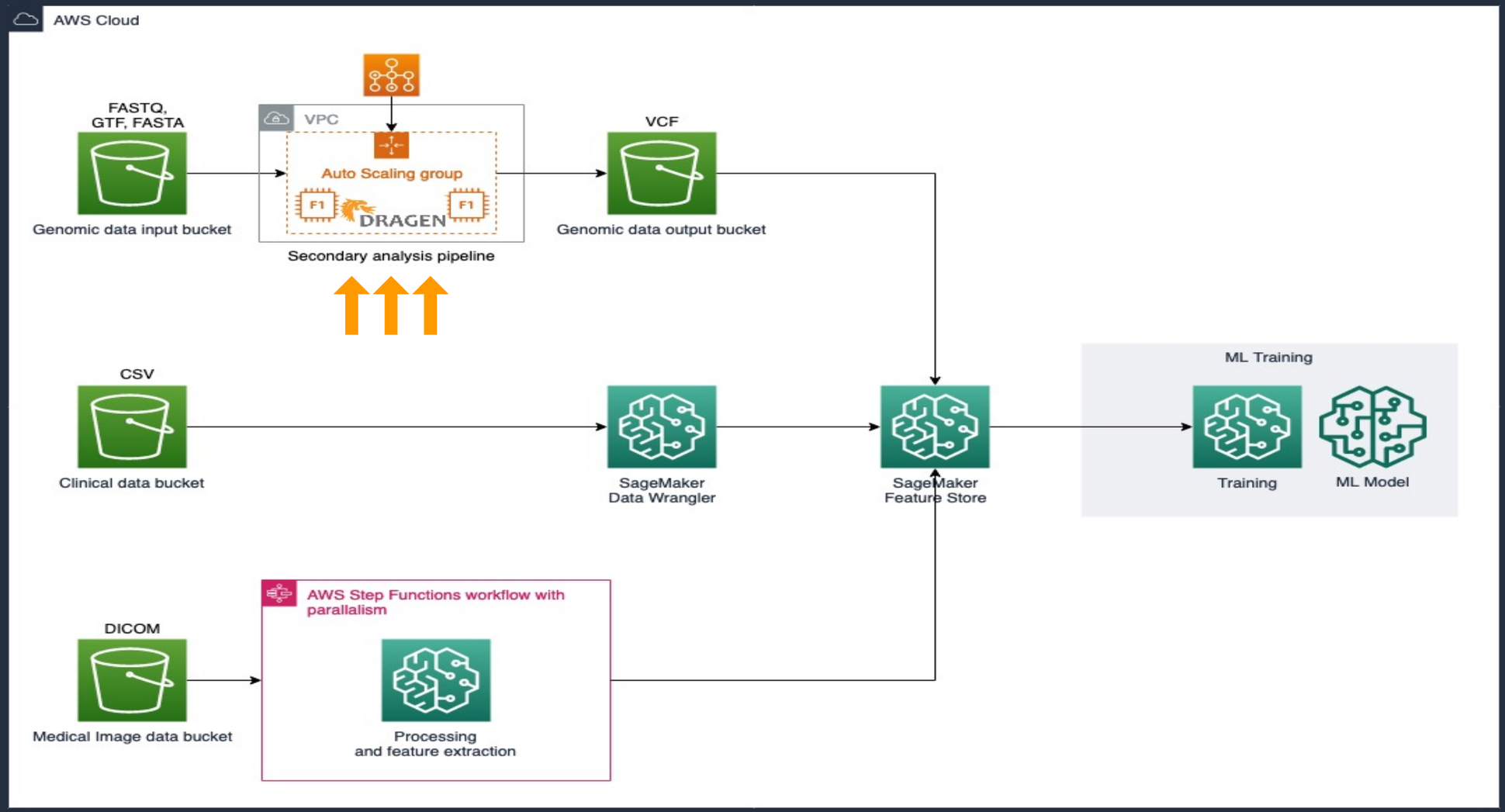
Workflow automation
and secondary analysis

Data aggregation
& governance

Interpretation & ML
for tertiary analysis

Clinical
translation

Sagemaker workflow for multimodal health data analysis



https://aws.amazon.com/blogs/industries/building-scalable-machine-learning-pipelines-for-multimodal-health-data-on-aws/
https://aws.amazon.com/blogs/industries/training-machine-learning-models-on-multimodal-health-data-with-amazon-sagemaker/

aws genomics

How AWS enables scalable genomics workloads

Scalable and secure

Reduce costs and improve turnaround time for genomic analysis

Best fit flexibility

Start building with AWS reference architectures, Amazon Genomics CLI, AWS Partner offerings

