Globus Flows and Compute:

Smart automation around data submission

Andy Boughton
(abought@globus.org)
APECx Demos – April 2025

globus

Globus Flows: Part of an integrated platform

Learn more:

https://www.globus.org/platform

https://docs.globus.org/api/flows/

https://globus-compute.readthedocs.io/en/latest/

Today's notes:

https://github.com/globus/apecx-demos/

Try it yourself

Get an endpoint: https://www.globus.org/globus-connect-personal

Library of public flows: https://app.globus.org/flows/library

Globus Flows:

Secure, managed automation of complex workflows at scale

globus

Big projects need automation

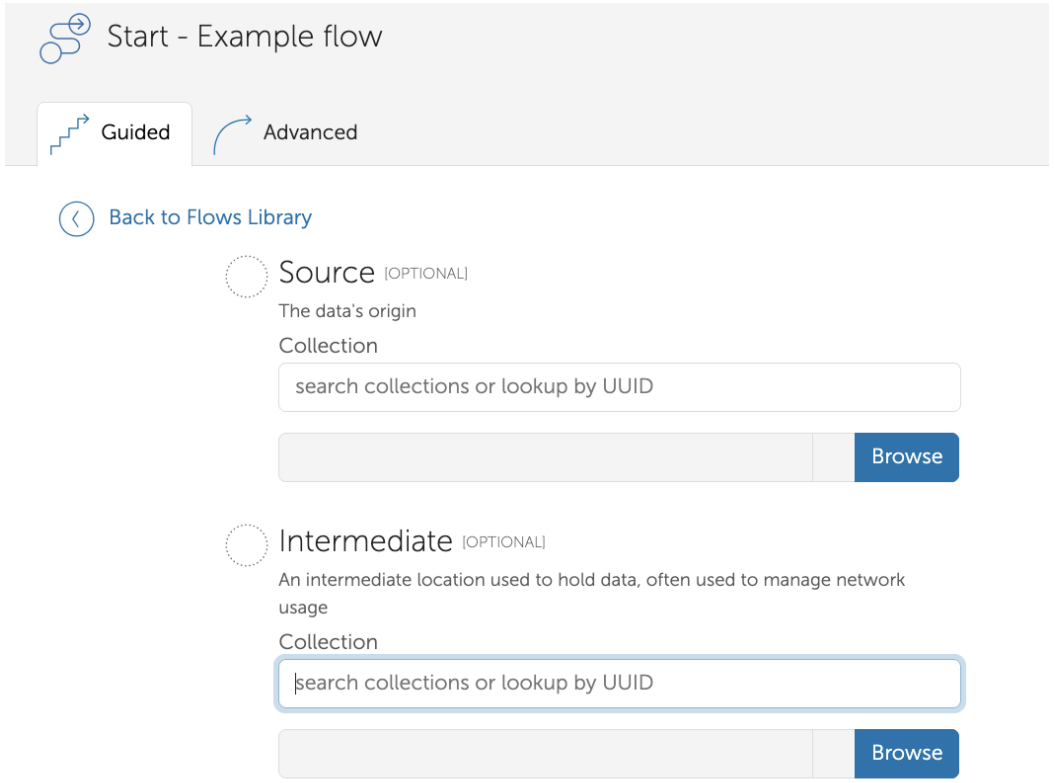
- **Globus Flows: automate common tasks**
- **Example use cases**
  - Ensure that users obey best practices for secure data movement
  - Define complex procedures such as multi-step uploads to a secure environment
  - Hide sensitive services from being directly invoked, except in a controlled way

