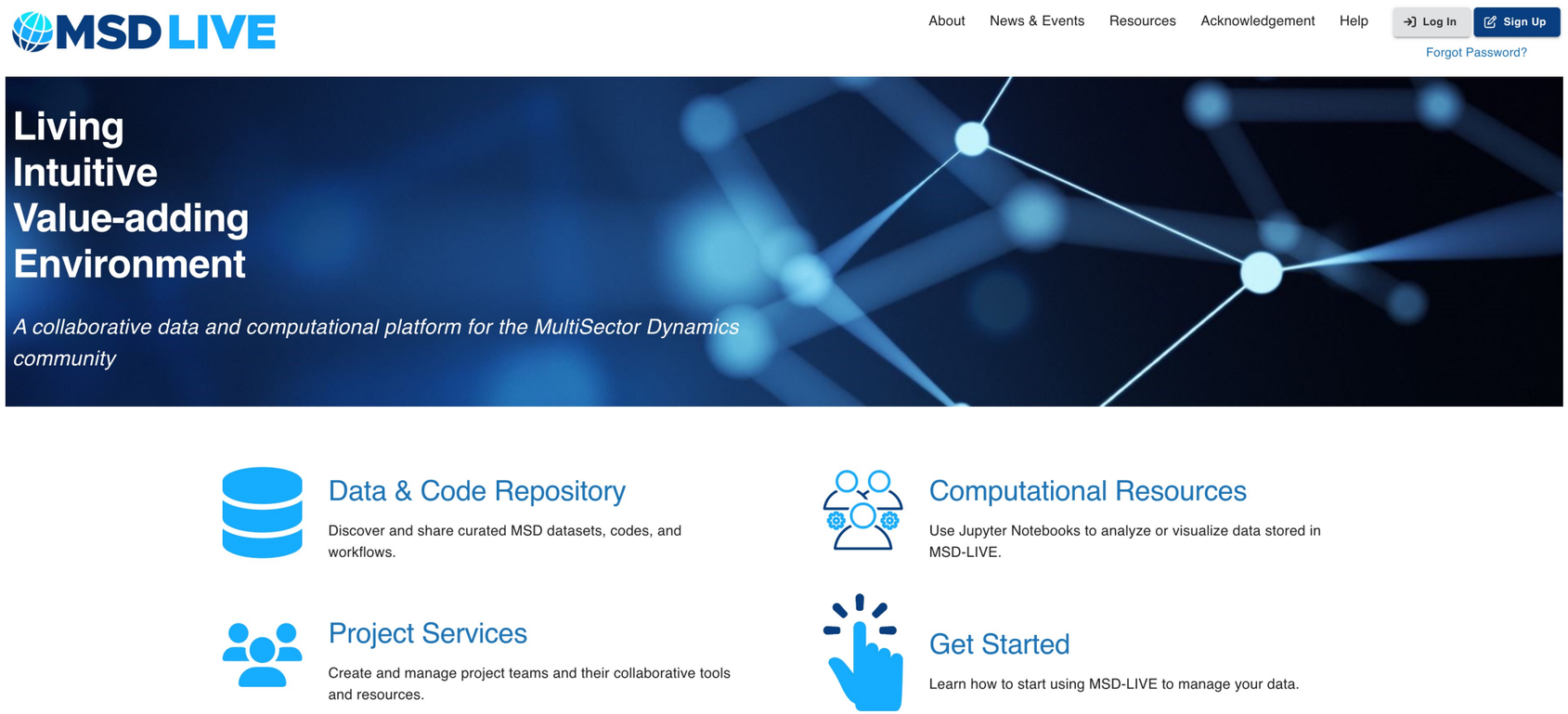
# FY23 Updates to the MSD-LIVE Data and Computational Platform

Casey Burleyson, Carina Lansing, Zoë Guillen, Matthew Macduff, Devin McAllester, and Jon Weers

