

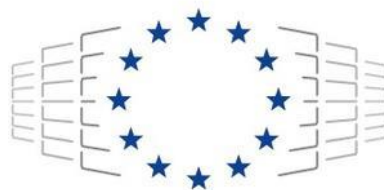
**HORIZON-EUROHPC-JU-2021-COE-01**



**The European Centre of Excellence for Engineering  
Applications**

**Project Number: 101092621**

**D4.5  
Data Management and Data Analytics in  
EXCELLERAT (Update)**



The EXCELLERAT P2 project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101092621. The JU receives support from the European Union’s Horizon Europe research and innovation programme and Germany, Italy, Slovenia, Spain, Sweden and France.

<b>Work Package:</b>	WP4	Workflow Development
<b>Author(s):</b>	Marvin Hubl Christian Gscheidle	SSC FHG
<b>Approved by</b>	Executive Centre Management	18.06.2025
<b>Reviewer</b>	Andrej Kosicek	ARCTUR
<b>Reviewer</b>	Claudio Arlandini	CINECA
<b>Dissemination Level</b>	Public	

Date	Author	Comments	Version	Status
2025-05-19	Christian Gscheidle	First Draft (Update, D4.5)	V1.1	Draft
2025-05-30	Christian Gscheidle Marvin Hubl	Second Draft	V1.2	Draft
2025-06-04	Christian Gscheidle Marvin Hubl	Final Draft	V1.3	Final

## List of abbreviations

API	Application Programming Interface
CFD	Computational Fluid Dynamics
DMD	Dynamic Mode Decomposition
DOI	Document Object Identifier
FTP	File Transfer Protocol
HDF5	Hierarchical Data Format, Version 5
HPC	High-Performance Computing
HTTPS	Hypertext Transfer Protocol Secure
InSiDS	In-Situ Data Analysis Toolbox
JU	Joint Undertaking
MPI	Message Passing Interface
OFTP2	Odette File Transfer Protocol 2
PCA	Principle Component Analysis
POD	Proper Orthogonal Decomposition
S3	Simple Storage Service
SFTP	Secure File Transfer Protocol
SME	Small or Medium-sized Enterprise
SVD	Singular Value Decomposition
SVM	Support Vector Machine
T	Task
WP	Work Package
YAML	Yet Another Markup Language

## Executive Summary

The document outlines key advancement in the “Data Management” (Task 4.4) and “Data Analytics” (T4.2) tasks of EXCELLERAT P2 for months 13-30. One focus is the further development of SCALES, a software toolbox which standardises the handling, storage, and exchange of data between High-Performance Computing (HPC) resources and users, with a focus on Small and Medium-sized Enterprises (SMEs). This tool automates secure data transfer and improves data traceability while reducing data volume through chunking and deduplication methods. In data analytics, new streaming routines and algorithms have been created to automate comparisons and enhance analysis efficiency during runtime. The integration of the In-Situ Data Analysis Toolbox (InSiDS) and SimExplore software allows for real-time extraction of features and facilitates comparison of results from multiple simulations. These advancements contribute to better data management and analysis capabilities, supporting collaboration among SMEs in the HPC environment and aligning with industrial standards for protecting intellectual property. Future work will focus on supporting the optimizing workflows developed in task 4.3, enhancing streaming algorithms and running performance and scalability tests of the algorithms involved.

## Table of Contents

1	Introduction .....	7
2	Technology basis.....	7
2.1	Data exchange and management with SCALES .....	7
2.2	Comparative analysis of simulations with SimExplore .....	9
2.3	In-Situ data analysis of CFD simulations with InSiDS.....	10
3	Data exchange and management for industrial engineering applications .....	11
4	Technical progress of Task 4.2 - Data Analysis.....	15
4.1	Data I/O .....	15
4.2	Streaming algorithms for modal decompositions.....	15
4.3	Workflow for comparative analysis .....	15
4.4	Meshing routines tailored for data analysis.....	17
5	Connection to EXCELLERAT P2 tasks and work plan .....	18
6	References .....	19

## Table of Figures

Figure 1: Scheme of SCALES as the technological basis.....	8
Figure 2: Workflow overview .....	9
Figure 3: Low dimensional representation computed with SimExplore of a total of 90 CFD simulations. Three distinct clusters of physical behaviour as well as a non-converged solution can be identified .....	10
Figure 4: Overview of the data model and user interface of InSiDS .....	11
Figure 5: Architecture for the data management tool SCALES in EXCELLERAT P2 .....	12
Figure 6: Functional extension of SCALES in EXCELLERAT P2 (part 1 of 3) .....	13
Figure 7: Functional extension of SCALES in EXCELLERAT P2 (part 2 of 3) .....	14
Figure 8: Functional extension of SCALES in EXCELLERAT P2 (part 3 of 3) .....	14
Figure 9: Workflow for the comparative analysis of transient CFD data .....	16
Figure 10: Generation of the analysis mesh based on a mesh morphing approach .....	17

## 1 Introduction

The Software development tasks for “Data Management” (T4.4) have been continued. There, an objective for the project’s end is the development of a software tool for standardised data handling, storage and exchange between HPC resources and users to support cooperative work from an industry perspective with a focus on small and medium-sized enterprises. Besides the consideration of essential requirements from industrial users for engineering applications, the tool shall support traceability of data in the storage environments respectively archives. This relates to information about what data are stored where and when, about the frequencies of data usages, the purposes of the data usages as well as about where the data are processed. For the objective of establishing a standard, pertinent recent industrial conceptions and standardisation results shall be integrated.

