

Globus Automate

Ian Foster, Ryan Chard, Kyle Chard

Argonne National Laboratory & University of Chicago

foster@anl.gov, rchard@anl.gov, chard@uchicago.edu

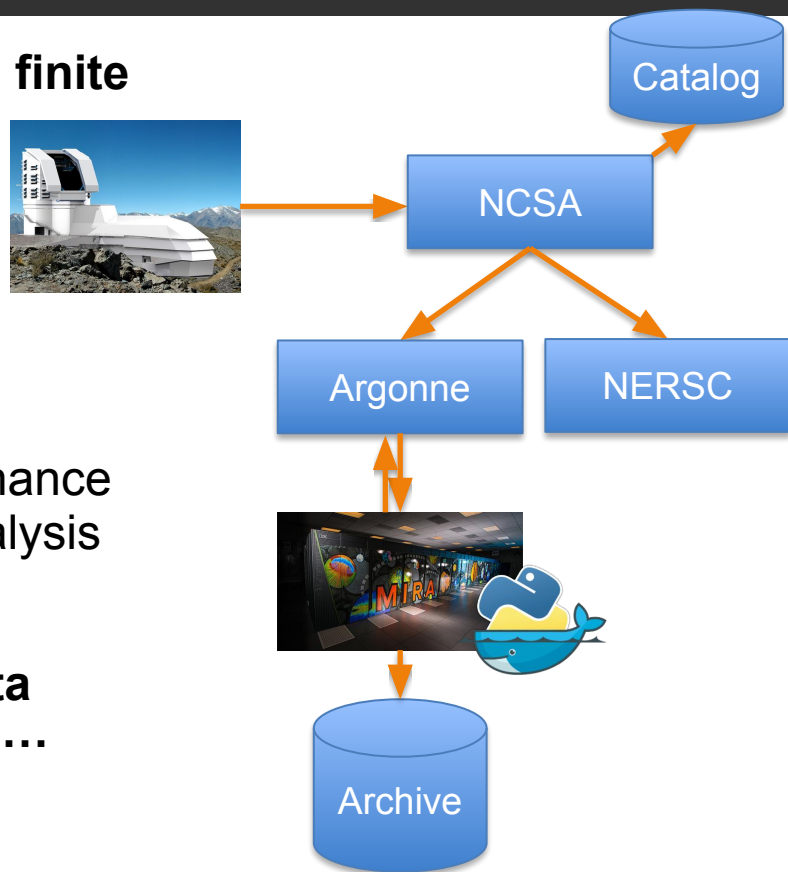
Data management challenges as volumes increase

Data volumes and velocities are overwhelming finite human capabilities

Scientific results are dependent on

- Data acquired at various locations/times
- Analysis processes executed on distributed resources
- Catalogs of descriptive metadata and provenance
- Dynamic collaborations around data and analysis

Best practices are often overlooked, useful data forgotten, errors propagate through pipelines, ...



LSST data distribution and analysis pipeline

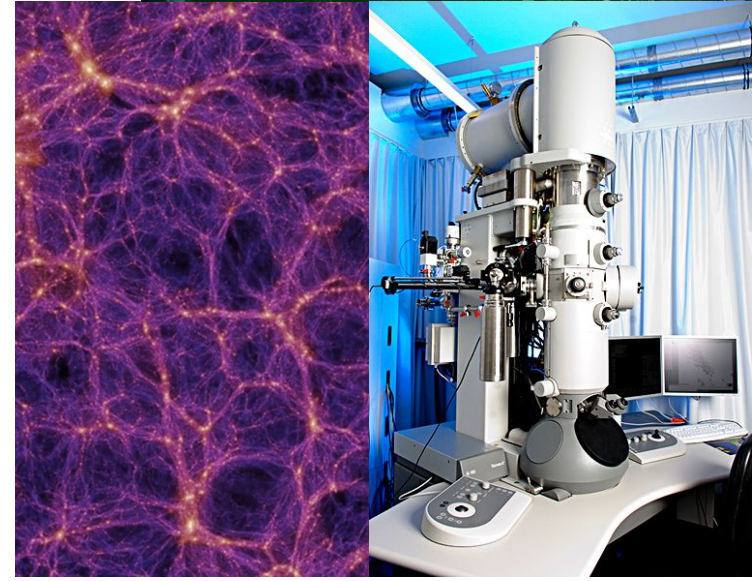
Experimental Science

Data management issues are particularly evident in large scale experimental science

Researchers are allocated short periods of instrument time

- Must maximize experiment efficiency and output data quality/accuracy

Inefficiencies mean less science is performed and researchers may have to wait months for another chance.

