ARCITECTA®
OPERATING SYSTEMS FOR META+DATA

RMACC

DATA ORCHESTRATION PLATFORM

Craig Vanderborgh | 2022

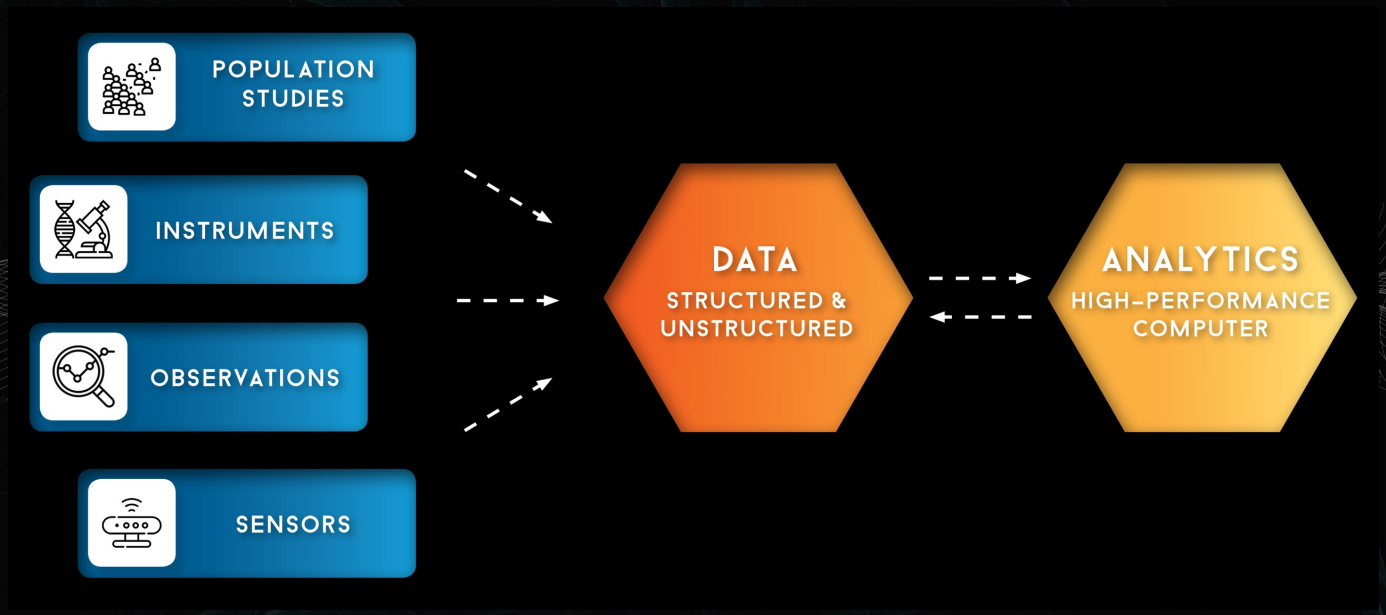
Why we need a Data Orchestration Platform

- Why we need HPC in the first place?
- The traditional focus has been very “Cluster Centric”, there are many pieces to the system
- Fundamental tenet is keep the Cluster busy but how? “Check in – Check out” of scratch storage is key.
- Focus has been on “job scheduling” not overall system throughput/performance and the human aspect. “Repeat the job from 14 months ago”.
- Organizing and moving Data Sets via a “Policy Engine” or “Data Movement Engine”
- Persistent and Archive Storage is an expensive part of the overall system, discuss ways to virtualize these pieces and avoid vendor ”lock-in”

RMACC

Why HPC?

WE ARE IN THE "DATA AGE" – DISCOVERIES ARE DATA DRIVEN

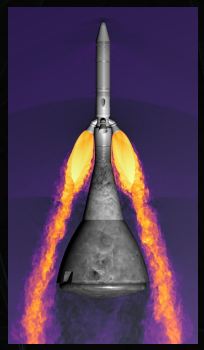


Collect everything, automate, identify patterns, gain insight.

DATA DRIVEN RESEARCH

Traditional view of HPC

Sci-Vis



Schedulers (SLURM, PBS..)
Clusters (Beowulf, Hybrid), Cloud
IO channels, fabric, interconnect

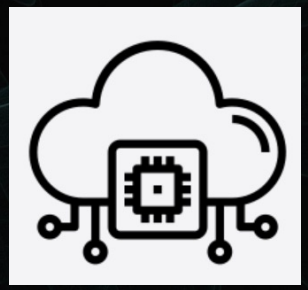
Scratch Storage

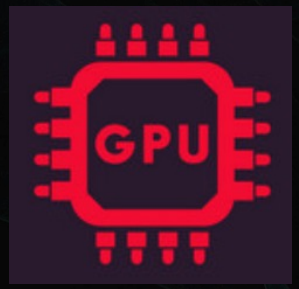
NAS: FLASH – HDD

HOT DATA

COLD DATA

Cloud/ Hybrid





Images courtesy of NASA

Where a Data Orchestration platform fits

HPC CLUSTER AND
“SCRATCH” STORAGE

VIRTUALIZED STORAGE –
SINGLE GLOBAL NAMESPACE

Images courtesy of NASA

Feed the Beast

- You spent a fortune on an HPC Cluster
- Almost as much on Network infrastructure
- Tons on storage
- Why?
- To refine computing models to ever-increasing fidelity, and develop sophisticated new ones, data staging is often imperative.
- Non-trivial - art/science called “Data Wrangling”

Tools for Data Wranglers:

- Storage Virtualization
- High-Speed Data Movers
- *Data-driven* workflows automate wrangling
- Large Scale Atomic Operations
- Data Mover provides EASY, secure access

Key points wrt the Data:

- Find the data (i.e. know where all the input datasets needed are)
- Move the data (Data needs to be staged on Scratch storage before the HPC scheduler can run the job). Afterwards the output jobs need to be moved off of Scratch and tagged so they can be found and associated with the input dataset.
- The ability to organize disparate data into organized Datasets or Collections.

Manage the data via metadata

Types of metadata

- "System Metadata"
  - File name, size, create, access, modify time, ownership permissions, etc.
  - In Unix/linux world this information comes from inodes and is often used by Backup or HSM software.
- Embedded File Metadata
  - Typically parsed out via MIME type.
- User Defined Metadata
  - This enables data life cycle management, notes, accounting.

