DDN® STORAGE

SUCCESS STORY

ACCELERATE: LIFE SCIENCES

Public Health England Pioneers Sequencing Services for Infectious Disease Intervention with DDN Storage Solutions

CHALLENGES

- A robust IT infrastructure with large-scale, high-throughput performance is essential to supporting PHE's new centralized sequencing service
- Requirement to turnaround results in hours instead of days to support the need to rapidly and effectively characterize pathogens
- To meet stringent sequencing and analysis demands, a complete data management lifecycle solution with multiple tiers of storage to deliver parallel I/O data throughput, a near-line archive and a private cloud for geographically distributed data access.

SOLUTION

A high-performance computing cluster for variant calling and de novo assembly paired with an SFA-based EXAScaler high-performance parallel storage solution. WOS and iRODs for collaboration between PHE sites and a near-line archive for re-analysis.

RESULTS

- 16x faster processing of genomic data
- Results can be delivered in hours instead of days

PUBLIC HEALTH ENGLAND (PHE) WAS ESTABLISHED IN APRIL 2013 TO CONSOLIDATE HEALTH specialists from more than 70 organizations into a single public health service. As an executive agency of the Department of Health in the United Kingdom, PHE's mission is to protect and improve the nation's health while reducing health inequalities.

PHE employs more than 5,500 people, most of whom are scientists and public health professionals focused on making the public healthier through research, analysis and guidance to government entities as well as supporting action by local government, the National Health Service and other organizations. For example, PHE has been closely monitoring the July 2014 Ebola Virus outbreak in West Africa, which is the largest known outbreak of this disease, to assess risk to the UK and ensure mechanisms are in place to detect and respond to any unusual infections nationwide.

PHE's MS bioinformatics unit has been involved in the establishment of a Next-Generation Sequencing (NGS) Service that provides the means to sequence the whole genomes of pathogens. This sequence can be used to characterize and type pathogens, which in turn can be used, for example, to identify and monitor outbreaks locally and nationally. The same sequence may also help scientists better understand the evolution of bacteria and viruses or predict trends in the patterns of antibiotic resistance. Analysis of pathogen genomes is performed by lots of different groups internationally, but PHE MS will be among the first public health organizations to perform this as part of its routine daily work. By taking advantage of high-performance computing and powerful parallel file system storage, PHE MS will be able to analyze multiple bacteria samples in parallel, which in turn will lead to faster delivery of results to PHE stakeholders such as hospitals.

THE CHALLENGE

To better support its NGS analysis service, PHE MS sought a High-Performance Computing (HPC) system that would enable simultaneous processing of hundreds of bacteria samples received from hospitals and other stakeholders. The objective of PHE's MS bioinformatics unit was to deploy the compute and storage resources needed to support the first centralized service for analyzing bacteria samples. The existing system did not have sufficient capacity, so the group set out to find a powerful platform to drive the new service.

To meet stringent sequencing and analysis demands across the complete research data lifecycle, PHE MS wanted a data management solution that would enable scientists to generate, analyze, archive and share massive amounts of genomics data. In particular, the ability to process multiple whole genome sequencing samples in parallel would enable PHE to meet the service needs of multiple centers, which is necessary to offer a service nationwide; the performance would be especially critical to monitor aggressive pathogens during a major outbreak or emergency response situation. For example, instead of running multiple laboratory tests, PHE MS can obtain a lot more information in one hit, which will help public health scientists to rapidly see how pathogens are related to one another based on the number of shared mutations, as well as if those mutations are responsible for drug resistance.

BUSINESS BENEFITS

- The ability to quickly and effectively characterize major pathogens by determining mutations and gene variants will prove pivotal in reducing major outbreaks and epidemics
- DDN is helping PHE fulfill its vision of speeding the identification and typing of bacteria and viruses
- Scientists across PHE MS will be able to access, analyze, and share large amounts of biological data