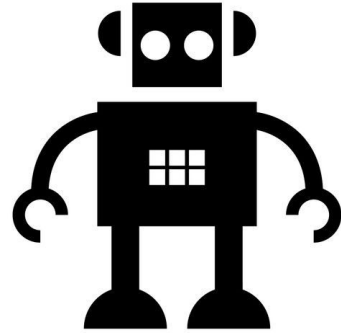


Automation

Goal: Automate data manipulation tasks from transfer and sharing to acquisition, publication, indexing, analysis, and inference

Requirements: A platform that...

- Can automate best practices (replicate, catalog, share)
- Is data driven -- responds as data are created
- Can be applied across arbitrary storage and compute infrastru
- Can be dynamically programmed to respond to new events
- Enable non-expert users to define automations



Approach: Compose and execute data manipulation flows through the Automate PaaS

Globus Automate

Built on AWS Step Functions

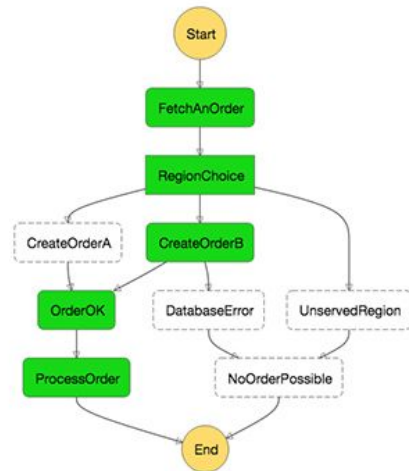
- Simple JSON-based state machine language
- Facilitates conditions, loops, fault tolerance, etc.
- Propagates state through the flow

