Digital Health: Changing the Way Healthcare is Conceptualised and Delivered
E. Cummings et al. (Eds.)

doi:10.3233/SHTI190787

Running Genomic Analyses in the Cloud

Conrad LEONARD[a], Scott WOOD[a], Oliver HOLMES[a], Nic WADDELL[a], Dominique GORSE[b], David P. HANSEN[c] and John V. PEARSON[a,1]
[a]*QIMR Berghofer Medical Research Institute, Brisbane, Australia*
[b]*QFAB Bioinformatics, Brisbane, Australia*
[c]*Australian e-Health Research Centre, Brisbane, Australia*

Abstract. Genomic testing is rapidly moving into healthcare practice. However it comes with informatics challenges that the healthcare system has not previously faced - the raw data can be hundreds of gigabytes per test, the compute demands can be thousands of CPU hours, and the test can reveal deeply private health-related information that can have implications for anyone related to the person tested. While not a panacea, cloud computing has particular properties that can ameliorate some of these difficulties. This paper presents some of the key lessons learned while deploying a set of genomic analyses on cloud computing for Queensland Genomics.

Keywords. Genomics, cloud computing, next-generation sequencing

Introduction

Queensland Genomics is a $25 million initiative created by the Queensland Government in 2016 to translate research genomics into healthcare practice. The first round of QGHA funding included clinical genomics demonstration projects and a number of capability workstreams that were tasked with providing infrastructure to support the clinical projects. The Genomic Information Management (GIM) capability team was responsible for providing informatics infrastructure which included bioinformatics, genome analytics and a genome data repository. In this paper, we present some of the lessons learned by the GIM team as they deployed a cloud-based genomic analysis platform for Queensland Genomics.

Genomic testing is related to but differs from genetic testing. Genetic testing involves testing of targeted regions of the genome and is a routine part of current clinical practice. The January 2019 version of the Australian Government Department of Health's Medicare Benefits Schedule (MBS) Book lists almost 50 genetic tests for which there is an MBS rebate. These tests include fluorescence *in situ* hybridisation (FISH) to look for large scale chromosomal rearrangements, examination of chromosomes by cytogenetics or genome-wide microarrays and characterisation of germline or somatic mutations in single genes that are implicated in some cancers and inherited diseases.

In contrast with genetic testing, genomic testing is not rebated under the MBS and is less commonly seen in clinical practice. Genomic testing uses a technology called next-generation sequencing (NGS) to look at the human genome at a much higher level of detail and can generate data that is thousands to millions of times larger than the data

[1] Corresponding Author: John V Pearson, QIMR Berghofer Medical Research Institute, 300 Herston Rd, Herston 4006, Queensland, Australia; E-mail: john.pearson@qimrberghofer.edu.au.

generated by genetic tests. Genomic testing is a loosely defined term that is often used to include disparate types of testing including determining the DNA sequence of the genome, cataloging the sequences of RNA transcripts (transcriptomics), and examining chemical modifications of some parts of the genome (epigenomics). For the purposes of this discussion, we will use the term genomic test only to refer to the determination of the DNA sequence of some or all of the genome in a given sample.

Genomic testing of DNA can be undertaken at a variety of scales and these are usually grouped into three categories – gene panels, exomes and genomes. Gene panels are the smallest scale of genomic test and are targeted at genes that are relevant to a particular disease, for example a cardiac panel would only include genes that have been implicated in cardiac disease and a metabolic panel would only include genes that have been determined to be relevant to metabolic disorders. Gene panels come in a range of sizes, but they typically include dozens to hundreds of genes and cover less than 0.1% of the whole human genome. An exome test sequences all of the approximately 20,000 genes in the human genome and covers 2% of the whole human genome. Whole genome sequencing (WGS) covers the entire human genome with the exception of some regions that are difficult to sequence.

Genomic analyses are also often not achieved by running a single piece of software but may instead involve running hundreds of different software jobs using dozens of different software programs and where outputs from one or more earlier jobs are used as inputs for later jobs. These complex meshes of interrelated units of work are called workflows and new computer languages are appearing that are designed to allow abstract representations of these workflows. Workflow Description Language (WDL) and Common Workflow Language (CWL) are new standards worth considering when developing workflows for genomic analyses.

1. Analysing Genomes in the Cloud

1.1. Adapting to the New Environment

There are a number of technical issues involved in migrating analysis workflows from on-premise computing infrastructure to cloud and there are also a number of basic operational processes that need to be re-thought. One of the first of these is paying for cloud use. Many organisations do not charge for internal use of on-premise resources. This is in contrast to cloud infrastructure where every operation has a cost. Cloud users must commit to careful testing of workflows as they will be charged for the resources used regardless of whether the workflow succeeds or not. Failing on the final operation of a long and complex workflow can be expensive as well as frustrating. Cloud workflow testing should be designed so that the smallest possible dataset that exercises all aspects of the analysis should be used for development until the workflow is ready to enter system testing.