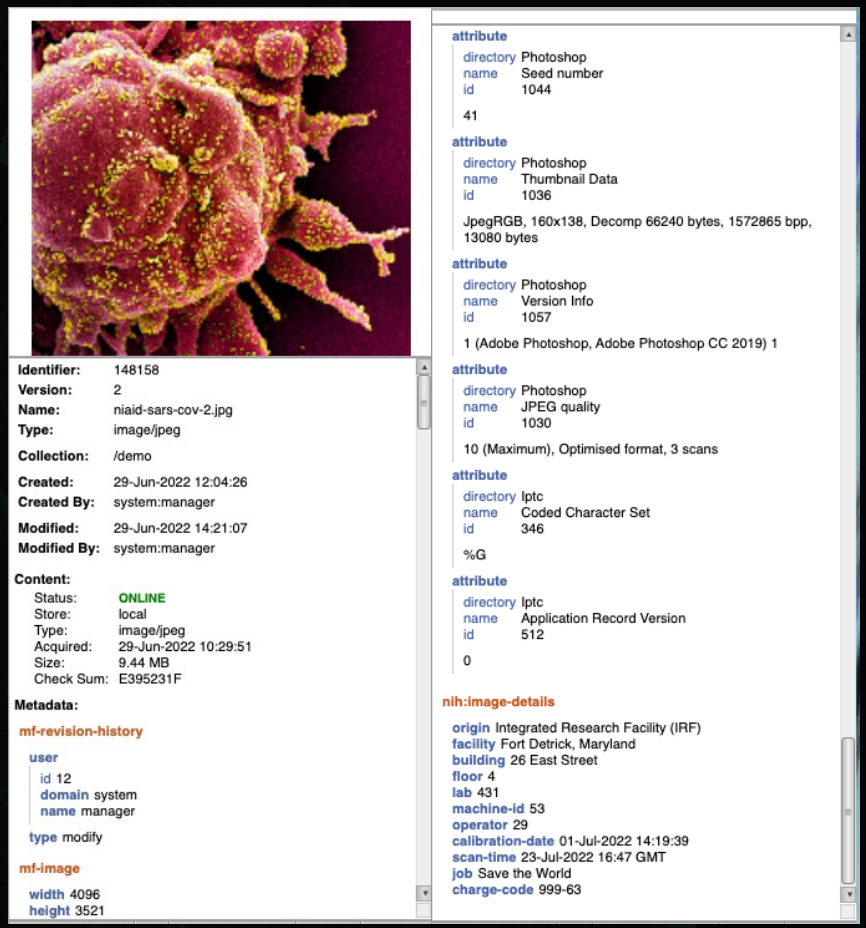


Image courtesy of NIH
National Institute of Allergy and Infectious Diseases (NIAID)

Move the Data very Quickly

- Most Complete Solution
- Best Software Architecture

HPC wire

- Arcitecta Recognized for Its Data Management Platform in Data Mover Challenge at SCA22


Mediaflux®
Livewire
• Highly Efficient
• Secure Collaboration
• Incredibly Fast
• Zero Error
• Low Overhead

Q: Does a system actually exist that can do all this?

A: Yes!

EVERYTHING IS A SERVICE

Mediaflux is an “operating system for data wrangling”

There are around 2000 built in services for every aspect of data management

Those can be extended with plugin services

Make calls from other systems to controlled APIs

Make calls to other systems

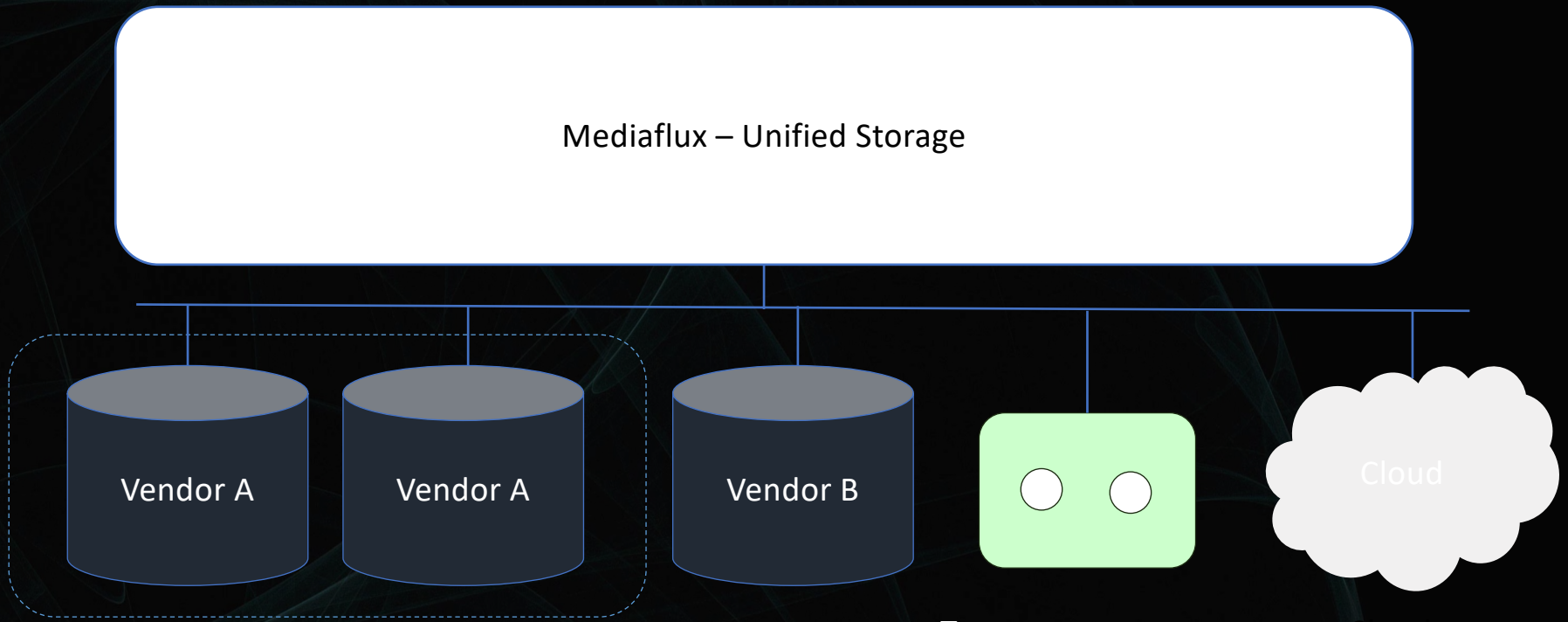
Integrate other storage, identity providers, sources, publication end-points, compute.

WITHIN A COHESIVE ENVIRONMENT

THAT CAN BE EXTENDED

INTERCONNECT DATA WORKFLOWS

Virtualized Storage ..



Disk

Tape:
- DMF
- Spectra Black Pearl

Cloud:
- S3
- Glacier
- Azure
- OpenStack
- …

Virtualized Storage ..

Grouping of storage into “pools” based on a policy:

- Least loaded
- Balanced
- Waterfall

Add / remove stores from pools at any time:

- Re-distribute data according to new policies

Virtualization allows:

- Heterogeneous storage to be grouped
- Transparently changing / evolving storage over time
- Elimination of storage silos!

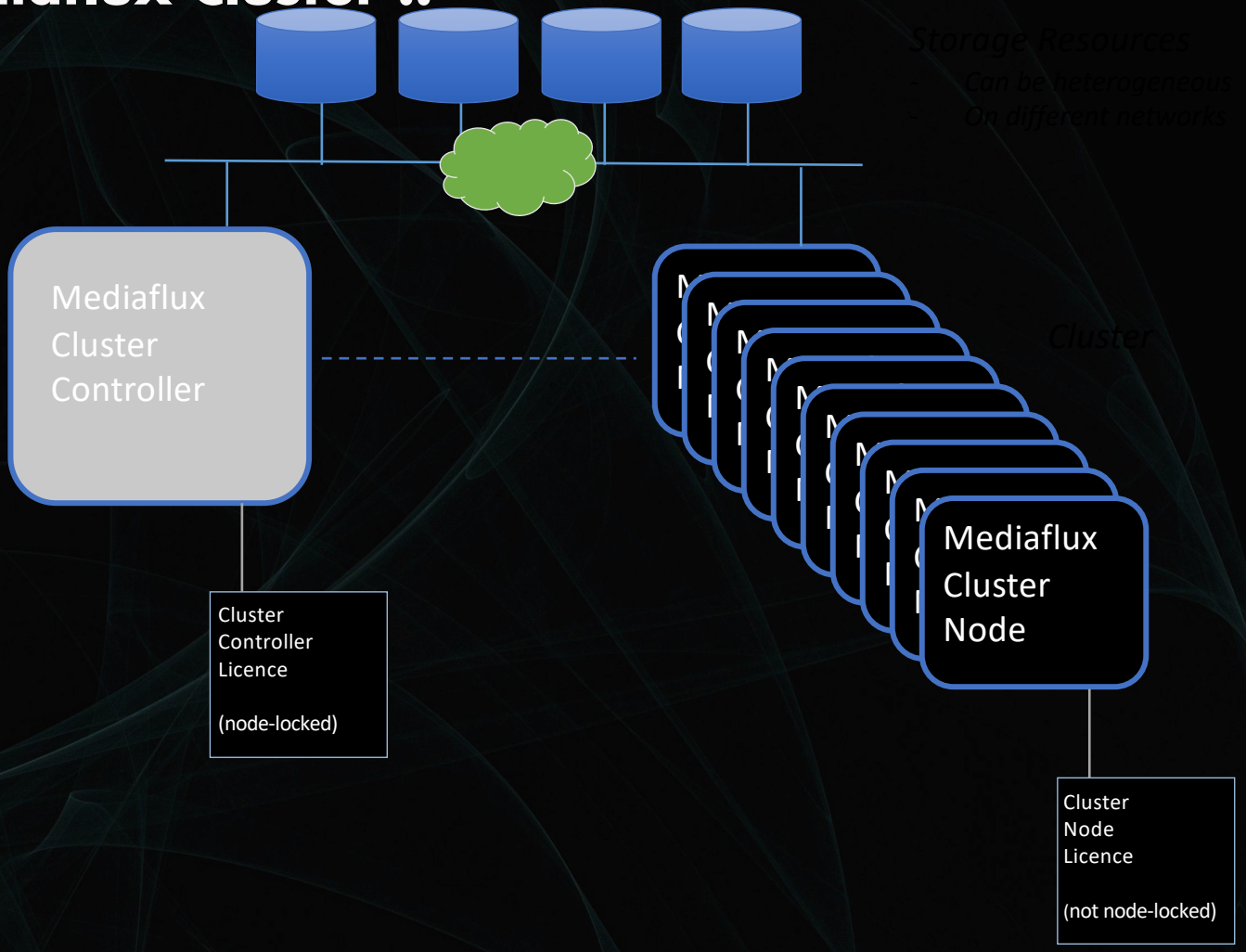
Clustered Mediaflux ..

