eFlows4HPC

Enabling dynamic and Intelligent workflows in the future EuroHPC ecosystem

D2.3 First version of Data Logistics

Version 1.0

Documentation Information

Contract Number	9555558
Project Website	www.eFlows4HPC.eu
Contratual Deadline	01.04.2022
Dissemination Level	PU
Nature	Other
Author	Jedrzej Rybicki (FZJ)
Contributors	Maria Petrova-El Sayed (FZJ), Chrisian Böttcher (FZJ), Jorge Ejarque (BSC)
Reviewer	Stephane Zeng (Atos)
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	Initial version
V0.2	Addressing first review comments
V1.0	Final format fixes

Table of Contents

1. Executive Summary

The Data Logistics Service (DLS) will facilitate the data movements required by the workflows developed by the Pillars. This document marks the availability of the first version of the service. It is integrated with TOSCA/YORC to implement the first project workflow, in which DLS is responsible for the data movement from a data repository to a processing facility and back.

The service is based on open-source Apache Airflow with some project-specific customizations. Within the service, data movements are formalized as pipelines implemented in Python. In the long term, this will enable the communities to contribute their own pipelines.

The automatic deployment of the service is done in a stable Cloud environment offered in FZJ enabling high service quality and short turn-around times for service extensions.

1.1. Introduction

The Data Logistics Service (DLS) will facilitate the data movements required by the workflows developed by the Pillars. Those complex workflows utilize heterogeneous resources ranging like HPC, HTC and/or Cloud to accomplish the scientific tasks (Figure 1). Clearly, such heterogeneous infrastructure requires efficient data movement between processing stations. To this end, eFlows4HPC implemented a first version of the Data Logistics Service, which will be extended as the project progresses.

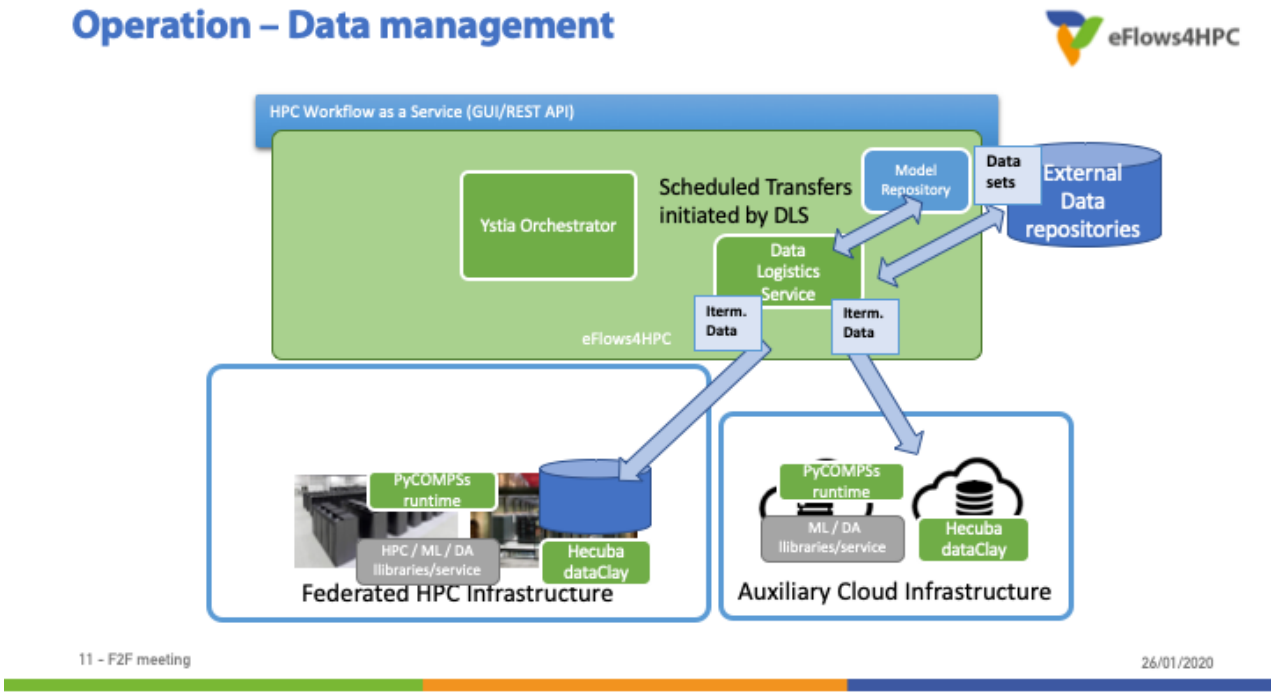

Figure 1 Architecture Overview

The first version of Data Logistics Service is based on Apache Airflow [1]. This is an open source software solution to programmatically author, schedule, and monitor data movement pipelines. This software was selected to capitalize on the partner's (FZJ) expertise. The initial evaluation showed that the data movement required in the project can be successfully implemented as

Airflow pipelines. Nonetheless, in the future, a migration to different technologies is also possible. A thorough evaluation of the ease of use and required functionality is ongoing. One candidate technology is Prefect [2], which might prove to be more flexible and user-friendly for the communities. The pipelines created for Airflow can easily be deployed to Prefect.

