Centre for Victorian Data Linkage

Virtual Machine access model

The CVDL data access model for linked data has changed

The Centre for Victorian Data Linkage has developed a secure, online platform, Victorian data Access Linkage Trust (VALT), for release and analysis of linked, de-identified data for approved projects. This platform has been in beta version for several months and has been used by selected linkage projects during the proof of concept phase. Feedback from our proof of concept users has been positive, indicating that the Virtual Machines (VMs) are powerful and suitable for undertaking complex analysis of large datasets.

Once the VALT environment is fully operational, release of de-identified linked data for approved research and government projects will progressively take place via this secure environment. The CVDL will provide researchers with access to de-identified linked data within the secure cloud environment with analysis undertaken via Virtual Machines.

Why is the CVDL introducing this secure environment?

Data linkage authorities in Australia and internationally already provide access to de-identified linked data through a secure platform instead of providing extracts directly to researchers, and others are currently developing similar secure platforms. The use of a secure platform such as VALT enables researchers to access linked de-identified data from their own computers while ensuring that the original data is securely held in the CVDL technical environment.

The CVDL now routinely links 35 plus health, human services, justice and education datasets, and provision of data within a secure environment contributes to public and stakeholder confidence in the security and appropriate use of the data.

The CVDL typically provides research extracts at a unit record level to enable detailed analysis. While these extracts are de-identified, there remains a risk of re-identification. The CVDL recognises that, depending on the number of linked datasets and data elements provided, data extracts may never be truly protected from re-identification.

It is expected that VALT will facilitate a higher level of compliance with privacy legislation and information management standards than has previously been possible. This reflects that the linked data asset held by the CVDL is currently much larger and more complex than in the early days of the CVDL operations, and is consistent with public expectations regarding security of detailed person-level data assets.

How secure is the platform?

The Microsoft Azure platform that the CVDL uses is very secure. The Azure cloud environment has been certified up to “Protected” by the Information Security Registered Assessors Program (IRAP) of the Commonwealth. The IRAP is an Australian Signals Directorate (ASD) initiative to provide high-quality information and communications technology (ICT) services to government in support of Australia's security.

How will I access the linked data?

Researchers with an approved CVDL's linkage project apply for access to a project VM by completing the Virtual Machine Application form (Appendix 1) and submitting the request to cvdl@health.vic.gov.au.

Separate virtual machines (data environments) will be set up for each approved project, and registered users provided with access to their project specific data via their project virtual machine. While multiple researchers can be registered to use a specific virtual machine, however, there can only be two concurrent users on each virtual machine.

The linked data released to the virtual machine may be at unit record level once approved by data custodians. Unless negotiated with the CVDL and agreed by data custodians, the standard CVDL's data de-identification and confidentialisation processes will apply to unit record data access via the virtual machine. These processes are described in the CVDL application form.

Researchers can analyse the data on the virtual machine, using available software such as R, Python and SQL. Researchers will also have access to Microsoft Office suite of products. Researchers may request to install preferred software such as Stata on a virtual machine on a BYO licence arrangement. Approved researchers will be provided with a log-on to access the virtual machine.

Note that Power BI and SAS software are not available in the virtual machine environment due to licencing limitations.

How do I remove my research results?

Researchers require authorisation from the CVDL for removal of research results and outputs from the Virtual Machine to ensure a sufficient level of aggregation has been undertaken to meet privacy and confidentiality requirements. In general, this will involve release of analysis, modelling, graphs and tables. As a guide, the CVDL requests that cells with counts below the threshold of 5 should be avoided by either combining categories or suppressing cells.

Approval may be provided for removal of unit record outputs from the virtual machine as an exception. Specific confidentialisation and release requirements should be raised with the CVDL as early as possible as part of the application process to enable discussion with data custodians.

Wherever possible, please provide three days' notice for removal of outputs.

Note that approval to remove data from the secure environment does not override the need for researchers to seek further approval from the CVDL and data custodians prior to publication of any results.

Is there a cost?

The CVDL is charged by Microsoft for cloud computing processing power that is used by the secure environment. Cost-recovery will therefore be applied to each project for the supply of the VM environment. The estimated monthly cost for use 24 hours, 7 days per week depends on the project size, CPU and memory, and are indicted in the table below.

The CVDL will estimate the cost of projects, based on the amount of cloud computer processing power required in consultation with the researcher at the time of application. Payment of the estimated cost will be required prior to data access being provided. Note that the cost is incurred only for the hours that the virtual machine is switched on, and the virtual machine should be turned off when not in use, otherwise costs will be incurred.

Finalised cost will be available at the end of the project. The CVDL will send an invoice reflecting the total cost and pre-paid amounts, any dues or refunds will be processed then.

Researchers applying for project funding should consider the estimated cost for accessing the linked data in their grant applications.

Machine Size	Estimated 24/7 monthly cost (including shared infrastructure cost 30%)	Estimated per hour cost (including shared infrastructure cost 30%)
D2s v3 (2 vCPU, 8 GB RAM)	$260	$0.35
D4s v3 (4 vCPU, 16 GB RAM)	$519	$0.72
F8s (8 vCPU, 16 GB RAM)	$1,078	$1.48
D8s v3 (8 vCPU, 32 GB RAM)	$1,361	$1.86
F16s (16 vCPU, 32 GB RAM)	$2,154	$2.95
D16s v3 (16 vCPU, 64 GB RAM)	$2,079	$2.85
D32s v3 (32 vCPU, 128 GB RAM)	$4,159	$5.69
The below SKUs are recommended by Microsoft.		
DS11_v2 (2 vCPU, 14 GB RAM)	$329	$0.46
DS12_v2 (4 vCPU, 28 GB RAM)	$657	$0.90
DS13_v2 (8 vCPU, 56 GB RAM)	$1,313	$1.79
DS14_v2 (16 vCPU, 112 GB RAM)	$2,624	$3.60
DS15_v2 (20 vCPU, 140 GB RAM)	$3,282	$4.50

To receive this document in another format, email cvdl@health.vic.gov.au

Authorised and published by the Victorian Government, 1 Treasury Place, Melbourne.

Except where otherwise indicated, the images in this document show models and illustrative settings only, and do not necessarily depict actual services, facilities or recipients of services. This document may contain images of deceased Aboriginal and Torres Strait Islander peoples.

In this document, 'Aboriginal' refers to both Aboriginal and Torres Strait Islander people. 'Indigenous' or 'Koori/Koorie' is retained when part of the title of a report, program or quotation.

Available at https://www2.health.vic.gov.au/cvdl