Provides Virtualized Storage at Scale

Capable of moving 10's of TB per hour – we have a customer doing that (13 cluster nodes)

Using Clustered Mediaflux

Mediaflux Cluster ..



Cluster Nodes

Characterized by capabilities:

- Resources
- Operation (read, write, i/o, service)
- Max-units
- Weighting

Chosen using a policy:

- first
- least-active-by-op
- least-active-by-resource
- least-active-by-resource-and-op [the default]
- round-robin

Cluster Nodes

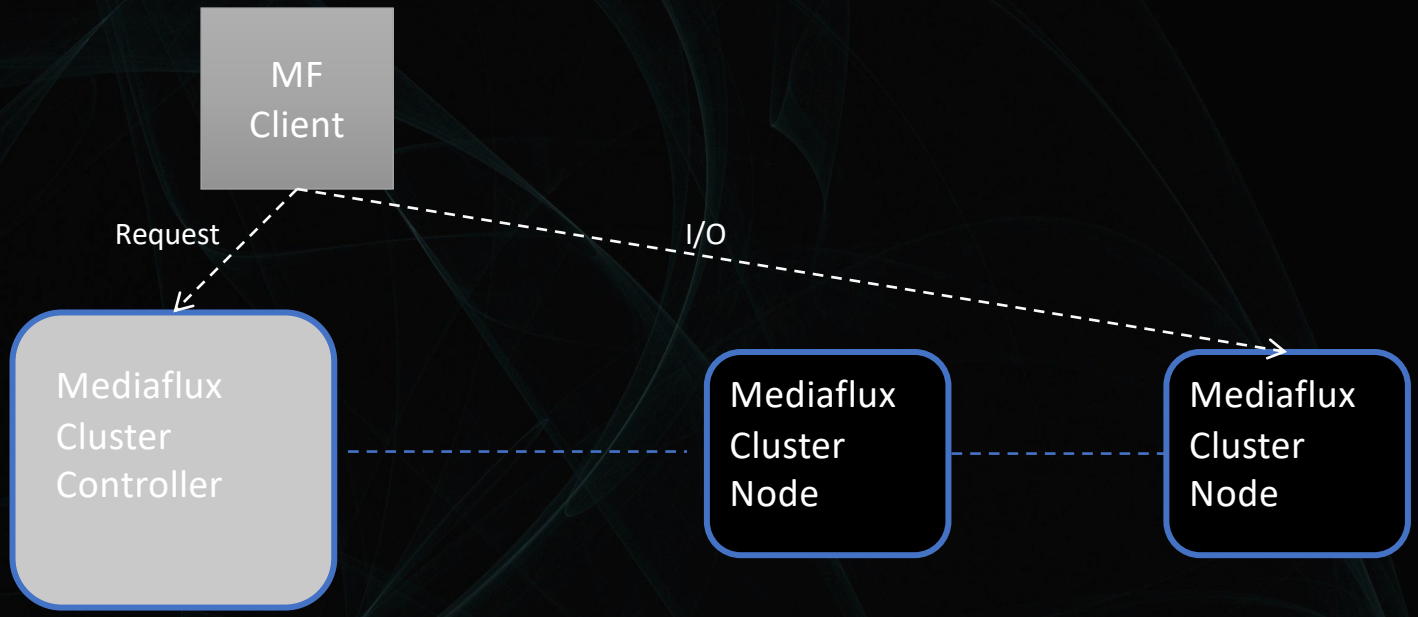
Operations:

- Read
- Write
- Copy
- Move
- Diff
- Generate checksums

Soon:

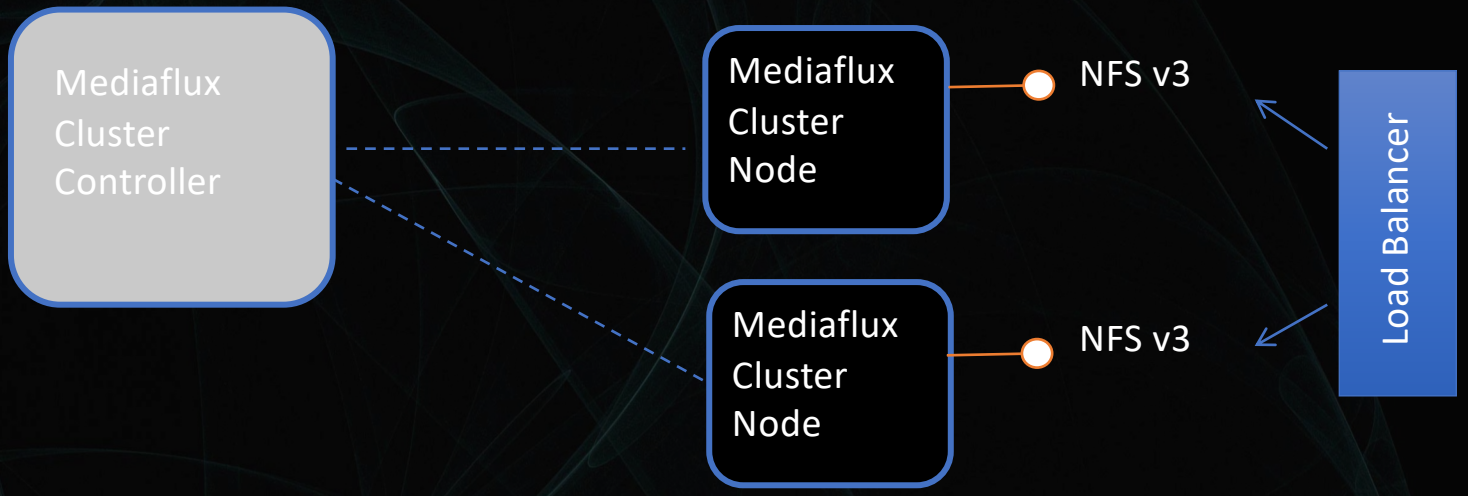
- Arbitrary computation (via plugin services)

