Virtual Geophysics Laboratory (VGL): Scientific workflows
Exploiting the Cloud

Josh Vote, Ryan Fraser, Terry Rankine
Ben Evans, Michael Chapman
Lesley Wyborn, Richard Chopping
Scientific workflow – Virtual Geophysics Laboratory (VGL)

- Scientific Workflow Engine (or Virtual Laboratory)
- Automates and massively expands Geophysicists computational capacity via the Cloud
  - Amazon – EC2 / S3 (and others using this interface)
  - OpenStack
- Collaboration between CSIRO, GA, NCI, Monash, UQ, and ANU
- VGL is just a pretty face
  - User Driven GUI
  - Leverages data providers and cloud technologies to do all the heavy lifting
- Non-Restrictive Open-source
AuScope Grid and SISS

• AuScope Grid: Charged with delivering software infrastructure to enable Geoscience Community
  – Establishing governance and sustainability of services
  – Delivering solutions to research and government organisations
  – Interoperability

• Spatial Information Services Stack (SISS)
  • An open-source, open-standards SDI
  • Achieves Interoperable data exchange
  • Standardises on 3 components – Format, Content, Tools
V(what)GL

• VEGL – Virtual Exploration Geophysics Laboratory
  – One primary science collaboration
  – One primary workflow
  – One collection of geophysical data sets

• VGL - Virtual Geophysics Laboratory
  – Collaboration with multiple partners
  – Supporting multiple workflows
  – New data sets
  – New data types
  – New Use – Not just exploration.
Geophysics (as seen by a software dev)

- Geophysics is taking physical measurements…
  - Magnetism over an area
  - Acceleration due to gravity over an area
- …Applying lots of mathematics…
- …to infer the structure of what is under the surface…
- It is not geology
  - Samples are never taken, only measurements

But -- Where is all the gold?
Our Geophysics Problem

• Measurements coming from the field are ‘raw’
  – Varying spatial reference systems
  – Noisy
  – Artifacts from collection process

• This data needs processing
  – From raw data to a data product

• Data products are valuable
  – They will be re-used and referenced repeatedly

• Processing is a time consuming process
  – Made worse by a purely manual workflow
The Past

• Compile raw data using proprietary FORTRAN
  – Also use software – Intrepid
• Transform to a regular grid using more software
  – MATLAB, Intrepid, ER Mapper, ESRI ArcGIS, QGIS
• Crop data spatially to suit final data product
  – eg: everything in Victoria
• Transform data into a file format that can be read by proprietary scientific code.
  – This is usually done with some handwritten python or c
  – There is no version control, code is often rewritten / redone
• Upload data to HPC
  – Manually enter input parameters/start job
Let’s map it out…

Hardcopy of data
- Get handed field data
- Download results

Intrepid
- Visualise data
- Crop data to area of interest

MATLAB
- Transform to a regular grid
- Reformat data for processing

SSH Client
- Configure job and start processing
- Upload data to NCI

Visualise data

Get handed field data

Crop data to area of interest

Transform to a regular grid

Reformat data for processing

Download results
There seems to be a problem…

- Reproducibility – there is none
  - What was inputted into your model?
  - What transformations occurred?
- It’s a manual process
  - Time consuming
  - Error prone
- Expensive
  - Licensing costs
Our solution

• Virtual Geophysics Laboratory
• Reproducibility
  – All input data is saved and then published with the final data product
• VGL automates portions of the workflow
  – Allowing scientists to focus on science
• Built entirely on open source tools
  – No licensing costs
Data Selection
Step 4: Define your job script.

Available Components
- Shell Components
- Python Components
  - VEGL Workflow Steps
    - Define Function Name
    - VEGL Step 1
      - Step 2
      - Step 3
      - Step 4
    - Step 5
    - Step 6
    - Step 7
    - Step 8
    - Step 9
  - Specify 'Main' Function
- File Commands
- Execute Commands
- Job Details

Current Script
- sim (Simulation Container)
  - DefinePythonFunc0 (Define Python Func
  - VEGLStep1-i1 (VEGL - Step1)
  - VEGLStep2-i2 (VEGL - Step2)
  - VEGLStep3-i3 (VEGL - Step3)
  - VEGLStep4-i4 (VEGL - Step4)

Script Source
#!/usr/bin/env python

# Please load the Job Options
import subprocess, csv, ...

# autogenerated function
def main():
    # -------------- VEGL...
    f = file(VEGLParams
    input_csv = csv.read...
    data = []
    lineCount = 0 # The
    for strX, strY, str...
        if lineCount >
            x = float(s...
            y = float(s...
            z = float(s...
Job Monitoring

Richard’s really big job

Description: This is a really big job that takes a long time (several days) to do anything
Submitted on: Thu Aug 04 2011 10:45:52 GMT-0800
Published provenance records
From this...

- Hardcopy of data
  - Get handed field data
  - Download results

- Intrepid
  - Visualise data
  - Crop data to area of interest

- MATLAB
  - Transform to a regular grid
  - Reformat data for processing

- SSH Client
  - Configure job and start processing
  - Upload data to NCI

- Visualise data

- Get handed field data

- Transform to a regular grid

- Reformat data for processing
Virtual Geophysics Laboratory

- Discover raw data
- Select spatial bounds
- Build “science” from existing libraries
- Run job
- Collect and publish results
VGL - Summary

• VEGL has been built for a Geophysics workflow
  – Its concepts can be re-used for other scientific workflows
• Is in the process of being deployed at GA
  – It can produce actual scientific data products
• Is capable of integrating with any SISS data provider
  – Or any provider that understands the OGC standards
• It’s built from many ‘generic’ components that can be repurposed
• Is just a pretty face
  – The power lies with the underlying services
  – These services are accessed using standardised protocols
Future Work

• VGL - Virtual Geophysics Laboratory
  – Collaboration with multiple partners
  – Supporting multiple workflows
  – New data sets
  – New data types
  – New Use – Not just exploration.

• Exploiting the generic
  – Modularising the workflow for general scientific usage

• Repurposing for other use cases – nature hazards, climate prediction, etc

• Commercial uptake

• Integration with other VLs to achieve ultimate aim…
Sustainable Energy Policy

Energy Exploration Integrated Virtual Laboratory

Fishery adaptation V. Lab

Virtual Geophysical Laboratory
Virtual Core Laboratory
Virtual Geodesy Laboratory

Virtual Climate Laboratory
Virtual Fisheries Laboratory

Virtual Laboratories

Virtual Libraries to Laboratories

Geophysics
Borehole data
Geodesy
Climate Modelling
Fisheries Monitoring

Societal Need
Integrated Virtual Labs
Virtual Libraries

Modelling & analytic tools
Thank you and for more information:

https://twiki.auscope.org/wiki/Grid/VEGLPortalDevelopment

CSIRO Earth Science & Resource Engineering
Josh Vote
Software Developer

Phone: +61 8 6436 8607
Email: josh.vote@csiro.au

Web: www.csiro.au
www.seegrid.csiro.au
www.auscope.org