Database Issues Pertinent to Analysis

RD Schaffer

Outline:

- Event model
- Event collections
- Issues
- Where do we stand today?

The Overall Event Model

We have what may look like a fairly traditional event structure:

Raw Data - ~1 MB on average

ESD - Event Summary Data - ~100 KB, reconstruction information

AOD - Analysis Object Data - <10 KB, primary set of info used for analysis

Event Tag - <100 B, first pass selection of events

Event Model, cont.

There are also a few constraints on these elements:

Raw Data:

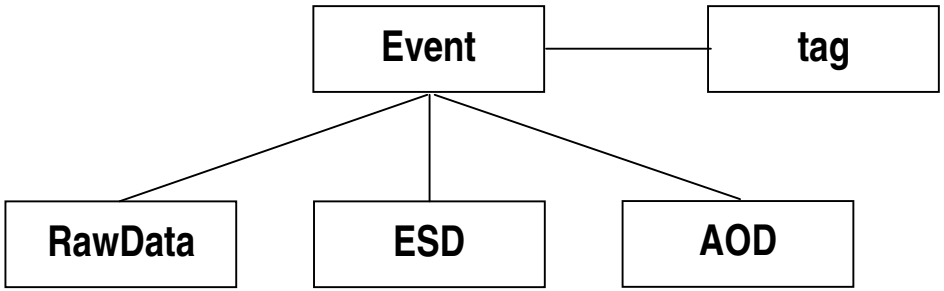
- the bulk will remain at CERN, although a “fraction” may travel outside (Region Center or Institute), e.g. selected events/samples

ESD

- Due to the above constraint, there should be enough detail here redo most of the reconstruction and to (re)generate the AOD.
- Large samples may not travel outside of RC’s, e.g. few TB

Event Model, cont.

A generalised event model will look like:



One should view the event from a logical point of view:

- Via an event (header) object, one can always access the rest of the event information by some form of navigation.

The actual “physical” organization of the different pieces will most likely be different, and may possibly depend upon the event type.

Event Collections

In the OO world, analysis of events will proceed by iterative looping over “event collections”:

One can view that a single collection “holds” a set of events within a particular context:

- a specific physics channel
- belonging to the overall system, physics group, or individual user
- containing priority access to a certain part of the event (I.e. tags, aod, esd, raw data)

Event Collections, cont

A possible analysis scenario could proceed as:

Loop over (a fraction of) an event sample:

- select events according to information in the tags
- navigate to the AOD and histogram various quantities, OR
- navigate to the AOD and select events for an OUTPUT event collection

Some Issues concerning the event and collections

A key point, of course, is to optimize access to the desired event information and as well the turn around time for typical analyses.

There a many handles to respond to this problem:

- Layering access to the different parts of the event
  - e.g. first use tags, then aod, then esd.
- Optimizing the physical layout of the different pieces:
  - organize the AOD objects “column-wise”, or
  - organize the *attributes* of the AOD “column-wise”
  - use various indexing schemes, e.g. for tags
  - recluster sparse events after selection

Some Issues, cont.

To be able to understand how to optimize, it will of course be necessary to understand:

- what are the various analysis scenarios for physics chanels from high pt physics to B physics
- what information and/or parts of the event need to be access for the analyses

So clearly a good round of discussions on requirements is needed.

Some Issues, cont.

Ultimately, I would hope that we can design an event structure which can allow efficient access to the event information needed for EACH analysis without having to store EVERYTHING in a combined ntuple.

Where do we stand?

We are of course fairly early in the game.

For the event model, what has been done up to now concerns access to raw data (digits) needed for the reconstruction developments.

- Access is possible from both zebra and objy

RD45 has developed a model for access to a set of generic tags.

- This has been integrated into HepExplorer, but can be used in other tools.

Where do we stand, cont.

Work has begun on defining the general tools for the intermediate pieces: ESD and AOD

- Once the infrastructure is available, we can begin to decide what goes into the different parts.

Other elements are available from RD45:

- Event collections
- Hierarchical/directory naming facilities
- New lightweight persistency prototype which can be used to storing, e.g. histograms.