

# EOSDIS Data Services In the Cloud

Earthdata Cloud Services Team

# Main Goals

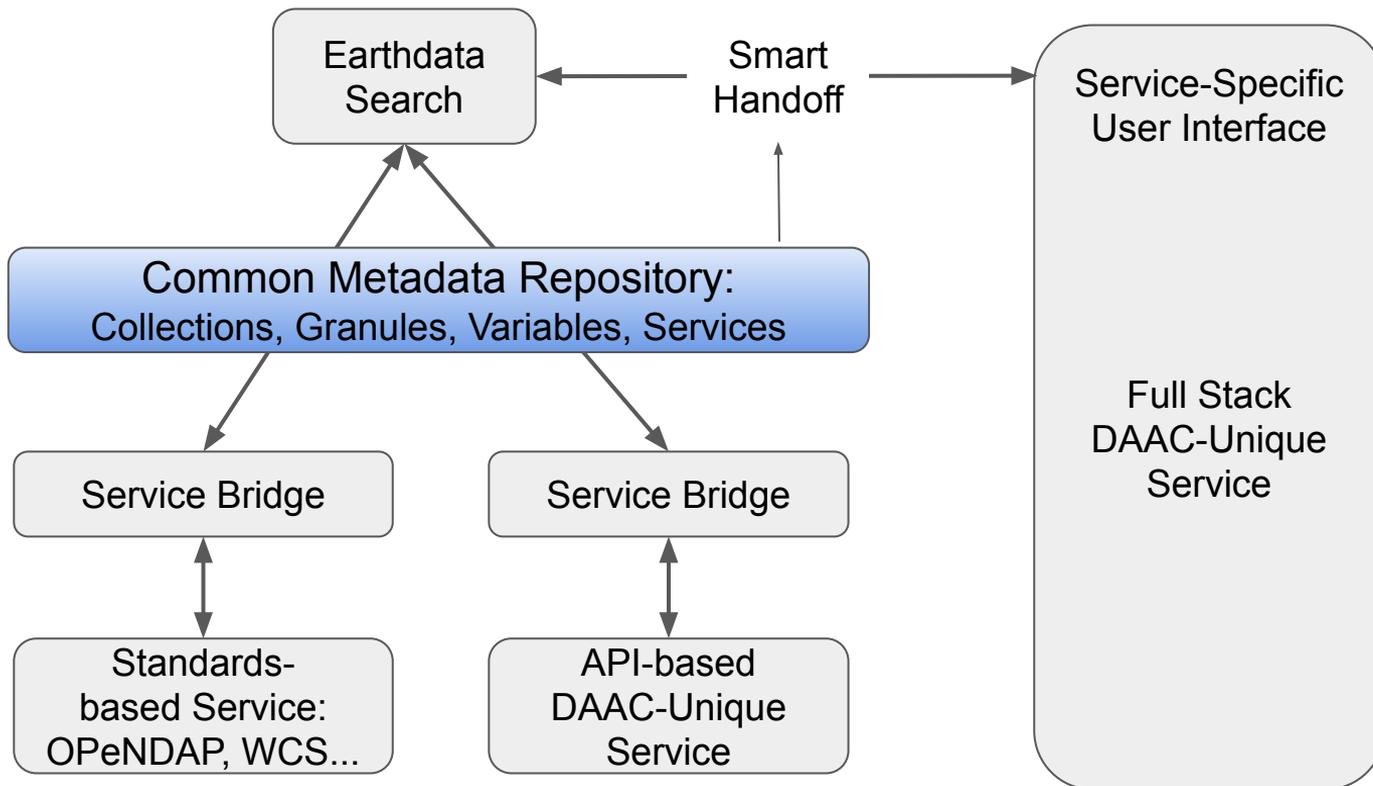
1. **Functionality:** Provide data transformation functions needed by end users to make data ready for analysis
2. **UX:** Enable a good user experience
  - a. Humans (UI)
  - b. Machines (API)
3. **Efficiency:** Leverage both DAAC and Enterprise assets



# Data Transformation Services in the Cloud

- Common Data Transformation Services
  - Subsetting: Variable, Spatial, Temporal
  - Reformatting: shapefile, etc.
  - Reprojection / Reprojection / Orthorectification
  - Spatial Aggregation: Stitching / Mosaicking
  - Time aggregation
- Dataset-Specific Preprocessing
  - Radiometric Terrain Correction (Synthetic Aperture Radar)
  - Geophysical Retrievals
  - Etc.

# The Big Picture: Connecting (All) the Dots



# Criteria for Enterprise- vs. DAAC-Supplied

1. Dataset specificity
2. Reuse of existing tools
3. Available staff

# What's So Special about Data Services in the Cloud?

## 1. New Challenges

- a. Access: data are in Web Object Storage (not filesystems) => sub-file access is complicated (=> CoG, Hyrax, Kita)
- b. Cost: more services =>
  - i. more processing => more money
  - ii. subsetting => less egress => less money

## 2. New Opportunities

- a. Scaling: faster services => synchronous => machine interfacing
- b. Virtual Co-Location: Mixed-Data Services + Mixed-Service Workflows
- c. Cheap Staging: analysis-in-place