Mediaflux Client Libraries are Cluster Aware



Combine with data store policies for storage virtualization

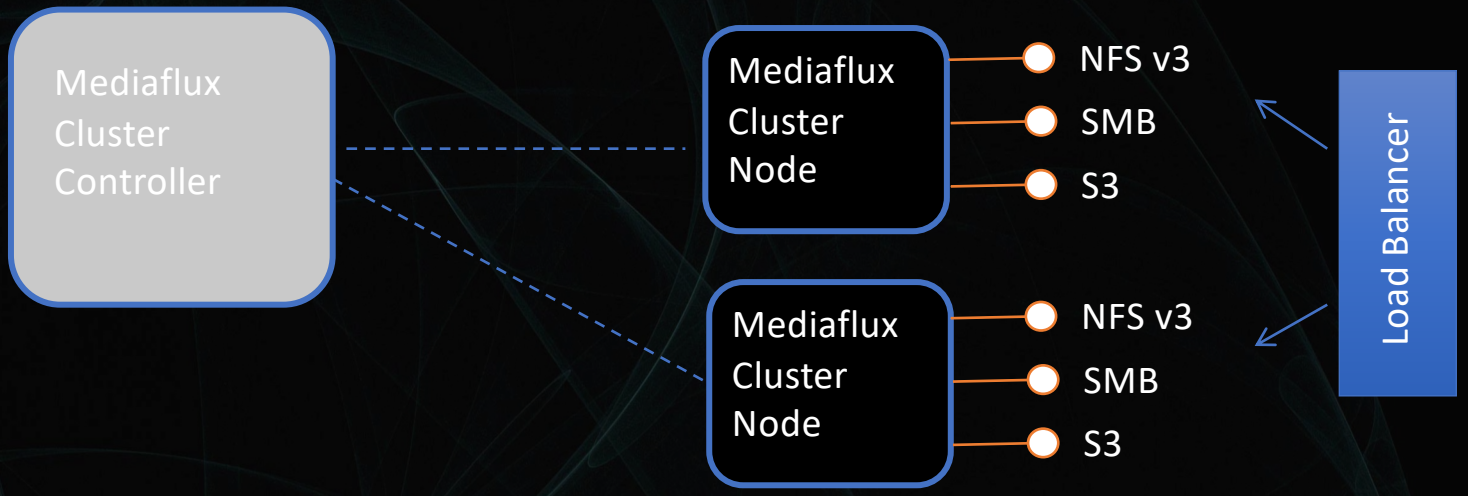
Cluster Nodes can be NFS v3 Servers



Load balance, or shard to each NFS server

Global file system space (via controller), I/O via cluster node

Cluster Nodes can also be SMB and S3 Servers

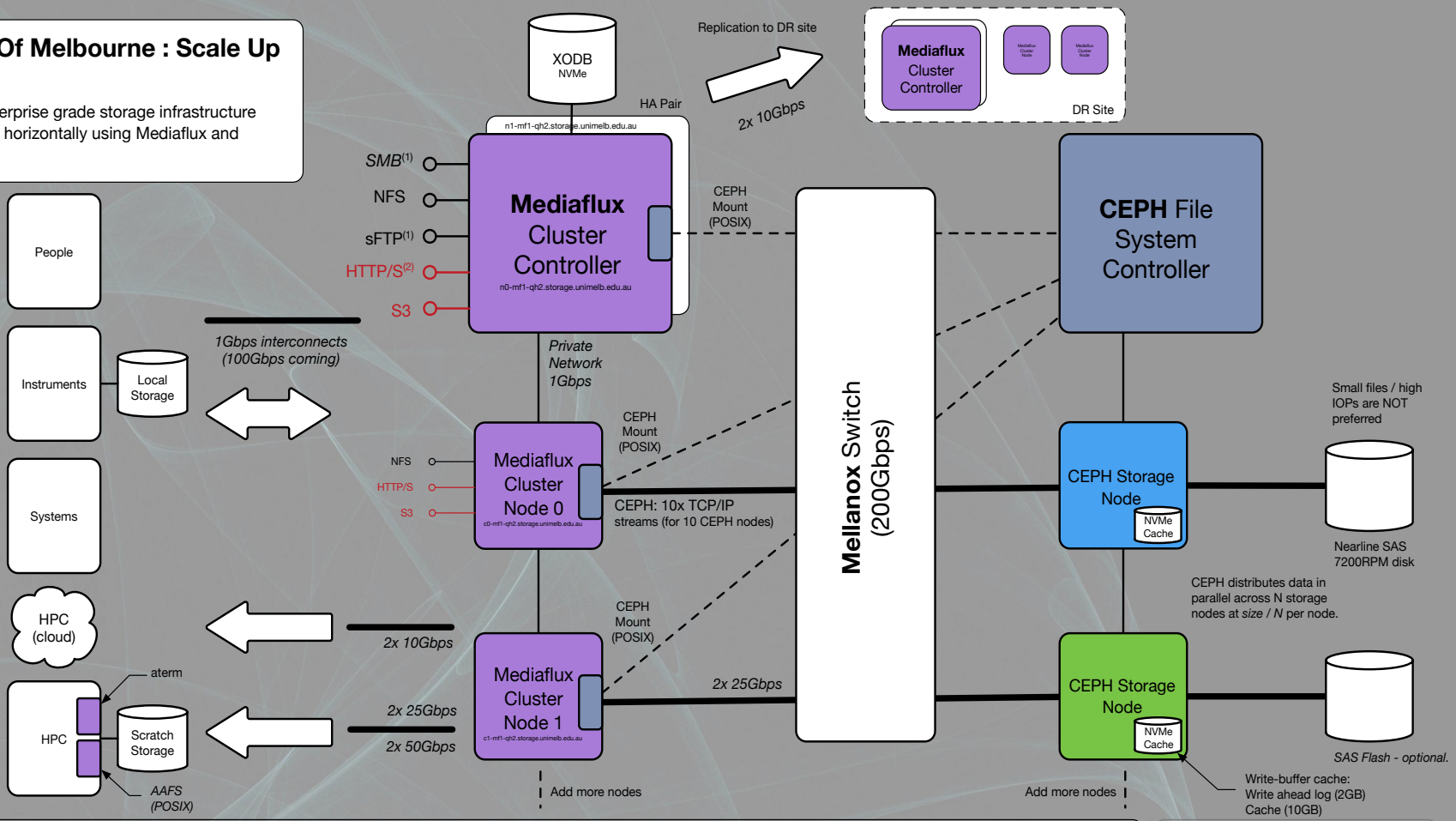


Load balance, or shard to each <protocol> server

Global file system space (via controller), I/O via cluster node

University Of Melbourne : Scale Up

Purpose:
To provide an enterprise grade storage infrastructure that will scale out horizontally using Mediaflux and CEPH.



How It Works:

1. CEPH: Will present 2x mounts: a) Flash and b) Nearline SAS. Policy above will drive usage.
2. Larger files are better - consider gene sequence example - need to bundle (at client, or server-side). If at server side, then need server side "scratch" area - that can be used for staging data and then creating a larger bundle which is written to the CEPH cluster.
3. CEPH: Large file limit: <= 4TB. Preferred max file size: 1TB (for bundling we would want to bundle to a much smaller number — say max 1 - 10GB, say 2GB — easily fits in NVMe 10GB cache)
4. CEPH: Object size = 4MB, stripe size=1MB, strip count=4
5. I-nodes: We could cluster / group files behind the scene to reduce i-node counts on the backend file systems. CEPH has the concept of active i-nodes (files in use - current limit 60 million) and inactive i-nodes.

Notes:

(1) - sFTP is not cluster aware. It is currently only supported by the controller. In future, it could be supported by the cluster nodes (requires work).

(2) HTTPS/S (using Mediaflux client) and S3 are cluster aware. HTTP/S for get (e.g. watching video can be readily made cluster aware).