PNNL is operated by Battelle for the U.S. Department of Energy

# The Vision for MSD-LIVE



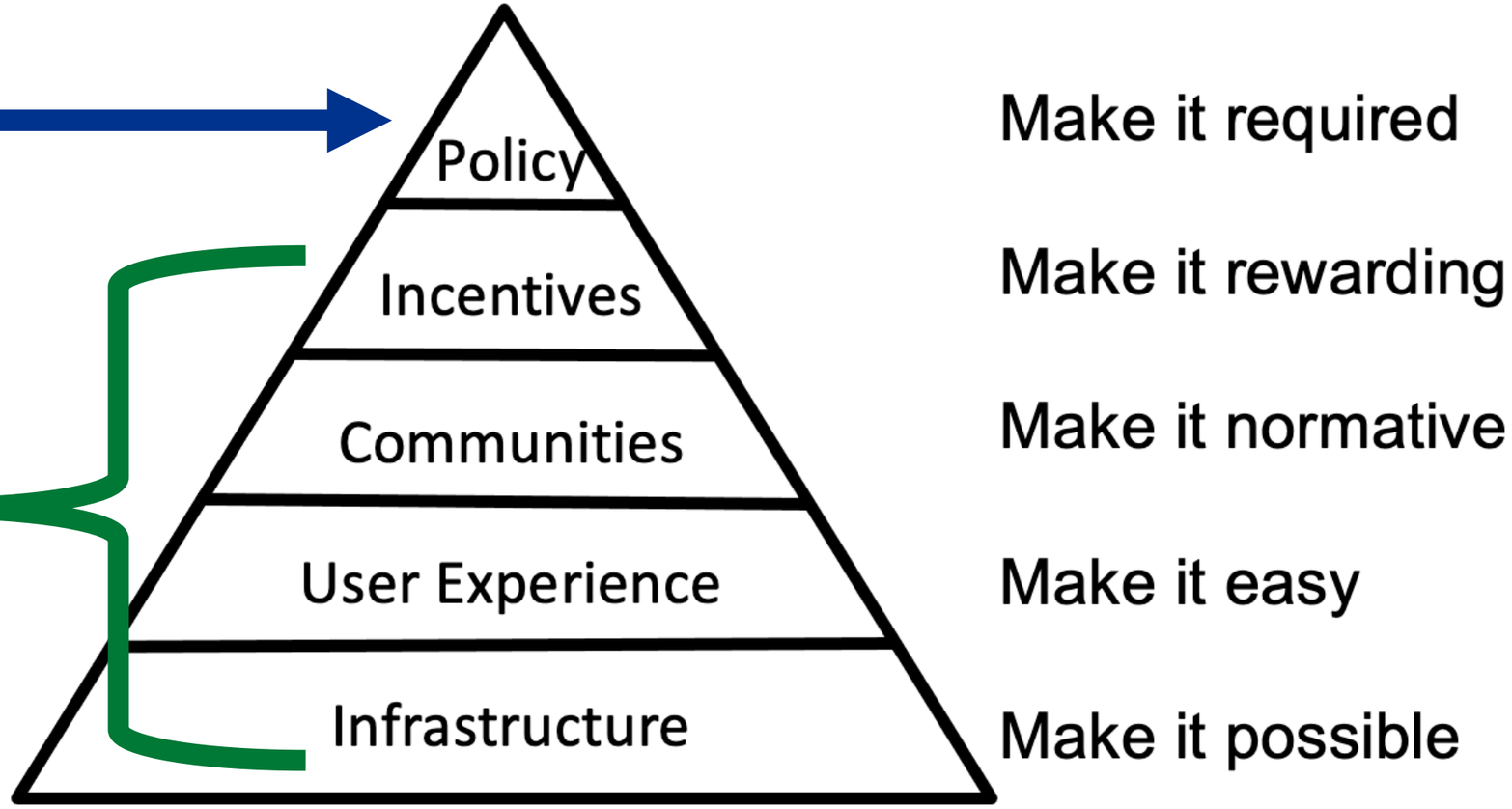

**Visit us at https://msdlive.org**

- A cloud-based data management system and advanced computing platform
- Will enable researchers to document and archive their data, run their models and analysis tools, and share data, software, and multi-model workflows
- A cornerstone capability of the MSD Community of Practice

# Facilitating Open MSD Science

Journals largely skipped to this end of the pyramid…

MSD-LIVE is about tackling these foundational elements of the pyramid…




*Conceptual diagram from Brian Nosek of the University of Virginia and the Center for Open Science*