One of the most important requirements from the industry for engineering applications is the economically motivated necessity to protect the knowledge on product development contained in the digital simulations models as intellectual properties. This requires specific solutions that affect the software architecture as well as the data exchange processes. Therefore, effort had been taken on elaborated software system designs as basis for exact specifications and ultimately for the goal-oriented implementation.

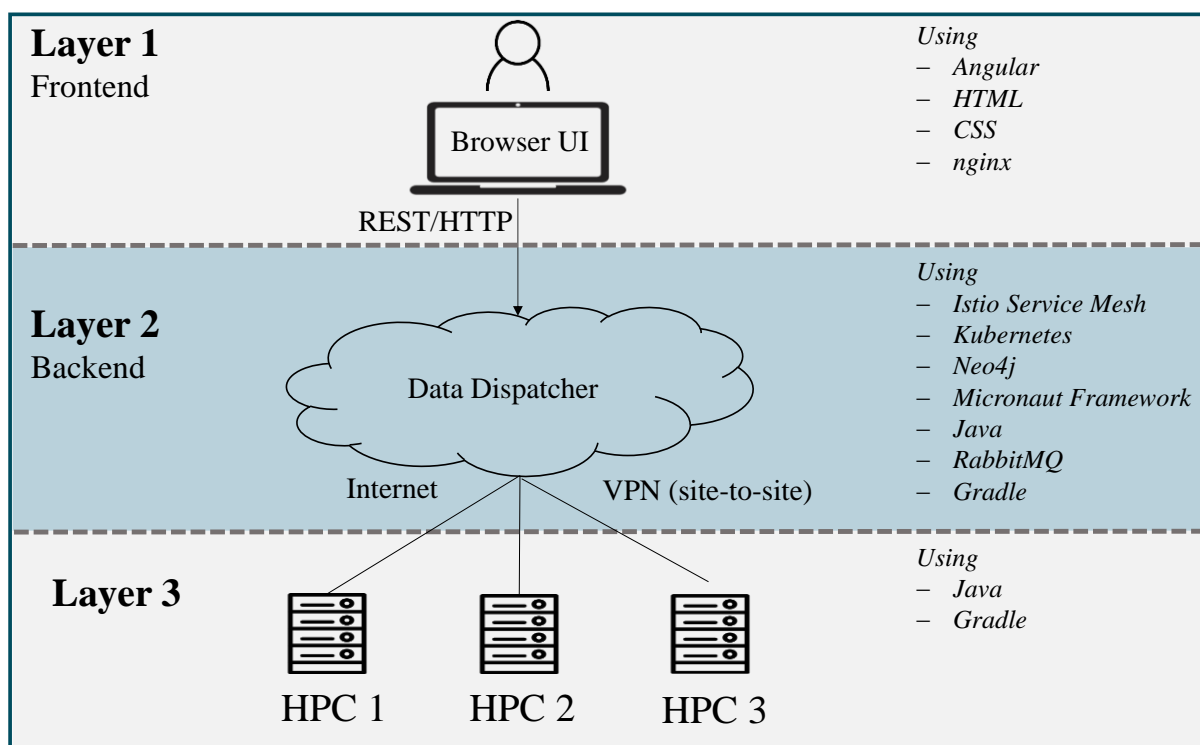
For “Data Analytics” (T4.2) the extraction of knowledge from large-scale engineering simulations poses several challenges for data analysis routines and workflows. To handle the large amounts of data being produced during the simulation, new streaming routines have to be developed that reduce data I/O and allow an initial analysis of the data already during runtime. When performing parameter studies of multiple simulations, a manual comparison of all data is oftentimes not feasible. Thus, smart algorithms that automate a comparison and highlight key features and differences in simulation is necessary. In EXCELLERAT P2, a focus lies on using unsupervised data-driven approaches that can structure data without the need of providing manual labels. Furthermore, in-situ algorithms are developed and implemented to extract important features and gain deeper insights into transient simulation and thus exploit the full potential of exa-scale systems.

## 2 Technology basis

### 2.1 *Data exchange and management with SCALES*

The technological development is based on the data exchange and management tool SCALES, continuing developments from the first phase of EXCELLERAT [1]. The platform is not only used for data processing, but also enables a safe and traceable, online data transfer between the data generators and several HPC centres represented in the EXCELLERAT project. This data transfer will be highly automated to avoid duplication of the transferred content. This approach reduces the amount of transferred data, which can ultimately save time and costs. The portal divides uploaded or stored data into chunks. Then hash values of the chunks are calculated and used as the basis for comparisons of data chunks and duplication discovery. Data chunks that had already been uploaded or stored do not need to be uploaded or stored again. The portal provides relevant HPC processes for the end users, such as uploading input decks, scheduling workflows, or processing HPC jobs.

The added value of the workflow portal in relation to exa-scale data are the topics such as data reduction, volume reduction and data compression of input and output data. In concrete terms, this means that the data becomes smaller and data transport becomes more manageable for the data transport. Figure 1 sketches a scheme of the logical components and used technologies for SCALES.