Data transfers should also be carefully monitored. As a general rule, uploading data to a cloud vendor is low or zero cost however downloading data or moving it between geographic locations can be very costly. Where possible genomic analyses should be moved to cloud compute resources in the same region as the cloud data storage rather than moving the data. For collaborators or clients who already host their datasets in the cloud, having them provide access to their data in place can be a powerful enabler by removing the need for data transfer. It is worth noting that in general the owner of the data is responsible for any charges accrued due to data transfer so any time you are going to process cloud data owned by others, it is wise to do careful testing beforehand to estimate costs and to have a discussion up-front about those costs and how they are going to be met.

Cloud adoption can also disrupt existing practices and require organisational change:

- Purchasing on-premise computing is a capital expenditure whereas cloud computing is an operational expense.
- Cloud must be paid for monthly, bills can vary widely from month-to-month, and planning can be difficult because billing occurs after the work is complete.
- Cloud computing may not easily map onto existing tender and procurement processes.

There can also be significant organisational financial benefits to cloud adoption including access to effectively unlimited informatics infrastructure without a requirement to purchase; cost savings through not having to buy infrastructure to cope with peak demand and then having it stand idle at other times; and minimizing opportunity costs by not having to buy infrastructure in advance of projected need.

1.2. Interacting with Cloud Systems

Cloud vendors generally provide multiple ways to interact with their system including a web-based console, a set of command-line utilities, and one or more application programming interfaces (API's) that can be used to create custom software that can interact with and manage services on the cloud platform. While the web console is the initial way that most cloud adopters will interact with their platform, across time it is generally sensible to migrate to the more complicated but much more powerful command-line and API interfaces.

1.3. Storage

The amount of data that is generated by genomics workflows can be enormous. A typical analysis workflow for a single WGS cancer subject (test vs control) will use and generate almost 1TB of data with a residual data footprint of 300-600GB depending on whether both the raw and processed data is retained. Fortunately, the amount of storage that cloud providers offer is effectively infinite however every gigabyte of storage that is used must be paid for.

Storage on cloud platforms can differ significantly from on-premise and may not be able to be presented to software applications in the same way. Cloud storage systems are typically object based, whereas on-premise systems are often block storage volumes. This may mean that a workflow will need to be modified to stage data files from an object store to a block store prior to working on it. Additionally, once data that is to be persisted has been generated, it needs to be pushed to the cloud storage service. This will often differ from on-premise workflows where the data inputs and outputs for a workflow are often on the same filesystem.

One benefit of the pay-for-what-you-use model on cloud is that users are more likely to refine the data that is kept from a genomics workflow in order to lessen the storage footprint and associated costs.

Cloud storage systems are typically more secure and resilient than on-premise systems however those benefits only apply if cloud users understand, incorporate and implement the security and resilience features offered by their cloud vendor. Cloud is no different than on-premise in that poor design leads to poor outcomes. Cloud platforms will often have both data replication and archiving built in and the replication may be across data centres in different locations to mitigate against local physical disasters. Many cloud providers have data centres in different countries and replicating data between these can be part of an organisation's back-up and disaster recovery plans. It should be noted that due to the highly personalised nature of genomic data, it is important to check that storing the data in another country will not infringe upon any legal, regulatory or compliance requirements.

1.4. Compute

Cloud providers generally offer a wide range of nodes with different compute, memory, storage and network specifications upon which to run your genomic analyses. This has the benefit of allowing you to choose the compute node that most closely matches the requirements of your workflow and depending on how your workflows are designed and run, can even allow you to choose different node types to run different work units within a workflow. Nodes can also usually be configured with a range of operating systems, which adds flexibility to your workflow design and allows you to use the most appropriate solution for each of your work units. There can still be difficulties with workflows that have been designed on-premise and have been optimised for the nodes available if there is no closely matching cloud node type.

Despite the many choices, there is often still a trade-off that must be made between cost and execution time when choosing a node for a particular task. An optimised node with many fast cores that has a high memory to core ratio may process a task much quicker than a standard node, but the costs in running the optimised node may be much higher than the standard mode, even though it processes the task more quickly. This trade-off is best addressed at the project level, where the priorities of the project dictate which path to go down

Genomic analyses can often take a long time from start to finish. When testing a workflow in the cloud, it is best to have a small data set so that any failures due to workflow modifications or hardware differences can be identified quickly. It is much better to fail early and inexpensively, than for a workflow to fail at a late stage when a lot of work has already been done and must be paid for. It should be noted that the successful completion of a workflow against a small dataset does not necessarily mean that a large, or real-life dataset will also be successful. There could be errors in the workflow that only present themselves when dealing with large data sets.

Workflow failures when running on cloud nodes can also often be trickier to diagnose than when running on-premise. Once an instance is no longer under your control (i.e. if it becomes unresponsive or fails) you may not have access to the information (log files etc) required to diagnose the problem. One way to mitigate against this is to initially configure cloud compute nodes so that they most closely mirror your on-premise setup and introduce new optimised cloud nodes only once the workflow has been successfully migrated in a non-optimised form.

