QUANTORI BioTechX

Efficient Data Management for CryoEM HPC in AWS Cloud

Quantori

is a Full-Service Scientific Informatics, Data Sciences, and Digital Solutions Provider for the Life Science and Healthcare Industries.

Using our domain knowledge and technical expertise, we develop cutting-edge data science, digital engineering, and technology platforms for biotech, pharmaceutical, and healthcare companies that accelerate drug discovery and improve patient outcomes.

Our innovative approach harnesses the power of data engineering and informatics, machine learning, emerging technologies, cloud, and HPC expertise to advance research and development and ultimately bridge the gap between meaningful data and patient success.

CryoEM Processing

Since 2017 CryoEM has become an important tool for drug discovery. The majority of large companies in life science joined the competition then. They built or adjusted existing computation centers to run CryoEM.

CryoEM not only requires having access to an expensive microscope to get RAW data but also infrastructure built for processing and storing the data.

Just entering the field requires significant investment, and small-mid-sized companies naturally select the cloud for CryoEM processing.

CryoEM Requirements

A single CryoEM run can capture thousands of images and can generate anywhere between 1TB to 15TB of raw datasets

Processing requires extremely fast parallel storage and scalable GPU compute close to the data

Processing of one sample takes days, and in some cases months till achieving a reasonable result

Data Flow

1

RAW data comes from CryoEM LAB into buffer storage:

- ○ Local storage close to microscope (large institutions)
- ○ **Amazon S3 bucket**
- ○ USB drive ☺

2

Data is copied over to shared storage connected to the cluster

3

Processing data is being generated during the run

4

Analysis is completed, data goes to archive (S3 deep glacier)

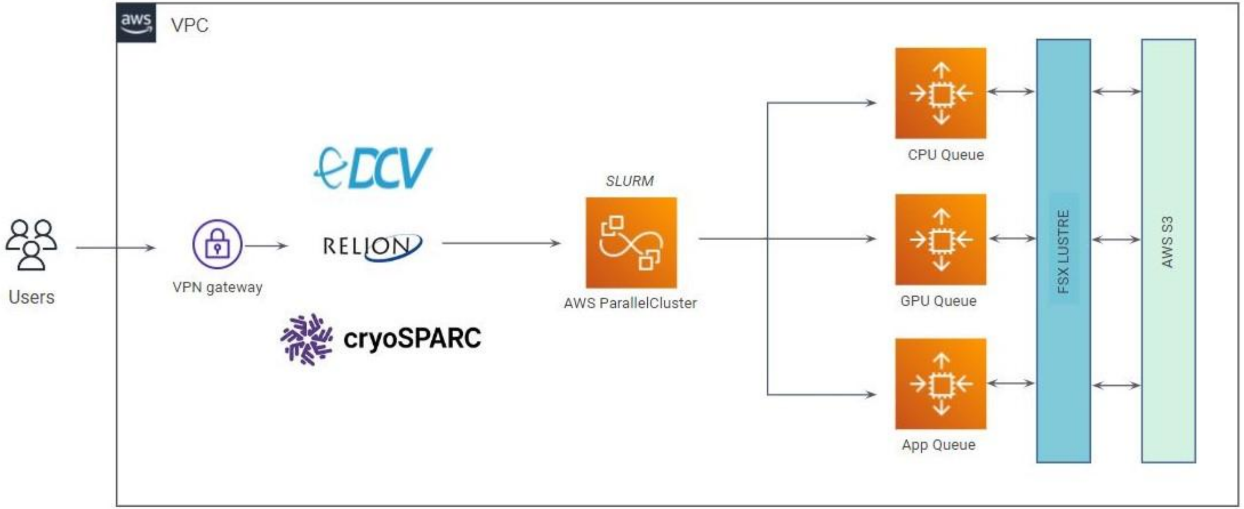
Approaches in AWS

Standard Approach

- Persistent GPU instances (or cluster) running 24/7
- Persistent storage used for RAW and intermediate computation data connected to GPU instances

Advanced Approach

- **Scalable GPU cluster**
- Persistent storage used for RAW and intermediate computation data connected to GPU instances



