Data Centre Technology to Support Environmental Science

Bryan Lawrence
Scientific Data Workflow

... with the role of “data centres” primarily in the purple boxes!
All of these states and activities require “data centre support” for in situ, upstream and downstream users!
Growing range of interacting communities

Many interacting communities, each with their own software, compute environments, observations etc.

Figure adapted from Moss et al, 2010
The Rise of Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

Surface Winds
Precipitation
Reflection and Transmission
Evaporation
Transpiration
Surface Temperature

Circulation
Surface Winds
Precipitation
Reflection and Transmission
Surface Temperature
Evaporation
Currents
Upwelling

Infiltration
Runoff
Nutrient Loading
Surface Temperature
Currents

Image: from J. Lafaille, 2006

http://www.bgs.ac.uk/research/environmentalModelling/home.html

Coulthard and Van De Wiel DOI: 10.1098/rsta.2011.0597
CMIP6 data volumes and data rates not yet known, but the European contribution to HiresMIP alone is expected to exceed 2 PB.

Sentinel 1A (2014), 1B (2016)
Sentinel 3A (2016) 3B (2018?)

Data rate: \( o(6) \) PB/year
The unorganised data deluge

Slide concept: Carole Goble via Liz Lyons
The consequences of data at scale - download doesn’t work!

Earth System Grid Experience:

Started with Individual End Users
- Limited resources (bandwidth, storage)

Moved to Organised User Groups
- Organize a local cache of files
- Most of the group don’t access ESGF, but access cache.

Then Data Centre Services
- Provide access to a replica cache
- May also provide compute near to data
- BADC, DKRZ, etc

Trend from download at home, to exploit a cache, to exploit a managed cache with compute!
Transforming data into information - Revisited

**Primary**
- "Raw-ish" data, as acquired.
  - (EO datasets, Model Runs, Instrument Datasets)
  - High Volume
  - Filesystem Only
  - Low Numbers of Expert Users

**Secondary**
- Organised Processed *Multiple Sources*
  - Medium Volume
  - API Accessible?
  - Medium Numbers of "Technically-Savvy" Users

**Tertiary**
- Information Products
  - Small Volume
  - Web Accessible?
  - Many Different Types of Users

This can be the same physical storage and compute infrastructure! With different individuals or groups responsible for the different “doors”.

Data Centre Technology to support Environmental Science
Bryan Lawrence - London, 13/10/2016
JASMIN — The Data Commons

- Provide a state-of-the-art storage and computational environment.
- Provide and populate a managed data environment with key datasets (the “archive”).
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.
JASMIN — The Data Commons

- Provide a state-of-the-art storage and computational environment.
- Provide and populate a managed data environment with key datasets (the “archive”).
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.

Platform as a Service
-----
We provide you the “Platform”; you can LOGIN and exploit the batch cluster.

Infrastructure as a Service
-----
We provide you with a cloud on which you INSTALL your own computing.

Software as a Service
-----
We provide you with REMOTE access to data VIA web and other interfaces.
The Data Commons

Bringing Computation to the Data

Virtual Environments

Summary

Hardware

JASMIN

- 16 PB of fast disk; 0.5 PB of bulk disk (for virtual compute); >30 PB of tape.
- 5000 compute cores (cluster and hypervisors); dedicated high memory and transfer machines.

- The Archive - curated data directly available to local compute.
- Group Work Spaces — fast storage with tape accessible via the “Elastic Tape” service.
- Generic Platform Compute — machines configured for generic scientific analysis and data transfer.
- Hosted Platform Compute — bespoke machines deployed in the “Managed Cloud”.
- Infrastructure Compute — private cloud portal and customised compute in the “Un-Managed Cloud”.
- Lotus Batch Cluster — managed cluster with a range of node configurations (processor and memory).
CEDA

Four internal data centres: [http://ceda.ac.uk](http://ceda.ac.uk)

Acquiring and Curating Data Archives

- Provides the initial mass for the “gravity well”, by feeding in both NERC and third party data products, available through the “back door”.
- An example of a tenant organisation in its own right, delivering services through the “front door”.
- Supports groups delivering customised services through “family doors”.

Other data centres could be tenants and contribute to the data commons in the same way.
The High Resolution Climate Modelling (HRCM) programme is a collaboration between the Hadley Centre (UK Met Office) and the NCAS Climate Directorate. The programme produces and uses hundreds of terabytes of data, with data stored on a JASMIN Group Work Space and Elastic Tape.