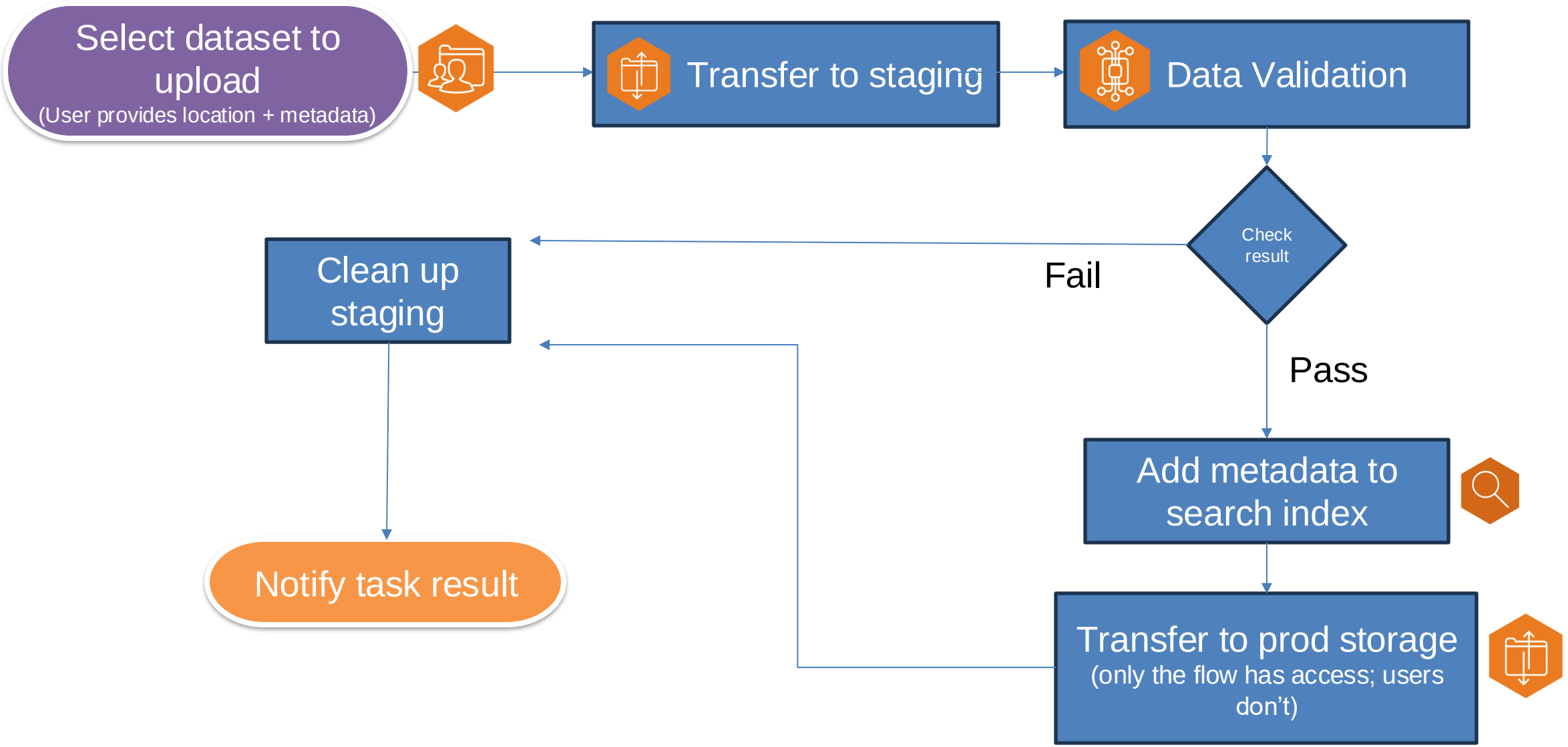
Imperative, distributed, and resilient

- **Distributed design allows leveraging many different services as *action providers***
- **Automation will retry and resume if a system component is unavailable**
- **Imperative flow syntax based on Amazon Step Functions**