Questions - for Scale-Up:

- **SLA**: maximum outage of 1/2 day?
- How many users? 100's - ~1000.
- How many instruments? 10's of (maybe 40)
- How many files? ~1 billion? files (starting at 10's millions). Maybe 100's millions.
- How much data? ~1PB (files or order of MBs per file)
- What scale per protocol? (percentages)

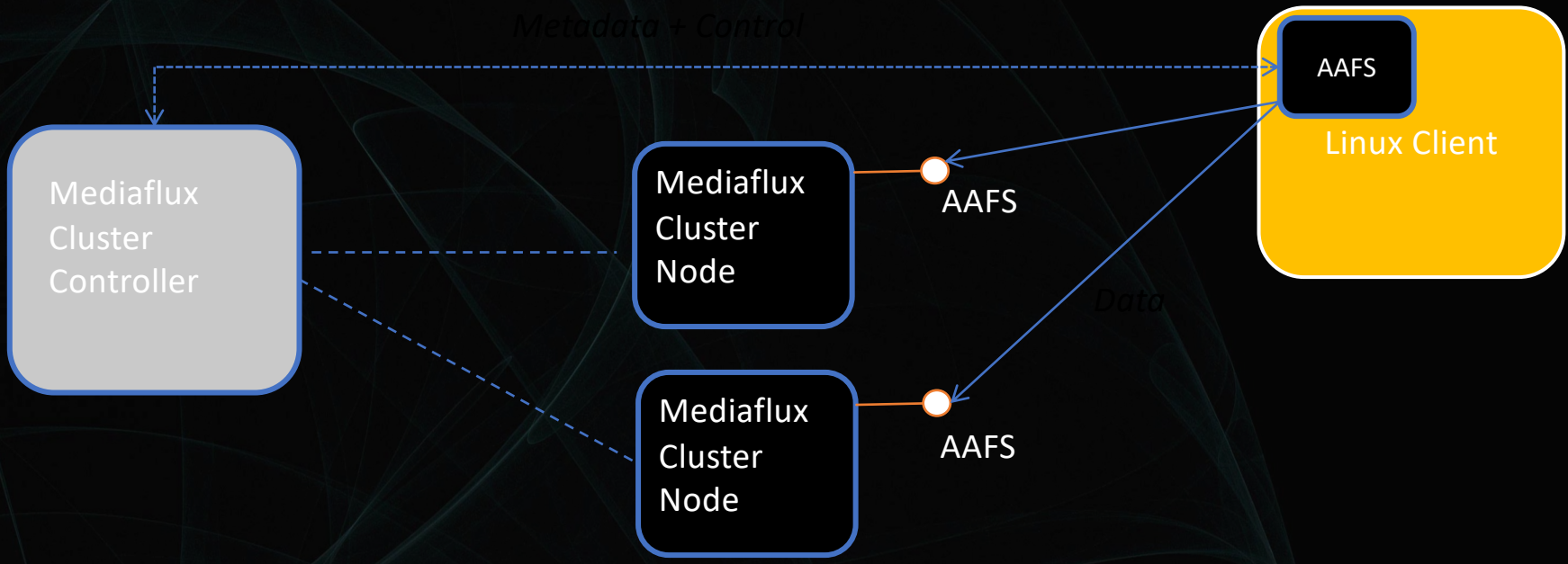
Hosts:

n0-mf1-qh2.storage.unimelb.edu.au [primary]
n1-mf1-qh2.storage.unimelb.edu.au [failover]
c0-mf1-qh2.storage.unimelb.edu.au [cluster node 0]
c1-mf1-qh2.storage.unimelb.edu.au [cluster node 1]

ARCITECTA®

June-2018

Arcitecta File System (Coming...)



- Mediaflux AAFS File-System client installed on Linux host
- Mounted as a POSIX file system
- Semantically aware (no attribute timeouts)

Processing Queues ..

Moving Data From A to B is not always simple …

- Move as much in parallel as possible
- Don't flood the networks
- Deal with "chaotic" tape / HSM systems
- Allow prioritization (and re-prioritization)

That's why we have processing queues:

- Allow concurrent resource operations, based on system characteristics
- Process data when it is available (e.g. on recall from HSM)
- Prioritization
- Transient or persistent
- Cluster aware

Cluster is Fault Tolerant

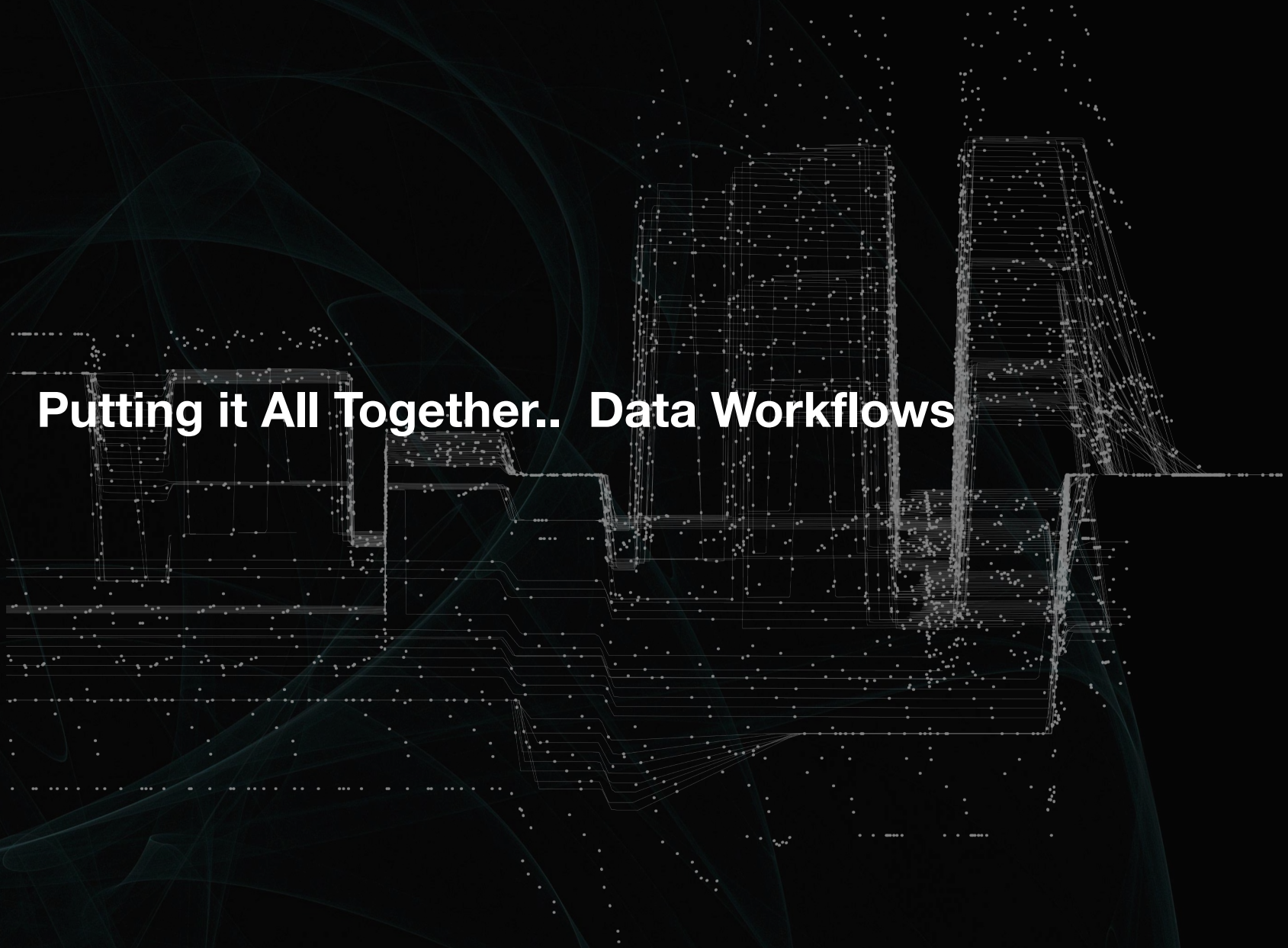
Failure by any node will be redirected to another node

Cluster nodes can be moved from one compute resource to another (not node locked)

Processing Queues ..

