Earth Science in the Data Era

How data is changing the way we do science
Summary

• Data & Science

• The Earth Science Case
Data & Science

Where are we?
Where to go next?
Data Grow & Storage Cost

IDC’s Digital Universe Study, sponsored by EMC, June 2011

http://www.mkomo.com/cost-per-gigabyte
Expected growth rate for data archive @ DKRZ (real case)

2009-2011 presented a lower growth. Reason: Data generation was more difficult than expected.
“[...] the number of files the datacenter will have to deal with will grow by a factor of 75, at least. Meanwhile, the number of IT professionals in the world will grow by less than a factor of 1.5”

And Security, of course...

• “[...] only about half the information that should be protected is protected.”

Scientific Community Growth

Figure C-8
First university degrees in natural sciences and engineering, selected countries: 1998–2006

Thousands

NOTE: Natural sciences include physical, biological, earth, atmospheric, ocean, agricultural, and computer sciences and mathematics.


Science and Engineering Indicators 2010

Figure C-9
Doctoral degrees in natural sciences and engineering, selected countries: 1993–2007

Thousands

NOTE: Natural sciences include physical, biological, earth, atmospheric, ocean, agricultural, and computer sciences and mathematics.


Science and Engineering Indicators 2010
In the past 30–35 years there have been no increases in the average annual number of published papers per scientist [in America in physics, astronomy, geophysics, mathematics, and chemistry]

Data storage costs are decaying exponentially
Data generation is growing exponentially
Data is undermanaged (e.g. security)
Scientists / IT personnel is barely growing
Scientists “productivity” remains constant
To think about

• Are current methods for processing the increasing amount of data good enough?
• Can we work as we do now with 100x more data?
• Can we do 100x more “science”?
• (Can we remember 100x more Acronyms?)
What is holding data growth?

- Expenditure
- Technology
- Or perhaps we are?
The generation of new information is limited by two key factors, by the incurring economic costs and by the capacity of the human brain to process and store data and information; the controlling agent needs to retain an overall understanding even when data is generated by semi-automatic processes.

Furthermore

• We have no “time” to process data at all...

Someone has to decide what isn’t important...
What now?

- Data generation will not slow down
- *We* can’t keep the pace much longer
- ... so science will have to go through a major “change” (again)
The three paradigms of science

• Thousand years ago – empirical science
• Last few hundred years – analytical science
• Last few decades – computational science (simulations)

• Now the next paradigm shift is due

  – cf. Thomas S. Kuhn: The Structure of Scientific Revolutions
The Fourth Paradigm: Data-Intensive science


- “The architecture for data-intensive computing should be based on storage, computing, and presentation services at every node of an interconnected network.” (p. 116)

- “We should aim to provide scientists with a cyberinfrastructure on top of which it should be easy to build a large-scale application capable of exploiting the world’s computer-represented scientific knowledge.” (p. 171)

- “Scientific publication will become a 24/7, worldwide, real-time, interactive experience.” (p. 225)
The Fourth Paradigm: Data-Intensive science

- "Scientific publication will become a 24/7, worldwide, real-time, interactive experience." (p. 225)


- 2926 Authors, 169 Institutions
The Earth Science Case

Data management in the CMIP5 project
Data in Earth Science

- Simulations
- Observations
  - Satellites
  - Sensor data
- ...

Estanislao Gonzalez MPI/DKRZ
Coupled Model Intercomparison Project 5 (CMIP5)

• Project aimed at allowing the comparison of global coupled models
Coupled Model Intercomparison Project 5 (CMIP5)

- Project aimed at allowing the comparison of global coupled models
- The main source of data for the Intergovernmental Panel on Climate Change (IPCC) Assessment Report
- Defines and standardizes vocabulary, experiments and data output required for data comparison
- In this iteration the expected audience is broader than the climate community (this implies more metadata!)
- This iteration also adds quality control
- ... and DOIs for data citation
Comparison to last iteration

<table>
<thead>
<tr>
<th>CMIP3</th>
<th>Data Volume</th>
<th>Models</th>
<th>Metadata</th>
<th>User Community</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>36 TB</td>
<td>25</td>
<td>CF-1 + IPCC-specific</td>
<td>Thousands of users; WG1, domain knowledge</td>
</tr>
<tr>
<td></td>
<td>83,000 Files</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Downloads: ~500GB/day</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CMIP5</th>
<th>Data Volume (expected)</th>
<th>Models</th>
<th>Metadata</th>
<th>User Community</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-10 PB</td>
<td>~35</td>
<td>CF-1 + IPCC-specific</td>
<td>10s of thousands of users; Wider range of user groups will require better descriptions of data, attention to ease-of-use</td>
</tr>
<tr>
<td></td>
<td>??? Files</td>
<td></td>
<td>Richer set of search criteria; Model configuration; Grid specification from CF (support for native grids)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Downloads: 10s of TB/day (+1Gbps)</td>
<td>More experiments; Increased complexity (ex: biogeochemistry)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