Infrastructure as code

Maximize results by minimizing operational overhead associated with infrastructure

Accelerate experimentation

Bioinformaticists and Data Scientists modernize and accelerate Genomic research and analysis

aws genomics

Resources

AWS for Health
aws.amazon.com/health

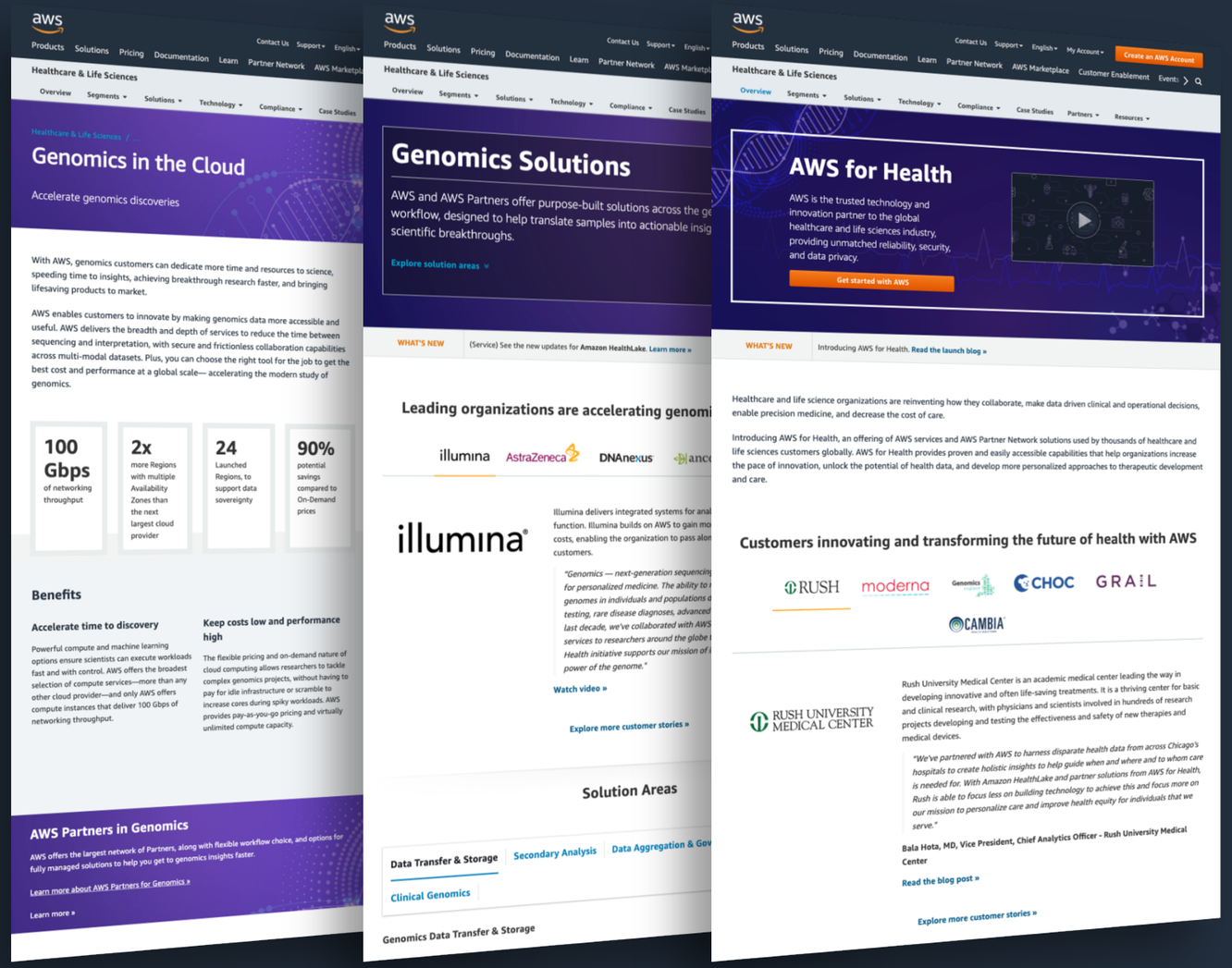
Genomics in the Cloud
aws.amazon.com/health/genomics

Genomic Solutions
aws.amazon.com/health/genomics/solutions

Guide to Genomics Workflows on AWS
docs.opendata.aws/genomics-workflows

AWS Marketplace
aws.amazon.com/marketplace

AWS Partner Network
aws.amazon.com/partners/find



Thank you!

aws genomics

Questions and answers

Appreciate your feedback here!

https://eventbox.dev/survey/WLPJJRF

aws genomics