The DLS pipelines, formalizing the data movements, are written in Python: a language that is commonly used in scientific applications, and thus lowers the entry barrier for the users. Given the ability to easily integrate with existing command-line scripts, the reuse of existing data movement solutions is possible.

1.2. Results and impacts

The creation of the first version of the Data Logistics Service involved the following steps:

- requirement analysis,
- architecture specification for a project-specific deployment,
- implementation of project-specific extensions of the service,
- creation of a project-wide instance of the service,
- implementation of initial data movement pipelines and their integration with data sources,
- establishment of a process to define, deploy, and run new pipelines,
- integration with a project-specific workflow execution system.

The architecture of the service, with deployment description can be found in the project repository. The development of the service takes place in the local GitLab repository, whereas the stable versions can be found in the eFlows4HPC GitHub repository [3]. The deployment of the instance is conducted automatically on the Cloud resources from HDF Cloud [4].

The implementation of the pipelines was, in the first project phase, guided by the minimal workflow use case defined in the project. The workflow is defined to address the requirements gathered from Pillars. It resembles the real use cases and will serve as bedrock for the implementation of the more complex Pillar workflows later in the project. The workflow comprises two data movements: data deployment from repository to HPC system (where data processing takes place) and the movement of the computation results to a data repository. As the first data repository to integrate, B2SHARE [5] was chosen and the processing took place on an HPC system in BSC. The minimal workflow is described in OASIS TOSCA [6] and executed YORC [7]. More details on the eFlows4HPC minimal workflow can be found in its documentation [8].

Data Logistics Instance is available at: https://datalogistics.eflows4hpc.eu.

The availability of the first version of the Data Logistics Service impacts the project in a positive way. The DLS integrated with Data Catalogue constitutes an important step in implementing FAIR [9] principles in the project. The visibility of the data sets used as well as the formalization of the data movements in form of pipelines adds to reproducibility of the research conducted in the project.

The implementation of the minimal workflow supported by the DLS reinforced the validity of the approach taken in eFlows4HPC and provide the users with a reference of how the implementation of their workflows with eFlows4HPC software stack can look like, and work.

2. Acronyms and Abbreviations

- CA – Consortium Agreement
- D – deliverable
- DLS – Data Logistics Service
- HPC – High Performance Computing
- HTC – High Throughput Computing
- HDF – Helmholtz Data Federation
- M – Month
- PM – Person month / Project manager
- OASIS – Organization for the Advancement of Structured Information Standards
- TOSCA – Topology and Orchestration Specification for Cloud Applications
- WP – Work Package
- YORC – Ystia Orchestrator

3. References

[1] "Apache Airflow," [Online]. Available: https://airflow.apache.org/. [Accessed March 2022].

[2] "Prefect," [Online]. Available: https://docs.prefect.io. [Accessed March 2022].

[3] "Data Logistics GitHub Reposiotry," [Online]. Available: https://github.com/eflows4hpc/data-logistics-service. [Accessed March 2022].

[4] B. Hagemeier, "HDF Cloud – Helmholtz Data Federation Cloud Resources at the Jülich Supercomputing Centre," *Journal of large-scale research facilities,* vol. 5, pp. 1-7, 2019.

[5] S. B. Ardestani, C. J. Hakansson, E. Laure and I. Livenson, "B2SHARE: An Open eScience Data Sharing Platform," in *IEEE 11th International Conference on e-Science*, Munich, Germany, 2015.

[6] OASIS, "TOSCA Standard," [Online]. Available: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca. [Accessed March 2022].

[7] "Yorc," [Online]. Available: https://yorc.readthedocs.io/en/stable/. [Accessed March 2022].

[8] eFlows4HPC, "Minimal Workflow Documentation," [Online]. Available: https://eflows4hpc.readthedocs.io/en/latest/Sections/04_Usage_Example.html . [Accessed March 2022].

[9] M. D. Wilkinson, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data,* vol. 3, no. 1, 2016.