Cloud platforms have transformed how software is developed and delivered

Software as a service: **SaaS**
(web & mobile apps)

Platform as a service: **PaaS**

Infrastructure as a service: **IaaS**
Cloud platforms have transformed how software is developed and delivered

Can we do the same for science?

• Identify cross-cutting capabilities required by many groups
• Define simple REST APIs for accessing those capabilities
• Operate high-quality, scalable, secure, performant cloud-hosted implementations
• Ensure persistence and evolution over time

In so doing, enable many scientists and tool developers to automate and outsource tasks that are not central to their core mission: thus reduce costs, increase quality, promote interoperability
What tasks?

- **Auth**: Manage identities, authentication, and authorization
- **Transfer**: Manage movement of files from A to B
- **Sharing**: Manage who can access data at a location
- **Publish**: Preserve, identify, describe, curate
- **Search**: Index and search data
- **Identifiers**: Assign identifiers to collections of files
- **Automate**: Organize sets of activities
- **Learn**: Discover, train, run machine learning models
- ...
An example -- NCAR RDA: climate data

- **Globus Auth** for single sign on & federated login
- **Globus sharing** for restricted data access
- **Globus transfer** for data movement
- **Globus management** for administrative tasks
The Modern Research Data Portal: A Design Pattern for Networked, Data-Intensive Science

https://docs.globus.org/mrdrp

The Modern Research Data Portal is a new design pattern for providing secure, scalable, and high performance access to research data.

GitHub Repo
provides code for the simple data portal that you can experiment with online
LEARN MORE

Example Data Portal
allows you to experiment with an example implementation of the design pattern
LEARN MORE

Code Walkthrough
provides a narrative description of the simple data portal code
LEARN MORE

A technical article describes this design pattern, reviews representative examples at research laboratories and universities (see below), and uses coding examples to demonstrate how Globus APIs can be used to implement a range of research data portal capabilities.

https://peerj.com/articles/cs-144/

LEARN MORE

Jupyter Notebook
demonstrates some Globus features described in the technical article
LEARN MORE
A key message: Outsource all that you can to cloud-hosted automation platforms

For example:

• Outsource responsibility for **determining user identities**
• Outsource **control over who can access** different data and services within the portal
• Outsource responsibility for **managing data uploads and downloads** between various locations and storage systems
• Leverage **standard web user interfaces** for common user actions
Automate and outsource: Publication and discovery

Move to permanent location
(or publish in place)
Compute and record checksums
Obtain and record metadata
Assign persistent identifier
Index for discovery

2 petabytes
100 Gbps
Globus APIs
Automate and outsource: Publication and discovery

Move to permanent location (or publish in place)
Compute and record checksums
Obtain and record metadata
Assign persistent identifier
Index for discovery

PETREL
Data Management and Sharing Pilot

Data Publication
Indexing

PETREL
Data Management and Sharing Pilot

2 petabytes
100 Gbps
Globus APIs

MDF
Index
117
Data resources indexed
>3.4M
Records

MDF
Publication
8
Repertories inventoried
~ 200
Datasets
~ 300 TB
Mass discoverable

61
Total datasets
29
Institutions
150
Authors
22
CHiMaD datasets
> 18 TB
Data Volume

materialsdatafacility.org
Automate and outsource: Publication and discovery

Web browse and search

Data Publication

Indexing

Programmatic access (REST, Python, Jupyter)

materialsdatafacility.org

2 petabytes
100 Gbps
Globus APIs

Programmatic access (REST, Python, Jupyter)

<table>
<thead>
<tr>
<th>elements</th>
<th>sources</th>
<th>my_ep</th>
<th>my_path</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Al&quot;</td>
<td>&quot;khazana_vasp&quot;, &quot;slusch&quot;, &quot;ab_initio_solute_database&quot;</td>
<td>&quot;b8ee7e5c-6d04-11e5-ba46-22000b92c6ec&quot;</td>
<td>&quot;/Users/ben/Desktop/blaiszik-macbookpro/dft_training_set&quot;</td>
</tr>
</tbody>
</table>

```python
mdf = Forge()
res = mdf.search_by_elements(elements=elements, sources=sources, limit=9999)
res, dest=my_path,
local_ep=my_ep, preserve_dir=True)
```
NSF supports development of new nationwide data storage network

The Open Storage Network will enable researchers to manage data more efficiently than ever before
Automate and outsource: End-to-end data pipelines

For each data, must apply quality control, assign identifiers, move to compute, extract features, eventually publish to public repository, ...

Building a different custom pipeline for every situation is impractical
Automate and outsource: End-to-end data pipelines

For each data, must apply quality control, assign identifiers, move to compute, extract features, eventually publish to public repository

Building a different custom pipeline for every situation is impractical

Automate: Trigger-action programming ("if this happens, then do that")

Outsource: Cloud-based trigger-action service for reliability, scalability, ease of use, security, sustainability
Automate and outsource: End-to-end pipelines with trigger-action programming

National Facility

- Local Storage and Compute
  - Quality Control
  - Assign Handle

Central Storage and Compute (CSC)
- Feature extraction
- Aggregate and convert format

Beamline Instrument

Globus Transfer

Archive
Automate and outsource: End-to-end pipelines with trigger-action programming

**National Facility**
- Local Storage and Compute
  - Quality Control
  - Assign Handle
- Beamline Instrument
  - Email / SMS notification

**Central Storage and Compute (CSC)**
- Feature extraction
- Aggregate and convert format

**Rules**
- IF new files THEN run quality control scripts
- IF quality is good THEN send email and transfer data to CSC

1. Globus Transfer
2. Archive
Automate and outsource: End-to-end pipelines with trigger-action programming

Rules
1. IF new files THEN run quality control scripts
   IF quality is good THEN send email and transfer data to CSC
2. IF new files THEN run feature extraction
   IF feature detected THEN transfer data to archival storage
   IF time since ingest > 6 months THEN publish dataset to Materials Data Facility
I reported on the work of many talented people

Thanks also to:
• Other members of the Globus team
• Participants in the Open Storage Network project
• Globus users around the world

We are grateful to our sponsors
Modern information technology is about leveraging platforms to outsource and automate

For example:

• Outsource responsibility for **determining user identities**
• Outsource **control over who can access** different data and services within the portal
• Outsource responsibility for **managing data uploads and downloads** between various locations and storage systems
• Leverage **standard web interfaces** for common user actions
• Outsource **data publication, search, pipeline management**, ...