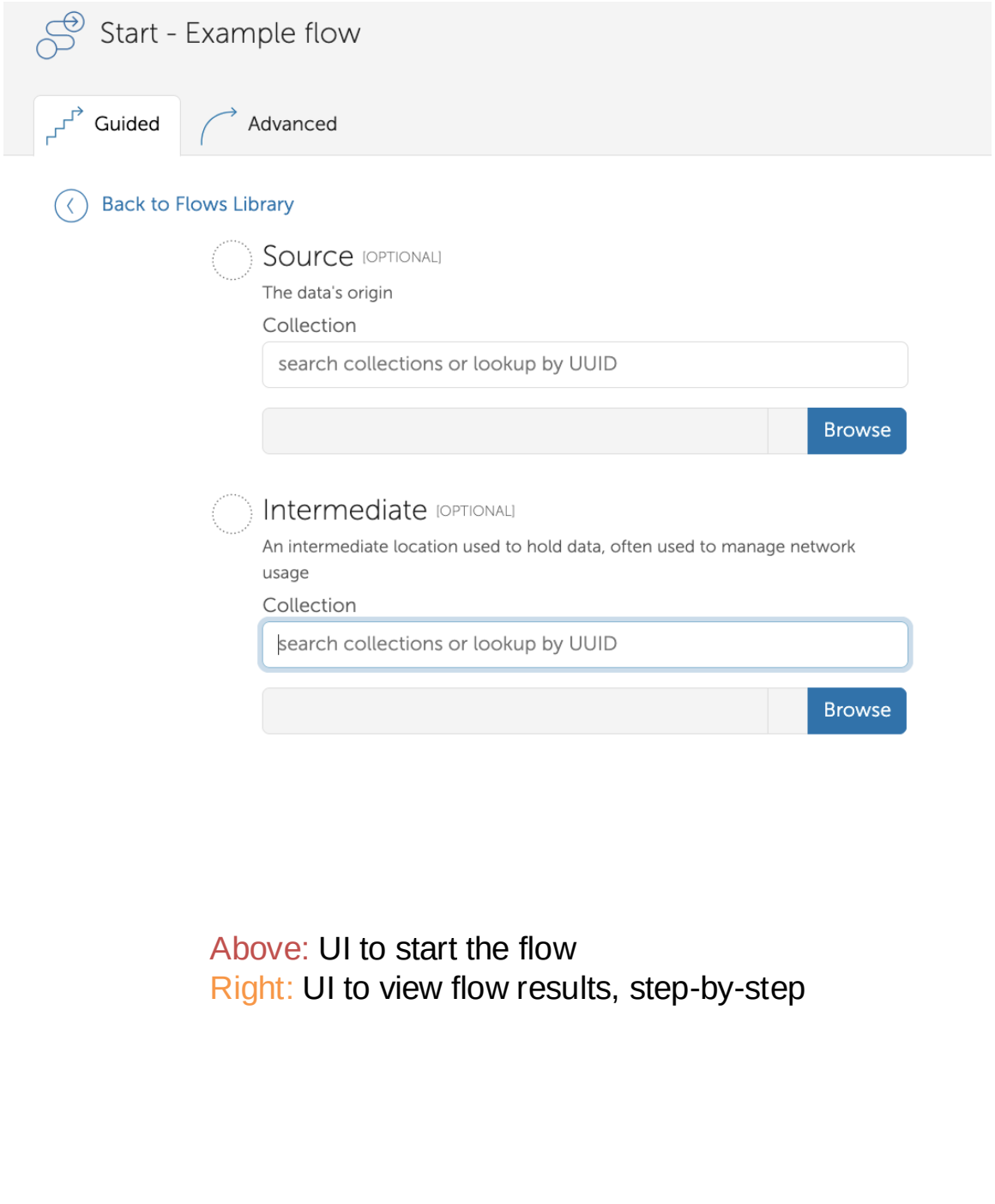
Integrates with Globus Services

- **Action providers can invoke many Globus Features**
  - Conditional logic can take steps based on results at each step
- **Auto-generated UI for running flows and selecting files**

