The NUS occupational health genome project powers breakthrough research in Singapore with AWS

NUS Business School is the business school of the National University of Singapore (NUS). It is consistently rated as one of the top business schools in Asia Pacific and worldwide. Its faculty conducts impactful research, and one such example is the NUS occupational health genome project, which uses insights from genetic and environmental analyses, gene mapping, and genetic pleiotropy to transform the understanding of the relationship between occupation and wellbeing—and to help improve the lives of people in Singapore and beyond.

The data processed in the project's studies involves large scale whole genome or exome sequencing processed datasets of up to 500,000 samples. They typically range from gigabytes to terabytes or even larger, and require a scalable, reliable, high performance computing (HPC) environment. The researchers used the National Supercomputing Center (NSCC) in Singapore and an in-house data center for this task. However, when issues began to arise accessing and storing large volumes of data and collaborating with other institutions, they decided to create a pilot collaboration with Amazon Web Services (AWS) to explore the use of the AWS Cloud to run genomic analyses.

AWS was selected because it offered the services, expertise, and track record of success the researchers needed. The AWS Cloud offered flexible, pay-as-you-go access to a spectrum of computing resources, and the ability to store the voluminous amount of genome datasets with additional meta data information. AWS Command Line Interface (CLI)

migrated the mass scale genomics datasets into the resilient storage environment. The researchers began to migrate data and workloads to AWS in February 2021. They used scalable high performance computing (HPC) environment to provision the optimal compute resources to handle the large scale genomics analysis more efficiently. Multiple AWS services were used, including Amazon EMR, Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), and Amazon FSx for Lustre. The researchers benefitted from constant interactions and a few deep dive sessions with AWS on how to get the most out of the AWS Cloud, as well as comprehensive, ongoing support services.

Today, the researchers are using the NSCC to store data and run most of the project's genomic analyses. The AWS Cloud, meanwhile, is helping them optimize their work with its better accessibility, particularly when the collaborators and students are overseas; better group management; parallel computing capacity; and data repository and web service capabilities. In addition, they are using AWS to run analyses of GWAS plots and genetic correlations using summary statistics.

Looking ahead, the researchers are working with AWS to optimize and build a reusable protocol to further streamline processes and to launch a dedicated website for collaborating and sharing results with other medical institutions and organizations around the world.