# Top-10 Most Pressing Use Cases

| Use Case | Name | Description |
|---|---|---|
| 1.1 | Find Data | Quickly and easily find datasets produced by other users and projects |
| 1.2 | Archive Data | Permanently archive small (<250 MB), medium (250 MB - 50 GB), and large (50 GB to 20 TB) final-form datasets and generate data Digital Object Identifiers (DOIs) in order to meet journal requirements for data sharing |
| 1.3 | Version Control | Use an intuitive web-based user interface to document and share versioned datasets and associate data with the code used to produce it |
| 1.4 | Training | Train new team members on MSD projects to effectively manage data and code and capture the institutional knowledge of members that leave a project |
| 1.5 | Control Access | Create and manage teams that cross institutions in order to quickly and easily grant access to data and code without having to obtain multiple sets of institutional credentials |
| 2.1 | Share Data | Share working datasets across multiple institutions collaborating on a project in real-time |
| 2.2 | Analyze Data | Create, run, and share Jupyter notebooks to analyze or visualize datasets in MSD-LIVE |
| 2.3 | Multi-Model Workflows | Create, execute, document, and publish multi-model workflows where component models run on computational resources at different institutions |
| 2.4 | ML/AI | Virtually assemble data lakes and execute ML/AI algorithms on large pools of data that may be physically located in different places |
| 2.5 | Move Code to the Data | Easily deploy models to run on new computational resources and make it easy for users to bring their code to the data when transferring data is infeasible |

# Milestones and Capabilities

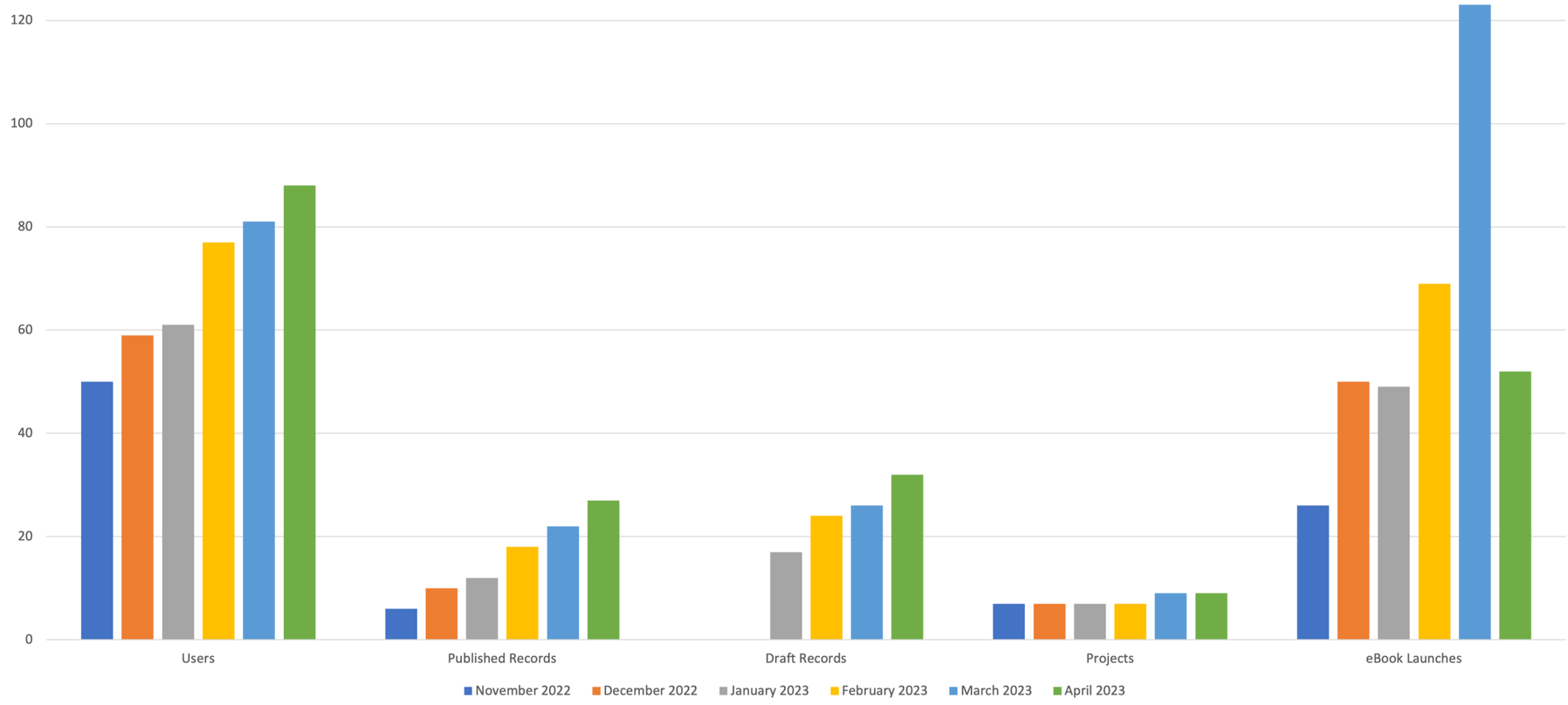
| Core Capabilities | Use Cases |
|---|---|
| Data Repository | Find Data |
| | Archive Data |
| | Version Control |
| | Training |
| | Control Access |
| Advanced Computing | Share Data |
| | Analyze Data |
| | Multi-Model Workflows |
| | ML/AI |
| | Move Code to the Data |