Are also used for scheduling arbitrary service execution (e.g. compute operations) …

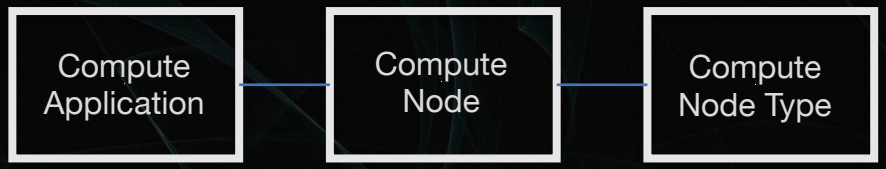
Putting it All Together.. Data Workflows



Compute Services Framework

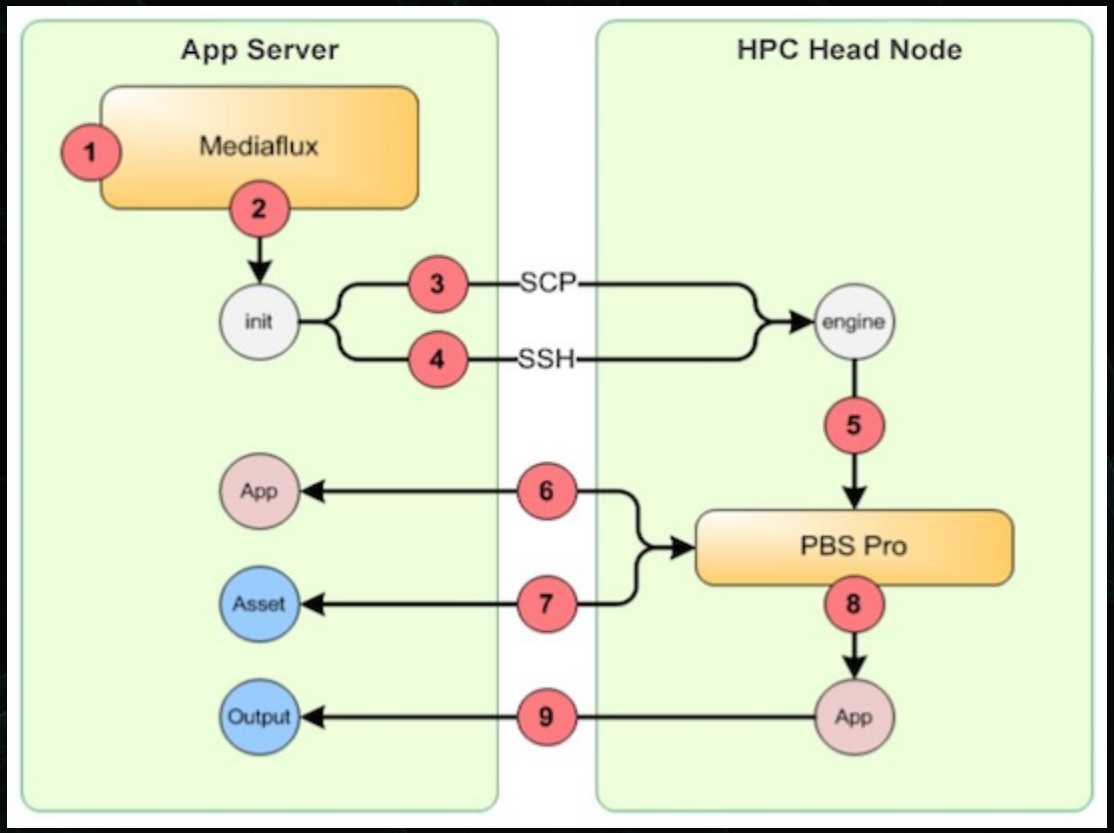
- Many use cases require external computation using applications and/or resources external to Mediaflux.
- Examples include HPC applications, workflow and processing pipelines – PBS Pro, SLURM,…
- Mediaflux Compute Services provides a framework for configuring, deploying and executing *compute applications*
- Once configured, a compute service may be utilized by any authorized application and/or user interface by calling a *single service within Mediaflux*.

Compute Services – 3 Key Components

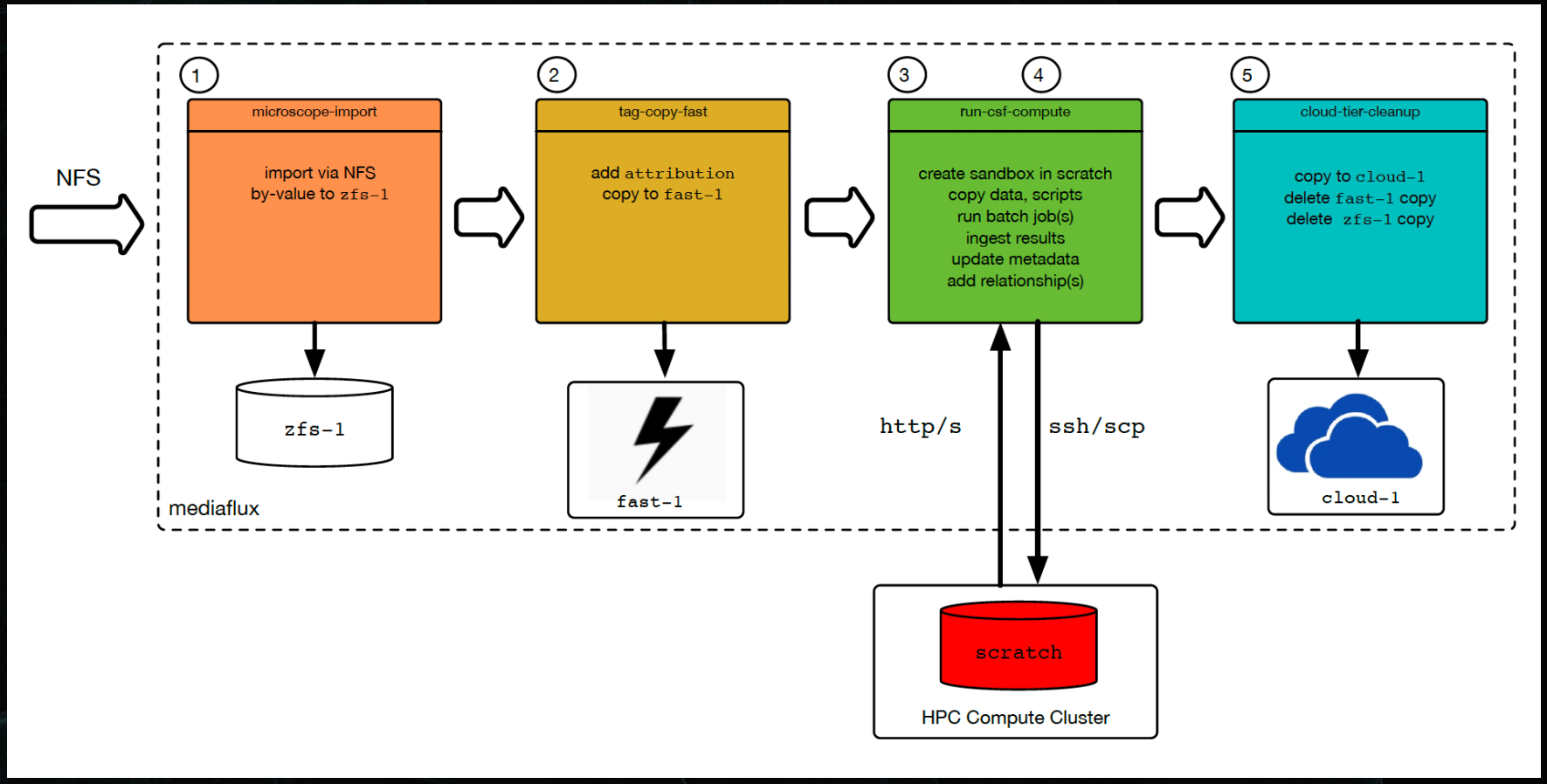


- *Compute Application:* represents an executable application. Encapsulated Details: MIME types, init scripts, etc.
- *Compute Node:*
  Represents host on which the Compute Application will run. EXAMPLE: head nodes in an HPC environment
- *Compute Node Type:* Encapsulates details of compute nodes and enables developers/administrators to create new types