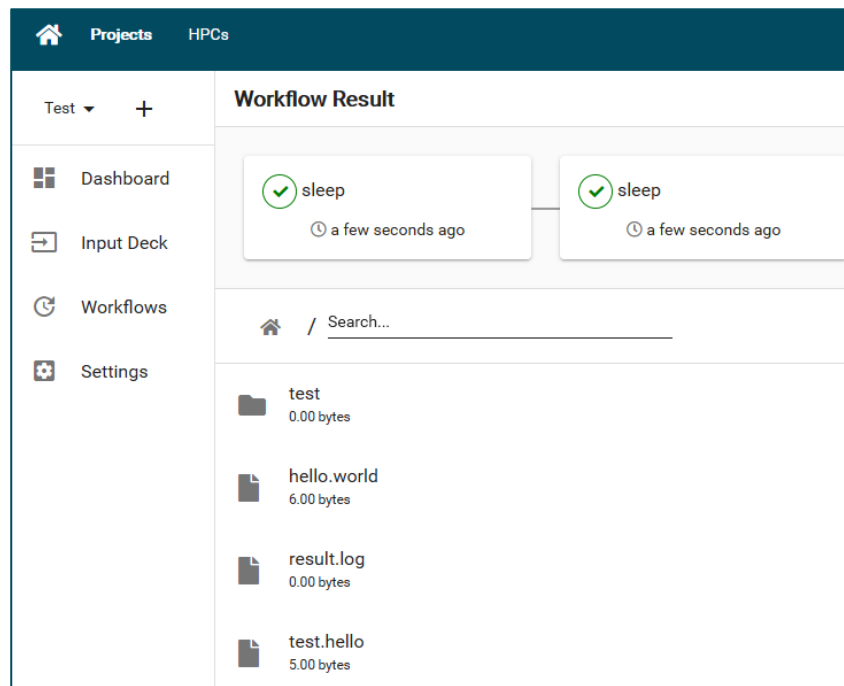


**Figure 1: Scheme of SCALES as the technological basis**

After a successful login, the user starts on a web-based dashboard page. To use the corresponding HPC resources, a connection to the cluster, on which the calculation is to be performed, is required.

The basic structure of the platform consists of projects corresponding to the workspaces on the clusters. It is mandatory to specify how long the retention period (1-30 days with a possible extension of three times, which corresponds to a total of 120 days) should be. Each project consists of a dashboard, an input deck, workflows, and project settings. Figure 2 shows an example frontend view of SCALES.

To prepare the simulation, data is created locally and can be uploaded to the corresponding cluster through the input deck menu. After the appropriate files have been selected, the user is asked on which machine these files should be uploaded. During the upload, the system checks whether there are identical file pieces, that are already in use and do not need to be uploaded again.



**Figure 2: Workflow overview**

A control file "excellerat.yaml" describes the workflows, how the simulation should look like as well as what should be done with it in the workspace. Additionally, you can specify scripts that may run in pre- and post-processing here. Those could be included in the YAML file as batch scripts or uploaded at the beginning and called in YAML. Around the control file the input data is added to run the simulation. The corresponding workflow can be scheduled and started. After the run has been successfully scheduled, each step is processed in the background, executed automatically and the user receives a browser pop-up notification at the end.

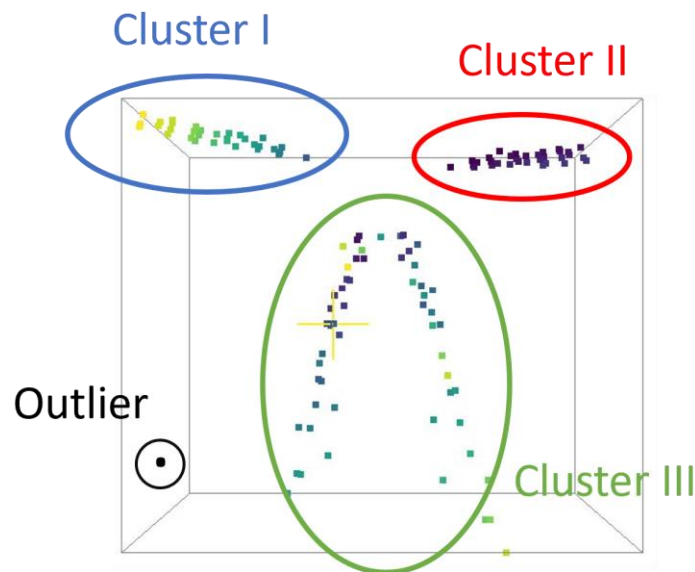
## **2.2 Comparative analysis of simulations with SimExplore**

SimExplore [2] is a software tool developed by Fraunhofer SCAI supporting the engineering design process by providing a comparative analysis of simulation results. It is available for academic or industrial customers via individual test or production licences upon request.

SimExplore allows a simple comparison of multiple simulations from parameter studies in a simple two- or three-dimensional representation. Similarities and differences between simulations and as well as simulation outliers can be identified and explored interactively. A focus is on data-driven features that enable an investigation of the global behaviour of the simulations. This provides deeper insights into the simulations beyond the information that can be extracted from local or integral quantities that are usually defined as the quantity of interest. To compute the low dimensional representations, SimExplore provides two different techniques of dimension reduction. First, Diffusion Maps can be applied to extract intrinsic structures in the high dimensional data that lie on a non-linear manifold by computing global geodesic or "diffusion" distances between data points. As a second approach, a spectral basis on the mesh can be computed based on the eigenvectors of the Laplace-Beltrami operator. This has the advantage that the basis does not depend on the data itself but only on the geometry and discretisation. Thus, it is suitable to be applied in a streaming or in-situ way during runtime of the simulation by projection new data snapshots onto the basis as soon as it is available.

SimExplore has successfully been applied to car crash simulations highlighting different crash behaviour based on changing geometrical or material properties of the mechanical structures. Furthermore, it can be used to show trends during runtime of the simulation highlighting bifurcations or unphysical behaviour. A particular advantage lies in the analysis of many simulations at once. While, in principle, this is also possible to do manually by a visual analysis of each simulation result, SimExplore provides all relevant information in one view with a couple of seconds. Figure 3 shows the representation of a sample dataset of 90 simulations that describe the behaviour of the flow through an aerodynamic nozzle with two inlets and varying inlet velocities. Here, each point in the low dimensional space represents one simulation. We can clearly observe three major physical modes that are present for different settings of the inlet velocities.

