Data Preservation at the Tevatron

Ken Herner, Bo Jayatilaka

Fermilab

21 March 2013

Reminder of goals

- The Run II DP Project aims for Level-4 preservation at both CDF and D0 (full analysis capability, including new MC generation), ideally until at least 2020
  - Project is not looking at access by general public
- Experiment-specific resources will shrink over next few years
- Challenge: Preserve full capability via modifying infrastructure/migrating to shared FNAL resources
  - Will require everything to run on Scientific Linux >=6

What is needed

- To perform a CDF or D0 analysis in the future (post-2015):
- Data access
- Batch facility
- MC generation tools
- “Standard” analysis framework(s)
- Concise and current documentation. Able to run everything without asking for help!

Job submission and data access

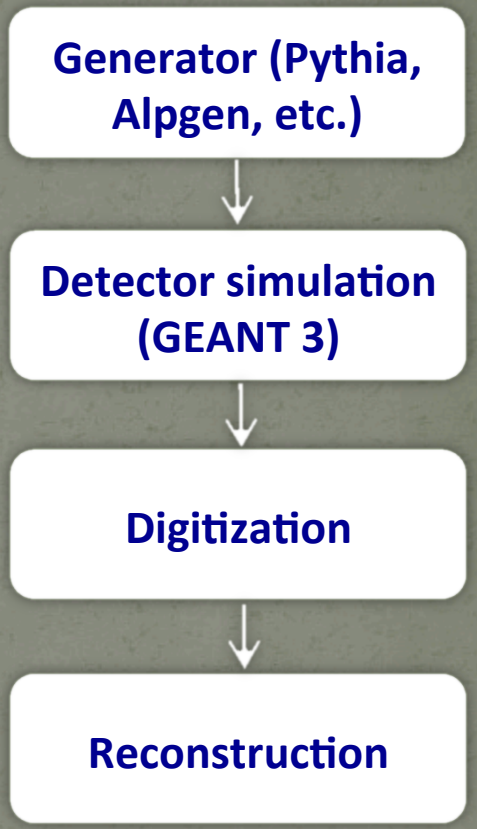
- D0 has PBS-based dedicated cluster for most user analysis jobs; capacity will steadily diminish and disappear in 2015
- D0 is examining altering infrastructure to either Grid-based submission system (similar to CDF, CMS, IF experiments) or D0 cluster-like VMs spawned on Fermilab cloud
  - Pros and cons to both
  - Expect to converge by end of this year
- Data access and file delivery to analysis jobs is via SAM (Sequential Access via Metadata)
  - Current system will work until 2015
  - Have already modified D0 software to use http-based SAM access developed for IF experiments (more detail In tomorrow's talk); CDF will do this shortly
- CDF making full copy of data at INFN
  - Enables easier access for European Grid nodes
- ~10 PB of "data" per experiment, migrating to modern tape technology

CVMFS

- CVMFS an attractive option for future job submission
  - Can't guarantee all grid nodes have D0 software installed
  - CVMFS will reduce load on FNAL resources (don't need to copy EVERYTHING over all the time)
- FNAL has set up a test server for D0
  - Expect good progress in the next few months
- D0 has a graduate student supported by DASPOS (@UW); will be involved in this effort
  - Good test case for users setting up software at home institutes
- CDF also exploring CVMFS

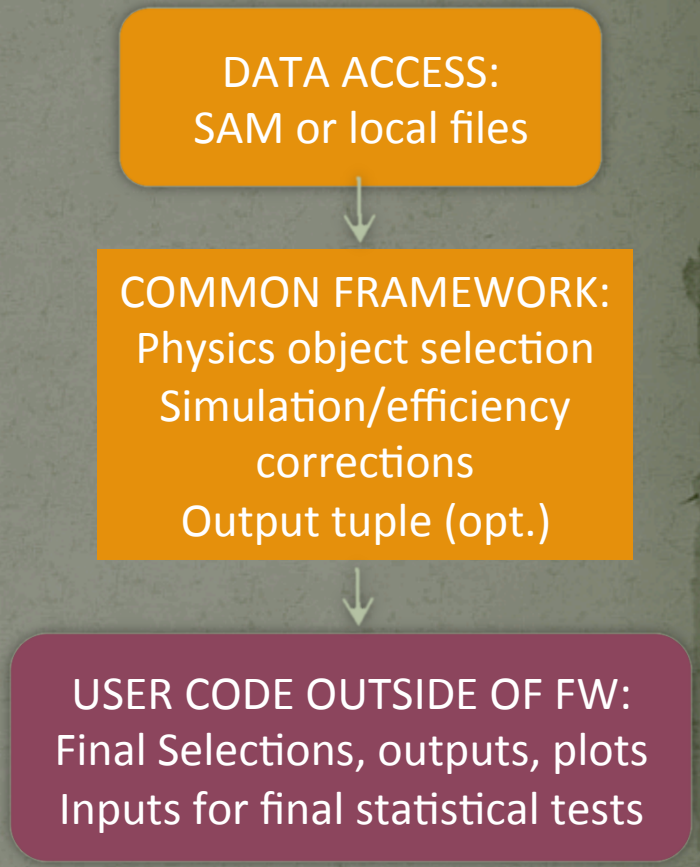
MC Software Chain Validation

- Four steps to MC production
- Critical to retain this capability
- Existing software has been verified to run on SL6 machines (within CDF and D0 release environments)
- Support for newer generators and/or PDFs available (can run GEANT and onward with any LHA-formatted generator output)
- Will need to ensure continued calibration DB access (Oracle DBs) throughout DP project (more Fri.)



Plans for D0 analysis software

- Two "main" frameworks (V+jets, b-physics)
- Plans to make sure they work
- Steps in orange: all code in CVS; DP project will guarantee that they work
- Plum: outside of project scope; has always been user's responsibility
- Have verified that current release works within SL6
- SW release usually 100-200 GB counting UPS products



Plans for CDF Software

- Developing a "legacy release" for DP use
  - Built on SL5 now; executable on SL6
  - Eventually built on SL6
  - Eliminate all pre-SL5 dependencies (compatibility libraries, etc.)
- Enable user to do everything that can be done now
- Aiming to drop ups products that are redundant with OS-supplied products (e.g. Python)
- Planning to finalize release in summer
  - Release is ~30 GB + UPS products when built

Documentation

- Want user to be able to run chain with only documentation as a guide
- Developing concise set of instructions on job submission and running a simple analysis through the common analysis tools
- Also need access to internal notes so that analyzers can look at previous analyses for insights
  - Internal Notes, Agenda server: D0 moves to long-term supported archives completed; CDF moves progressing well
- Detector/online info: Migrating logbooks and DBs to supported software (read-only in some cases), underway
- More detailed analysis tests
  - Validation analyses: work with physics groups to provide step-by-step (extensible) instructions on running from beginning to end
- Internal Mailing lists/group discussions/minutes: catalog everything to be saved, work with FNAL listserv admins to make sure everything is ported to any future system (probably read-only)
- Wiki/Twiki pages: convert to static pages once need for write access is gone

Summary

- Goal of the Run II DP Project is to ensure preservation of not only the data itself but analysis capability
- Must preserve everything that a user needs now, and document everything in a clear, concise way
- CDF and D0 modifying job submission and data access infrastructures for long-term support
- Modifying code bases as needed and ensuring smooth performance at least through 2020 (end of SL6 support)