Data Workspaces

At a Glance

DWS: A framework for managing experiments, enabling collaboration and reproducibility

Availability: See https://dataworkspaces.ai

Contact: Rupak Majumdar rupak@mpi-sws.org and Jeff Fischer jfischer@mpi-sws.org at the Max Planck Institute for Software Systems

Modern scientific workflows can be very complex, involving many data sources, software components, and partial results. At the same time, many scientific workflows are not automated and incur significant manual effort or depend on brittle, one-time, scripts. As a result, scientists and data professionals have issues with managing experiments, collaboration, and reproducibility.

Data Workspaces (DWS) is an open source framework for managing scientific data and automating experiment workflows. Data Workspaces maintains the state of a science project, including data sets, intermediate data, results, and software. It supports reproducibility through snapshotting and lineage tracking and collaboration through a push/pull model inspired by version control systems for code.

The goal is to provide the reproducibility and collaboration benefits with minimal changes to your current projects and processes.

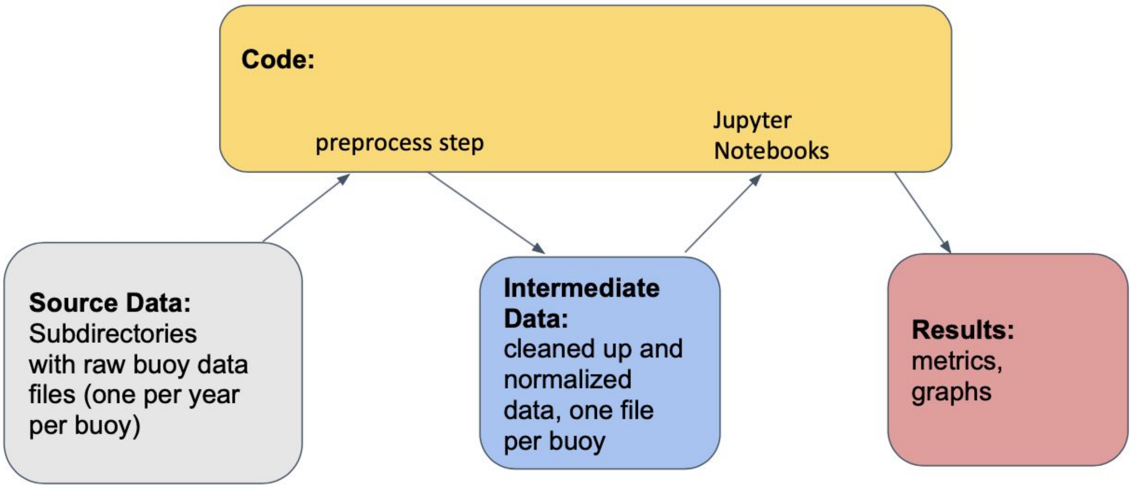
Capabilities: Data Workspaces lets you:

1. Track and version all the different resources for your science project from one place.
2. Automatically track the full history of your experimental results and generate relevant reports summarizing the results.
3. Reproduce any prior experiment, including the source data, code, and configuration parameters used.
4. Go back to a prior experiment as a “branching-off” point to explore additional permutations.
5. Collaborate with others on the same project, sharing data, code, and results.
6. Easily reproduce your environment on a new machine to parallelize work.
7. Publish your environment for others to download and explore.

Example Scenario:

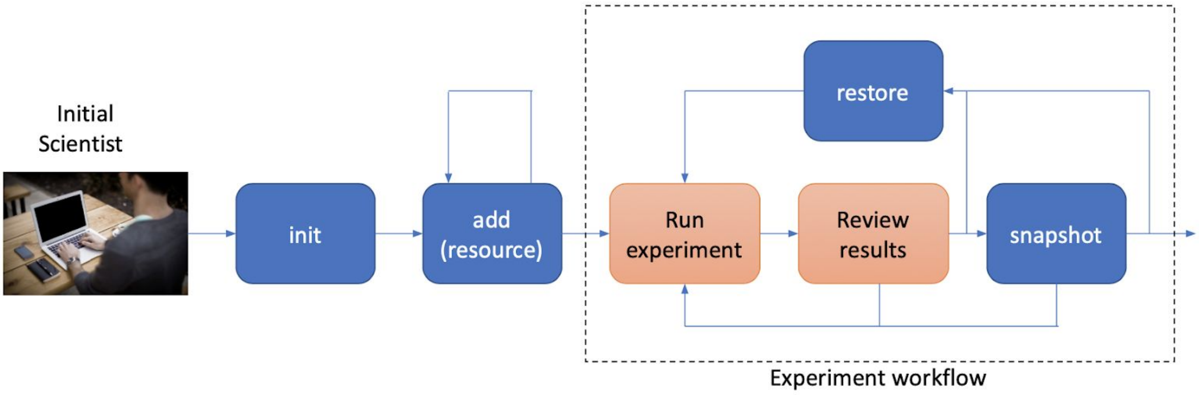
To get a sense of how DWS can benefit a project, we will use an example data analysis of historical temperature data from ocean buoys. This data has been captured for almost 50 years and is available online in text file format from various government and research organizations (e.g. the National Buoy Data Center). A Data Workspace for the analysis of this data uses the following resources:

