**Advanced Data Set Composition with RNTuple**

RNTuple [1] is ROOT’s next-generation columnar data format and I/O subsystem, aiming at:
- Less disk and CPU usage;
- Efficient use of modern hardware and object stores;
- Modern and robust interfaces.

To streamline analysis and reduce the need for storage resource, experiments may offer common, centrally produced data sets containing only a fraction of all available physics objects, while still covering up to 80% [2, 3] of the analysis use cases.

Data set joins open up the possibility for the remaining use cases to use custom event data without having to duplicate the common data.

---

**Joins with RNTuple**

Joining a primary and a secondary RNTuple requires the construction of an index for the secondary data set. This index maps one or more field values to entry numbers and is used to find entries corresponding to the ones in the primary data set.

**Questions to consider** include:
- How to handle one-to-many relations?
- What is the best representation for the index?
- Should the index be stored (and if so, how)?

The RNTupleProcessor is a work in progress, aiming to:
- Provide a low-level interface to iterate over events;
- Enable joins and vertical concatenations;
- Serve as a backend for RDataFrame and other analysis frameworks.

A crucial design aspect are the composition rules that determine the order in which data sets are processed (i.e., horizontally, vertically or a combination). This will significantly influence processing performance.

Other questions to consider include:
- How to make the composition rules transparent to the user?
- How to deal with data distributed across multiple files?
- How to deal with distributed analysis?

---

**Contact and more information**

florine.de.geus@cern.ch
https://root.cern

---

**References**

