Suggestions for the rephrasing of the PhD topic

What we had collected before:

- Reconnection feature for Jupyter notebooks and interactive distributed analysis: it is currently possible to run an analysis interactively from a Jupyter notebook and transparently distribute its computations via analysis frameworks that support Dask as a backend (RDataFrame, coffea). However, Jupyter notebooks that are closed while they were running some computation are unable to reconnect to the underlying process when they are opened again. In order to support analysers launching the execution of their analysis, then closing their browser and opening it again later, a Jupyter extension to reconnect to the ongoing computation would need to be developed. This would allow users to know what cell of their notebook were running and whether it finished executing.

- Live plots with distributed analysis: running an interactive analysis distributedly can take some time and resources. Sometimes, it might be useful to know what partial results are being generated in the distributed execution (e.g. histograms) and display them live in the notebook, so that the analyser can decide to cancel such execution if the (partial) results are already not satisfactory. This feature would need to be implemented in an analysis framework such as RDataFrame or coffea.
- Random points:
    - Sustainability: REANA etc.

Original description of the thesis topic:

Project title:

End to End Scalable Concurrent Analysis Environments
Project type:
PhD topic
Doctoral advisor:
Prof. Dr. Philipp Neumann
DESY advisor:
Yves Kemp, Christian Voss
CERN advisor: To be named

Project description:

Modern data science has embraced interactive notebooks (e.g., Jupyter) because they enable researchers and analysts to efficiently explore data, visualize results, and iteratively refine their analyses. For large-scale HEP workflows, scaling from a single-node

notebook to multi-node clusters remains complex. Tools like ROOT RDataFrame, Coffea, AnalysisX, Spark, and Dask offer promising solutions but need cohesive integration and user training.

In the foreseeable future the use of accelerators, namely GPUs, will become essential for many analyses. The shared usage of GPUs offered at a facility and the transition to specialised resources for ML, such as ml.cern.ch, is only in its infancy and far from seamless integration.

Within the HEP community several attempts at providing analysis specific services are being made. The functionality is often very similar, but a transition between systems isn't possible without additional work. This limits the sharing of expertise and code between teams depending on different resource providers.

Objectives:

1. Seamless Jupyter Integration: Enable "one-click" scaling from local notebook prototypes to large distributed clusters (NAF at DESY, pilot services at CERN). This includes the use of resources such as GPUs.
2.HEP-Specific Library Integration: Evaluate, refine, and package libraries like Coffea, RDataFrame or ServiceX for in-memory HPC settings. (discussion on what means In Memory ?? Spark ??)<--- needs clarification ….is this Columnar ??? CLARIFICATION!!!!!
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. (lots of this has been done) .. HPC integration (toke VRE and SWAN).. Not working with HT Condor … Understand what the problem is (Maarten??)
• 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
• Streamlined HEP analysis processes, promoting reproducibility and efficiency.
Strengthened synergy between DESY, CERN, and the global HEP analysis community.

Common approach for the user facing aspects .