- Source Data: The original data files collected from the National Buoy Data Center
- Code: Software to preprocess data and to run analytics
- Intermediate Data: Space to store intermediate results of the analyses
- Results: The results of the analyses


Code:
preprocess step
Jupyter Notebooks
Source Data: Subdirectories with raw buoy data files (one per year per buoy)
Intermediate Data: cleaned up and normalized data, one file per buoy
Results: metrics, graphs

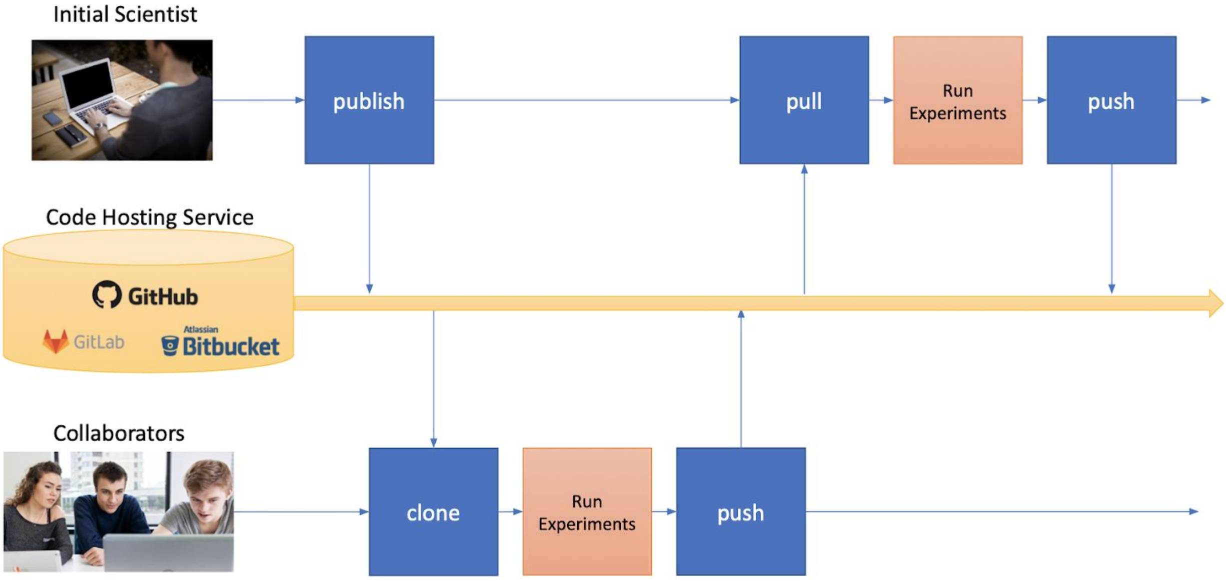
Initial Workflow

Under Data Workspaces, the workflow for this project would involve the steps shown below. We first create an initial (empty) workspace, and add the four kinds of resources to the workspace. Note that resources such as source data and code may reside in local file systems or in their own local or remote repositories. For example, data can reside in a database, an NFS server, or an Amazon S3 bucket and code can reside in its own version control repositories. Once the resources are added to the workspace, they are transparently managed by DWS. Thus, the scientist can run one or more experimental workflows on the data, and can take snapshots to precisely track the data, code, and parameters used to obtain a certain result. Later, the state of the system (or individual resources) can be restored to the settings of a previous snapshot. DWS can generate reports of the snapshots and experimental results.



Collaboration Workflow

Further, a workspace can be published in a repository and shared with collaborators. DWS consolidates the different resources required to replicate the workspace on a different machine, and allows collaborators to share workspaces, experiments, and snapshots. The workspace, including dependencies, can be replicated and executed on a local machine as well as on a hosted service.

