

# Event Data Model

ATLAS Software Workshop  
Reconstruction Task Force Session

David Adams

BNL

March 5, 2003



David Adams  
**BROOKHAVEN**  
ATLAS NATIONAL LABORATORY

# Contents

What is EDM?

Persistency

Data access

Granularity

Using the granularity

- File placement
- Dataset/Event collection
- Object replication
- Incomplete objects
- Virtual data

Conclusions



# What is EDM?

EDM (Event Data Model) is the schema for event data including

- Detector (raw data)
- Simulation
  - Generated particle, secondaries, hits, ...
- Reconstruction
  - Clusters, tracks, jets, physics objects
- Summary (AOD)
  - Event attributes
  - For rapid event selection and most analysis



# Persistency

The EDM is persistent to record the state of the processing to enable

- Easy verification of intermediate results
- Different processes to handle different stages of processing
  - Different executables
  - Different sites
- Redoing a later stage of processing without redoing earlier stages
  - Refit tracks without repeating track finding



# Data access

Access time for the data of interest is often the determining factor in the time required to carry out a job

- Limiting how often reconstruction can be repeated as algorithms evolve
- Limiting how much data a physicist can examine during an analysis session
  - Thus limiting discovery potential

# Data access (cont)

Time for data access depends on

- Catalog(s)
  - Time required to find the data
- Location
  - Nearby data is accessed more quickly
  - Replicate to optimize for different locations
- Placement
  - Data may be accessed more quickly if it is close to other data that has already been accessed
    - > E.g. for fixed-size files, minimize the # files required to access the data of interest



David Adams

**BROOKHAVEN**  
NATIONAL LABORATORY

Event Data Model

RTF session

March 5, 2003 6

# Granularity

The granularity for the data affects these features in competing ways

- Finer granularity allows
  - more precise placement and
  - more replication (datasets are smaller)
- Coarser granularity
  - reduces catalog time by reducing the number of entries

# Granularity (cont)

There are some natural boundaries for the components of the data model

- Event (beam crossing) ID
- Category (raw, MC, reco, summary)
- Algorithm output
  - An EDM object is associated with one algorithm instance (see figure)

Finer divisions are possible

- Algorithm may produce two EDM objects
- No reason to split data that is accessed jointly



David Adams

**BROOKHAVEN**  
NATIONAL LABORATORY

Event Data Model

RTF session

March 5, 2003 8

# Using the granularity

The persistency service should provide the following to allow use of the granularity:

- File placement
- Dataset/Event collection
- Object replication
- Virtual data
- Incomplete objects

# File placement

Place data that is likely to be used together in the same file(s), excluding other data

- Exclusive event streaming
  - E.g. data with  $>3$  jets, 2 leptons and  $mET > 200$  go into one stream (series of files)
- Content placement
  - Summary data in one file stream
  - Reco data in another
    - > Or tracking in one and
    - > Jets in another



David Adams

**BROOKHAVEN**  
NATIONAL LABORATORY

Event Data Model

RTF session

March 5, 2003 10

# Dataset/Event Collection

Specify a collection of data objects that belong together for analysis

- Explicit object identifiers (event collection)
- List of files (DC dataset)
- Filter another dataset
  - Event selection
  - Content selection
    - > E.g. summary data
    - > Or tracking data
- Merge two or more datasets



# Object replication

Placement cannot be optimized for all production and analysis.

Replicate objects of interest for samples that are to be processed more than once

- Event selection
  - E.g. Higgs candidate events
- Content selection
  - Only summary data or
  - Only tracking data

# Incomplete objects

If only part of an object is of interest, we may replicate that part and discard the rest.

- Assume object is a container
- Copy must be smart enough to use original indices for retained objects and return error if an absent object is requested
- E.g. keep
  - Tracks with  $p_T > 5 \text{ GeV}/c$ 
    - > Might keep partial info for other tracks
  - Clusters used on tracks
- Special instrumentation of data objects



David Adams

**BROOKHAVEN**  
NATIONAL LABORATORY

Event Data Model

RTF session

March 5, 2003 13

# Virtual data

Virtual data is (re)generated on demand

- Record algorithm and its input data objects rather than data
- On demand for data, produce data by applying algorithm to input data objects
- Produced data should be very reproducible
- E.g. clusters created by unpacking raw tracking data and applying a simple clustering algorithm
- Requires good provenance info



# Conclusions

Data access can be optimized by

- Choosing an EDM with the appropriate granularity
- Creating a persistency system which takes full advantage of this granularity

Requirements for reconstruction

- OO EDM with appropriate granularity
- Special instrumentation for incomplete objects
- Provenance for virtual data



David Adams

**BROOKHAVEN**  
NATIONAL LABORATORY

Event Data Model

RTF session

March 5 , 2003 15