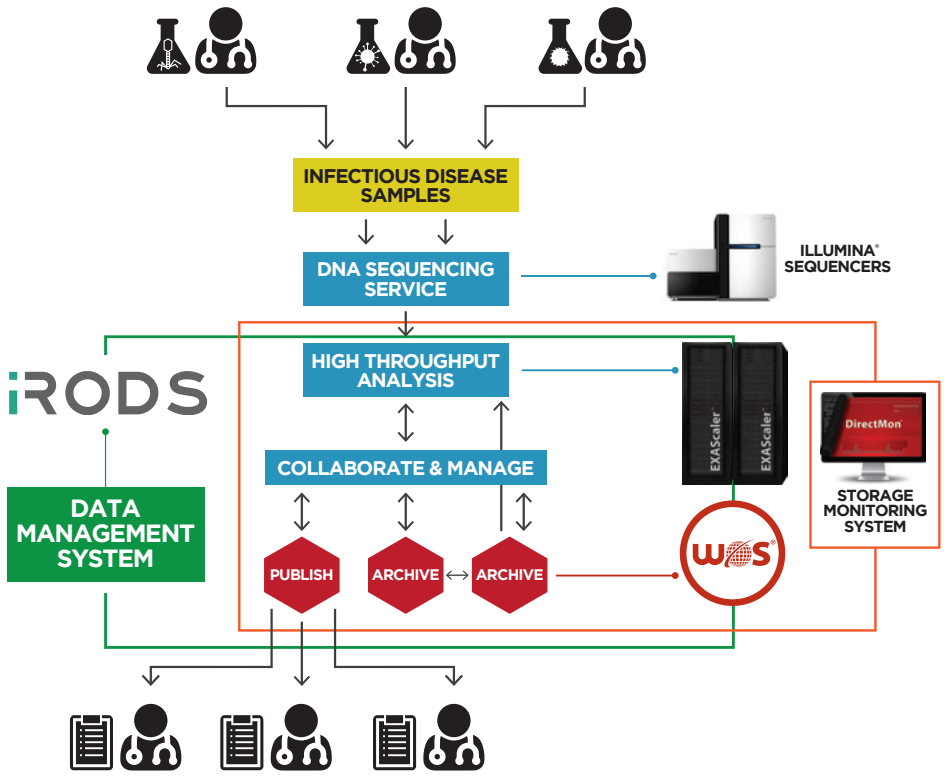
PHE MS conducted a series of feasibility studies to determine the complexity and cost of delivering massive throughput and highly parallel access to storage resources. The team also sought a three-tier storage strategy, encompassing high-performance, parallel I/O; an active archive for storing up to six months of research; as well as cloud-based storage to facilitate collaboration with other researchers worldwide.

PHE MS looked at several technologies to assess which platform would deliver the robust I/O performance, massive scalability and reliability required to process whole genome sequences. Additionally, there was the desire to connect multiple sites seamlessly to expedite sequencing analysis while bolstering resiliency by replicating data between multiple PHE sites. Finally, PHE wanted to provide easy access to an active storage archive to facilitate data sharing with internal, external and private resources around the world. To accomplish this, the team needed a solution that would help organize, share, protect and preserve its vital scientific data. The creation of a private cloud environment, where researchers could access geographically dispersed and replicated file data, would increase productivity ultimately.

THE SOLUTION

To implement the ideal solution, PHE partnered with OCF, a UK-based integrator with a strong partner ecosystem and proven HPC expertise. Storage was a critical part of the decision process in terms of performance and resilience. It was complicated to take a position on storage, as the organization had to have the right mix of software to cost-effectively achieve the required performance. The complexity of the software factored into the final decision, as well as how the system would be maintained, expanded and administered.

In particular, the choice of Lustre as the underlying file system was given careful consideration. The team was looking for open source software to prevent the technology stack used from being locked into a single vendor, in line with UK government policy. For that reason, Lustre was appealing even though PHE lacked onsite expertise to implement and manage the system. Looking to the future, PHE also saw a trend where Lustre was being used more by the Top500® supercomputers globally. As a result, it was felt that this approach would best meet the institution's needs, despite the learning curve required for deployment and support.



TECHNICAL BENEFITS

- Highly parallel I/O enables PHE to process and analyze hundreds of DNA samples simultaneously
- PHE gains seamless long-term data access through WOS active archive and offsite backup and data lifecycle management using iRODS
- PHE is better positioned to meet the stringent demands of its bioinformatics analysis workflows, using best practice tools for variant analysis, mutation detection and de novo assembly

The PHE team spoke with colleagues, including researchers at the Wellcome Trust Sanger Institute, another UK-based genome research powerhouse, which deployed more than 22PB of EXAScaler™, DDN's high-performance Lustre® storage solution. Wellcome Trust stressed the need for a storage platform with lots of parallel I/O. The requirement to stream data in parallel also was key.
The next step was determining whether EXAScaler was best suited to power PHE's Lustre parallel file system. As reducing the complexity and learning curve associated with Lustre could be mitigated by choosing the right storage partner, DDN's Lustre leadership, significant knowledge and industry experience were major drivers in making the decision.

PHE noted DDN's strong participation in the scientific community, which gave the company a perspective well beyond its own products. Because of DDN's capacity and ability to talk about issues that are highly important to the scientific community, PHE MS gained a lot of insight into other technologies and applications that weren't part of DDN's core products, but were essential to the research process and desired collaboration with other scientists and researchers.

