DataFlow - A Data Bridge to Air-Gapped Instruments

Greg Shutt\textsuperscript{1}, Marshall McDonnell\textsuperscript{2} and Suhas Somnath\textsuperscript{3}

\textsuperscript{1}Software Services Development Group, National Center for Computational Sciences
\textsuperscript{2}Applications Engineering Group, Computer Science and Mathematics Division
\textsuperscript{3}Data Lifecycle and Scalable Workflows Group, National Center for Computational Sciences

ORNL is managed by UT-Battelle LLC for the US Department of Energy

DOE Data Days
June 1, 2022
Challenges in Everyday Research

• Observational instruments produce large datasets
• Instruments are often off-network
• Large datasets need strong compute resources for analysis
Today’s Solution

- USB drives for data movement
- Personal laptops for data storage and analysis
- Unnecessarily long experiment iterations
- Researchers need simple data movement and management
DataFlow

**Two primary components:**

1. Web application software
2. Hardware bridge
DataFlow

Observational Science Facility

Instrument  Control Computer

Instrument  Control Computer

Data and Computational Facility

DataFlow server

Storage

Cloud

Compute
DataFlow

- No upper limit on file sizes
- User authentication provided by Globus
- Instrument and storage can be in different facilities
DataFlow

Two methods of interaction

1. Web-based GUI
2. API with Python library
DataFlow Demo

https://dataflow.ornl.gov

Welcome to DataFlow

DataFlow is a tool that enables scientists to capture scientific metadata and transmit data via a simple web interface to a designated, centralized data storage resource (CADES Home by default).

Once data is at a centralized location, users can move data to other resources; use a scientific data management system to share, manage, and organize data; use connected cloud and high performance computing resources for large scale data analytics; publish data and many more things. DataFlow can also be deployed in proximity to scientific instruments if the instrument control computers are not connected to the internet and such data transmission capability is desired at these computers (without connecting the computers to the broader network) and/or when consistent / high data transfer rates are desired.

Initial DataFlow screen
DataFlow Demo

Login via Globus
DataFlow Demo

Recently uploaded datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>My New Dataset</td>
<td>May 20, 2022 at 9:26 am</td>
</tr>
<tr>
<td>Test</td>
<td>May 18, 2022 at 10:33 am</td>
</tr>
<tr>
<td>Test</td>
<td>May 18, 2022 at 9:52 am</td>
</tr>
<tr>
<td>Dataflow</td>
<td>Nov 2, 2021 at 1:35 pm</td>
</tr>
<tr>
<td>Test2</td>
<td>Nov 2, 2021 at 12:59 pm</td>
</tr>
<tr>
<td>Test</td>
<td>Nov 2, 2021 at 11:28 am</td>
</tr>
</tbody>
</table>
DataFlow Demo

Creating a new dataset
DataFlow Demo

Newly-created dataset, ready for upload
DataFlow Demo

File successfully uploaded
API Benefits

- Enable automated data upload for long-running experiments
- Monitor and automatically upload all new data
from ordflow import API

# create an example dataset
response = API(api_key).dataset_create(
    "My new dataset with nested metadata",
    metadata={"Sample": "PZT",
               "Microscope": {
                               "Vendor": "Asylum Research",
                               "Model": "MFP3D"
                           },
               "Temperature": 373
    }
)
DataFlow API Demo

```python
4 from ordflow import API
5 # upload a file to the dataset
6 response = API(api_key).file_upload("/Users/g5s/Downloads/olivia.jpeg", dataset_id)
```

Uploading a file
DataFlow API Demo

Dataset and file created
Future Work

• Additional adapters: S3, DropBox, OneDrive
• Adding hooks to integrate into workflow systems
Contact

For more information, please contact:

Greg Shutt shuttgl@ornl.gov

Suhas Somnath somnaths@ornl.gov

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.