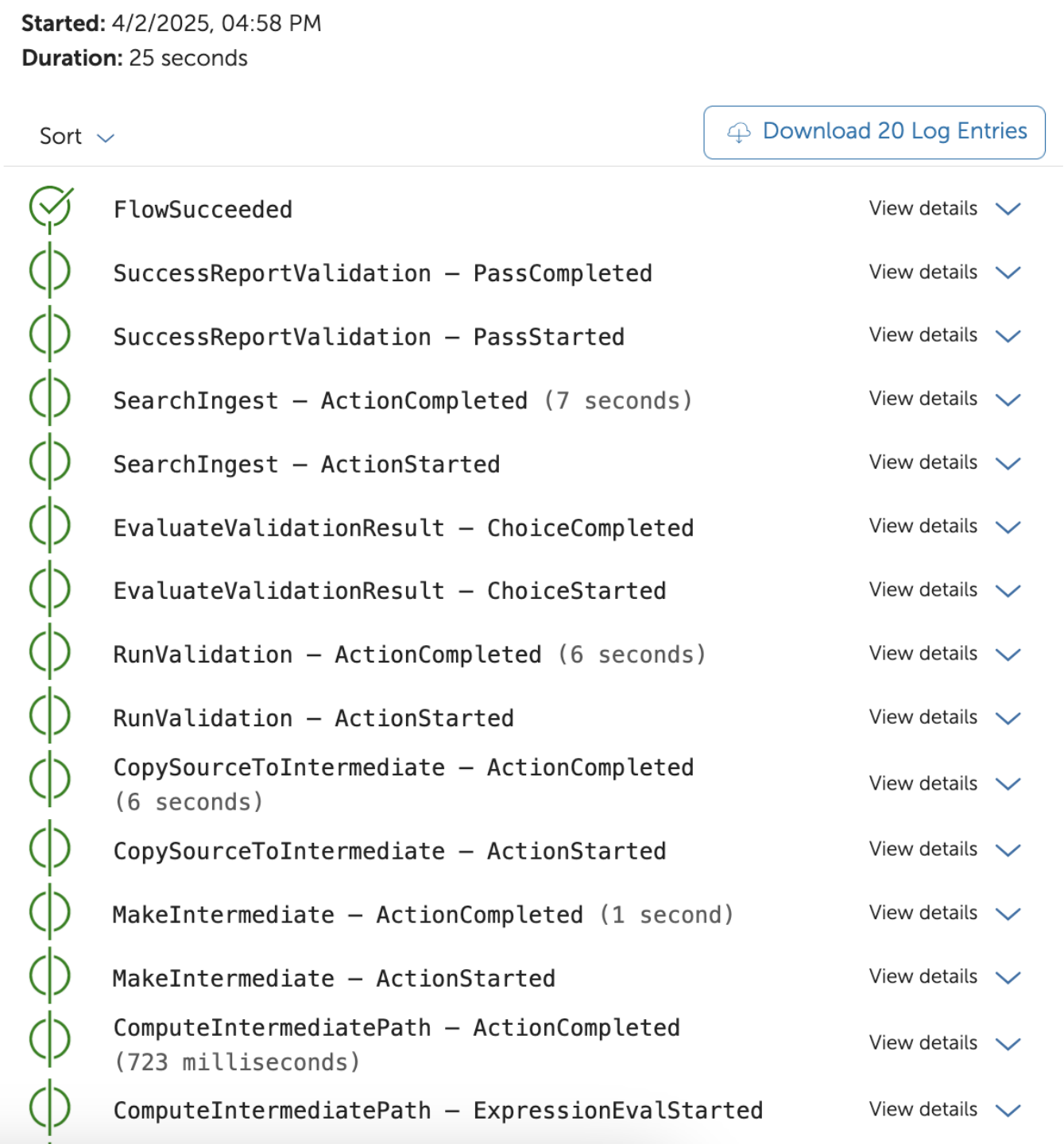




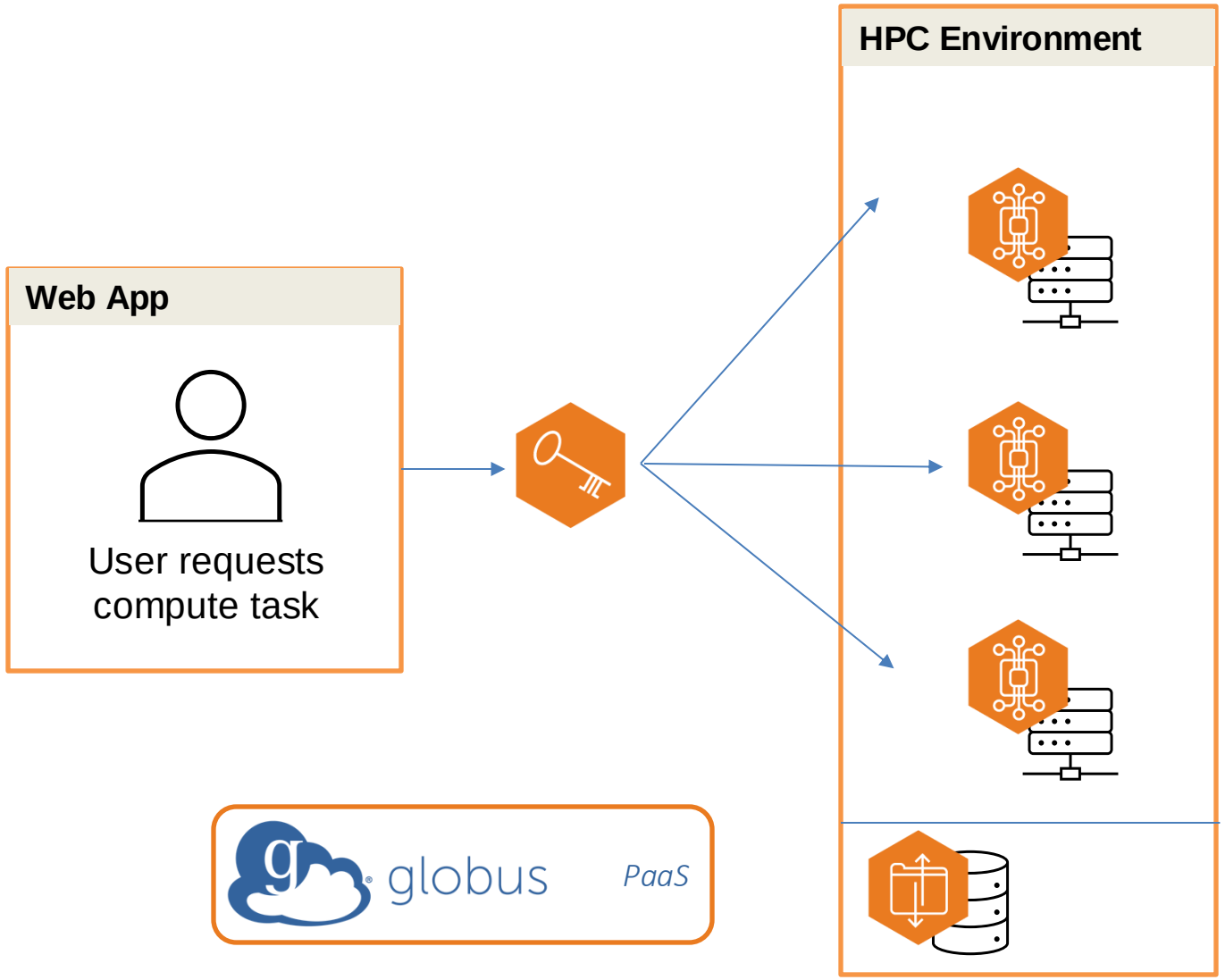
Transfer + Compute + Search



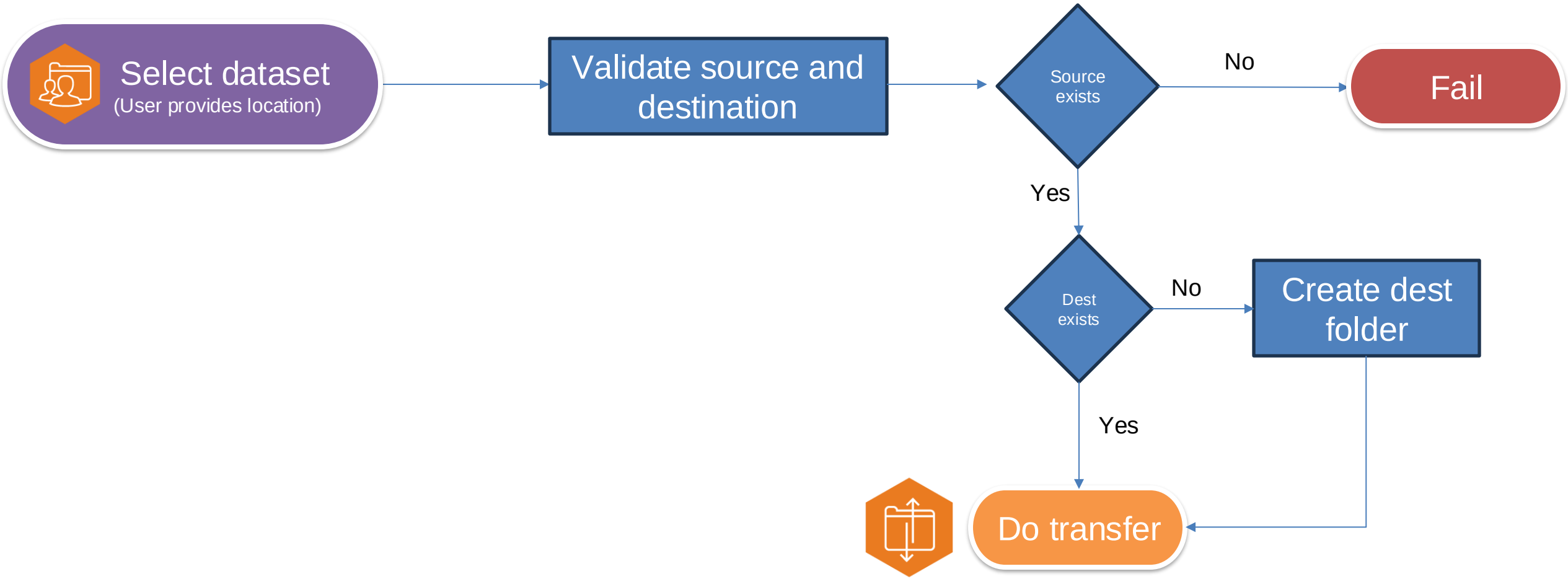
Above: UI to start the flow
Right: UI to view flow results, step-by-step