In EXCELLERAT P2, SimExplore will be extended to handle data from large scale Computational Fluid Dynamics (CFD) simulations. A focus lies on parameter studies that are performed in order to investigate the influence of changes in the flow boundary conditions or geometrical bounds. Here, the complex and chaotic nature of turbulent flows poses additional challenges for data analysis routines. Furthermore, multiple simulations with a high temporal and spatial accuracy can produce enormous amounts of data that induce a further need for efficient and scalable data analysis and data compression workflows. Therefore, SimExplore is combined with an in-situ data analysis that can compute suitable data driven features during runtime of the simulation, thus reducing the amount of data even before it is written to disk.



**Figure 3: Low dimensional representation computed with SimExplore of a total of 90 CFD simulations. Three distinct clusters of physical behaviour as well as a non-converged solution can be identified**

### ***2.3 In-Situ data analysis of CFD simulations with InSiDS***

The In-Situ Data Analysis Toolbox for Simulations (InSiDS) is a software tool aiming at academia and research providing a prototyping environment for the development of efficient and scalable streaming algorithms. The core library consists of a data model in Python to represent flow data from CFD simulations with the specific needs of efficient, streaming and distributed data analysis algorithms. In contrast to representations that are common for visualisation tasks, the data model represents temporal snapshots in a batch wise manner. This allows a trade-off between memory consumption and efficiency while computing data-driven features during the simulation's runtime. The user interface is aligned with the Application

Programming Interface (API) of Scikit-learn [3], thus enhancing a simple integration of custom algorithms with existing processing routines and models from widely used data analysis tools. A high-level parallel execution of algorithms is provided by running local operations simultaneously on all Message Passing Interface (MPI) ranks. For low level parallelisation, mpi4py can be exploited to express communications between MPI ranks, such as needed for a parallel computation of modal decompositions. To store simulation raw data and data analysis results, a file-format is provided based on HDF5 that fulfils the specific needs for high performance data analytics. The data is stored in a temporal batch-wise and spatial block-wise manner. The latter can either hold raw data from one MPI rank of the simulation or local patches of structured data analysis results. To access simulation data during runtime common in-situ interfaces are exploited that are known from the visualisation community, such as ParaView Catalyst [4] and that are already implemented in many simulation codes. For a simple integration with existing simulation workflows, InSiDS provides user interfaces in three different settings: i) as a plugin for ParaView [5] ii) as a stand-alone python tool that can, for example, be used in a IPython server for parallel batch processing iii) as a Catalyst Plugin for in-situ execution. Figure 4 gives an overview of the data and execution model, as well as the user interfaces of InSiDS.

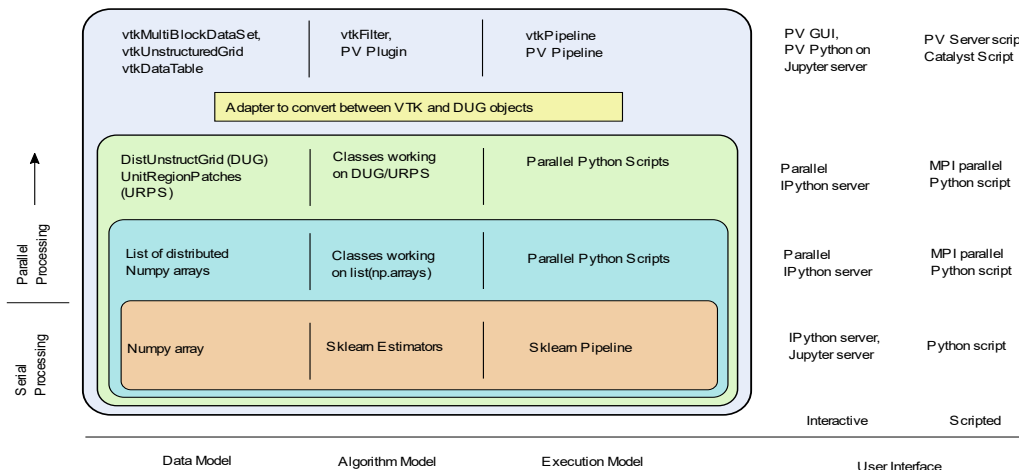


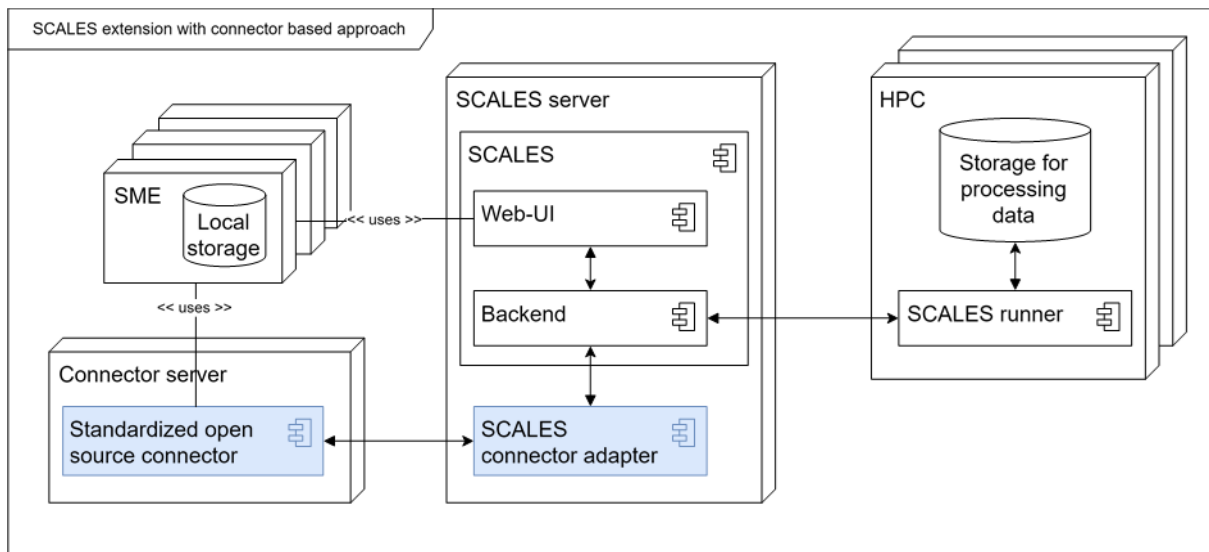
Figure 4: Overview of the data model and user interface of InSiDS