## Strengths

Scaling

S3 costs

Virtual co-location

## Weaknesses

Egress cost

Storage access granularity

## Opportunities

Faster services

Synchronous machine interface

Mixed-data services

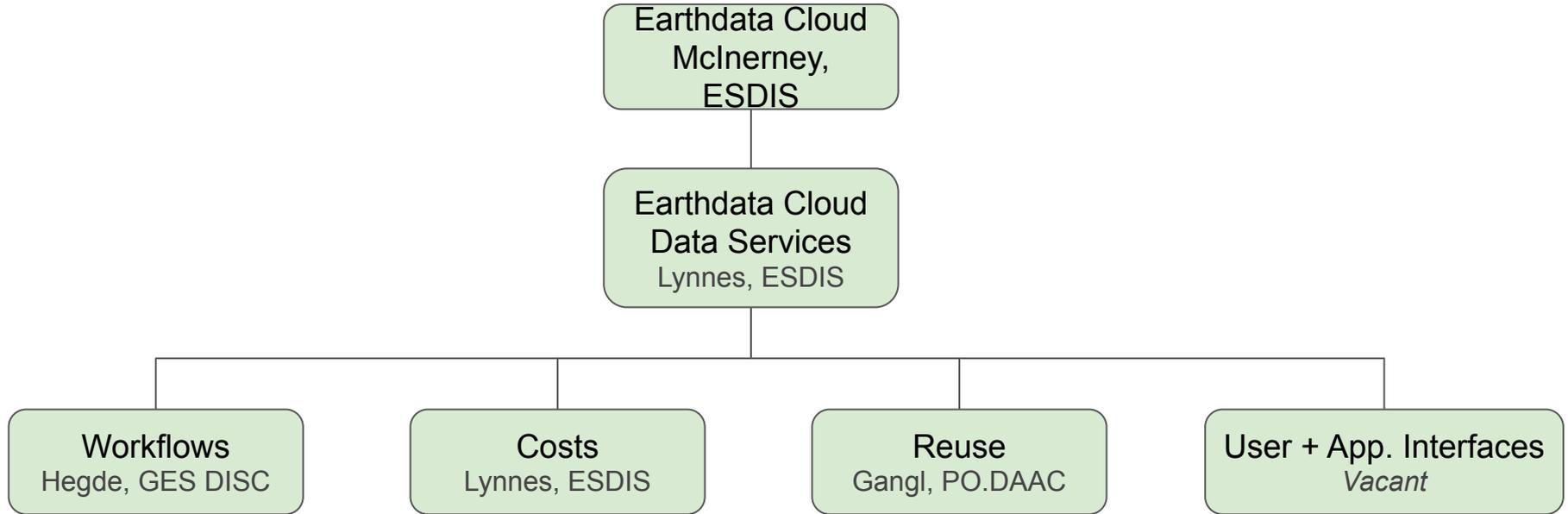
Mixed-service workflows

Analysis-in-place

## Threats

Unbounded demand-driven cost

# Cloud Data Services Subgroups





# Reuse

- Need to identify prioritized functionality for re-use
  - Lots of options from subsetting to reformatting to regridding
- Modes of reuse intrinsically tied to costs
  - SAAS vs Containers vs Source Code
- Adaptation of existing on-prem capability
  - scalability



# Workflows

## Goals:

1. Best practices for running multi-step processes in the cloud
2. Evaluate applicable workflow specifications/standards

## Deliverables:

1. Guidelines for picking a workflow-based solution vs a serverless event-driven solution
2. Best practices for workflows in the cloud (AWS specifically)



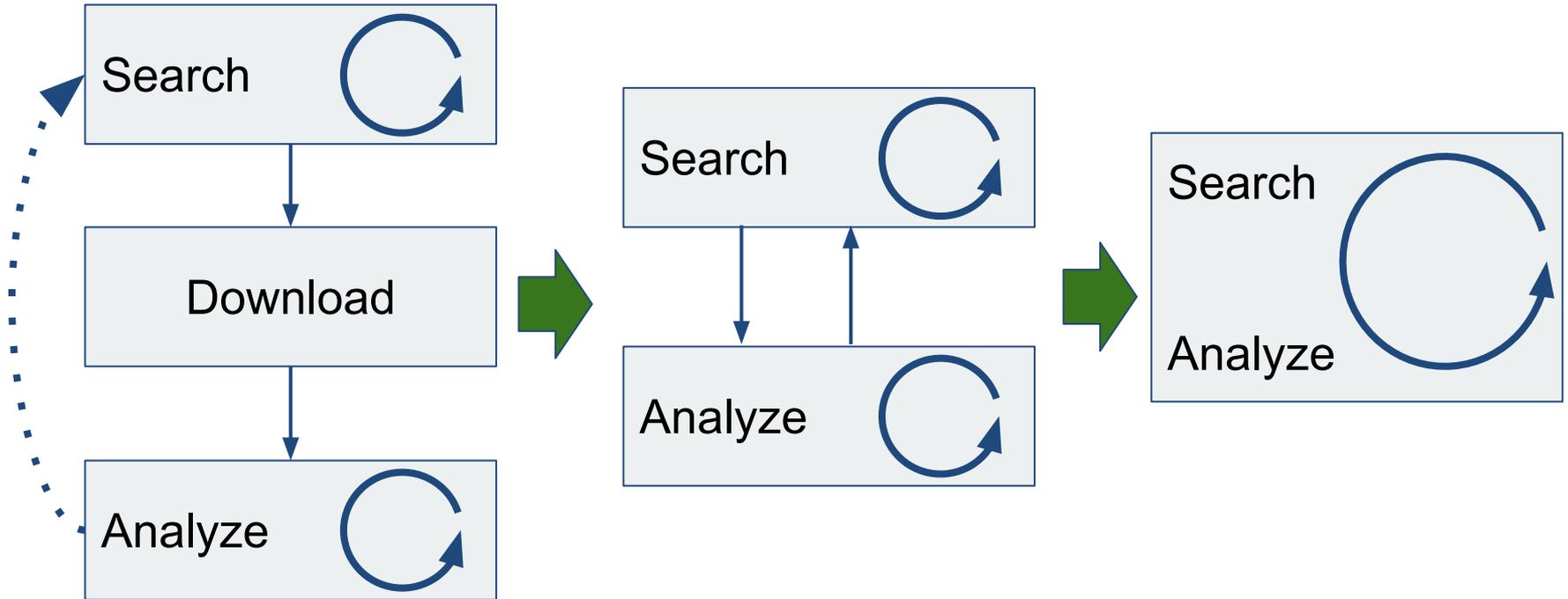
# Managing Cost

- Managing cost is HARD for data services!
  - Egress vs. Processing vs. Storage
  - Easy Calls:
    - Promote subsetting
    - Promote analysis “in place”
  - Harder tradeoffs
    - How much to do for the user?
    - How much to cache?
- New Tasks
  - Developing *cost-effective* data transformation capabilities
  - Monitoring ongoing expenditures vs. budget



