

# Ghub Pegasus WMS Workflows

- Ghub scientific gateway is built on the HUBzero platform for scientific collaboration.
- Ghub provides you and other members of the Ghub community a single community place, accessible from anywhere in the world via a web browser, and hosts analytical tools, data, and other shared resources.
- Who develops the analytical tools on Ghub? You and others member of the Ghub community.
- Here we outline the procedure for developing a Pegasus WMS computational workflow analytical tool on Ghub.

- Computational workflows are a formalization of the manual computational workflow job steps that a scientist performs to obtain scientific results.
- A Workflow Management System (WMS) comprises software that manages the distribution and execution of computational workflows.
- The Pegasus WMS was selected as the best general purpose WMS to provide the structured platform required for implementing computational workflows on the HUBzero platform [1, 2, 3].

# Example Pegasus WMS Architecture

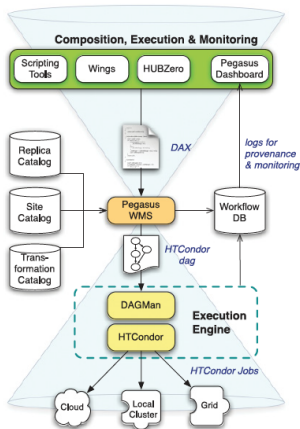


Figure 1: Pegasus WMS Architecture Example[4]

- HTCondor is a workload scheduling system for computational jobs.
- HTCondor provides a job queuing mechanism, scheduling policy, priority scheme, resource monitoring, and resource management.
- DAGMan is a HTCondor tool that allows multiple jobs to be organized as a workflow, represented as a directed acyclic graph (DAG) in which the nodes represent computational tasks and edges represent the dependencies of those tasks [5].
- DAGMan automatically submits workflow jobs such that certain jobs need to be complete before other jobs start running.
- DAGMan provides the workflow engine for Pegasus.

- Pegasus requires three information catalogs to plan a workflow. These are the Site Catalog, the Transformation Catalog and the Replica Catalog.
- The Site Catalog describes the site where the workflow jobs are to be executed. For Ghub, workflow jobs are executed on CCR's UB-HPC cluster general-compute nodes by interfacing with CCR's SLURM (Simple Linux Utility for Resource Management) Workload Manager.
- The Transformation Catalog describes the launch scripts invoked by the workflow jobs. The launch scripts loads and then runs executable modules available on CCR. Execution of workflow jobs on UB-HPC requires access to executable modules available on CCR.
- The Replica Catalog tells Pegasus where to find the input files required by the workflow jobs.

- Pegasus comprises a set of system components which can be invoked via command line tools or an API interface.
- On Ghub, a Jupyter Notebook provides the interface which allows the user to interface with Pegasus, via Pegasus Python API commands, to create a YAML formatted file defining the workflow jobs, the input and output for the jobs, and the job dependencies.
- The Ghub submit command line tool allows Pegasus jobs to be executed remotely.
- The Ghub submit command line tool is used to plan and launch Pegasus workflow jobs via the Pegasus kickstart process.

- Pegasus processes the YAML file, resolving data and software locations and all required data movements, and creates the required information catalogs and compiles a DAG file. DAGMan, as directed by the DAG file, orders the jobs according to their dependencies, and submits the jobs ready for execution to HTCondor.
- SLURM provides the framework for queuing jobs, allocating compute nodes, and starting the execution of jobs. When a SLURM job execution completes, the final status of the finished job is returned to Pegasus.



# Pegasus WMS Workflow Process Steps I

---

- Step 1: Get user's workflow parameters.
- Step 2: Create a workflow using Pegasus API commands.
- Step 3: For each executable required by the workflow:
  - Create a launch script for the executable.
  - Add the launch script to the Transformation Catalog.
- Step 4: For each parallel run of the launch script(s):
  - Create a job for the launch script.
  - Add arguments for the job based on user's parameters.
  - Add inputs and outputs for the job.
  - Add dependencies for the job.
  - Add the job to the workflow.
  - Add the inputs for the job to the Replica Catalog.

# Pegasus WMS Workflow Process Steps II

---

Step 5: Create the YAML file for the workflow.

Step 6: Plan and submit the YAML file.

Step 7: Wait for the workflow to complete.

Step 8: View and analyze the workflow's output.