Fan out across many workers

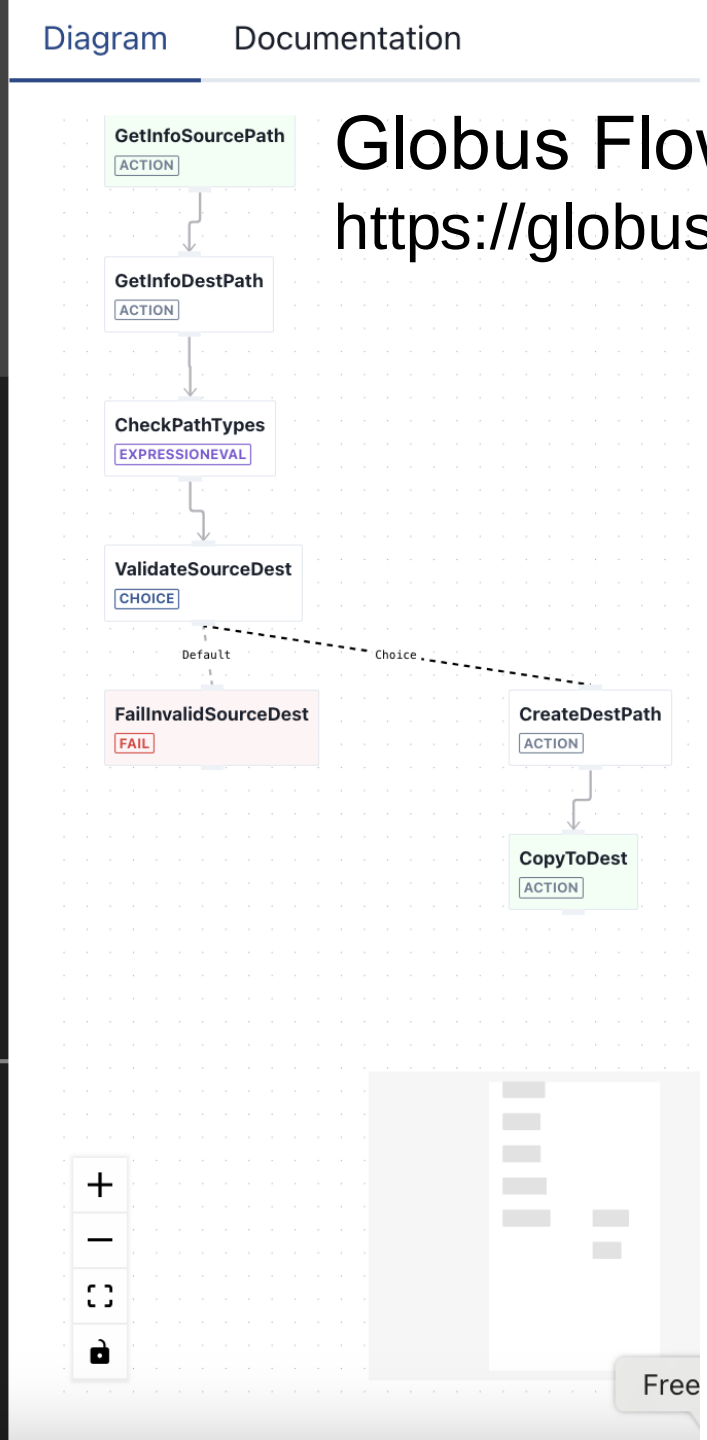


- **Allow users to request big compute from a website**
- **Scale compute separately from webapp**
- **Enable remote access to compute resources, without SSH**