28~280x
## CMIP5 current status

<table>
<thead>
<tr>
<th>Summary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modeling centers</strong></td>
<td>32</td>
</tr>
<tr>
<td><strong>Models</strong></td>
<td>59</td>
</tr>
<tr>
<td><strong>Experiments</strong></td>
<td>109</td>
</tr>
<tr>
<td><strong>Data nodes</strong></td>
<td>22</td>
</tr>
<tr>
<td><strong>Gateways</strong></td>
<td>5</td>
</tr>
<tr>
<td><strong>Datasets</strong></td>
<td>51907</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>1,304.04 TB</td>
</tr>
<tr>
<td><strong>Files</strong></td>
<td>3,054,622</td>
</tr>
</tbody>
</table>

*Data Accessible from Federated System – Not all data has been published yet - 18.6.2012*
CMIP5 Data Requirements

- Integrated model metadata (data generation)
- Fast data search
- Distributed user management
- Security & Licensing
- Versioning – QC/DOI
- Notification System
- Long Term Archival
- Replication
1st Solution

A two-systems solution:

- **Data node**: data access management (PCMDI)
- **Gateway**: user management, search, metadata (NCAR)
• Gateway is a monolithic (was closed-sourced) system
• Search (3Store) did not scale well
• NCAR had no resources to solve this in time, and though manifesting the intention to open the code, this was delayed for months
• Basically many communicational issues
• A group of institutions got together and decided to join efforts in solving these shortcomings
• The ESGF group was born and a parallel development started
A single module-based System. It has 4 modules at this time:

- **Data**: same as the older data node
- **Computation**: allows visualization and basic computation operators at the data node
- **Index**: user interface and distributed search
- **Idp**: Identity provider for managing authentication/authorization at a federation level
ESGF P2P Federation

CMIP5 / IPCC AR5

ESGF Federation

Search
Security
Compute
Data
Manager

Users
ESGF P2P Architecture
ESGF P2P Components

**Index**
- Solr
- esgf-seach
- esgf-web-fe

**Idp**
- MyProxy
- esgf-idp

**Data**
- Thredds
- Postgres
- esgcet (publisher), cdat, esgf-orp

**Compute**
- ferret
- LAS

**ESGF Node Manager**
- Java, Git, Ant, Tomcat, curl, openssl
- esgf-node-manager, esgf-dashboard
ESGF P2P Search
ESGF P2P Search

• Solr Backend
• esgf-search
  – provides standardized API
    • query=temperature&model!=MPI-ESM-LR&...
    – returns Solr results (XML/JSON) or a bash script for simple downloading
• esgf-web-fe (front end)
  – Implements GUI for faceted search
  – Displays CIM metadata (link to external source)
• No security is required for searching, basic concept is:
  – Metadata is free and unsecured, data isn’t
ESGF P2P Security
• esgf-idp
  – OpenId provider (authentication)
  – Attribute service (authorization)
  – Registration service (Joining a group, license, etc)

• MyProxy
  – automates generation of short lived X509 Certificates (72Hs) for non interactive system access

• esgf-orp (on the data side)
  – Set of Tomcat filters used for securing resources and handling authorization & authentication (X509, OpenID)
ESGF P2P Data + Compute
ESGF P2P Data + Compute - LAS
ESGF P2P Data + Compute

• Thredds Data Server (TDS)
  – Hosts metadata via xml catalogs
  – Allows HTML browsing
  – Manages file access

• GridFTP
  – Preferred method for downloading

• esgcet
  – Set of tools to handle publication (versioning, metadata extraction, catalog generation, etc)
ESGF P2P Data + Compute

• LAS
  – Data visualization
  – Simple analysis tools (min, mean, max)
  – Data comparison
ESGF P2P Manager
ESGF P2P Manager - Dashboard

[Image of the ESGF P2P Manager Dashboard, showing various maps and statistics related to network availability and services.]
ESGF P2P Manager

- esgf-node-manager
  - Registry (of all available services)
  - P2P protocol (peer discovery, etc)

- esgf-dashboard
  - Visualization of node status and metrics
ESGF P2P Interactions
Thank you.