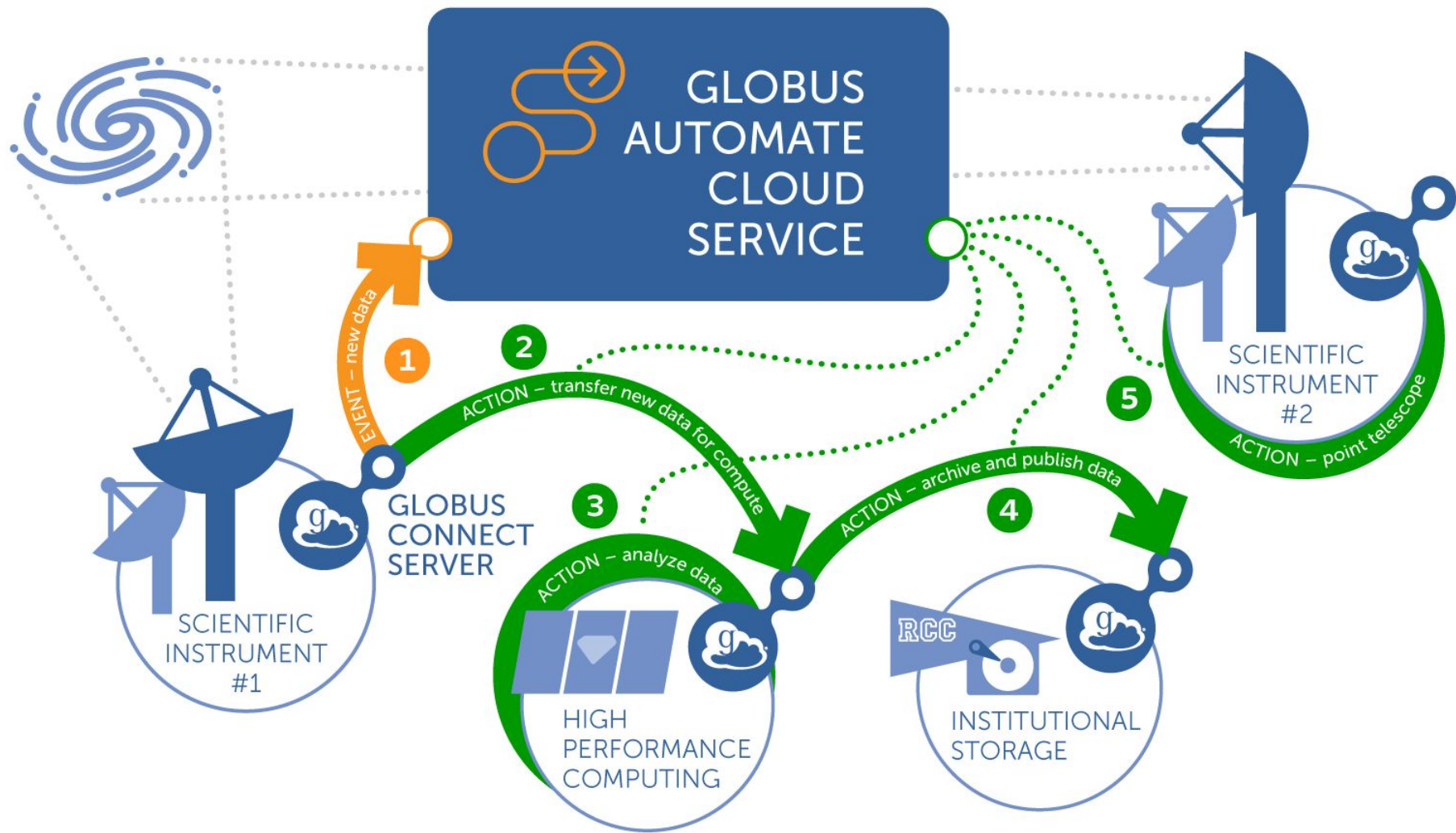


Standardized API to integrate custom event & action services

- Actions can be either synchronous or asynchronous
- Custom Web forms can halt flows for user input

Actions are secured with Globus Auth





Automate Prototype: The Service

Define JSON flows that step between action services and describe JSON doc of default input data

- Definition based on AWS state machine language

Associate a trigger condition -- event data is passed in when executed

We provide a polling SFN activity you can use to halt a flow until an action_id has completed

Automate Prototype: Actions

Any service can expose the Action API

- /automate/v1/action/run, status, cancel, introspect, ...
- .../status used to enable polling
- We give the service an action_id on invocation

When registering an action we make an internal lambda function that calls your service's url

- Makes an ARN for it and maps to a user-friendly name for use in flows

Introspect tells us what input the action accepts -- used during flow creation

The action can then be stepped to in a flow

Some Actions

Auth



User login

Secure service interactions

App identity and interactions

Identifier



Manage namespace

Mint DOI

Search



Catalog

Faceted search

Search query

Transfer



File operations

Transfer data

Set permission

Execute



Remote execution

Secure connections

Self optimization

Automate Prototype: Events

Any service can expose the Events API

- /automate/v1/event/register, poll, introspect, ...

Automate polls each event interface and adds responses to a reliable Simple Queue Service queue