# User and Application Interfaces

*Moving toward Search-Analysis Convergence*





# Timeframe for First Outputs

Reuse: June 3

Workflows: June 30

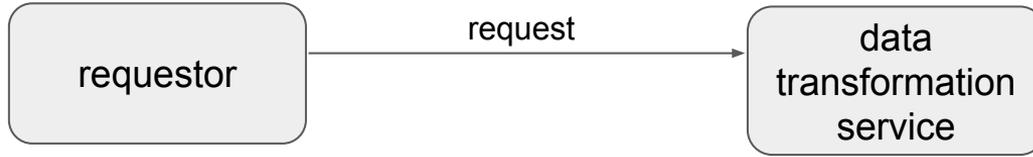
Costs: June 30

User + Application Interfaces: June 30

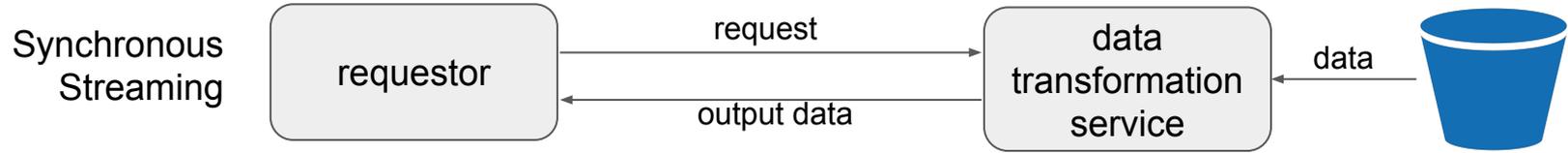
# Backup Slides

# Service Fulfillment Modes

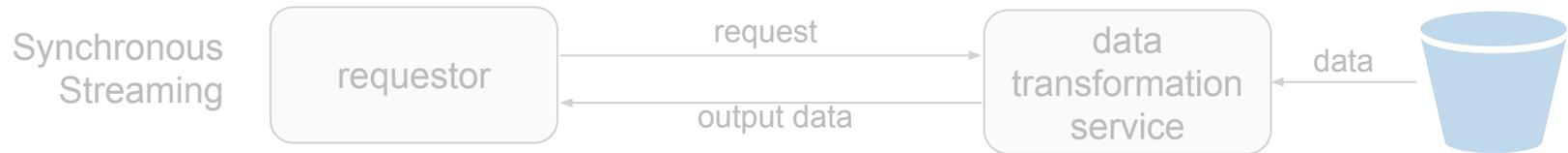
Synchronous  
Streaming



# Service Fulfillment Modes



# Service Fulfillment Modes

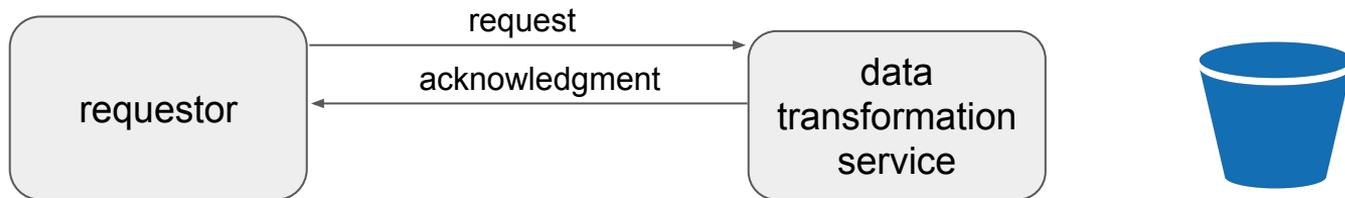


# Service Fulfillment Modes

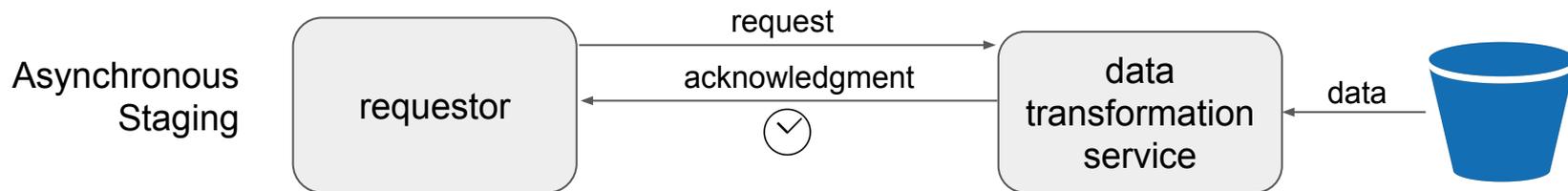
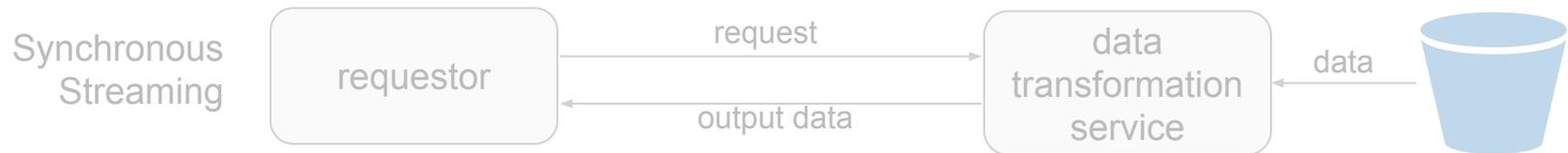
Synchronous  
Streaming



