Project title: **Columnar Workflow Enablement and Data Formats**

Project type: **PhD topic**

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

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

CERN advisor:

Project description:

HEP analyses often rely on ROOT files optimized for classic batch workflows. Transitioning to in-memory frameworks (Spark, Dask) and common data formats like Parquet allows for efficient columnar queries, but bridging these paradigms requires tool and format adaptations.

Objectives:

1. **Mapping ROOT to Columnar Formats:** Investigate conversion strategies from ROOT to Parquet or other columnar representations suitable for Spark/Dask.
2. **Performance Benchmarks & Optimization:** Evaluate I/O throughput, memory usage, and query performance across different data layouts, cluster configurations, and hardware.
3. **Integration & Compatibility:** Ensure tight integration with HPC resource managers and existing IDAF data architectures, while maintaining backward compatibility for legacy HEP workflows.

Methodology & Work Plan:

- **Data Format Analysis:** Compare existing solutions (ROOT I/O, RDataFrame, Parquet) regarding performance and usability in HPC contexts.
- **Prototype Conversions & Tools:** Develop or adapt open-source tools (e.g., from OSG or WLCG) to facilitate seamless transitions between formats.
- **Extensive Benchmarking:** Use realistic Belle II, ATLAS, CMS, or other HEP workflows to test throughput, memory footprints, and interactive latencies on Spark/Dask clusters.
- **Optimization & Deployment:** Propose best-practice guidelines, assist with production rollout on DESY's IDAF, validating real user needs.

Expected Outcomes & Impact:

- Clear recommendations on data formats for in-memory HPC analyses.
- Improved user workflows, letting physicists handle large datasets interactively.
- Contributions to the broader HEP community through open-source tools and best-practice documentation.