Cloud compute services often allow for the running of containerised workflows. The advantages here include the ability to design workflows that are not only cloud agnostic but that can also be run on-premise. This is a significant benefit for organisations pursuing hybrid private-public cloud strategies where they will run work on-premise and burst out to cloud when their internal resources are fully utilised.

1.5. Data Transfer

Data transfer in an out of cloud storage is often overlooked as a critical part of cloud migrations however it is at the moment of moving data between cloud and on-premise that data can be most at risk as it moves from one security environment to another and this must be carefully considered in any cloud implementation project.

The actual data transfer can be done using commercial cloud-aware software, using software developed in-house using the vendor's APIs, or using the vendor command-line tools, for example: the Amazon Web Services Command Line Interface or 'awscli'. The awscli tool must be installed on a system where you have access to the data to be uploaded and you then use awscli commands to upload the files through your internet connection to the cloud storage system where you want the data to reside.

Data transfers can be a significant cost, and this must be factored into any cloud data storage strategy. In general, data transfers into a cloud storage system are zero or low cost however all subsequent transfers may attract a cost. Transfers between a cloud vendor's services are usually free as long as the transfer stays within the same location however transfers out to an internet location or to a vendor's storage in a different location can be very costly.

For larger data sets, vendors may have storage appliances that can be physically shipped to your premises, loaded with data and couriered back to the vendor who will upload that data to your cloud storage. Amazon Web Services (AWS) offers this facility – their data transfer appliance is called Snowball and it comes in 50 TB and 80 TB sizes. One of the principal advantages of an appliance is that copying files locally may be much faster than network transfers depending on your organisation's internet connection. Appliances are not allowed to cross national boundaries, so they cannot be used as a mechanism to bypass the high costs that apply to data transfers between vendor storage systems in different locations.

Data transfer is one of the areas where cloud can initially seem more complicated than on-premise processes where data can be placed on encrypted physical hard drives and shipped by courier. On receipt of an encrypted drive, the transfer process is as simple as plugging in the drive and using the decryption password to copy the files. While the on-premise process can seem simpler, a properly automated cloud data transfer can be as simple and has the advantage of not involving physical drives that can be lost, delayed, misdirected or stolen. It should also be noted that a key cloud advantage is that it can make data transfer disappear entirely. Two cloud-based organisations could negotiate a set of policies that made it possible for them to securely share data in the cloud so that it is never transferred at all.

1.6. Security

To a large degree the challenges involved in and principles for managing security of genomic data are shared between on-premise and cloud environments with only the implementations differing.

Genomic data can include raw and analysed sequencing data, genomic variants detected by the sequencing, and protected health information (PHI) including patient details and clinical data. Some of this data is only required to be accessible for a short period and can then be archived however some needs to be online and accessible indefinitely.

For sensitive data and the services that process it, we need to control:

- Access — traffic control; can you drive right up to my house from the highway or only from my driveway?
- Authentication — once you are at my house, are you who you say you are?
- Authorisation — given you are who you say you are, what actions are you allowed to perform?

Table 1 shows a comparison of some of the possible security practices that could be used to address basic security concerns on-premise and on the cloud. The names of the services will differ between cloud vendors but the same basic principles will apply.

Table 1. An example comparison of security practices for on-premise vs cloud computing on Amazon Web Services (AWS)

Process	On-premise computing	Cloud computing
Access	Organisation firewall	Virtual Private Cloud (VPC) and security groups
Authentication	Active Directory (AD) / Lightweight Directory Access Protocol (LDAP)	Identity and Access Management (IAM) users
Authorisation	AD/LDAP groups mapped to Linux security groups, database roles and filesystem Access Control Lists (ACLs)	IAM groups, roles and policies
Encryption at-rest	GNU Privacy Guard (GPG) file encryption / Linux Unified Key Setup (LUKS) filesystem encryption	Enable encryption at the level of Amazon Simple Storage Service (S3) buckets
Encryption in-flight	Secure Sockets Layer (SSL) / Transport Layer Security (TLS) everywhere	Secure Sockets Layer (SSL) / Transport Layer Security (TLS) everywhere

A key part of security is managing risk and there are a number of strategies that can be used including:

- Ask for, accept and store only the data you need - don't take on unnecessary privacy liabilities.
- Always use the principle of least privilege - give each class of operations staff the permissions they need to accomplish their work, and no more.
- Create an audit trail — use named accounts and log access to sensitive information.
- Perform security audits and reviews regularly and preferably use an unrelated external organisation.

2. Conclusions

Migrating genomic analyses to the cloud is a non-trivial task however there are significant potential benefits if the process is carefully planned and managed. Cloud provides access to informatics infrastructure as and when needed, there is no need to purchase ahead of projected use and if you use less resources than expected, you don’t pay for what you don't use. Managing cloud storage, compute, and security can involve a steep learning curve for users and organisations that are not familiar with cloud however the issues are not fundamentally different that those associated with on-premise infrastructure. Genomic testing and cloud computing are no longer technologies of the future. They are in common use and they have some substantial synergies that are there to be taken advantage of by forward-thinking health care organisations.

Acknowledgement

Funding for this project was provided through Queensland Genomics, Queensland Health, and the Queensland Government.