eosc | Federation

Scientific use case:
Federating CERN’s REANA pipelines

eosc Node | CERN Physical Sciences & Engineering

eosc Node | Finland

eosc Node | Italy

eosc Node | Poland

eosc Node | SURF The Netherlands

EOSC Node | European Commission

Tibor Simko works at CERN’s Department of Information Technology, where he develops tools to advance Open Science, particularly for particle physics research. He leads initiatives such as the CERN Open Data Portal, which disseminates over 5 petabytes of particle physics data, and the Reana reproducible analysis platform, a computing engine designed to accompany data repositories.

The science case focuses on enabling **near-data computation**—sending computational workflows to where large scientific datasets are stored, rather than transferring massive volumes of data to the researcher. The use case demonstrates this concept through particle physics—a field that generates enormous data volumes—but it is applicable to many other domains, including astronomy and life sciences. The project aims to show how researchers can execute their analyses directly at the data source, using **REANA**—CERN’s Reproducible research data analysis platform—to manage containerized workflows across federated computing resources.

Problem addressed

Modern scientific research increasingly depends on the analysis of **massive, complex datasets** that cannot feasibly be downloaded, transferred, or replicated by individual researchers. This challenge is particularly evident in particle physics, where experiments such as those at the Large Hadron Collider generate tens of petabytes of data annually. Similar problems are emerging across other fields, including astronomy, life sciences, and climate science. Addressing these challenges requires a paradigm shift from traditional data movement towards “near-data computation,” where **computational workflows are executed close to the data’s physical location**. There is also a broader need to enable secure, efficient computation across distributed infrastructures and to support sensitive or closed datasets without compromising privacy or legal restrictions.

Technical solution

The solution centres on federated, declarative workflows executed close to the data. Researchers use workflow languages (e.g., CWL, Snakemake) and annotate them with “hints” that direct specific computation steps **to particular data locations within the EOSC Federation**. **REANA orchestrates these workflows** in containerized environments, enabling **seamless execution across multiple computing nodes**. The system supports cross-node workflows, where analyses can span multiple data sources (e.g., simulation data in one country, experimental data in another). Security is a key focus, with mechanisms to vet workflows and prevent misuse. This approach also supports future integration with AI tools and GPU resources for advanced analyses.

Scientific outcomes

One key application of this approach is the reinterpretation of experimental data to **search for phenomena beyond the Standard Model**, such as dark matter, neutrino mass origins, or matter–antimatter asymmetries. By enabling theorists to run their calculations where experimental data are stored, the system fosters **deeper collaboration between theory and experiment** and accelerates scientific discovery. Additionally, the ability to orchestrate workflows across federated nodes enhances interdisciplinary applications, including machine learning model training, sensitive data analysis, and real-time data processing for **next-generation instruments**. Demonstrator projects, such as reproducing aspects of the Higgs boson discovery at CERN, showcase the potential for education, outreach, and **large-scale collaborative science**. Benchmarking studies also demonstrate significant cost and efficiency benefits of near-data computation compared to traditional methods.

Added value of EOSC

The EOSC Federation provides the **federated environment and shared standards** that make this approach possible. By offering unified authentication, resource sharing, and policy frameworks, EOSC enables seamless execution of distributed workflows across national infrastructures. This fosters **data sovereignty** by keeping computation and storage within European resources, while also lowering entry barriers for a broader research community, from educators to data scientists. Ultimately, this federated, near-data computing model represents a **critical evolution in the way science is conducted**, transforming data-intensive science into a more collaborative, efficient, and open enterprise—laying the groundwork for discoveries that were previously beyond reach.