### 3 Data exchange and management for industrial engineering applications

For use cases comparable to cooperative work in engineering applications, literature on state of the art describes reference solutions that are being adopted by the industrial practice, primarily by the automotive and increasingly by the aerospace industry [6–10]. An advantage of the referenced approach is the fulfilment of industrial information control objectives by a granular automatable trust mechanism applicable on each exchanged data set. A significant progress compared with former solutions for applications with high information security requirements is that the approach operates based on a software technological so-called connector. The connector provides a standardised gateway and does not require a deep integration in proprietary corporate IT infrastructures but only a loose coupling. Furthermore, the connector

does not prescribe a certain data transfer protocol, but allows by design to integrate best suited protocols, e. g. HTTPS, FTP, SFTP, S3, OFTP2. The mode of action of the connector will be designed to be transparent to end-users (i. e. *not* visible). The end-user is, e. g., an engineer of an SME. Simulation data are, for instance, 3D models of an airfoil respectively the derived boundary conditions for computational fluid dynamic simulations, cf. EXCELLERAT P2 use case 6 [11].

Figure 5 depicts an expedient software architecture based on a comparative evaluation of different design alternatives. The components that are highlighted in blue establish a data transfer connection with respect to the requirements of industrial users. To this end, open-source software components can be integrated that have been developed as reference implementations supported by the EU in research and development projects for industries with important cooperative engineering value creation. [9, 10].



**Figure 5: Architecture for the data management tool SCALES in EXCELLERAT P2**

Figure 6 – Figure 8 show the functional extension of SCALES implementing the objectives for task T4.4 according to the grant agreement:

- Development of a data management (software tool)
- for standardised cooperative work and sharing of data
- from an industry perspective with a focus on SME.
- Enhancement of data lifecycle management:
  - Where is data stored or archived?
  - Where is data processed?
  - To what purpose is data?
  - How often is data used?
  - How old is data?

Users can select, defined data categories, when having uploaded the simulation input deck (Figure 6). These categories reflect the purposes to what the data are simulated and are meaningfully use case classes. Predefined are the EXCELLERAT P2 use cases. However, after the project phase it shall be possible that users define her own use case categories. Using

industry-wide accepted use case categories, e. g. determined in shared semantic models from industry interest groups (where in some of them e.g. the partner SSC is directly and actively involved) will enhance suitability for industry users.

Afterwards, the further data life cycle information is compiled (Figure 7). The data storage and processing locations are initially selected by the users in SCALES when choosing an HPC resource creating a corresponding workspace. The creation date is the time stamp of the data uploads. The deletion date is given as the resulting date specified by the retention period. The status provided information on whether the data are planned to be upload, uploaded and available for simulation execution on the HPC machine or whether the simulation is scheduled (Figure 8), running, interrupted, cancelled or finished respectively.

This lifecycle information is made persistent via SCALES independently for both sides, the user, and the HPC centre and provides it like a lightweight receipt. The receipt is secured against changes by cryptographic hash functions. Each change in the data life cycle leads to a new life cycle information receipt, made persistent for both sides (user and HPC centre). With this mechanism the data can be traced back during its the life cycle, including the information if respectively when data had been finally deleted. The traceability of the data “liveness” status, accounts for the desire of industrial users to know where their data – comprising their engineering knowledge and strategic information as valuable intellectual properties – are, which decreases perceived barriers to use HPC resources being outside of the users’ direct scope of control.