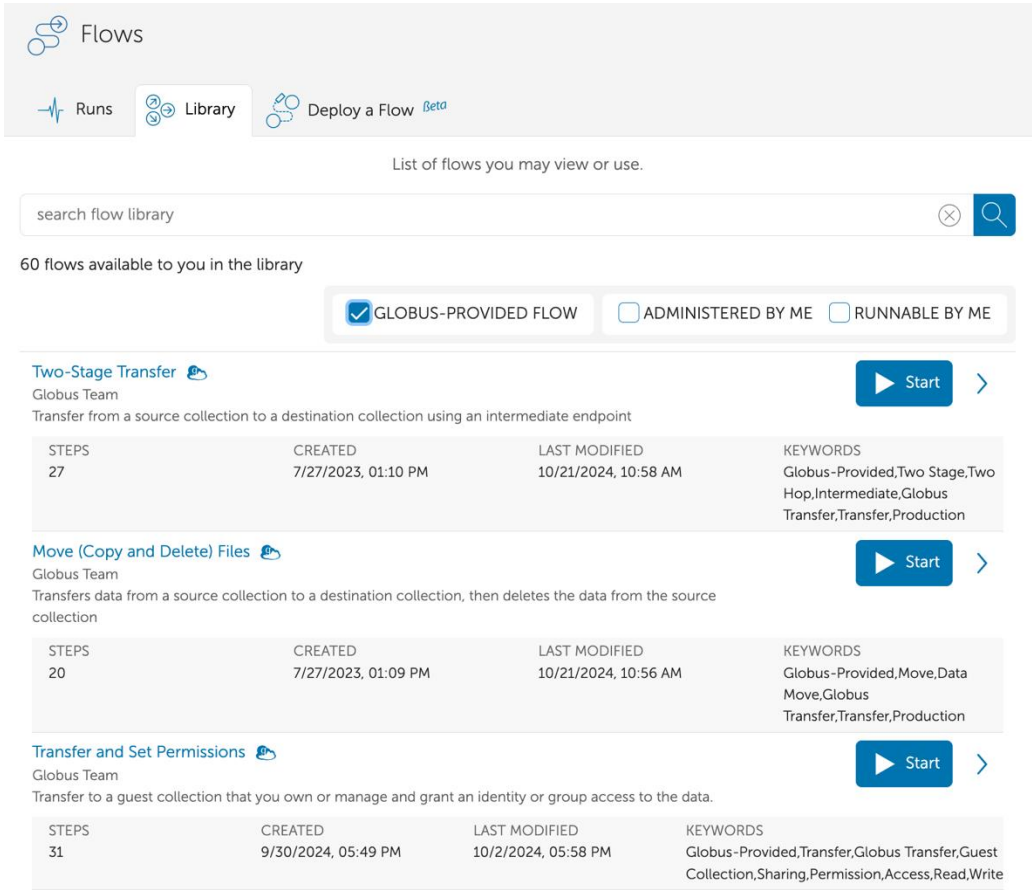
Process Automation: Ensure that data upload complies with internal policies

Demo repo: *flows-compute/flow/data/transfer_with_filters*



Globus Flows IDE (βeta)

https://globus.github.io/flows-ide



Globus Flows Library

https://app.globus.org/flows/library

Documentation and Tooling

- **Docs:**
  - https://docs.globus.org/api/flows/
- **Tools for writing flows:**
  - https://globus.github.io/flows-ide
  - https://gladier.readthedocs.io/

When to use Globus Flows

Common use cases

- **Manage complex permissions models (such as submission to data repository)**
- **Enforce best practices (such as filtering out sensitive data)**
- **Perform actions after transfer**
  - Add to search index
  - Data validation (quick compute)
  - Notify curators

Not aimed at these scenarios

- **Tasks focused on compute, rather than data management**
- **Very big workflows with many complex steps**
- **Makefile-like “only rebuild what changed” behavior**

Snakemake

nextflow

Globus Compute:

Add custom FaaS steps into your workflow

globus

Function as a Service (FaaS)

- **User-defined functions that can be run on a remote environment**
- **Enable access to compute resources without using SSH**
- **Retrieve results in workflows, notebooks, etc**

```
def process_images(input_path=None, result_path=None):

    import os
    import glob
    from PIL import Image

    files = (file for file in glob.glob(os.path.join(input_path,'*')) \
        if os.path.isfile(os.path.join(input_path, file)))

    if not os.path.exists(result_path):
        os.makedirs(result_path)

    for file in files:
        image = Image.open(file)

        # Generate thumbnail
        image.thumbnail((200, 200))

        # Save thumbnail image
        image.save(f"{result_path}/thumb_{os.path.basename(file)}")
```

https://github.com/globus/globus-jupyter-notebooks/blob/master/Compute_Introduction.ipynb

FaaS as an interface to the advanced computing ecosystem

Still need…

- **Single interface**
- **Homogenous execution environment**
- **Transparent and elastic execution**
- **Integrated with data management**

Example: Automating cryoEM flows

Globus Flows

Transfer

Transfer raw files

Compute

Launch analysis job

Carbon!