Asynchronous  
Staging



# Service Fulfillment Modes

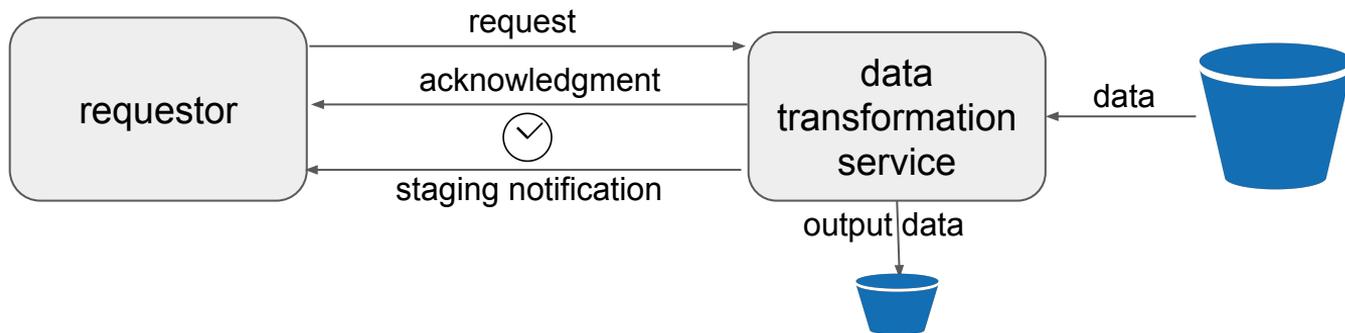


# Service Fulfillment Modes

Synchronous  
Streaming



Asynchronous  
Staging

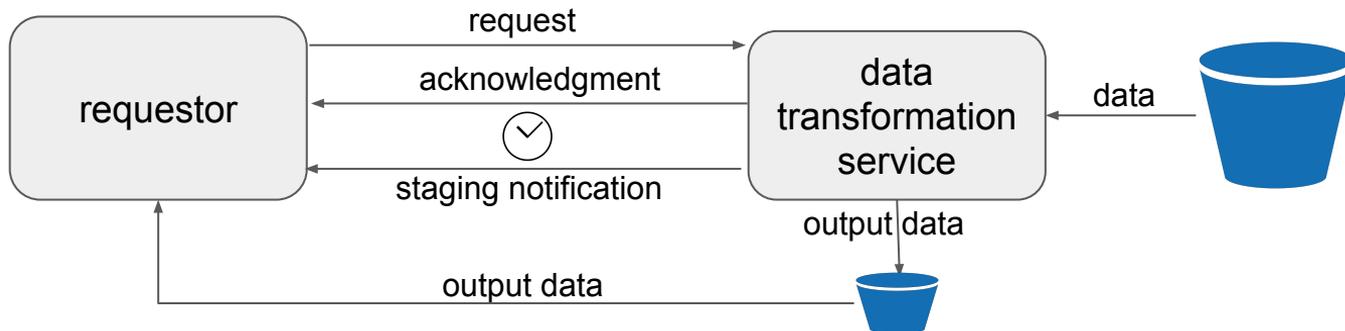


# Service Fulfillment Modes

Synchronous  
Streaming



Asynchronous  
Staging

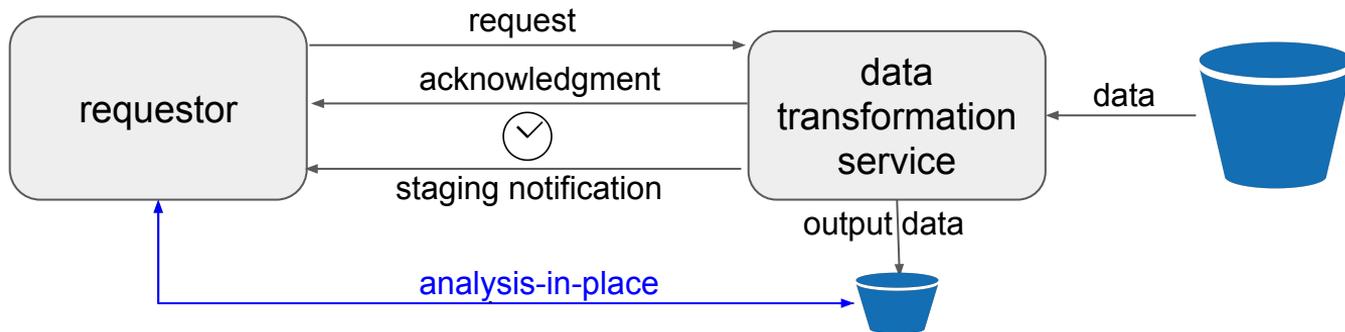


# Service Fulfillment Modes

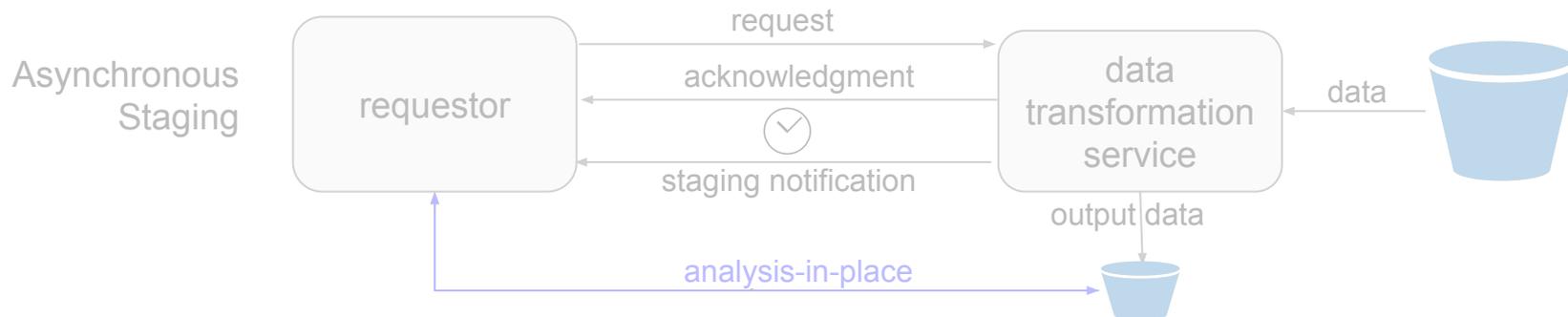
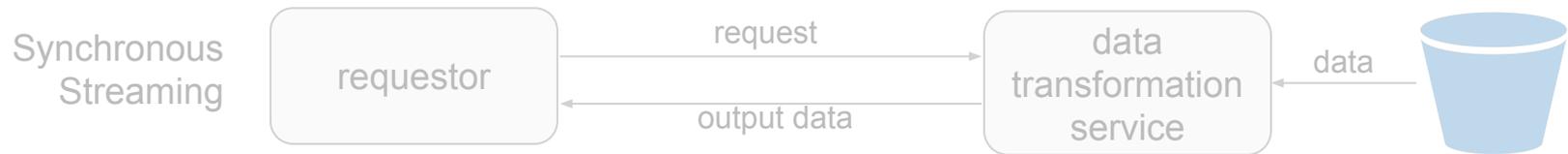
Synchronous  
Streaming



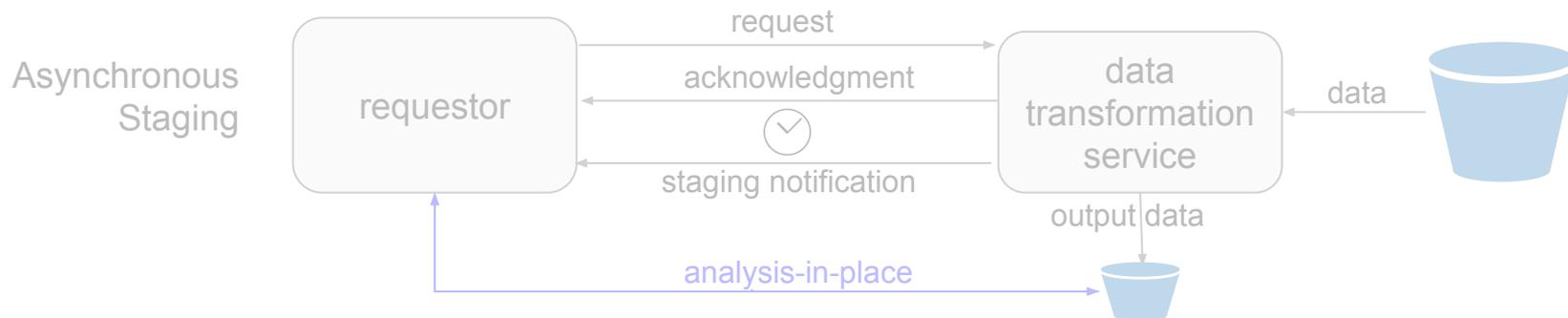
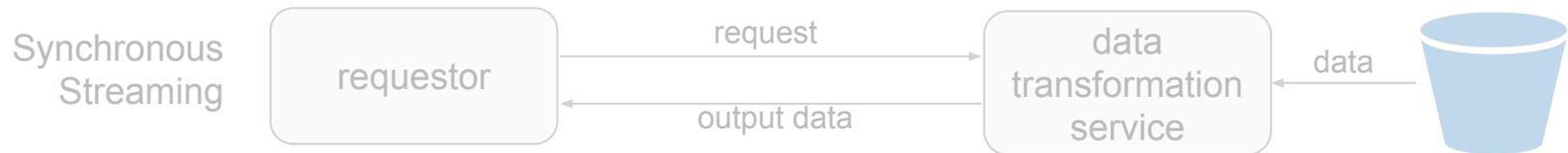
Asynchronous  
Staging



