

FACILITATING MEDICAL BREAKTHROUGHS

How Genomics England stores and provides access to petabytes of data

When you've been given a mandate to undertake a population-scale genome sequencing process by 2023, reliable and scalable data storage is absolutely crucial. With the future of medicine at stake, Genomics England chose Nephos to help them scale to extreme performance and capacity.

Huge data requirements for genome sequencing







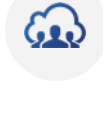
Genomics England Ltd. is owned by the UK Department of Health and Social Care. It was founded to help further medical research and improve patients' lives, by sequencing 100,000 genomes of National Health Service (NHS) volunteer patients and making the resulting data available to 3,000+ researchers worldwide.

After sequencing 100,000 genomes and generating 21 petabytes of data, Genomics England was then tasked to continue sequencing as many NHS patients' genomes as possible by 2023. The expected 140 petabytes of genome data must be stored safely in a single storage system, as researchers need to access the entire data set in randomised ways.

Challenge: Poor performance and limited capacity

The 100,000 Genome Project was supported by a scale-out network-attached storage (NAS) solution. By the project's end, this had reached its storage node scaling limit, and performance was compromised when the system was near capacity. There was no disaster recovery strategy as backing up all 21 petabytes was unfeasible. If a major disaster occurred, critical data would only be available on tape, and would require a lengthy restore process.

Results Achieved:

- 
Increase in performance
- 
Performance expected to continue increasing
- 
No more limits on capacity scaling
- 
Decrease in storage cost per genome (96% expected by 2023)
- 
Provision for fast restoration of critical data should a major disaster occur
- 
Robust authentication to protect against security threats
- 
Integration with a public cloud for computational elasticity.

"For the next phase of this project, we needed something much more scalable – an infrastructure that could grow to hundreds of petabytes. Our existing solution couldn't

Why Genomics England chose Nephos

Genomics England knew that they couldn't continue on with the solution they had and were looking for a new, innovative approach. They hired Nephos Technologies to undertake a strategy and architecture consultancy engagement to review their current requirements and future plans - and map these against possible options in the market. It was Nephos' independent approach to evaluating the possible options, as well as the recommended solution, that gave Genomics England the insights to make this critical decision confidently.

Nephos' ability to provide recommendations and help deploy and manage the chosen solution made the whole process from idea through design to a live system very smooth. Nephos contributed knowledge, support and expertise at every project stage - and have supported Genomics England in ensuring the vendors meet their commitments.

Resilient, scalable, high performance infrastructure achieved

The new infrastructure Nephos set up for Genomics England features WekaFS™ software, as well as Western Digital object storage and Mellanox high speed networking.

A fully parallel and distributed file system, WekaFS™ includes both high-performance flash technology and cost-effective disk storage. Data and metadata are distributed across the entire storage infrastructure to ensure massively parallel access to NVMe drives. Data is seamlessly tiered from flash to disk, resulting in extremely efficient storage media optimisation.

The initial deployment consisted of two tiers: The primary tier of 1.3 petabytes of high-performing NVMe-based flash storage supports the working data sets. The secondary tier of 14 petabytes of object storage provides a long term data lake and repository. Within the first 18 months, this infrastructure had already scaled by 2x. Each tier will continue to scale independently over time.

A new solution would have to:

- Scale capacity in a single space to store 140 petabytes of data
- Improve storage performance to accelerate innovation
- Include a strategy to protect valuable national data from a disaster
- Be easy to manage while delivering the enterprise features needed
- Accommodate tight budget constraints.

Next steps

This new data infrastructure has opened up interesting opportunities for Genomics England. As more research comes online, they plan to integrate a public cloud for computational elasticity. Weka's solution will enable them to burst to the cloud and provide fast, accessible on-demand computing resources for thousands of researchers around the world.

Nephos's relationship with Genomics England has expanded over time, with Nephos now providing a fully managed service on the environment. Nephos is also helping the client look to the future of data security and delivery, including data governance, data cataloging and metadata management.