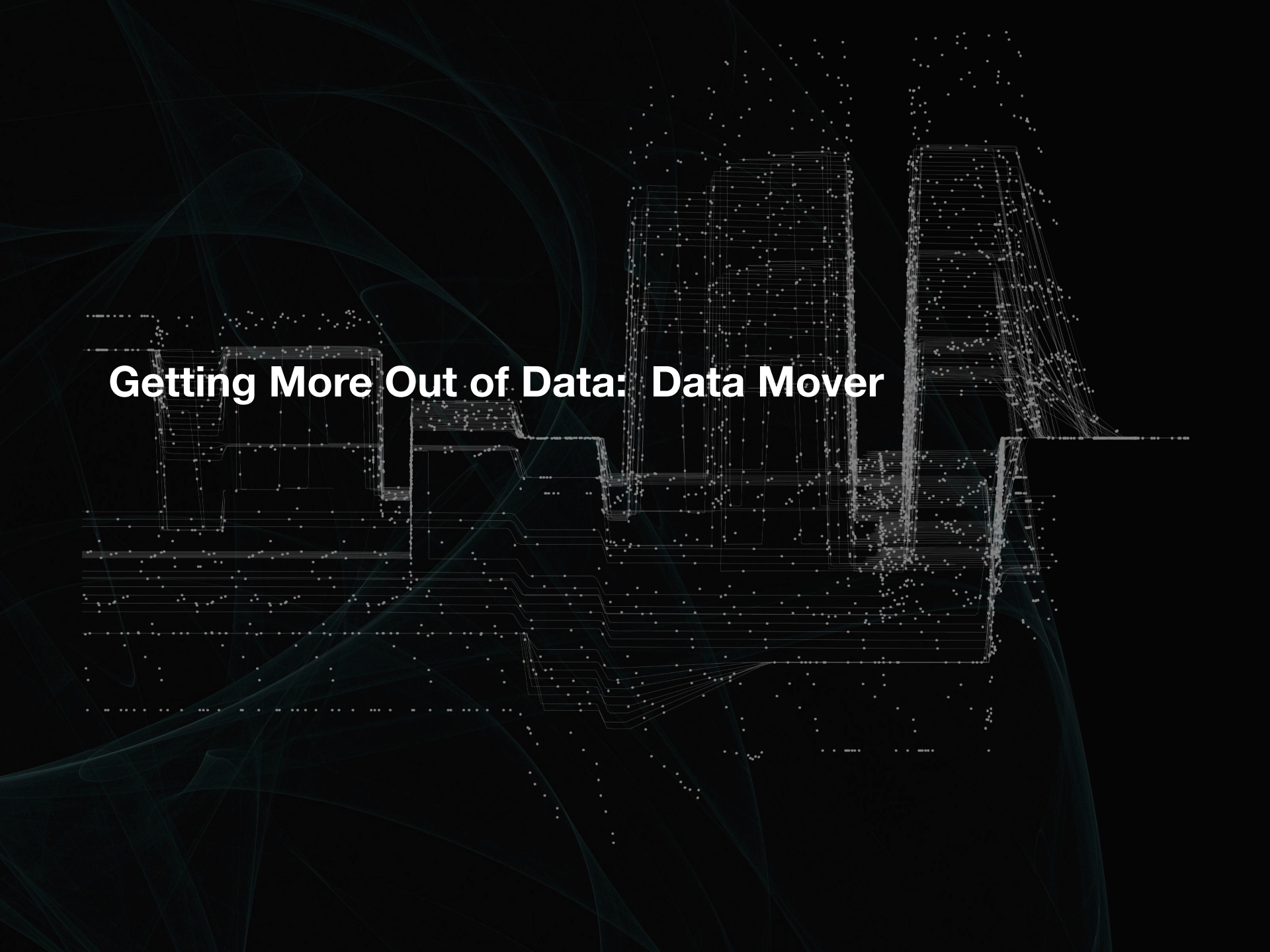
Example Compute Services Workflow



Putting it all Together…
Data-Driven HPC Workflow



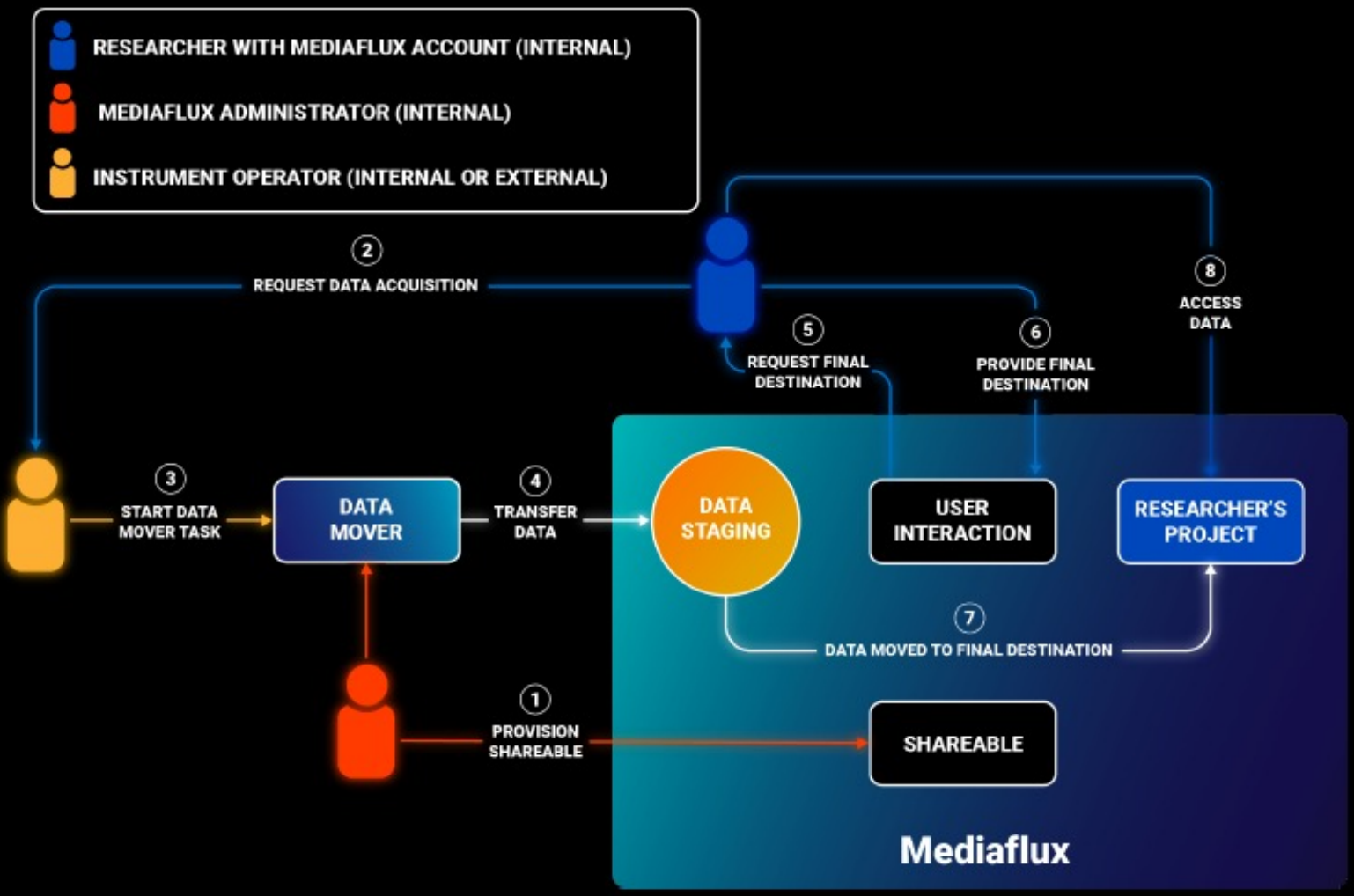
Getting More Out of Data: Data Mover



Data Mover

- Inconspicuous – but there when you need it!
- Support for Windows, Linux and Mac
- Enables “kiosk” operation *and* ad-hoc data movement for a single user
- Intuitive and simple – easy for both technical and non-technical staff
- Secure, fast and reliable transmission for any size datasets
- Ability to add *arbitrary* additional metadata on upload
- Upload/Download histories – know who downloaded/uploaded what and when
- Easily Pause/Resume
- Built-in integrity checking ensures *reliable* data transmission
- Move data *without copying*

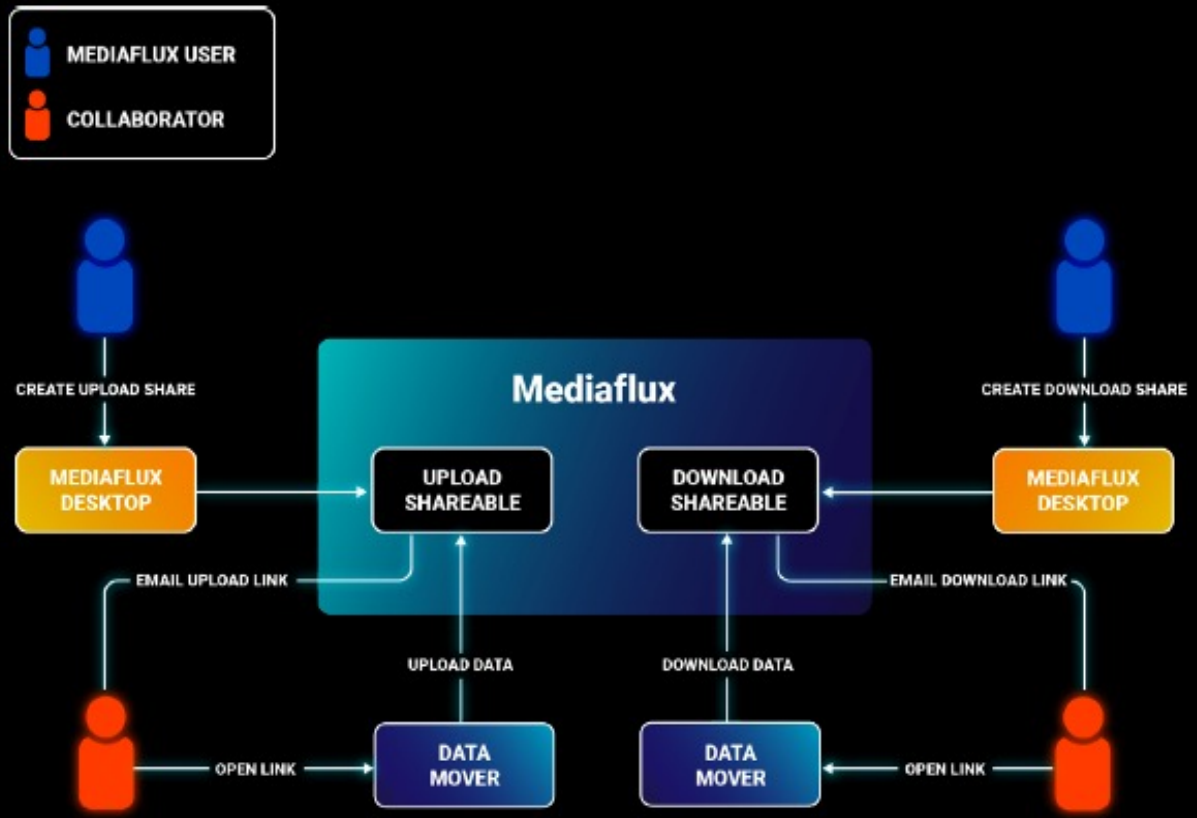
Instrument Integration



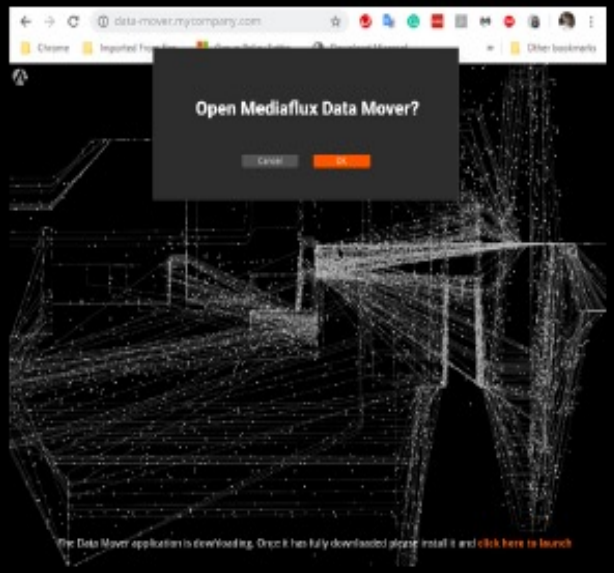
Upload & Download