= in v1 release  = in development

- Feb-22: Beta version of the data repository released for testing by our stakeholder group.
- Jun-22: The "*Addressing Uncertainty in MultiSector Dynamics Research*" eBook was migrated to MSD-LIVE. The eBook contains on-demand, interactive tutorials for common UC/UQ workflows.
- Aug-22: v1 of MSD-LIVE released to the MSD research community.
- Jun-23: v2 release includes better ways to move large data in and out of MSD-LIVE.
- Jun-23: Used to facilitate training sessions during the GCAM annual meeting. Participants can run the hector, stitches, and xanthos models in real-time on the AWS cloud.

# Usage Statistics

- 9 registered projects
- 88 registered users
- 27 published datasets and 32 open draft datasets
- 186+ Tb of total data
- 52 usages of the MSD UC eBook Jupyter notebooks in the last 30 days



# v2 Released Today!

- In v2: Data repository updates
  - Dedicated storage buckets by project
  - Ability to upload nested folder hierarchies instead of zip files
  - Upload via the command-line in addition to the web portal

---

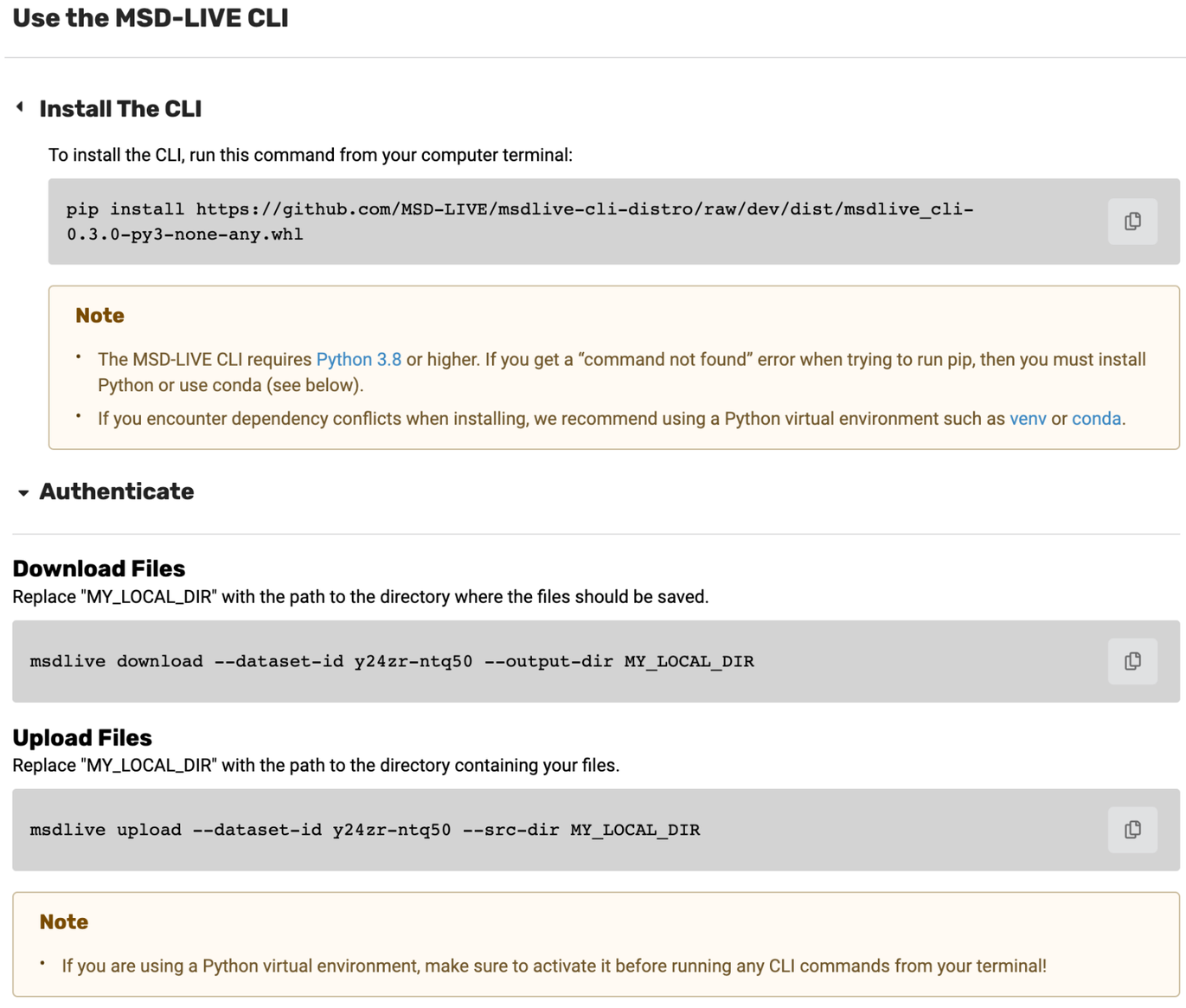
- Coming in v3: Compute services
  - Ability to create a Jupyter notebook to analyze or visualize data
  - Ability to manage collection of Jupyter notebooks (e.g., GitHub integration)
  - Allow notebooks to be directly connected to MSD-LIVE datasets
  - Automated support for training and demonstrations using Jupyter notebooks