# Ghub Exercise1

---

- The procedure to implement a simple pipeline Pegasus workflow is demonstrated with the `ghub_exercise1` project.
- The `ghub_exercise1` project's workflow executes Python scripts on CCR to analyze time series data from experiments files contained within predetermined UB CCR's mapped collection's modeling groups.
- Clone the `ghub_exercise1` on Ghub and follow the steps outlined by the `ghub_exercise1` README.md file.

# Ghub Exercise 1 Simple Pipeline Workflow Architecture

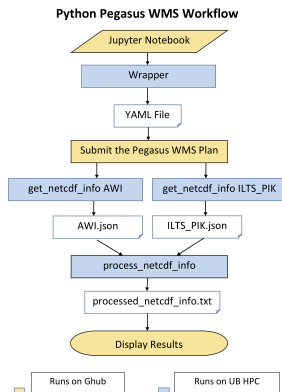


Figure 2: User Interface

# Ghub Exercise 1 Output

```
Pegasus workflow in progress. This should take approximately 30 minutes...
/projects/grid/ghub/ISMIP6/Projections/Reprocessed/CMIP6_Archive_Final/AIS
['AWI', 'ILTS_PIK']
self.parent:
self.tooldir: /home/thehub/renettej/AAA_notebooks/ghub_exercise1
self.bindir: /home/thehub/renettej/AAA_notebooks/ghub_exercise1/bin
self.datadir: /home/thehub/renettej/AAA_notebooks/ghub_exercise1/data
self.workingdir: /home/thehub/renettej/AAA_notebooks/ghub_exercise1
self.rundir:
self.folder: /projects/grid/ghub/ISMIP6/Projections/Reprocessed/CMIP6_Archive_Final/AIS
self.modeling_group_list: ['AWI', 'ILTS_PIK']
len(self.modeling_group_list): 2
self.maxwalltime: 30
Wrapper_5_0_1...
tooldir: /home/thehub/renettej/AAA_notebooks/ghub_exercise1
python_launch_exec_path: /home/thehub/renettej/AAA_notebooks/ghub_exercise1/remotebin/pythonLaunch.sh
Run 38429 registered 1 job instance. Mon Oct 9 10:35:24 2023
Run 38429 instance 1 released for submission. Mon Oct 9 10:35:54 2023
(413190.0) DAG Submitted at WF-ccr-ghub Mon Oct 9 10:35:57 2023
(413190.0) DAG Running at WF-ccr-ghub Mon Oct 9 10:36:33 2023
(413190.0) DAG Running at WF-ccr-ghub Mon Oct 9 10:42:08 2023
(413190.0) DAG Running at WF-ccr-ghub Mon Oct 9 10:48:08 2023
(413190.0) DAG Running at WF-ccr-ghub Mon Oct 9 10:54:49 2023
(413190.0) DAG Running at WF-ccr-ghub Mon Oct 9 11:00:09 2023
(413190.0) DAG Done at WF-ccr-ghub Mon Oct 9 11:08:09 2023
Workflow completed successfully

Workflow elapsed time: 32.83235842784246 minutes
```

Figure 3: Workflow Output

# Ghub Exercise 1 UB CCR Scratch Directory

Remote site: /panfs/panfs.cbils.ccr.buffalo.edu/scratch/grp-ghub/ghubjobs/1696898124\_00038429\_01/scratch