- Events processed by lambda functions

Integrates Ripple/Dash as an event source

- Can be driven by file events

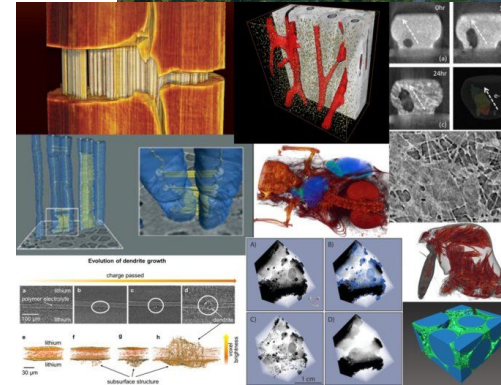
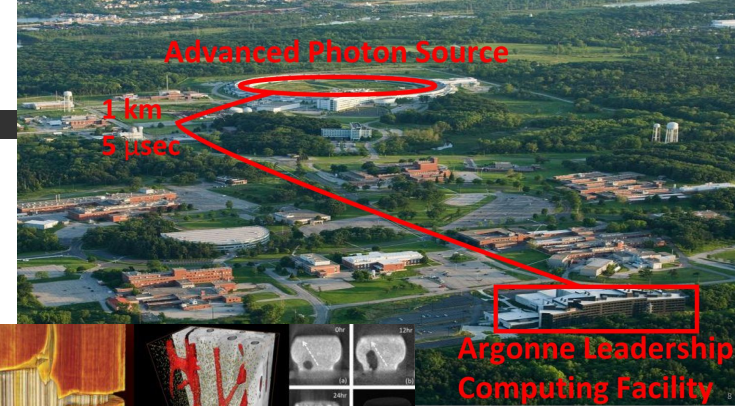


Use Cases

- Neuroanatomy (APS)

- Advanced Light Source

- Data Publication



Globus Data Publication

Globus simplifies the publication and discovery of research data. Use Globus to describe, curate, and preserve data at desired levels of durability. Make your data easily accessible to fellow researchers and other interested parties who can search and browse published datasets.

Click here to learn how to publish data. More information on how Globus data publication works is available here.

Try a free trial of Globus data publication.

Communities

Choose a community to browse its collections.

BD2K Data Repository Big Data to Knowledge Centers
Globus
Materials Data Facility Community for the Materials Data Facility Collaboration
National Data Service
RDCEP Center for Robust Decision Making on Climate and Energy Policy

Discover

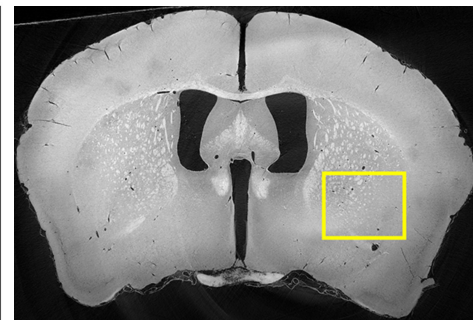
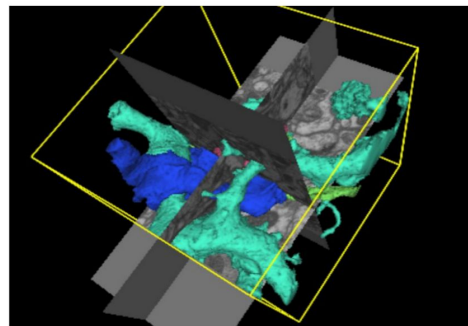
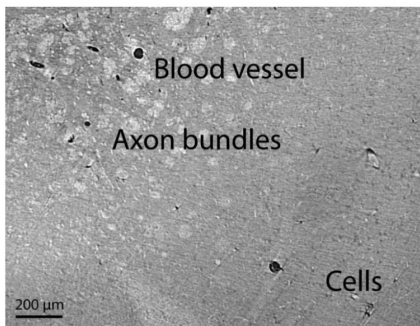
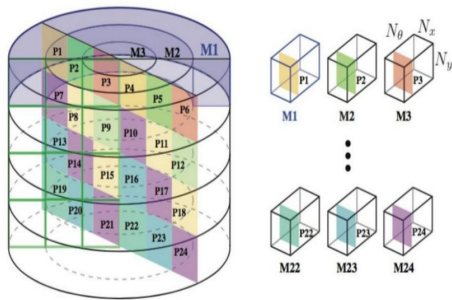
Author	Subject	Issue Date
GGCHI	Agricultural Impacts	2018
McInerney, D. J.	Climate Change	2017
Moyer, E. J.	Climate Impacts	2016
Sun, S.	CO2 effects	2015
Moyer, E. J.	Farm system models	2014
Schwarzwald, K.	Food Security	2013
Zhorin, V.	Model	
Hersam, Mark C.	Intercomparison	
Voorhees, Peter W.	climate	
	cmip5	



