eFlows4HPC

Enabling dynamic and Intelligent workflows in the future EuroHPC ecosystem

D2.1 Data Catalogue

Version 1.0

Documentation Information

Contract Number	9555558
Project Website	www.eFlows4HPC.eu
Contractual Deadline	31.08.2021
Dissemination Level	PU
Nature	Other
Author	Jedrzej Rybicki (FZJ)
Contributors	
Reviewer	Enrique Quintana & José Flich (UPV)
Keywords	Data Management, Storage, Implementation

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955558. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Germany, France, Italy, Poland, Switzerland, Norway.

Change Log

Version	Description Change
V0.1	Start of the document
V0.2	Reviewed by UPV
V1.0	Suggestions applied

Table of Contents

1. Executive Summary

The following describes the Flows4HPC Data Catalogue. This service will provide information about data sets used in the project. The Catalogue will store information required to access particular data along with additional metadata. This information will be used throughout the project, mainly by the Data Logistics Service.

The Data Catalogue was implemented with FastAPI, a modern Python framework allowing for flexibility, extensibility, and quick deployment. The Data Catalogue offers an API well-document in the (industry-standard) Swagger/OpenAPI format as well as a web-based GUI.

The primary use case for Data Catalogue is to store information about the dataset which can be then used by the Data Logistics Service to facilitate the required data movements. Secondly, the Catalogue will improve the visibility of the datasets created and used in the Project, enabling possible reuse and collaboration in spirit of FAIR data principles. To this end, the Data Catalogue offers a possibility to describe the items with a rich set of metadata.

2. Introduction

The eFlows4HPC Data Catalogue will store information about the datasets created and used in the project. Its primary goal is to facilitate the data movements in Data Logistics Service (DLS).

The Data Catalogue service was developed in Python using FastAPI [1]. The source code of the service can be found in the Project's GitHub repository [2]. The repository includes documentation and architecture overview.

The Data Catalogue service is deployed on the Helmholtz Data Cloud (HDF) infrastructure [3] and can be accessed through the following link:

https://datacatalog.fz-juelich.de/

The main function of Data Catalogue is to list storages. These can either be sources of data or targets to upload the data. The listing of storages is available to all users without a need to authenticate. Only addition, modification, and removal of the items require authentication. Each item in the list has a name, an associated URL, and can include some additional metadata. The internal data format is JSON allowing for further extensions of the model. Such extensions are expected during the integration with other services and can result in further specification of the expected Data Catalogue functionality. One of the main advantages of using FastAPI for the implementation is the availability of the API documentation in Swagger/OpenAPI format [4]. This is also crucial for easy integration of the Data Catalogue with other eFlows4HPC services.

The creation of the software as well as a running instance constitutes the Deliverable D2.1. Through the rest of the project the instance will be maintained and populated with the relevant data.

3. Acronyms and Abbreviations

- HDF – Helmholtz Data Federation
- DLS – Data Logistics Servicer
- FAIR – Findability, Accessibility, Interoperability, and Reuse

4. References

[1] "FastAPI," [Online]. Available: https://fastapi.tiangolo.com . [Accessed August 2021].

[2] "eFlows GitHub Repository," [Online]. Available: https://github.com/eflows4hpc/datacatalog/tree/stable-0.16 . [Accessed August 2021].

[3] B. Hagemeier, "HDF Cloud – Helmholtz Data Federation Cloud Resources at the Jülich Supercomputing Centre.," *Journal of large-scale research facilities,* vol. 5, no. A137, pp. 1-7, 2019.

[4] "Data Catalog API specification," [Online]. Available: https://datacatalog.fz-juelich.de/docs . [Accessed August 2021].

[5] "Swagger," [Online]. Available: https://swagger.io/resources/open-api/ .