- Coming in v3: Deployment environment
  - Protocols in place for tracking cost and usage by project

Green = Completed; Orange = In Progress; Red = Not Started

# New Command-Line Interface

- Python package that installs directly from pip
- Simple one-line commands for uploading and downloading large datasets
- Utilizes AWS-native file transfer protocols on the backend
- Fast, easy to use, and allows for easy recovery from interrupted transfers



## Use the MSD-LIVE CLI

### Install The CLI

To install the CLI, run this command from your computer terminal:

```
pip install https://github.com/MSD-LIVE/msdlive-cli-distro/raw/dev/dist/msdlive_cli-0.3.0-py3-none-any.whl
```

**Note**

- The MSD-LIVE CLI requires Python 3.8 or higher. If you get a "command not found" error when trying to run pip, then you must install Python or use conda (see below).
- If you encounter dependency conflicts when installing, we recommend using a Python virtual environment such as venv or conda.

### Authenticate

### Download Files

Replace "MY_LOCAL_DIR" with the path to the directory where the files should be saved.

```
msdlive download --dataset-id y24zr-ntq50 --output-dir MY_LOCAL_DIR
```

### Upload Files

Replace "MY_LOCAL_DIR" with the path to the directory containing your files.

```
msdlive upload --dataset-id y24zr-ntq50 --src-dir MY_LOCAL_DIR
```

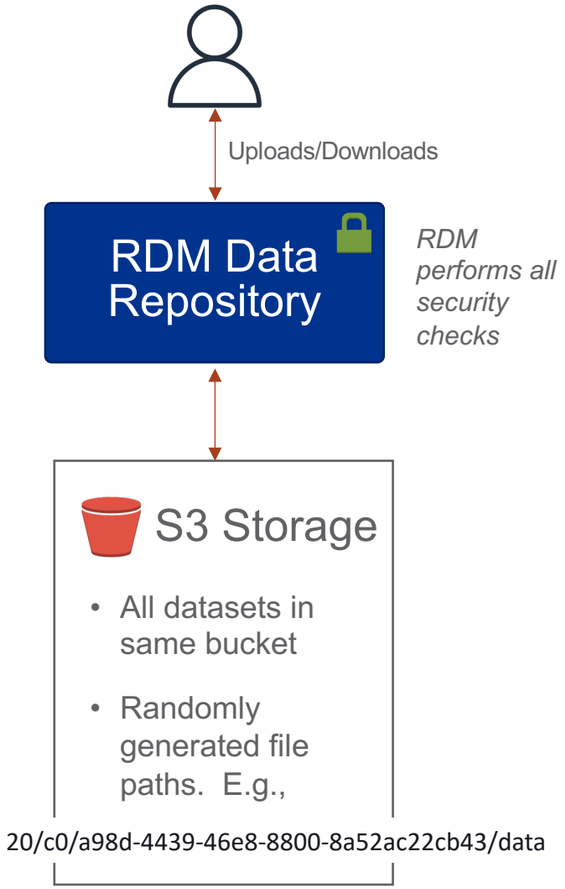
**Note**

- If you are using a Python virtual environment, make sure to activate it before running any CLI commands from your terminal!

# Changes to MSD-LIVE File Storage

## Before

- Uploads go through RDM = 2 hops!
- Network interruptions not recoverable
- Folders not supported
- Not accessible outside RDM
- Unreadable file names




## After

- Uploads direct to S3
- Uploads can recover from network interruptions
- Folder hierarchy and file names preserved
- Directly accessible by users and other AWS compute resources