The use of the JASMIN LOTUS batch cluster has

- enabled routine tracking of tropical cyclones from model simulations (50 years of N512 data can now be processed in one day with just 50 jobs).
- vastly sped up key analyses: e.g. calculation of eddy vectors has been reduced from 3 months to 24 hours with 1600 batch jobs.

For more details contact: Prof P.L Vidale (NCAS, University of Reading) or visit https://hrcm.ceda.ac.uk/research/
Near-Real Time volcanic plumes on JASMIN

- Real-Time (NRT) observations of atmospheric disturbances such volcanic plumes of ash and SO2 are increasingly important, especially with respect to air travel.
- A volcanic SO2 monitoring website has been launched displaying near real time (NRT) data from both IASI instruments within 3 hours of measurement.
- The unique relationship available on JASMIN between data archive and data processing facilities is invaluable for this work.

More details: Elisa Carboni (University of Oxford) or visit http://www.nrt-atmos.cems.rl.ac.uk/
Virtual Research Environments on JASMIN hosted cloud

Thematic Exploitation Platforms for ESA
CCI Open Data Portal for ESA
MAJIC interface to JULES model
EOS Cloud — Desktop-as-a-Service for Environmental Genomics
Hosted Ipython Notebooks
NERC Environmental Workbench

Data Centre Technology to support Environmental Science
Bryan Lawrence - London, 13/10/2016
Forestry TEP

▶ A one-stop shop for forestry remote sensing services for the academic and commercial sectors.

▶ Offers access to pre-processed satellite and ancillary data, computing power, and software access and hosting.

...built by VTT Technical Research Centre & Arbonaut (FIN), CGI IT & STFC (UK), and Spacebel (BEL).

CEDA is supporting the Forestry and Polar TEPS on the JASMIN un-managed cloud.
CCI Open Data Portal for ESA

The Climate Change Initiative

- Exploiting Europe’s EO space assets to generate robust long-term global records of essential climate variables such as greenhouse-gas concentrations, sea-ice extent and thickness, and sea-surface temperature and salinity.

- The CCI Open Data Portal is hosted on JASMIN and exploits a near complete copy of the CCI datasets held in the CEDA archive.
MAJIC: Managing Access to JULES in the cloud

- **JULES** is a community land surface model incorporating processes such as surface energy balance, the hydrological cycle, carbon cycle, dynamic vegetation etc.

- **MAJIC** provides a web portal running in the un-managed cloud which allows users to configure JULES to run on the JASMIN/LOTUS batch cluster and return results.

Welcome to MAJIC

Managing Access to JULES in the Cloud

MAJIC is a web application that allows users to run the JULES land-surface model within a high performance computing environment over the web. You can create new model runs, easily customise some of the main JULES parameters, and view and download results for new and existing model runs.
Summary

- Key role of data centres in the scientific workflow, dealing with the range of data from primary data to tertiary data, from expert users to consumer.
- Underlying trends: more data (volume and variety), more communities, and (more complexity of workflow).
Summary

- Key role of data centres in the scientific workflow, dealing with the range of data from primary data to tertiary data, from expert users to consumer.
- Underlying trends: more data (volume and variety), more communities, and (more complexity of workflow).
- Data gravity is “a thing”! Users value having “other” data with “their” data — provided there is adequate compute and storage available.
- Data gravity leads to “data lakes”. With a data lake, it’s possible to have a range of entrances\(^1\), from a front door for consumers to back doors for data experts.

\(^1\) Yes, I know a lake with doors is approaching an oxymoron!
Summary

- Key role of data centres in the scientific workflow, dealing with the range of data from primary data to tertiary data, from expert users to consumer.
- Underlying trends: more data (volume and variety), more communities, and (more complexity of workflow).
- Data gravity is “a thing”! Users value having “other” data with “their” data — provided there is adequate compute and storage available.
- Data gravity leads to “data lakes”. With a data lake, it’s possible to have a range of entrances\(^1\), from a front door for consumers to back doors for data experts.
- JASMIN provides a suitable environment for a “data commons”, already supporting a range of data centres and users exploiting a range of “doors”: from bespoke portals to batch cluster based data analysis.
- There is a strong argument that NERC should aggregate more of its data into the common environment (but perhaps not all, e.g. JASMIN won’t offer commercial levels of service that some applications such as BGS commercial might need).

\(^1\) Yes, I know a lake with doors is approaching an oxymoron!