1696898124\_00038429\_01

1696898124\_00038429\_01

scratch

Filename	Filesize	Filetype	Last modified	Permissions	Owner/Group
..					
packages		Directory	10/09/2023 13:40:45	drwxrwsr-x	ghub grp-ghub
__time_results.00038429_01	60	00038429..	10/09/2023 13:56:46	-rw-rw-r--	ghub grp-ghub
__timestamp_finish.00038429_01	11	00038429..	10/09/2023 13:56:46	-rw-rw-r--	ghub grp-ghub
__timestamp_start.00038429_01	11	00038429..	10/09/2023 13:56:05	-rw-rw-r--	ghub grp-ghub
__timestamp_transferred.00038429_01	11	00038429..	10/09/2023 13:56:05	-rw-rw-r--	ghub grp-ghub
00038429_01_413194.slurm	1,876	slurm-file	10/09/2023 13:36:41	-rw-r--r--	ghub grp-ghub
00038429_01_413195.slurm	1,999	slurm-file	10/09/2023 13:39:12	-rw-r--r--	ghub grp-ghub
00038429_01_413196.slurm	2,024	slurm-file	10/09/2023 13:39:17	-rw-r--r--	ghub grp-ghub
00038429_01_413198.slurm	2,018	slurm-file	10/09/2023 13:55:48	-rw-r--r--	ghub grp-ghub
AIS_AWI.json	108,270	json-file	10/09/2023 13:53:04	-rw-rw-r--	ghub grp-ghub
AIS_AWI.txt	91,281	txt-file	10/09/2023 13:53:04	-rw-rw-r--	ghub grp-ghub
AIS_ILTS_PIK.json	66,976	json-file	10/09/2023 13:47:57	-rw-rw-r--	ghub grp-ghub
AIS_ILTS_PIK.txt	57,535	txt-file	10/09/2023 13:47:57	-rw-rw-r--	ghub grp-ghub
chmod	58,584	File	11/16/2020 17:24:58	-rwxr-xr-x	ghub grp-ghub
chmod_pythonlaunch_ID0000001_0.stderr	0	stderr-file	10/09/2023 13:37:04	-rw-rw-r--	ghub grp-ghub
chmod_pythonlaunch_ID0000001_0.stdout	3,788	stdout-file	10/09/2023 13:37:04	-rw-rw-r--	ghub grp-ghub
get_netcdf_info.py	9,497	py-file	10/09/2023 13:36:26	-rwxr-xr-x	ghub grp-ghub
process_netcdf_info.py	4,316	py-file	10/09/2023 13:36:31	-rw-r-----	ghub grp-ghub
processed_netcdf_info.txt	3,518	txt-file	10/09/2023 13:56:46	-rw-rw-r--	ghub grp-ghub
pythonlaunch	1,137	File	10/09/2023 13:36:26	-rwxr-xr-x	ghub grp-ghub
pythonlaunch_ID0000001.stderr	0	stderr-file	10/09/2023 13:39:44	-rw-rw-r--	ghub grp-ghub
pythonlaunch_ID0000001.stdout	13,133	stdout-file	10/09/2023 13:53:04	-rw-rw-r--	ghub grp-ghub
pythonlaunch_ID0000002.stderr	0	stderr-file	10/09/2023 13:39:44	-rw-rw-r--	ghub grp-ghub
pythonlaunch_ID0000002.stdout	15,176	stdout-file	10/09/2023 13:47:57	-rw-rw-r--	ghub grp-ghub
pythonlaunch_ID0000003.stderr	0	stderr-file	10/09/2023 13:56:05	-rw-rw-r--	ghub grp-ghub
pythonlaunch_ID0000003.stdout	14,854	stdout-file	10/09/2023 13:56:46	-rw-rw-r--	ghub grp-ghub

Figure 4: UB CCR Scratch Directory



Michael McLennan, Steven Clark, Ewa Deelman, Mats Rynge, Karan Vahi, Frank McKenna, and Derrick Kearney. Bringing Scientific Workflow to the Masses via Pegasus and HUBzero. *SemanticScholar*, 2013.



Michael McLennan, Steven Clark, Ewa Deelman, Mats Rynge, Karan Vahi, Frank McKenna, Derrick Kearney, and Carol Song. HUBzero and Pegasus: integrating scientific workflows into science gateways. *Wiley*, 2014.



Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J. Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira da Silva, Miron Livny, and Kent Wenger. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*, 46:17–35, 2015. ISSN: 0167-739X. DOI: <https://doi.org/10.1016/j.future.2014.10.008>. URL: <https://www.sciencedirect.com/science/article/pii/S0167739X14002015>.



C. S. Liew, M. P. Atkinson, M. Galea, T. F. Ang, P. Martin, and J. I. V. Hemert. Scientific workflows. *Computing Surveys*, 49(4):1–39, 2016. DOI: <https://doi.org/10.1145/3012429>.





Ryan Mitchell, Loïc Pottier, Steve Jacobs, Rafael Ferreira da Silva, Mats Rynge, Karan Vahi, and Ewa Deelman. Exploration of Workflow Management Systems Emerging Features from Users Perspectives. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4537–4544, 2019. DOI: [10.1109/BigData47090.2019.9005494](https://doi.org/10.1109/BigData47090.2019.9005494).