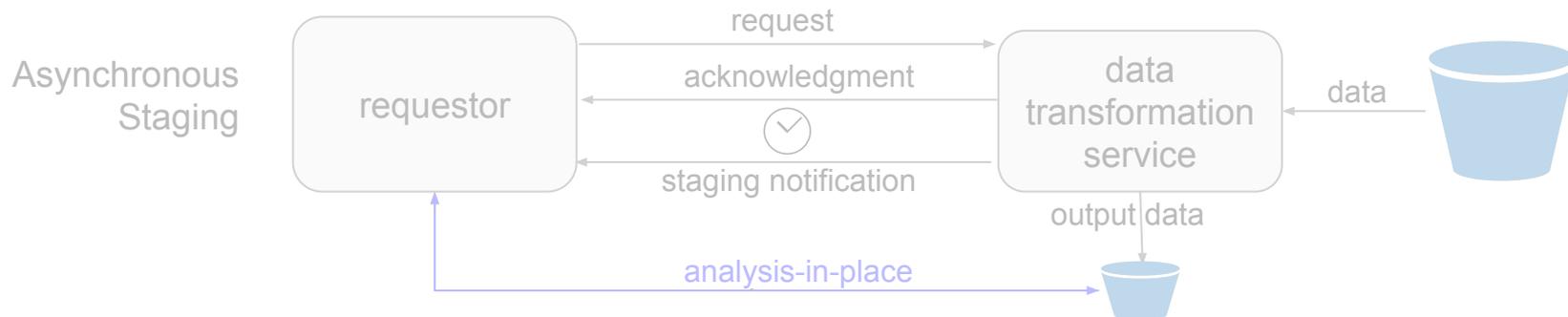
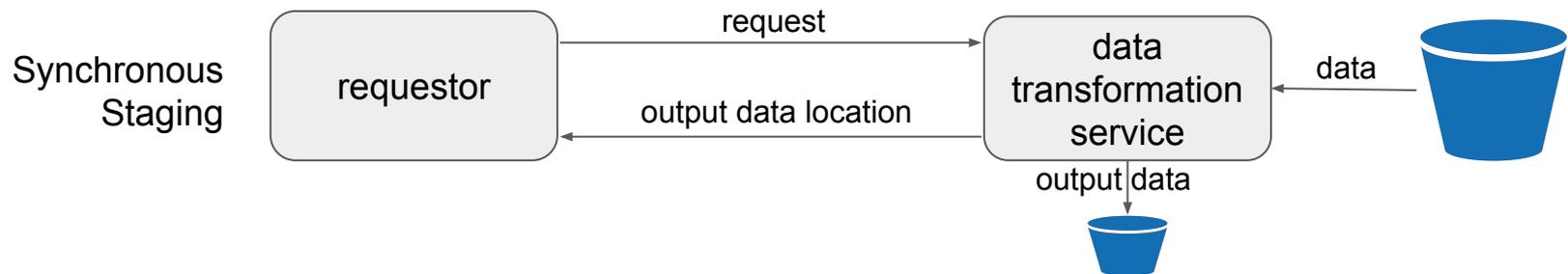
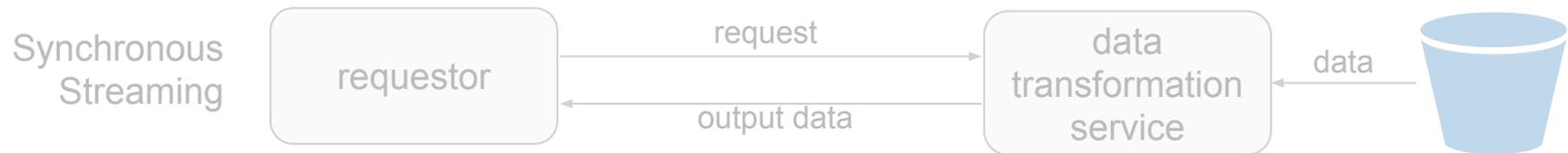
# Service Fulfillment Modes



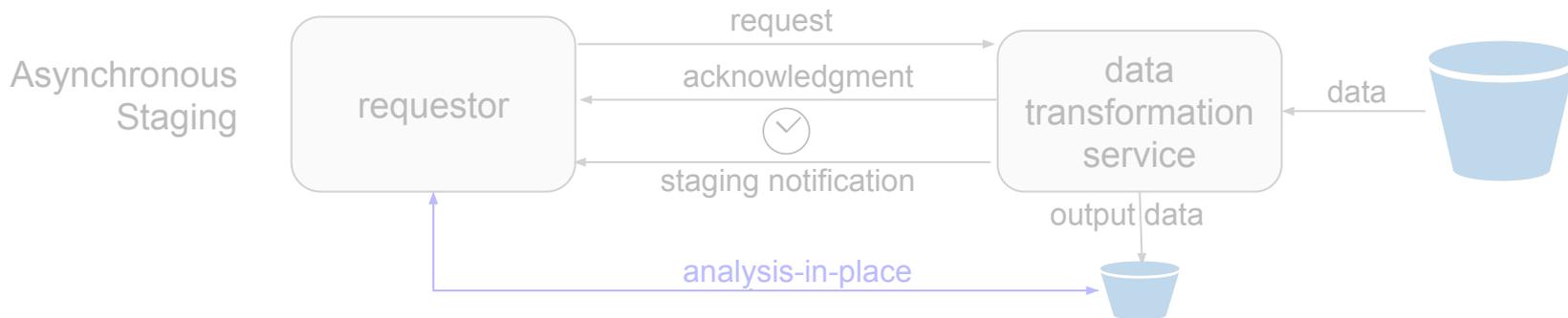
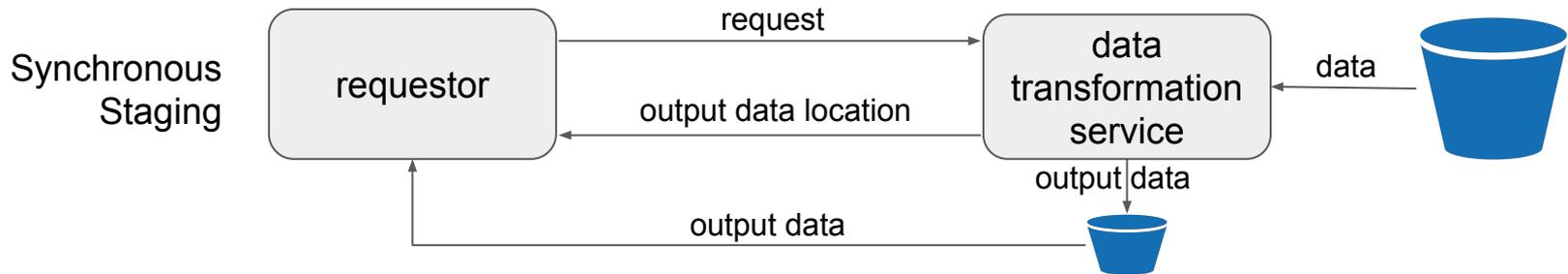
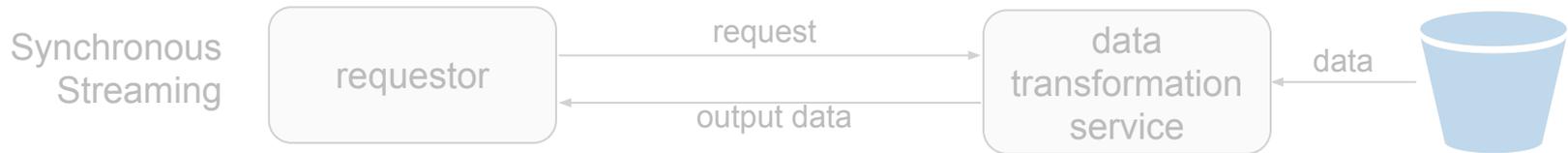
# Service Fulfillment Modes