Issues with Persistent Storage

1 **Expensive (~$1000/month for 10TB sample) or slow**

2 **Constantly growing**

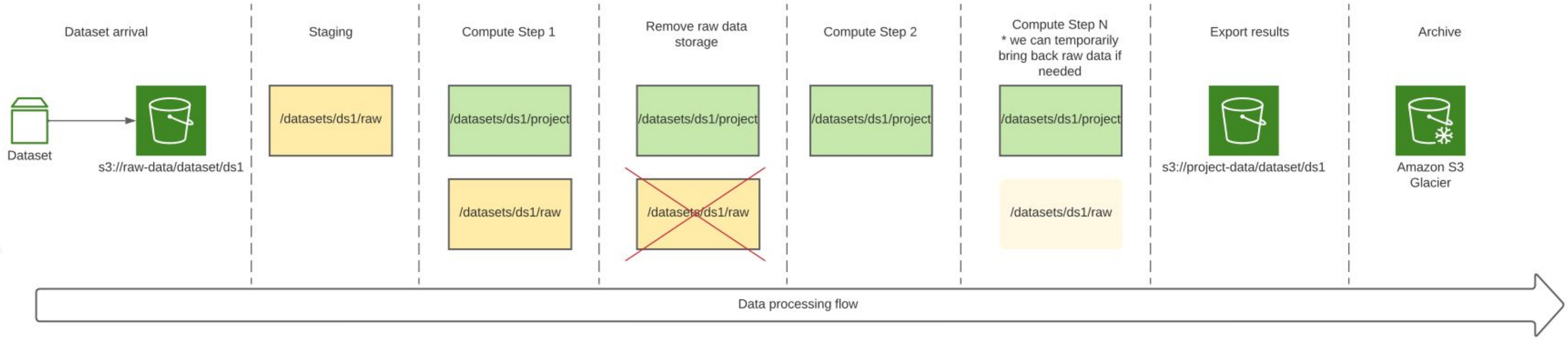
3 POSIX filesystem – data is not organized and tagged:

- Multiple copies
- “Trash” data

4 Data catalog - is hard to implement due to the nature of the filesystem

Dynamic Filesystems Approach in AWS

1. One centralized source – Amazon S3
2. Data can be natively tagged in Amazon S3 as it is object storage
3. Parallel filesystems are being created and removed on-demand
4. Fixed low price for Amazon S3. Fast and expensive filesystems are created on-demand only when needed



Pros and Cons

+ Pros

- Pay for what you use – **significant price reduction**
- Data is tagged naturally when enters the system
- Data lifecycle can be applied easily
- Data is isolated on the filesystem level – performance increase

– Cons

- Development required to adjust to specific data flow

Cryo-EM Platform

BioTechX QUANTORI

Challenge

The client needed an infrastructure with scalable high-speed storage to transform the Cryo-EM datasets into high-resolution protein structures. For this goal Quantori implemented a virtual GPU/CPU hybrid cluster with a multi-tier storage.

