
Contribution ID: **34** Type: **poster**

# A Data Skimming Service for Locally Resident Analysis Data

*Wednesday, 5 September 2007 08:00 (20 minutes)*

A Data Skimming Service (DSS) is a site-level service for rapid event filtering and selection from locally resident datasets based on metadata queries to associated "tag" databases. In US ATLAS, we expect most if not all of the AOD-based datasets to be be replicated to each of the five Tier 2 regional facilities in the US Tier 1 "cloud" coordinated by Brookhaven National Laboratory. Entire datasets will consist of on the order of several terabytes of data, and providing easy, quick access to skimmed subsets of these data will be vital to physics working groups. Typically, physicists will be interested in portions of the complete datasets, selected according to event-level attributes (number of jets, missing E_t, etc) and content (specific analysis objects for subsequent processing).
In this paper we describe methods used to classify data (metadata tag generation) and to store these results in a local database. Next we discuss a general framework which includes methods for accessing this information, defining skims, specifying event output content, accessing locally available storage through a variety of interfaces (SRM, dCache/dccp, gridftp), accessing remote storage elements as specified, and user job submission tools through local or grid schedulers.
The advantages of the DSS are the ability to quickly "browse" datasets and design skims, for example, pre-adjusting cuts to get to a desired skim level with minimal use of compute resources, and to encode these analysis operations in a database for re-analysis and archival purposes. Additionally the framework has provisions to operate autonomously in the event that external, central resources are not available, and to provide, as a reduced package, a minimal skimming service tailored to the needs of small Tier 3 centers or individual users.

**Primary author:** MAMBELLI, Marco (University of Chicago)

**Co-authors:** MALON, David (Argonne National Laboratory); CRANSHAW, Jack (Argonne National Laboratory); GIERALTOWSKY, Jerry (Argonne National Laboratory); EDWARD, May (Argonne National Laboratory); GARDNER, Robert (UNIVERSITY OF CHICAGO)

**Presenter:** MAMBELLI, Marco (University of Chicago)

**Session Classification:** Poster 2

**Track Classification:** Distributed data analysis and information management