Case 1: Neuroanatomy

UChicago's Kasthuri Lab study brain aging and disease

- Construct connectomes -- mapping of neuron connections
- Use synchrotron (APS) to rapidly image brains (and other things)
- Given beam time once every few months
- Generate segmented datasets/visualizations for the community
- ~20GB/minute for large (cm) unsectioned brains
- Perform semi-standard reconstruction on all data across HPC resources



Neuroanatomy Overview

APS



1. Imaging



2. Acquisition



3. Pre-processing



ALCF



4. Preview & Centre



5. User validation & input



6. Reconstruction



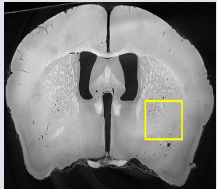
JLSE



7. Publication



UChicago



8. Visualization



Neuroanatomy Automation

Auth



Get
credentials

Transfer



Transfer
data

Execute



Run job

Transfer



Transfer
data

Web
form



User input

Search



Ingest

Share



Set policy

Identifier



Mint DOI

Describe



Get
metadata

Execute



Run job



Automate

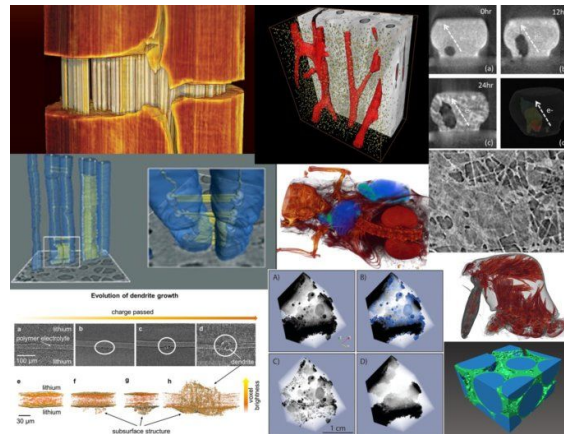
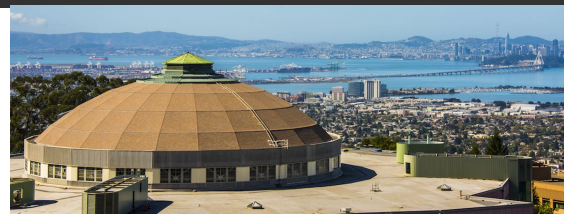
Case 2: Advanced Light Source