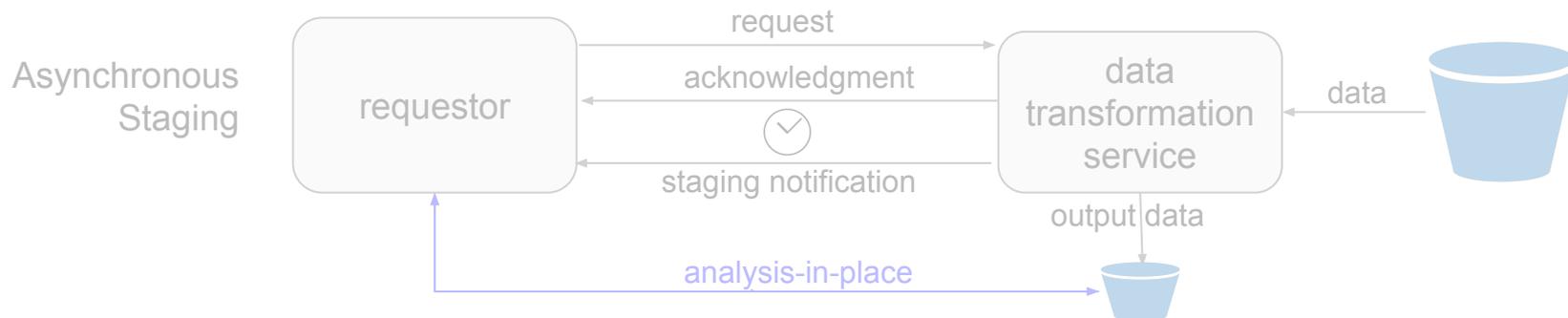
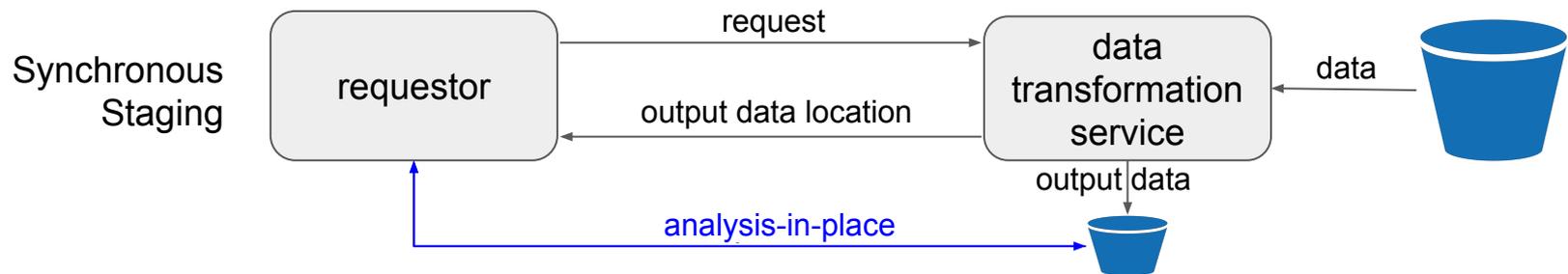
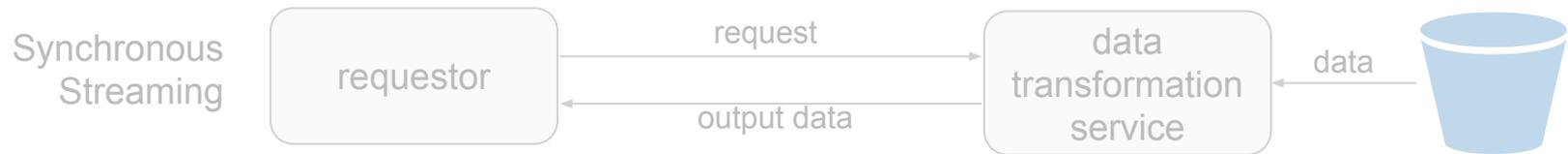
# Service Fulfillment Modes



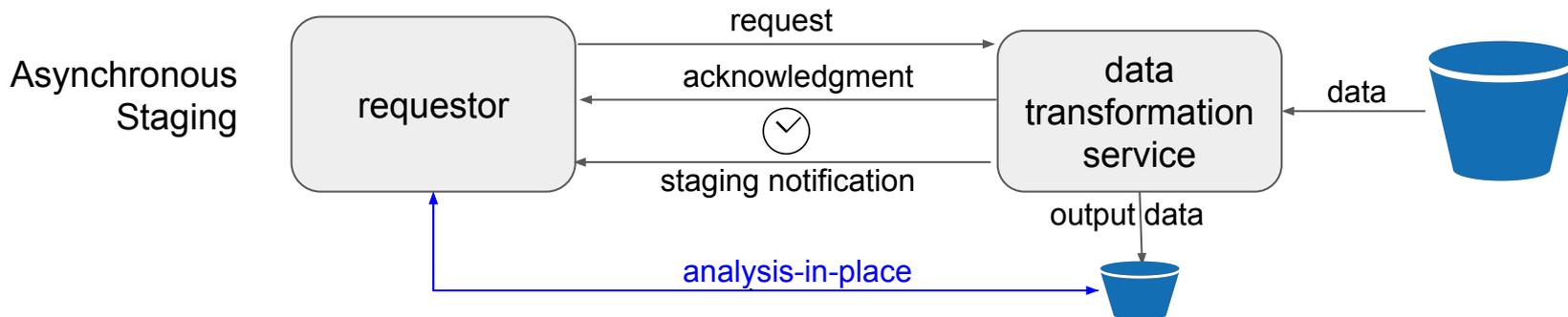
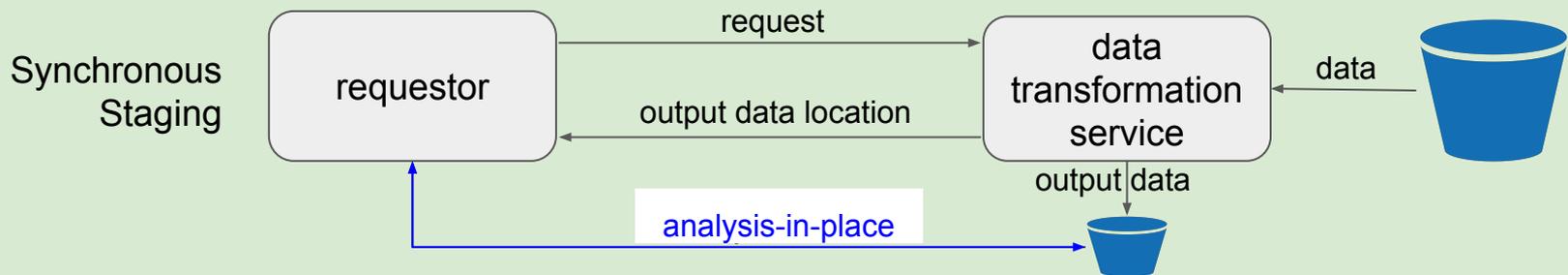
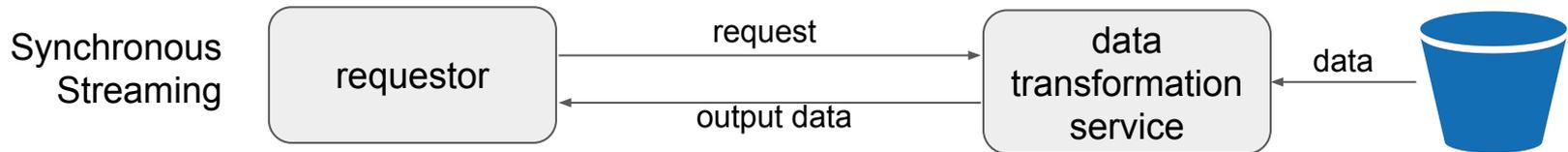
# Service Fulfillment Modes