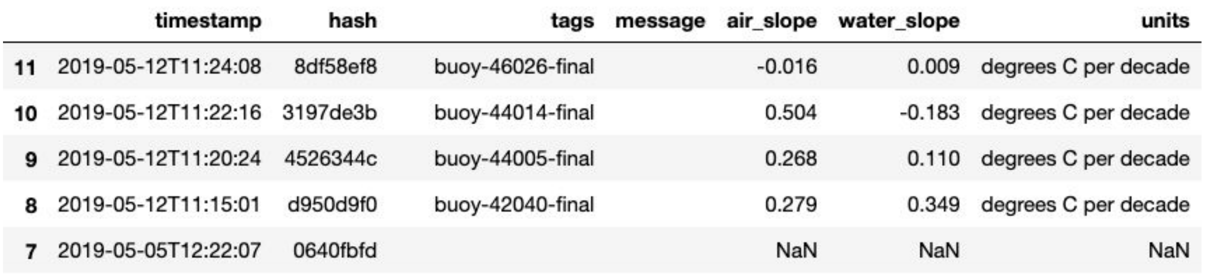


Reporting

As the project progresses, various reports can be generated to show:

- The history of snapshots taken of the workspace (across all the collaborators) and the key metrics for each snapshot
- Detailed parameters and results of a given snapshot
- Detailed lineage data for the current state of the workspace or any past snapshot

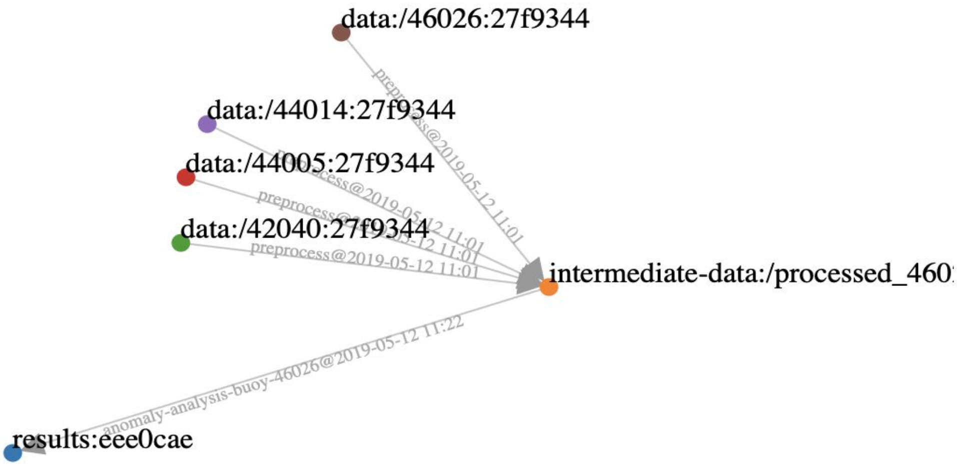
For example, here is a snapshot history report from the Buoy Data Analysis project:



	timestamp	hash	tags	message	air_slope	water_slope	units
11	2019-05-12T11:24:08	8df58ef8	buoy-46026-final		-0.016	0.009	degrees C per decade
10	2019-05-12T11:22:16	3197de3b	buoy-44014-final		0.504	-0.183	degrees C per decade
9	2019-05-12T11:20:24	4526344c	buoy-44005-final		0.268	0.110	degrees C per decade
8	2019-05-12T11:15:01	d950d9f0	buoy-42040-final		0.279	0.349	degrees C per decade
7	2019-05-05T12:22:07	0640fbfd			NaN	NaN	NaN

The report shows some previously snapshots of experiments. The tags are readily accessible names which can be used to get back to a specific state later. Each snapshot reports some metrics from the experiments. The "NaN"s indicate a failed state that was saved for later scrutiny.

Each snapshot keeps track of the state of the resources (as hashes), and also lineage information showing which resources were used in obtaining a result. Here is a visualization from the lineage report for snapshot 'buoy-46026-final" of the same project:


data:/46026:27f9344
preprocess@2019-05-12 11:01
data:/44014:27f9344
preprocess@2019-05-12 11:01
data:/44005:27f9344
preprocess@2019-05-12 11:01
data:/42040:27f9344
preprocess@2019-05-12 11:01
intermediate-data:/processed_460
anomaly-analysis-buoy-46026@2019-05-12 11:22
results:eee0cae

The nodes in this graph represent specific file paths within resources (e.g. data:/46026) and the edges represent step executions which read-from and write-to resources. The sequences of letters and numbers following the resource names (e.g. 27f9344) are the hashes of the resource content that were captured during snapshots.

Project Status:

Data Workspaces as been released as open source software[1] and is available for use today. The project has a small user community centered around the Max Planck Institute for Software Systems and is looking for new users. We are happy to answer questions and discuss how Data Workspaces can be incorporated into your workflow.

[1] https://github.com/data-workspaces/data-workspaces-core