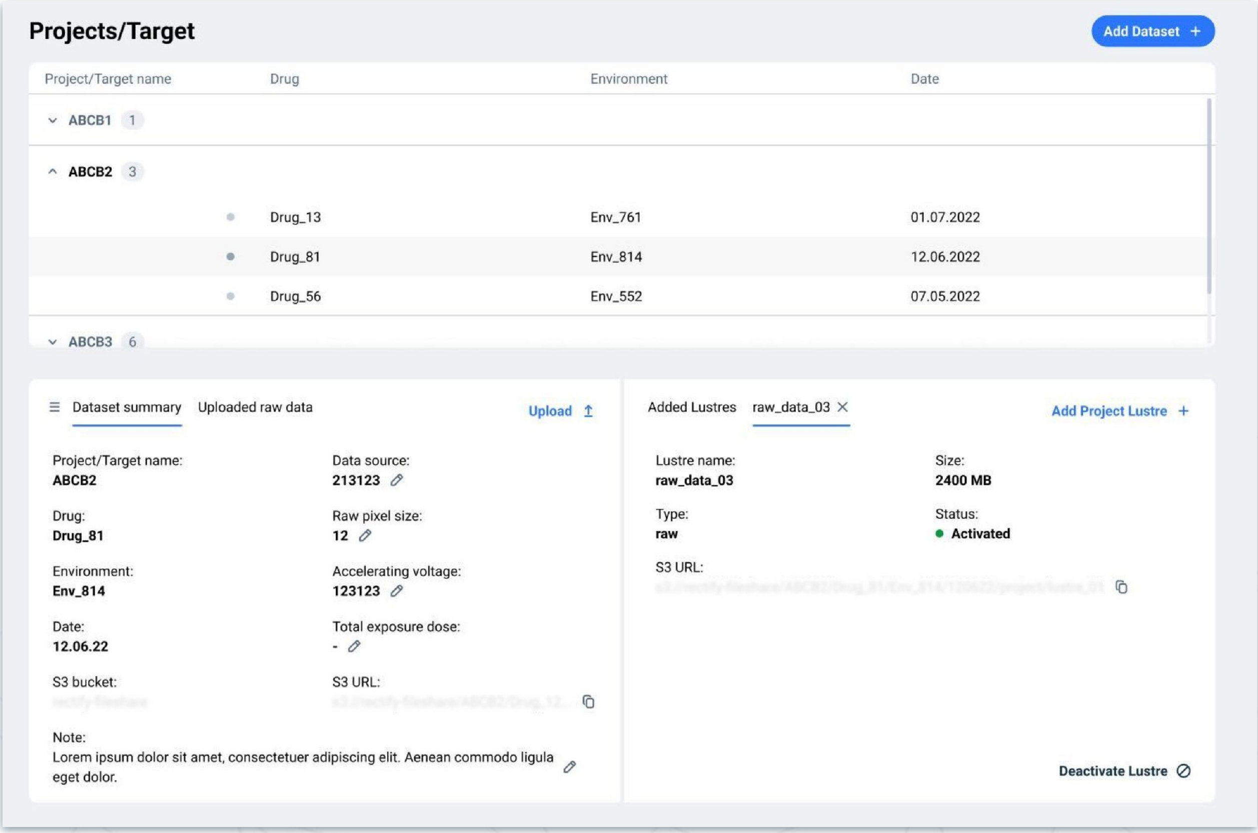
Solution

- Hybrid of Scientific platform & Cryo-EM accelerator.
- AWS cloud-based.
- Run Cryo-EM pipeline on multiple samples at one time.
- Integrated RELION and CryoSparc with the HPC scheduler.
- Results are available for post-processing via console/Python Notebooks/ Rstudio.
- LustreFSx and AWS S3 for scratch and long-term storage.

Benefits

- Platform performance benchmarks fit customer expectation and official RELION/CryoSPARC benchmarks.
- Ability to create/remove Lustre FSx partitions for RAW data significantly reduces storage and platform support costs.
- Research time is reduced due to the availability of results immediately after the computation.
- Batch processing gives a way of the pipeline’s steps automation – which increases efficiency and reduces costs.

Cryo-EM Platform



Projects/Target

Project/Target name		Drug	Environment	Date
ABCB1 1				
ABCB2 3				
	●	Drug_13	Env_761	01.07.2022
	●	Drug_81	Env_814	12.06.2022
	●	Drug_56	Env_552	07.05.2022
ABCB3 6				

Project/Target name:
ABCB2

Drug:
Drug_81

Environment:
Env_814

Date:
12.06.22

S3 bucket:
[redacted]

Data source:
213123

Raw pixel size:
12

Accelerating voltage:
123123

Total exposure dose:
-

S3 URL:
[redacted]

Note:
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor.

Lustre name:
raw_data_03

Type:
raw

S3 URL:
[redacted]

Size:
2400 MB

Status:
● **Activated**

Testimonial

BioTechX QUANTORI

> “*We are very pleased with the results using Quantori's solution. Being able to process multiple CryoEM datasets in parallel makes handling our constant influx of data possible. We are no longer limited computationally, but as it should be in biochemistry.*
>
> *Whenever we run into an issue or have additional features that we would like to be added, the people at Quantori have been great, and I often get a response from Quantori, if not a solution, in about a day. Overall I recommend Quantori for their knowledgeable, professional staff capable of providing solutions cost-effectively and efficiently.*

Bharat Reddy

Senior Scientist

Rectify Pharmaceuticals

QUANTORI

quantori.com