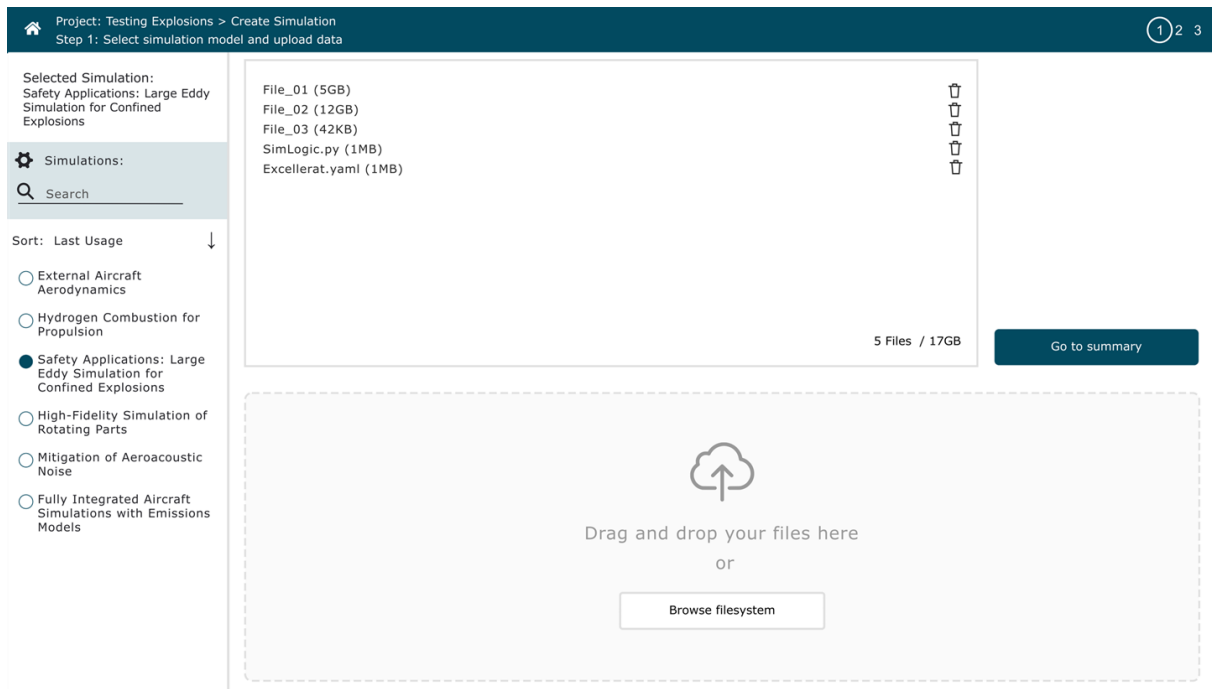


Figure 6: Functional extension of SCALES in EXCELLERAT P2 (part 1 of 3)

Projekt: Testing explosions > Create Simulation  
Step 2: Summary and confirm
1 2 3

Selected Simulation:  
Safety Applications: Large Eddy Simulation for Confined Explosions

HPC:  
HPE Cray EX4000 (Hunter)

Data information:  
5 Files / 17GB

Delete files after:  
5 days

Uploader name:  
Marvin

### Life cycle information

Information:  
The summary of the lifecycle information can be [downloaded now](#) or looked up later in the project history. Please note that only meta information is saved in the life cycle information and not the version of the data, so please be aware that a version recovery is **not** supported.

File	Storage location	Processing location	Data purpose	Creation Date	Deletion Date	Status
File_01 (5GB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	Today	2025-May-11	Upload planned
File_02 (12GB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	Today	2025-May-11	Upload planned
File_03 (42KB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	Today	2025-May-11	Upload planned
SimLogic.py (1MB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	Today	2025-May-11	Upload planned
Excellerat.yami (1MB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	Today	2025-May-11	Upload planned

Upload files
Upload files and start simulation

**Figure 7: Functional extension of SCALES in EXCELLERAT P2 (part 2 of 3)**

Projekt: Testing explosions > Create Simulation  
Step 3: Finalize
1 2 3

Selected Simulation:  
Safety Applications: Large Eddy Simulation for Confined Explosions

HPC:  
HPE Cray EX4000 (Hunter)

Data information:  
5 Files / 17GB

Delete files after:  
5 days

Uploader name:  
Marvin

### Life cycle information

Information:  
The summary of the lifecycle information can be [downloaded now](#) or looked up later in the project history. Please note that only meta information is saved in the life cycle information and not the version of the data, so please be aware that a version recovery is **not** supported.

Filename	Storage location	Processing location	Data purpose	Creation Date	Deletion Date	Status
File_01 (5GB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	2025-May-06	2025-May-11	Available Upload: 100%
File_02 (12GB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	2025-May-06	2025-May-11	Available Upload: 100%
File_03 (42KB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	2025-May-06	2025-May-11	Available Upload: 100%
SimLogic.py (1MB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	2025-May-06	2025-May-11	Available Upload: 100%
Excellerat.yami (1MB)	/ws/testing-explosions/	HPE Cray EX4000 (HUNTER)	Safety Applications	2025-May-06	2025-May-11	Available Upload: 100%

Status:  
Files available ✓  
Simulation scheduled ✓

Information:  
The files are available now and your simulation is scheduled. You can check your current simulation in the project see [Workflow](#).

Create new simulation
Back to project

**Figure 8: Functional extension of SCALES in EXCELLERAT P2 (part 3 of 3)**

## **4 Technical progress of Task 4.2 - Data Analysis**

### **4.1 Data I/O**

A data file format based on parallel HDF5 was developed that meets the specific requirements of high-performance data analytics. Temporal data is stored in a batch-wise manner as many data analysis algorithms are run over multiple time-steps while the entire simulation dataset would be too large to be held in memory at once. To allow a parallel execution of algorithms, the data is distributed and stored in spatial patches. The latter can either comply with the local data on the individual MPI ranks or represent local ‘regions’ of data analysis units. In addition, this allows a flexible selection of the number of MPI ranks for data processing. The file-format can hold time-dependant, block-structured or unstructured mesh with cells of arbitrary number of edges. The data on the mesh can be represented as point or cell-centred data with one or multiple components. The latter, again, is a specific requirement for data analytics results, such as spatial modes that can come along with multiple hundreds of components per snapshot.

Furthermore, routines to convert the data analysis results into the Animator’s A4DB file format was implemented. This allows to read data into further analysis tools, such as SimExplore, for a comparative analysis of multiple simulation runs.

### **4.2 Streaming algorithms for modal decompositions**

Modal decompositions, such as Proper Orthogonal Decomposition (POD) or Dynamic Mode Decomposition (DMD), can be used to extract dynamical features from turbulent flow data in order to get better insights into the physical mechanisms involved in the flow. However, when running large scale CFD simulation, it is not feasible or even possible to store all relevant simulation data to disk before running a modal decomposition due to limitations in data I/O. Thus, a common approach is the computation of these modes during runtime of the simulation while the data is still in memory of the HPC system. As part of InSiDs, a parallel and streaming algorithm to compute a Singular Value Decomposition (SVD) was implemented based on a split-and-merge approach. It exploits the already existing distribution of the simulation data over multiple MPI ranks by running local operations on each rank. Thus, the communication between ranks is reduced to a minimum during runtime. After the end of the simulation, the coefficients of the local matrices are collected into a global matrix in order to finalise the global SVD of the flow field. Based on the SVD, either a POD or DMD can be derived after runtime of the simulation or whenever needed to monitor sample convergence of the computation.