# Service Fulfillment Modes

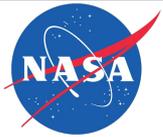


# Service Fulfillment Modes



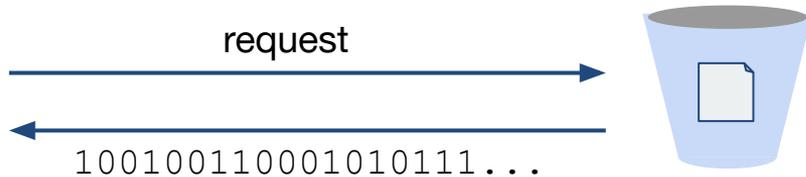
# Pros and Cons of Different Modes

Synchronicity	Data Flow	Examples	Pros	Cons
Synchronous	Stream to client	OPeNDAP	<ul style="list-style-type: none"><li>• Easiest machine interface</li></ul>	<ul style="list-style-type: none"><li>• Fast service reqt</li><li>• Data egress</li><li>• Single-file-out mode only</li></ul>
Synchronous	Stage to S3	WCS 1.1 “store=true”	<ul style="list-style-type: none"><li>• Easy machine interface</li><li>• Handles multiple files</li><li>• <a href="#">Analysis-in-place</a></li></ul>	<ul style="list-style-type: none"><li>• <i>Really</i> fast service reqt</li></ul>
Asynchronous	Stage to S3	HITIDE AppEEARS	<ul style="list-style-type: none"><li>• Unlimited number of files</li><li>• <a href="#">Analysis-in-place</a></li></ul>	<ul style="list-style-type: none"><li>• Hard machine interface</li></ul>



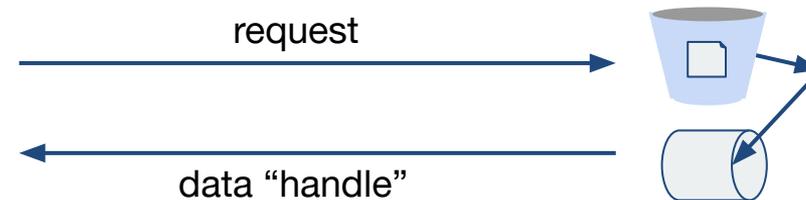
# User Interaction Patterns

synchronous streaming



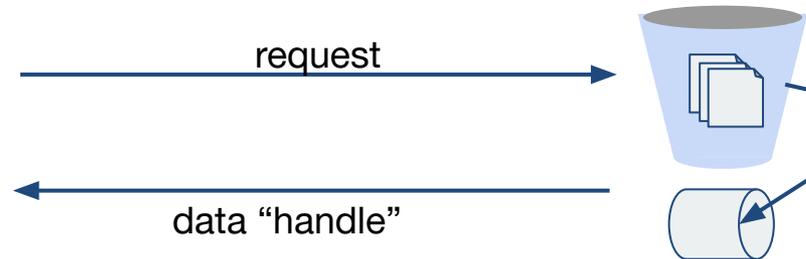
subsetting 1 file

synchronous staging



preprocessing 1 file to Analysis-Ready Data

asynchronous staging

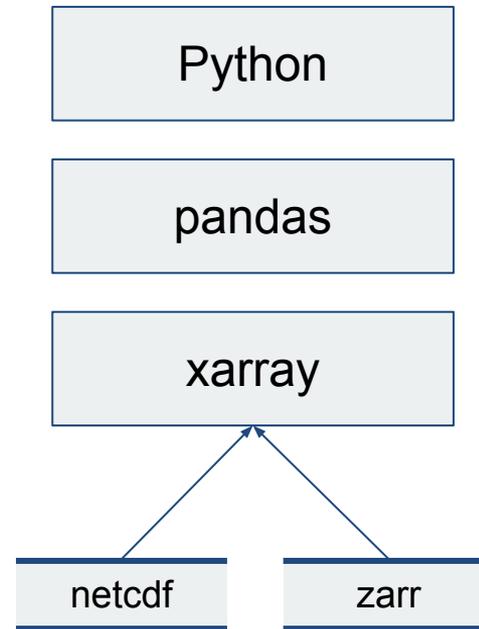


aggregating many files



# Interfaces: User vs. Application

The screenshot shows the NASA Earthdata Search web application. The browser address bar contains the URL: `search.earthdata.nasa.gov/search?m=0.070312510.1406251211010%2C2`. The page header includes navigation links like 'EOSSDIS', 'NASA', 'Software', and 'Find a DAAC'. A search bar at the top contains the text 'Type any topic, collection, or place name'. On the left, a sidebar lists 'Browse Collections' and 'Features' (Map Imagery, Near Real Time, Customizable). The main content area displays a satellite map of the Middle East and surrounding regions. Below the map, it shows '6027 Matching Collections' and sorting options: 'Sort by: Usage', 'Only include collections with granules', and 'Include non-EOSDIS collections'. Two collection entries are visible: 'NCEP/DOE Reanalysis II, for GSSTF, 0.25 x 0.25 deg, Daily Grid V3 (GSSTF\_NCEP) at GES DISC' and 'MODIS/Aqua Near Real Time (NRT) Aerosol 5-Min L2 Swath - 3km'.





