Data Format Working Group

ADM Meeting

H. Greenlee

Apr. 16, 2004

Working Group Members

H. Greenlee (chair)

S. Protopopescu

F. Deliot

S. Kulik

A. Lyon

G. Watts

Charge

The lack of a common root-based data format has been one of the problems we have faced in our data handling and physics analyses. This problem not only causes confusion and wastes computing resources, but is also a major source of duplication of effort. To address this issue, we have formed a data format working group with the following members:

Herb Greenlee (Chair), Frederic Deliot, Slava Kulik, Adam Lyon, Serban Protopopescu, Gordon Watts

We ask the working group to

- review currently available root-based data formats and associated analysis algorithms, understand the rationales, pros and cons of each data format;
- develop and implement a root-based data format incorporating desirable features of existing root-based formats and analysis tools, taking into account the needs of algorithm developments and physics analyses as well as the computing resources (storage, access time, how it scales with large dataset etc.) required for analyses.
- Our goal is to make this new format a common root-based format. We plan to produce them centrally. We request the working group to present a plan by May 1, 2004 and to complete its work by June 30, 2004.

We thank all working group members for agreeing to help us on this important project and ask all of you to help them to make it a success.

Gustaaf, Harry and Jianming

Essentials

- Data Format Working Group web page:
  - http://www-d0.fnal.gov/Run2Physics/working_group/data_format/
  - Linked from D0 Physics page.
- Archived mailing list:
  - d0dfwg@fnal.gov

The Current Analysis Environment

- After a painful transition, everyone in D0 is now basing analysis on thumbnails.
- Most people converting tmbs to tuples or trees.
  - tmb_tree.
  - top_tree.
  - Many others.
- Common Sample Group.
  - Tmb fixing/skims.
  - Standard object id./corrections (d0correct).

Problems with Current Analysis Environment

- Growing data set. Thumbnail getting larger (slower & more unwieldy) in p17 (tmb++).
- Some algorithms do not run in d0reco & results can not be stored in thumbnail data tier. Difficult & costly to support these algorithms for different analysis formats. No way to put these algorithms into d0correct, say.
  - b-tagging.
  - Vertexing.
- Duplication of effort wasting human & computing resources.

Proposed Solution to Analysis Problems

- Develop common root-based analysis format (tree/tuple).
- Central production of root files.
- How this will help (it is hoped):
  - Most people won't have to use thumbnails directly, but can use root files (even if they only use them to make new root files). But thumbnails will still be produced by d0reco and stored in sam.
  - Common format for storing output of high-level algorithms.

Physics Group Comments

- Some people said tmb_tree is too slow.
  - Maybe they are trying to read all branches every event?
- Some people said they wanted edm-like interface (edmroot mentioned). Others want simple column-wise ntuple.
- Decoupled from D0 software environment (can run on disconnected laptop, etc.).
- Fast skimming.
- Better documentation.

Design Issues

- Interface.
  - Class-based (like tmb_tree) vs. flat (simple scalars & arrays).
- Portability.
  - Linkage to d0library.
- Self-describing.
  - Browsable?
  - MakeClass?

Design Issues (cont.)

- Completeness.
  - How much of thumbnail to include?
- Customizable/extensible.
  - Adding/dropping branches.
- Performance.
  - Speed.
  - Size.
  - Skimming.

Design Issues (cont.)

- High-level/post-tmb algorithms.
  - Certified corrections (d0correct).
  - Primary vertices using beam position database.
  - Secondary vertices.
  - B-tagging.
  - Resonances.
- Need better support for reading root files in sam.

Status & Plans

- Data Format Working Group is currently reviewing existing analysis formats.
- We are currently soliciting collaboration input.
  - Physics groups polled, results on web page.
  - Open discussion at next week's CSG meeting (Wednesday, 10:30).
  - Send comments to d0dfwg@fnal.gov.
- First goal is to produce document with concrete design proposal.
  - Charge specifies deadline of May 1 for plan.
- Implentation follows.
  - Charge specifies deadline June 30.