Project title: **End-to-end In-memory Analysis Environments**

Project type: **PhD topic**

Doctoral advisor: **Prof. Dr. Philipp Neumann**

DESY advisor: **Yves Kemp, Christian Voss**

CERN advisor:

Project description:

Modern data science has embraced interactive notebooks (e.g., Jupyter) for exploration, visualization, and iterative analysis. For large-scale HEP workflows, scaling from a single-node notebook to multi-node clusters remains complex. Tools like Coffea/Casa, AnalysisX, Spark, and Dask offer promising solutions but need cohesive integration and user training.

Objectives:

1. **Seamless Jupyter Integration:** Enable “one-click” scaling from local notebook prototypes to large distributed clusters (NAF at DESY, pilot services at CERN).
2. **HEP-Specific Library Integration:** Evaluate, refine, and package libraries like Coffea/Casa or AnalysisX for in-memory HPC settings.
3. **User Engagement & Training:** Create training materials, tutorials, and feedback loops to ensure that both novices and advanced users effectively adopt in-memory columnar workflows.

Methodology & Work Plan:

- **Requirement Gathering & User Studies:** Collaborate with Belle II and other HEP groups to understand typical analysis pipelines, friction points, and feature requests.
- **Notebook-to-Cluster Infrastructure:** Integrate Spark, Dask, and HPC security layers (Kerberos, token-based auth) into user-facing notebook environments.
- **User Training & Support Framework:** Develop specialized training modules, hands-on sessions, and continuous feedback channels.
- **Iterative Improvement & Community Outreach:** Participate in WLCG, OSG, and HPC user groups to share experiences, gather best practices, and ensure broad alignment.

Expected Outcomes & Impact:

- Lower barriers to advanced HPC resources, especially for users transitioning from single-core notebooks.
- Streamlined HEP analysis processes, promoting reproducibility and efficiency.
- Strengthened synergy between DESY, CERN, and the global HEP analysis community.