### **4.3 Workflow for comparative analysis**

A data analysis workflow for the comparative analysis of CFD simulation bundles was derived. The analysis of transient flow field from multiple highly resolved CFD simulations can be difficult due to chaotic turbulent structures in the flow. An overview of the entire data analysis workflow for the flow around two cylinders is shown in Figure 9. As a first step, we compute the DMD of the flow field to gain spatial structures that correspond to fixed frequencies in time. They represent important global, dynamical mechanism in the flow while reducing the chaotic parts. Next, a spectral basis is computed based on the computational mesh that the modes are projected onto. Compared to cartesian coordinates, the spectral coordinates are more suitable to represent the periodic like structures of the modes and allow an easier arrangement of the modes according to their similarities. Based on the spectral coordinates, dimension reduction

algorithms, such as the Principal Component Analysis (PCA) or Diffusion Maps, can be applied to reduce the formal dimension of the data and identify intrinsic structures in the data. Eventually, with the help of clustering algorithms, e.g., KMeans or DBScan, groups of similar characteristic physical behaviour can be labelled. This also enables the identification of outliers in multiple simulation runs that can be an indicator for non-converged solutions or erroneous numerical settings. The individual clusters of similar behaviour are derived from a global analysis of the flow field. This leads to broader understanding of the results and identification of additional effects that are often not visible in the physical-based objective function derived from integral or local properties of the flow. In a final step, a relationship between the input variables, such as the flows boundary conditions, and the clusters of similar behaviour can be computed by training a classifier, e.g., a Support Vector Machine (SVM). It predicts the boundaries of the cluster and can be applied to classify new data points into one of the groups based on the input variables only - without the need to run additional, expensive CFD simulations.

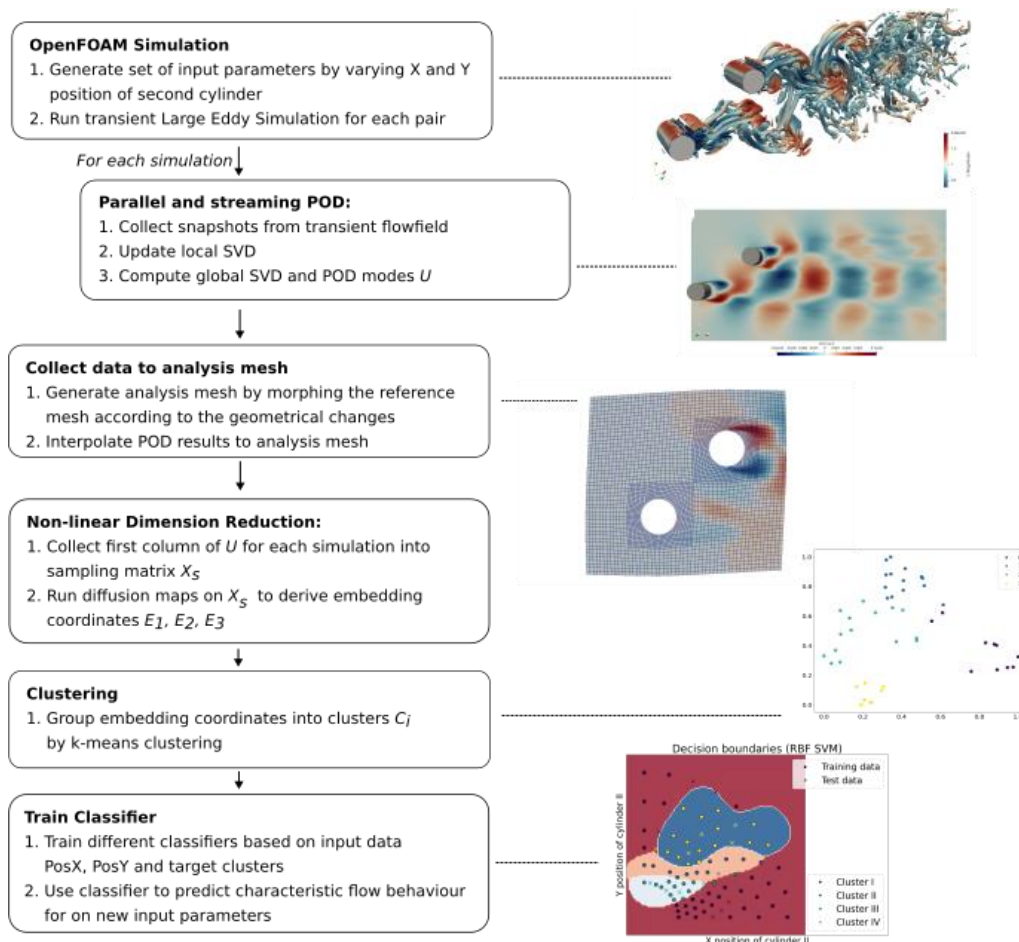
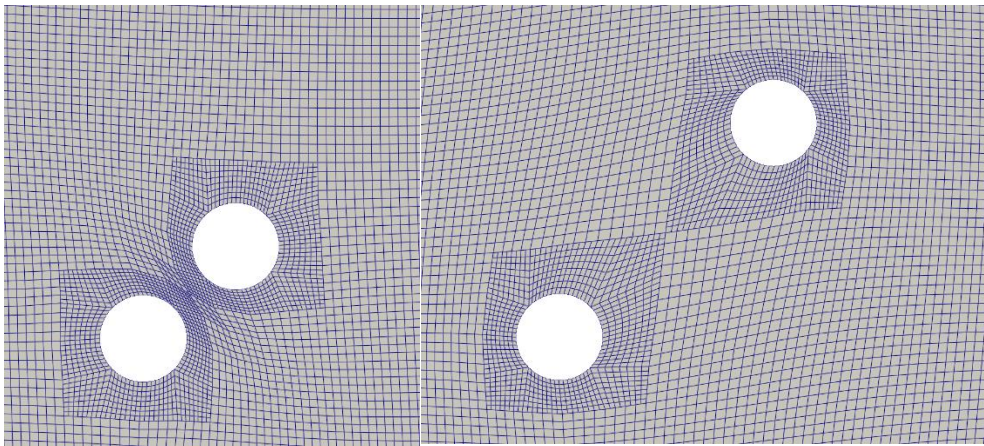