# Interface Convergence in Jupyter

The screenshot shows the NASA EarthData Search web interface. The browser address bar displays a search URL. The interface includes a search bar with the text "Type any topic, collection, or place name". Below the search bar is a map of the Middle East and North Africa region. The map shows various countries and their abbreviations. Below the map, there are search filters and results. The search results section shows "6027 Matching Collections". The first result is "NCEP/DOE Reanalysis II, for GSSTF, 0.25 x 0.25 deg, Daily Grid V3 (GSSTF\_NCEP) at GES DISC". The second result is "MODIS/Aqua Near Real Time (NRT) Aerosol 5-Min L2 Swath - 3km".

Jupyter

Python

pandas

xarray

netcdf

zarr



# User-Application Interface Convergence in Jupyter

The screenshot displays the Earthdata Search interface within a JupyterLab environment. The browser address bar shows the URL `https://che-k8s.maap.xyz/dashboard/#/ide/mdebelli/all-extensions-working`. The interface includes a sidebar with navigation options such as Dashboard, Workspaces (3), Stacks, Factories, Administration, and Organizations. The main content area features a search bar with the text "Type any topic, collection, or place name" and a "Show Tour" button. Below the search bar, there is a map of Gabon and a list of 11 matching collections. The first collection is "GPM\_3IMERGH1 v05 - NASA/GSFC/SED/ESD/GCDC/GESDISC". Below it, two "AfriSAR" collections are listed with details about granules and data sources.

**11 Matching Collections**

Sort by: **Relevance**  Only include collections with granules  Include non-EOSDIS collections

Tip: Add [+](#) collections to your project to compare and download their data. [Learn More](#)

**GPM\_3IMERGH1 v05 - NASA/GSFC/SED/ESD/GCDC/GESDISC**

**AfriSAR: Rainforest Canopy Height Derived from PolInSAR and Lidar Data, Gabon**

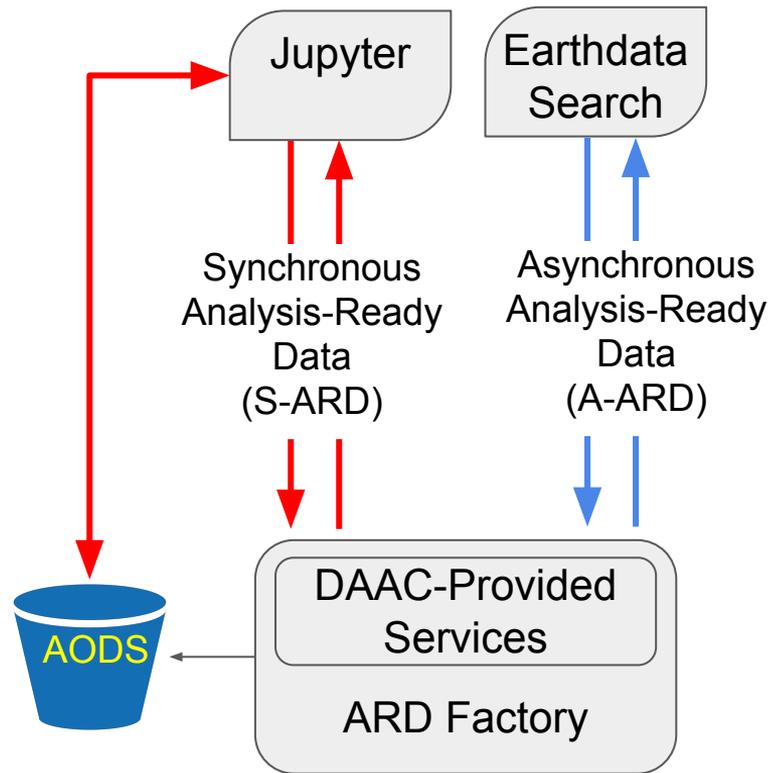
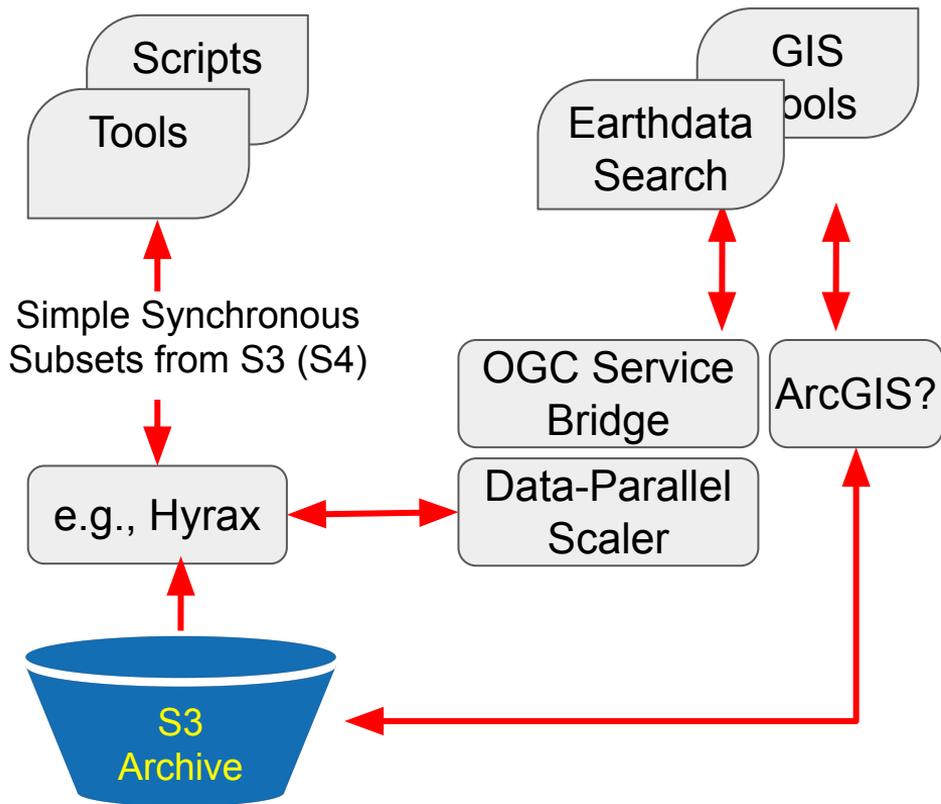
34 Granules - 2016-02-27 to 2016-03-08 - This dataset provides estimates of forest canopy height and canopy height uncertainty for study areas in the Pongara National Park and the Lope National Park, Gabon. Two canopy height products are included: 1) Canopy height was derived from multi-baseline Polarimetric Interferometric...

**PolInSAR\_Canopy\_Height\_1589 v1 - MAAP Data Management Team**

**AfriSAR: Canopy Structure Derived from PolInSAR and Coherence TomoSAR NISAR tools**

51 Granules - 2016-02-25 to 2016-03-08 - This dataset contains forest vertical structure and associated uncertainty products derived by applying multi-baseline Polarimetric Interferometric Synthetic Aperture Radar (PolInSAR) and Polarimetric Coherence Tomographic...

# High Level Architecture (?)



# Analysis in Place

