# Scalable High Demand Analytics Environments with Heterogeneous Clouds

K. Woods, R. Clegg, R. Millward, Tessella Ltd, Abingdon, UK
F. Barnsley, C. Jones, STFC/RAL, Didcot, UK

## Problem

World-leading research facilities, such as Diamond Light Source, ISIS neutron and muon source and Central Laser Facility (CLF) generate large and increasing ***volumes of data***. The variety of scientific techniques employed by researchers requires a lot of ***different but consistent analysis environments***. This is compounded with the challenges of a highly ***un-predictable demand*** and the desire to run an analysis close to a scientist's home institute.

## The Ada Lovelace Centre (ALC)

The ALC, part of the Science & Technology Facilities Council (STFC), has a remit to transform the use of real time data processing, computer simulation and data analytics to deliver more effective research at UK national facilities. The ALC provides on-demand, data analysis, interpretation and analytics services to scientists using UK research facilities.

The Ada Lovelace Centre and Tessella have developed a set of tools to facilitate the ***scaling of computing infrastructure*** in order to respond flexibly to variations in demand. This is primarily done by ***provisioning on to multiple clouds***, whether other institutions or commercial cloud providers, and then orchestrating analysis environments across clouds.

## External Cloud Provisioning

**Ansible** (via **AWX**) running in the home environment provisions the network infrastructure first in an external cloud. Libraries within Ansible abstract away a lot of the cloud specifics. For security, public and private subnets are built with associated internet and NAT gateways. A **bastion** server is used to ensure secure provisioning from the home environment.

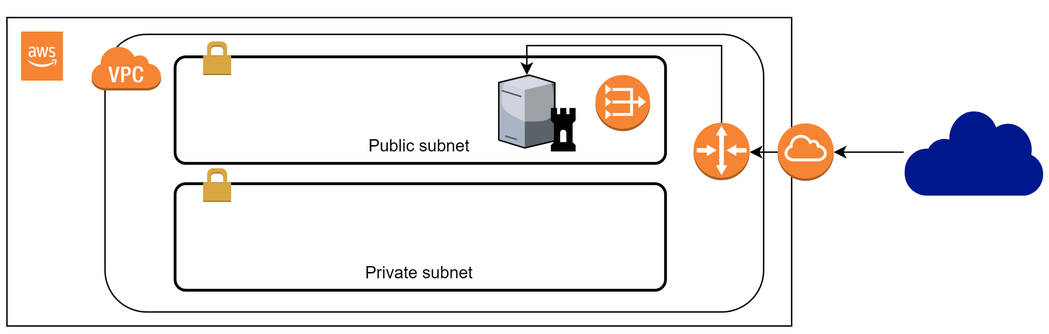


**Figure 1:** First step, provision network infrastructure

Once the network is in place then infrastructure servers are provisioned through the bastion, for example a **Squid proxy**, a **CernVM file system proxy** and a ***Data Movement System***. Communication back to the home network is through an **OpenVPN** client server pair.

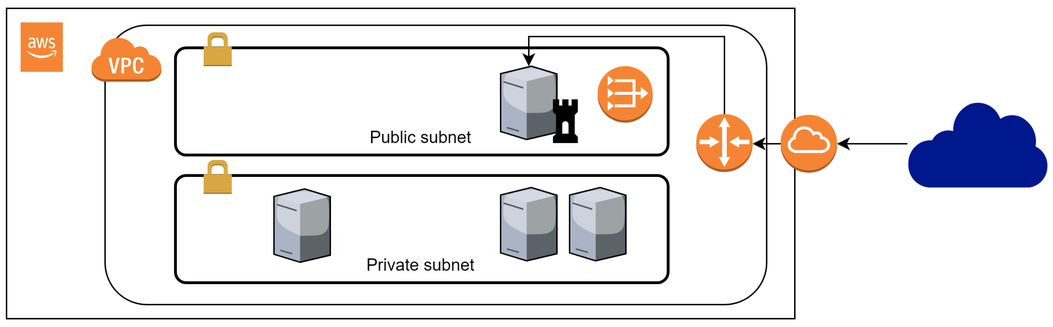


**Figure 2:** Second step, provision private infrastructure servers

With all the supporting systems in-place then analysis VMs can be dynamically provisioned into the public subnet as and when needed using the virtual machine manager.