As a result, PHE deployed DDN's SFA10K® EXAScaler storage appliance with the Lustre parallel file system and 300TBs of high-performance storage. Together with OCF, the team worked to implement DDN's SFA high-performance storage system to support the massive amounts of research data that would be generated by their NGS service. Shortly after this deployment, PHE selected DDN's WOS® Object Storage platform with 360TB of capacity to serve both as an active archive for valuable public data and as an offsite backup for its primary sites.

The idea was to implement a system, along with a scientific lifecycle application, integrated Rules-Oriented Data Management System (iRODS), which would make accessing PHE's sequencing data and analysis as easy as possible. As it was critical for internal and external scientists to quickly and easily interact and collaborate, PHE decided to invest in a proper scientific system designed for researchers, and decided DDN WOS with iRODS was the best fit. DDN's involvement with the iRODS consortium was an added plus since the company's ability to embed iRODS within the storage system ultimately will make it easier for researchers to view, manage, access and share data.

THE BENEFITS

Since deploying its high-performance DDN storage platform, PHE has been piloting a new centralized service as one of several alternative models for delivery of whole genome sequencing and analysis that offers the promise to better support public health interventions. The initial phase of the sequencing service relies on two Illumina HiSeq and two MiSeq sequencing machines to generate massive amounts of DNA sequence data from diverse bacteria and virus samples. Previously, PHE relied on dedicated scientific workstations, which could process about a dozen samples at a time. Now, the central bioinformatics core group takes advantage of robust parallel I/O performance to process 192 samples simultaneously—in approximately the same time it took previously to process 12 samples. The real game changer is the ability to process 200 samples in parallel without any performance hits. With a computing cluster and DDN's high-performance parallel storage, results from 200 samples can be turned around in hours instead of days.

During the high season for certain viruses and bacteria, PHE could need to process thousands of samples weekly, so the team currently is building the system to respond to those kinds of numbers. PHE already has processed more than half a million jobs, many of which have been part of the sequence analysis of over 10,000 samples. The plan is to continue to scale so PHE will be able to turnaround sequencing and bioinformatics analysis to customers in around four days. This wasn't possible before using the workstation-based infrastructure.

PHE also is better positioned now to meet the stringent demands of its bioinformatics analysis standards. The team uses BWA, Bowtie and several variant callers for mapping data to reference sequences, variant analysis and mutation detection, which is a key PHE workflow. To perform de novo assembly of new organisms, the team also uses Velvet and Spades. PHE uses Galaxy to provide a web interface for interacting with the HPC cluster and for training purposes. In supporting the scientific lifecycle, PHE also is building a solution that will be able to help people around the world access this data.

Seamless access to data through DDN WOS and iRODS is another overarching benefit of the new system. DDN WOS provides active archive and availability, while iRODS delivers visibility and workflow

management across a wider variety of heterogeneous data sources and instruments. With WOS, metadata can be assigned that will make it easier to locate critical sequence data and associated metadata wherever they reside. In fact, DDN's highly parallelized metadata search will empower researchers to efficiently query one or more indexed fields of user-specified tags across all the data storage tiers available in PHE MS.

In managing its scientific data lifecycle, PHE will use WOS to retain data as long as possible in an active archive that will allow researchers to continue to seamlessly access and use the data. Even after an old project has reached the stage in its lifecycle where it has been archived, the data can be easily accessed for re-analysis, whether because protocols have changed or for broader population analysis of earlier individual samples. The organization will also have a private cloud leveraging WOS versatility to serve as an offsite backup replicating data between different geographic locations, both for disaster recovery and research collaboration. The flexibility of iRODS in conjunction with WOS supports PHE's model to make appropriate taxpayer-funded data accessible publicly, perhaps through data repositories such as the short read archive at the European Bioinformatics Institute.

The ability to share access to microbial genome data could play a vital role in rapidly monitoring the spread of a potentially harmful variant of a pathogen and providing crucial clues to phenotypes such as drug resistance, which may lead to public health interventions that would limit further spread. DDN is helping PHE provide the infrastructure that underpins its vision to use genomics primarily to protect and improve public health as well as making data available that may help the organization and its partners better characterize the relationship between pathogen and disease. With data from PHE's centralized sequencing service stored on hardware provided by DDN, the organization will be able to ask questions that could not be asked before and get answers that can lead to new scientific breakthroughs.

ABOUT DDN®

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. For more than 15 years, DDN has designed, developed, deployed and optimized systems, software and solutions that enable enterprises, service providers, universities and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud. Organizations leverage the power of DDN technology and the deep technical expertise of its team to capture, store, process, analyze, collaborate and distribute data, information and content at largest scale in the most efficient, reliable and cost effective manner. DDN customers include many of the world's leading financial services firms and banks, healthcare and life science organizations, manufacturing and energy companies, government and research facilities, and web and cloud service providers. For more information, visit our website www.ddn.com or call 1-800-837-2298.

v1 (3/15)