Correct, classify, …

Compute

Extract metadata

Search

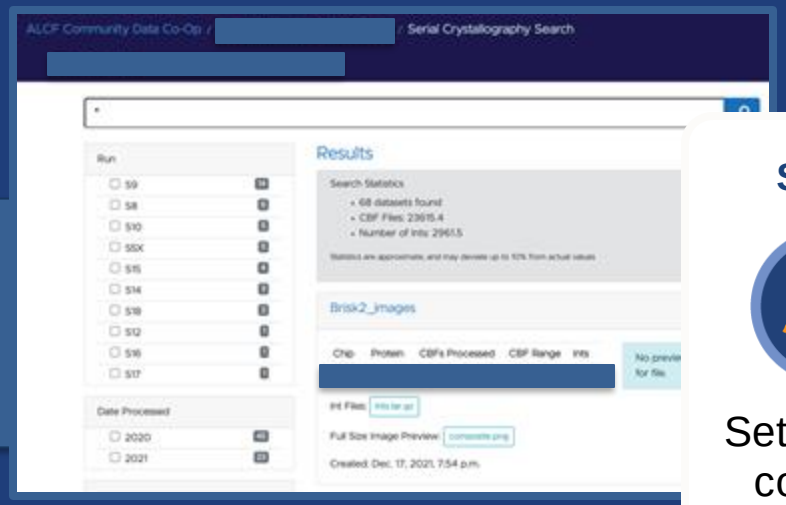
Ingest to index

Transfer

Move final files to repo

Share

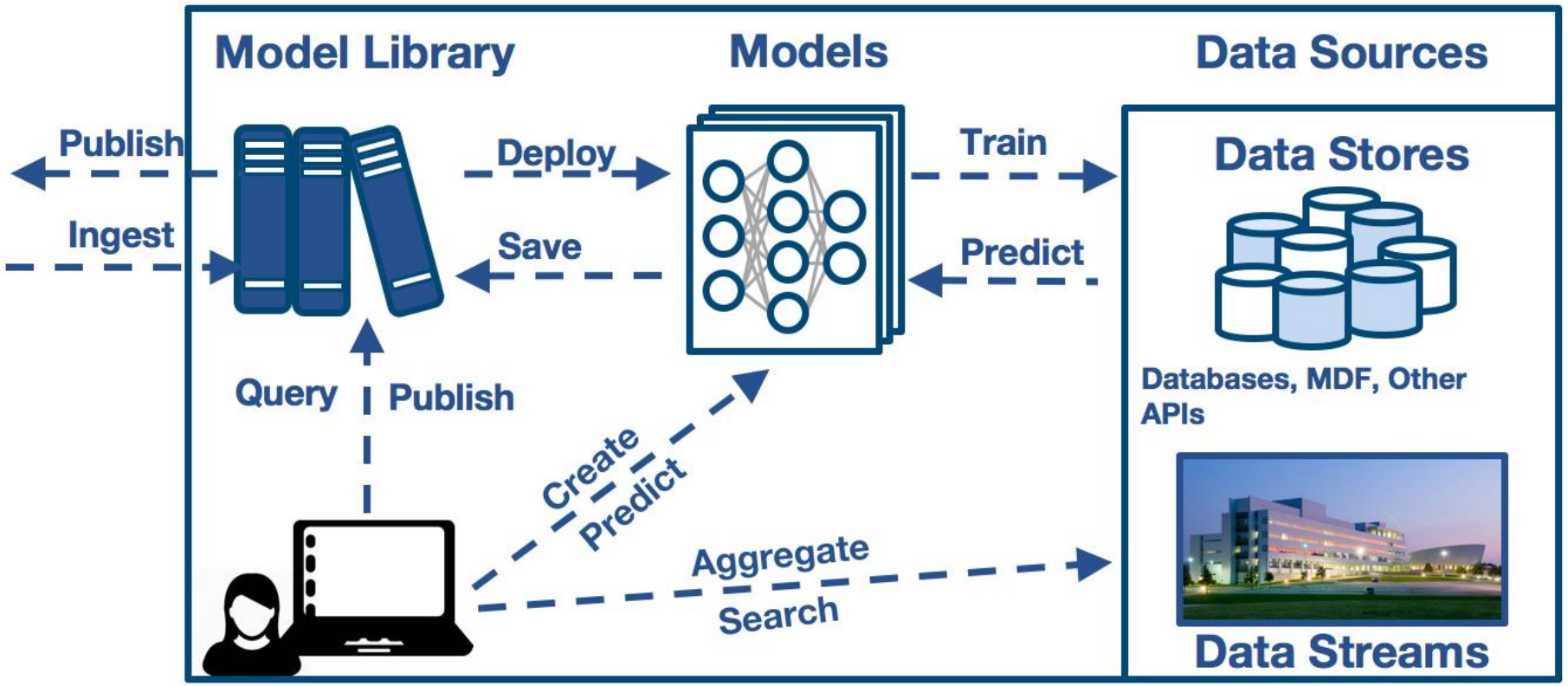
Set access controls

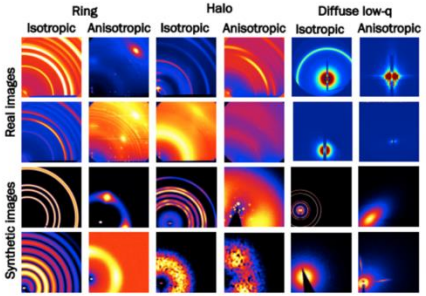


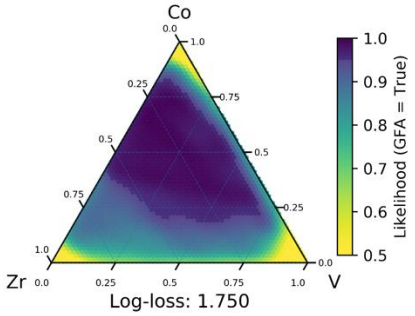
Integrates with Globus Flows

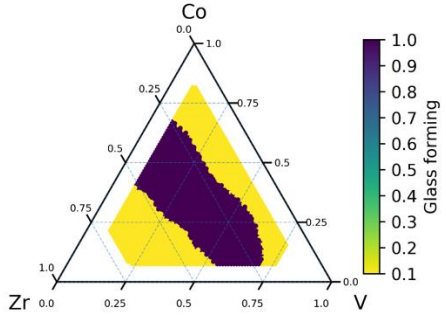
- **Can add custom data processing or validation into an upload step**
  - Verify that data is AI/ML ready before adding to catalog
- **Can use workflows to initiate compute functions in a controlled way**
- **Flows can directly use the output of this function in following steps**

