Virtual Geophysics Laboratory

Scientific workflows exploiting the cloud
Ryan Fraser, Terry Rankine, Lesley Wyborn, Joshua Vote, Ben Evans...

Presented by Robert Woodcock

October 2012
Gather data, process it, publish results
Simple, isn’t it?

Geo-information

- bedrock
- surficial
- mineral
- geophysical
- geochemical
- geochronologic
- hydrogeological

"I think you should be more explicit here in step two."
Recognise the complete picture....

Model, Simulate & Analyse 2/3/4D

Integrate 2/3/4D

Access

Acquire

Store & Manage
Introducing The Virtual Geophysics Laboratory
Data discovery

Layers discovered via remote registries

Layers consist of numerous remote data services
Data discovery

Some data services support reformatting e.g. CSV, NetCDF, GeoTIFF
Some data is only registered with flat files
Powered by the Spatial Information Services Stack

Common Platform

- Environment, Water
- Marine
- Groundwater
- Geology
- Geophysics
What just happened?

User Selecting Data

Get Available Datasets

VGL

Registry of data

Self Describing Service

Data Request
(Built but not executed)

Data Services

Cloud Compute

Cloud Storage
A variety of different scientific codes are already available in the form of “Toolboxes”
Data processing

Further input files can be uploaded.
Data processing

The steps so far have been building an environment to run a processing script...

Either write your own...

...or build from existing templates
What just happened?

User Running Job

Start processing

Perform data processing

Cloud Compute

Cloud Storage

Processing script/small input files uploaded

Download big data sets

Upload processing results

Script/user input files

Registry of data

Data Services

VGL

Virtual Geophysics Laboratory | Robert Woodcock
Successful jobs can have their entire process captured in a ISO 19115 ‘provenance record’
What just happened?

Publish the job’s process and results

Cloud storage will persist the final artefacts

What are the job input/outputs?

Monitoring Jobs

VGL

Registry of data

Cloud Compute

Cloud Storage

Data Services
Still under construction
What’s left?

• BYO cloud allocation
  • Users should be able to authorise VGL start jobs using their compute/storage resources.

• Confidential Data
  • How do you get access to ‘restricted data’ in a secure manner?
  • Where can you store the results? (geographical restrictions)

• Massive Horizontal Scaling
  • What’s the best way to set up a truly elastic pool of CPU’s for jobs to utilise?

• A Common Processing Services Platform – SISS like?