Figure 9: Workflow for the comparative analysis of transient CFD data

#### **4.4 Meshing routines tailored for data analysis**

A major challenge when comparing multiple simulations occurs for parameter studies with changing geometrical boundaries. This raises two problems. First, the computational mesh changes for each simulation run in the number of mesh points and their location, thus finding the correspondence between nodes is not trivial anymore. Second, there will be regions in one simulation that are not part of the flow field in other simulations and vice versa. Thus, there is no data available at these locations for comparison. To overcome both, a meshing routine is implemented that is based on a coarse reference mesh with hexa-elements only. This reference mesh will be morphed alongside the geometrical variation to create an individual data analysis target mesh for each simulation run. Finally, a structured mesh is placed into each element. The resolution of the local mesh can be chosen in a flexible manner to match the required accuracy based on the flow solution or its gradients. To guarantee a proper resolution of the geometrical bounds, a mapping of the local boundary meshes to the geometry is implemented. An example for the data analysis mesh generation is given in Figure 10.



**Figure 10: Generation of the analysis mesh based on a mesh morphing approach**

## 5 Connection to EXCELLERAT P2 tasks and work plan

The development described above is connected to other tasks within EXCELLERAT P2 as follows:

Service provisioning in task 5.1, training and education in task 5.2 and further applications in task 5.4 are based on solutions developed in tasks 4.2 and 4.4 concerning data analysis, data management and data transfer issues. In particular, a training series on data analytics for engineering data using machine learning is developed in collaboration between task 4.2. and task 5.2.

Since the development in task 4.4 integrates pertinent emerging technologies for data management, results of task 4.4 are closely connected to business development in task 6.3 and market exploitation in task 6.4. A regular exchange with industrial partners supported by Work Package (WP) 5 aligns the technical developments in WP4.

In the remaining time of the project, the focus of task 4.2 will be put on supporting the development of an optimisation workflow that is applied to Use-Case UC3 in task 4.3. Here, data analytic routines can be used to speed up optimisation by defining regions in the design space of similar physical behaviour. This can be exploited, e.g., by building multiple, local and more accurate surrogate models during the optimisation and thus reducing the number of simulation runs needed. A second focus will lie on further developing streaming algorithms for modal decompositions as well as investigating efficiency and scalability of the algorithms based on large-scale CFD test-cases.

## 6 References

- [1] EXCELLERAT P2, “Success Story: Enabling High Performance Computing for Industry through a Data Exchange & Workflow Portal”, online, <https://www.excellerat.eu/success-story-enabling-high-performance-computing-for-industry-through-a-data-exchange-workflow-portal/>, 28<sup>th</sup> May 2025.
- [2] SimExplore, online, <https://www.scai.fraunhofer.de/en/business-research-areas/numerical-data-driven-prediction/products/simexplore.html>, 30<sup>th</sup> May 2025.
- [3] Scikit-learn, online, <https://scikit-learn.org/stable/>, 30<sup>th</sup> May 2025.
- [4] ParaView Catalyst, online, <https://kitware.github.io/paraview-catalyst/>, 30<sup>th</sup> May 2025.
- [5] ParaView, online, <https://www.paraview.org/>, 30<sup>th</sup> May 2025.
- [6] H. Haße, H. van der Valk, F. Möller, B. Otto, “Design Principles for Shared Digital Twins in Distributed Systems,” *Bus Inf Syst Eng*, vol. 64, no. 6, pp. 751-772, 2022. DOI: 10.1007/s12599-022-00751-1.
- [7] S. Scheider, F. Lauf, F. Möller, B. Otto, “A Reference Systems Architecture with Data Sovereignty for Human-Centric Data Ecosystems,” *Bus Inf Syst Eng*, 2023. DOI: 10.1007/s12599-023-00816-9.
- [8] B. Otto, “A Federated Infrastructure for European Data Spaces,” *Communications of the ACM*, vol. 65, no. 4, pp. 44-45, 2022. DOI: 10.1145/3512341.
- [9] Eclipse Tractus-X “Connector KIT”, online, <https://eclipse-tractusx.github.io/docs-kits/category/connector-kit/>, 28<sup>th</sup> May 2025.
- [10] Aerospace-X “Building a digital ecosystem for an efficient and sustainable Aerospace Supply Chain”, online, <https://www.aerospace-x.net/en.html>, 28<sup>th</sup> May 2025.
- [11] EXCELLERAT P2 “UC-6”, online, <https://services.excellerat.eu/en/use-cases/active-control-for-drag-reduction-of-transonic-airfoils/>, 28<sup>th</sup> May 2025.