A shareble has a link that can be emailed to someone else and it will allow that person to either:

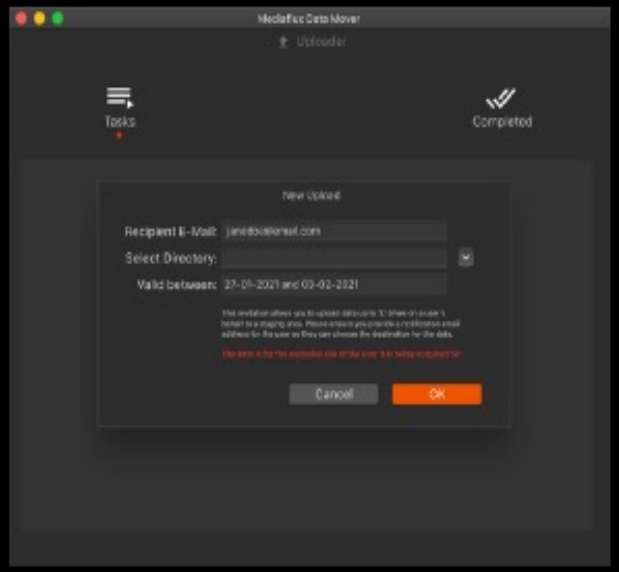
1. Download the shared content to their computer using the Data Mover, or
2. Upload data to the user's Mediaflux space with the Data Mover.



Link Workflow



1.
A shareable data link is sent that directs the user to a page that determines whether the Data Mover application is installed, and if not then directs them to a distribution site to download and install it.



2.
Once the application is installed, the user can launch the upload invitation attached to the smart link, which includes a token which provides access to the location on Mediaflux that content can be downloaded from or uploaded to.

3.
The progress of data being transferred can be seen in the Data Mover application. Any errors encountered during transfer will be logged for each task, making it easy for end-users to check errors associated with a specific upload.