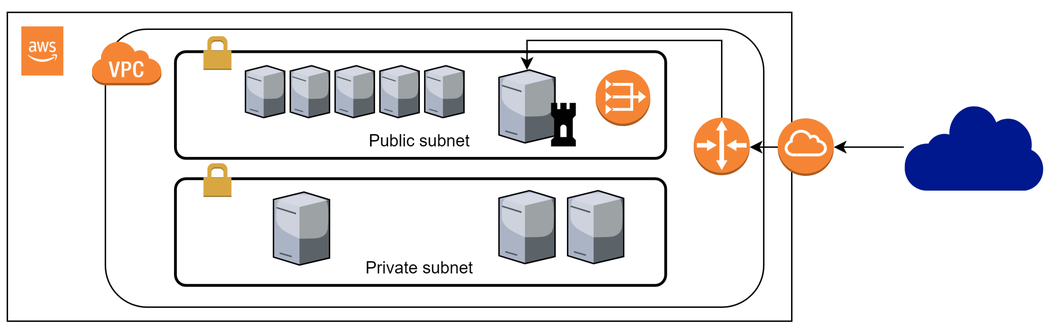


**Figure 3:** Third step, dynamically provision analysis VMs

## Supporting Utilities

Other supporting utilities integrate the external cloud resource into the wider data analysis capability.

- The ***Virtual Machine Manager*** (***VMM***) orchestrates multiple pools of virtual machines (VM) across clouds and utilises **libcloud** to make it cloud agnostic. It is responsible for provisioning all the types of VMs that are needed as well as managing their lifecycles.

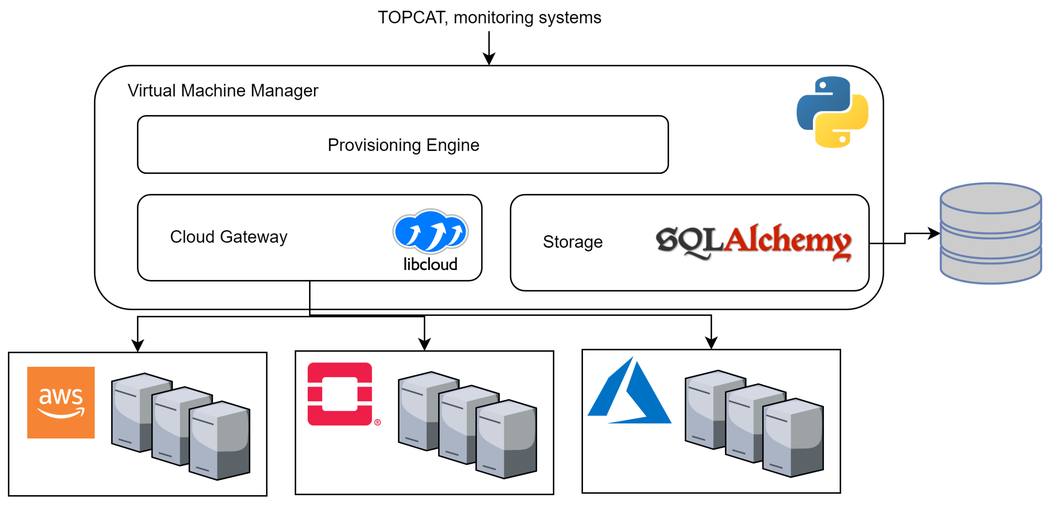


**Figure 4:** VMM conceptual model

- The ***Data Movement System*** (***DMS***) is responsible for moving data closer to the environments intelligently in order to balance cost of storage versus speed of access.

## Operational Benefits

The design of the provisioning components and supporting utilities provides several operational benefits:

- A faster response for users as environments are ready to go when needed.
- Consistent environments across heterogeneous clouds.
- More efficient system administration as all processes are automated via Ansible and lower overhead as cloud-specific knowledge is encapsulated in code.
- Better monitoring and analysis of usage through the information in the ***VMM***.
- Uniform security with respect to analysis environments across multiple clouds.

## Conclusion

We have created components that provide the foundation to provision infrastructure across different clouds as well as operating in heterogeneous clouds at once. These can then be used to support scalable, high-demand analytics environments for users.

*We acknowledge funding from UK Research and Innovation | STFC (UK SBS IT 18160).*