Reconstructions

- Move data to NERSC
- Submit batch reconstruction
- Return result to users

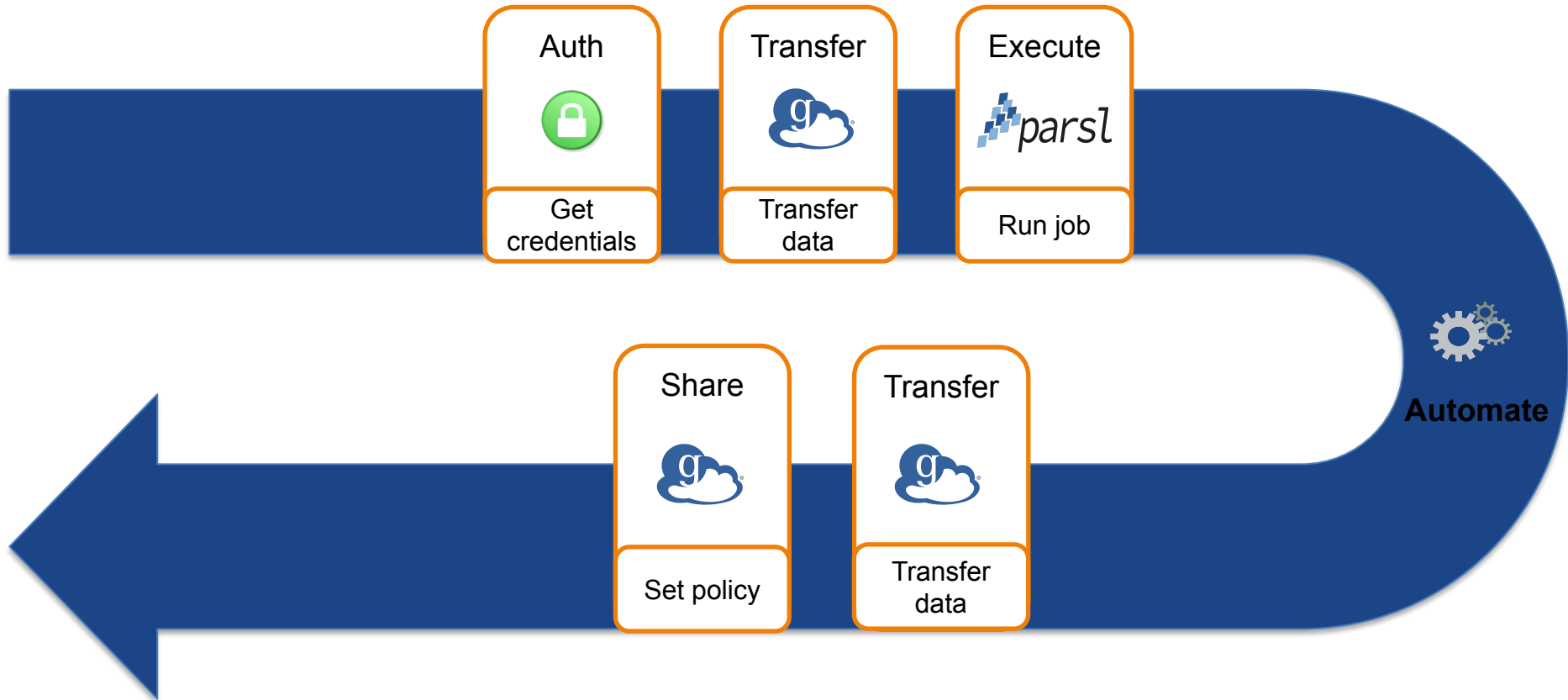
Requirements

- Plug in different tool, endpoint, allocation



Leverage multiple compute resources (NERSC, local, AWS)

ALS Automation



Case 3: Data Publication



Citable Data

Standard metadata,
persistent identifiers,
durable storage



Institutional Data

Many domains,
custom metadata,
locally managed storage



Community Data

Agreed schema,
larger datasets,
fine grained metadata

Globus Publication v1

- Cloud-based web app
- BYO storage & in-place publication
- User-managed collections
- Select pre-defined schema
- Handle, DOI persistent identifiers
- Adoption since 2015:
 - >2000 users, >600 datasets

The screenshot displays the Globus Data Publication web application. At the top, there's a navigation bar with the Globus logo and links for 'Manage Data', 'Publish', 'Groups', 'Support', and 'Account'. Below this is a search bar and a secondary navigation bar with links for 'Browse & Discover', 'Data Publication Dashboard', and 'Communities & Collections'.

The main content area is titled 'Globus Data Publication' and includes a brief description of the service. To the right, there's a 'MDF CONNECT' banner with the text 'It has never been easier to share your data with the community. Deposit data once, send to partner services.' and a 'Become a Contributor' button.

Below the banner, there's a 'Discover' section with a table of communities and a list of authors. The table has columns for 'Author' and 'Subject'. The authors listed are GGCMI, McInerney, D. J., Moyer, E. J., Sun, S., and Moyer, E. J. The subjects listed are Agriculture, Climate, and CO2 e.

At the bottom, there's a 'Find and Share Canadian Research Data' section with a search bar and a 'Deposit Data' button. Below this, there's a 'Filter Results' section showing a bar chart of data range and a list of search results. The search results show 62 results found, sorted by relevance, and a list of 1 to 35 of 62 results. The results include entries for 'Open Data Canada' and 'Parks Canada'.

Publication v2 via Automate

- Decompose Globus Publish v1 into platform services
- Allow for flexible re-composition and adaptation pf services
- Enable extension and enhancement