An ecosystem of AI/ML models: DLHub, Garden.ai, etc









garden https://thegardens.ai/

Limits

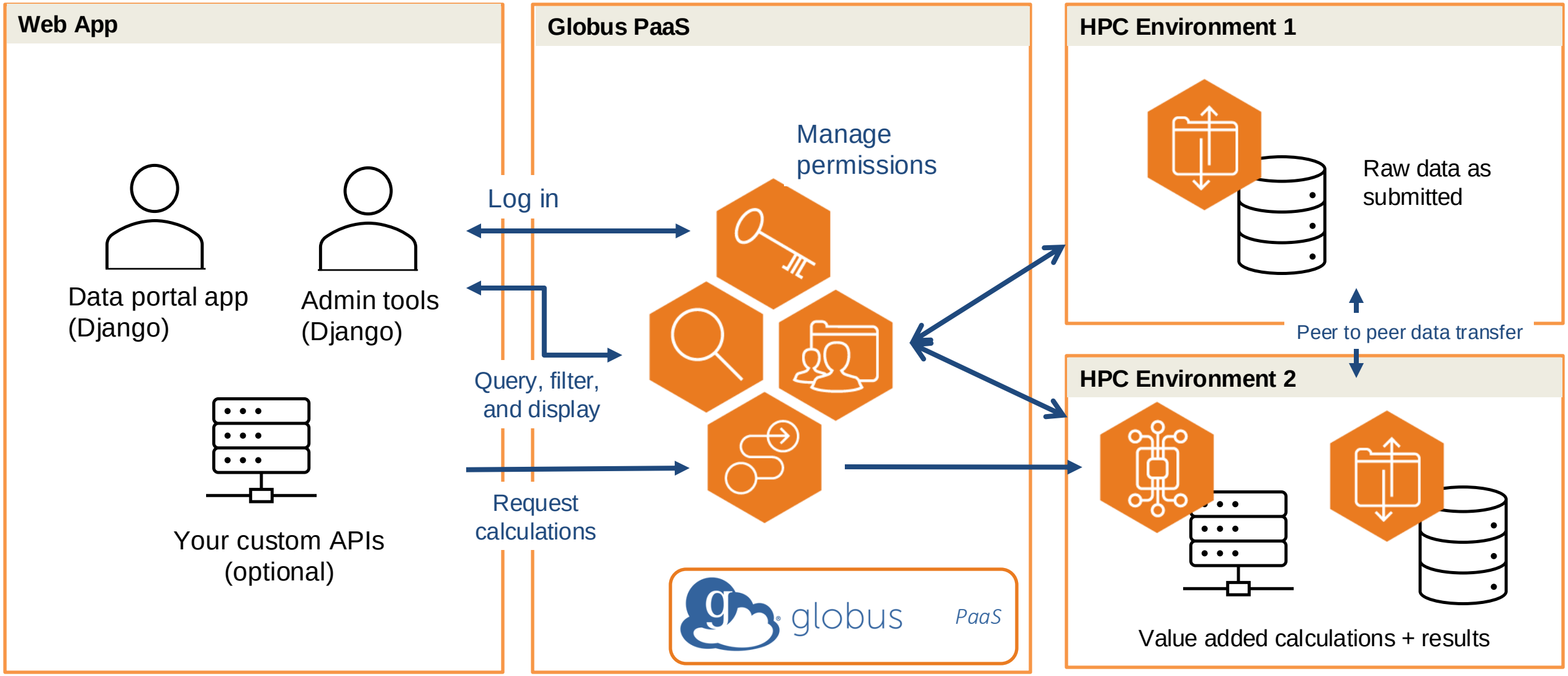
- **Not tightly integrated with Globus Storage/Transfer**
- **More complex functions may take sysadmin help to deploy the first time**
  - Not endpoint-agnostic
  - Pick a few well-defined AI/ML models
- **Compute continues to evolve to support new use-cases**

Recap

User stories recap

- **Transfer to get files in**
- **Flows to provide automation + data validation**
- **Compute to provide access to advanced resources**
- **Search index + DGPF UI to make data discoverable**
- ***…and a whole lot of science from you***

Portal Platform: Recap



Discussion questions

(…To be continued at ANL April Gathering!)

What do we want to automate?

- **Data deposition: teams with sensitive data management concerns**
- **Data validation: basic checks of common data types**
  - What teams are developing these pieces?
  - What is the path to integrating with a data repository for AI/ML ready models?
- **Data curation / required approval**
  - How can curation flows be integrated with data ingest?

APECx Models and Methods-v1-shared (Jan 2025, Yoon et al.)
APECx data catalog (ARPA-H team. March 2025): internal communication (VaxCo + UW)

What capacity is needed?

- **Compute can provide access to big compute resources, but requires setup**
  - Are there methods / tools we want to expand access to?
- **Are people interested in collocating compute + data?**
  - This may require advance planning, but could leverage a strength of the data repository

Who needs access?

- **Which team(s) would need to use compute? What environments need access?**
- **Touchpoints: Who is developing the data harmonization and curation? How do we integrate them with the catalog?**