Object storage model for CMS data

Nick Smith (FNAL)
On behalf of the CMS Collaboration

FERMILAB-POSTER-24-0303-CMS-CSAID

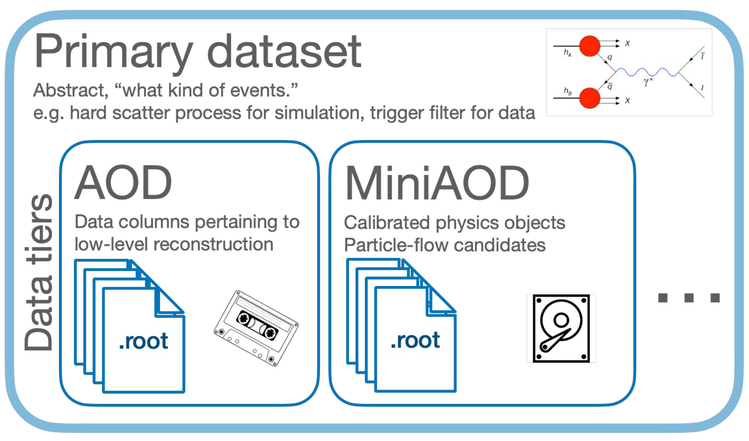
Background

In CMS, data access and management is organized around the data-tier model: a static definition of what subset of event information is available in a particular dataset, realized as a collection of files. In previous work[1], we have proposed a data management model that obviates the need for data tiers by exploding files into individual event data product objects. In this work, we estimate the potential savings in data volume based on user analysis patterns.

CMS Data Model

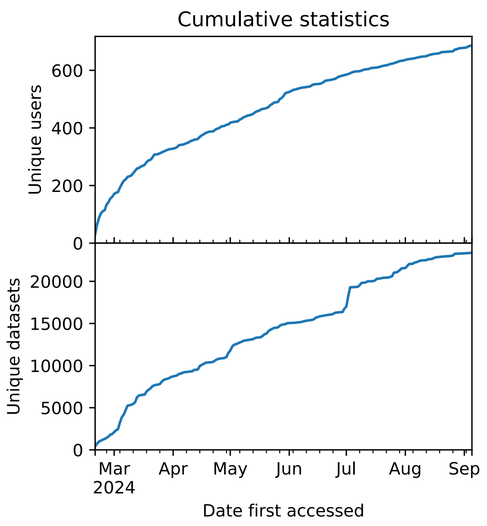
In CMS, event data is stored in datasets, which are collections of ROOT files. The content of a given dataset depends on its data *tier.* Through the CMS Remote Analysis Builder (CRAB) system, users can run CMS framework jobs accessing any data tier available on disk for a given dataset. For analysis tiers (*AOD), automatic recall from tape is possible.

Depiction of tiered data model in CMS, showing typical quality of service (disk or tape) for the largest analysis data tiers, AOD and MiniAOD. Each ROOT file stores all data columns in the tier, for a subset of all events in the primary dataset.



Within the CMS framework, only the subset of *data products* necessary for the analysis algorithm are read from the storage system. Each data product may be a complex C++ type. Though the ROOT TBranch representing the on-disk data layout may be split for better compression, it is either read entirely or not at all. The framework records which branches are accessed in each job.

CRAB per-branch popularity source

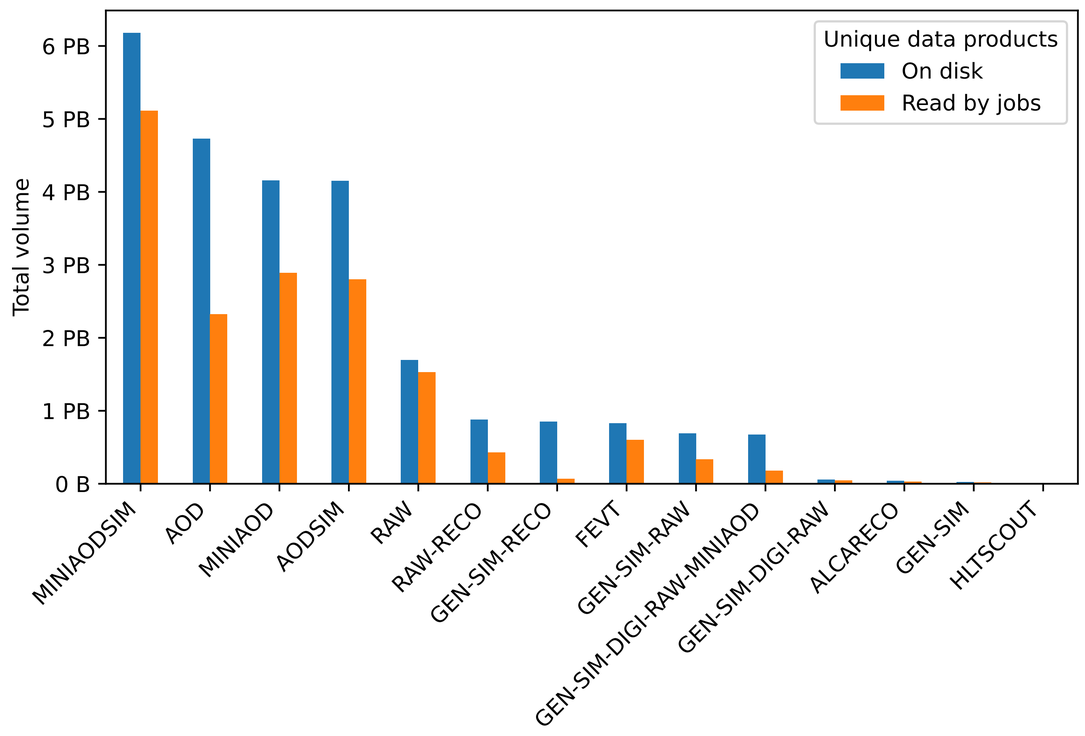


CMS users collectively submit an average 1200 workflows per day through the CRAB workflow management system, accessing over 20k unique datasets over the last 6 months, as shown left.

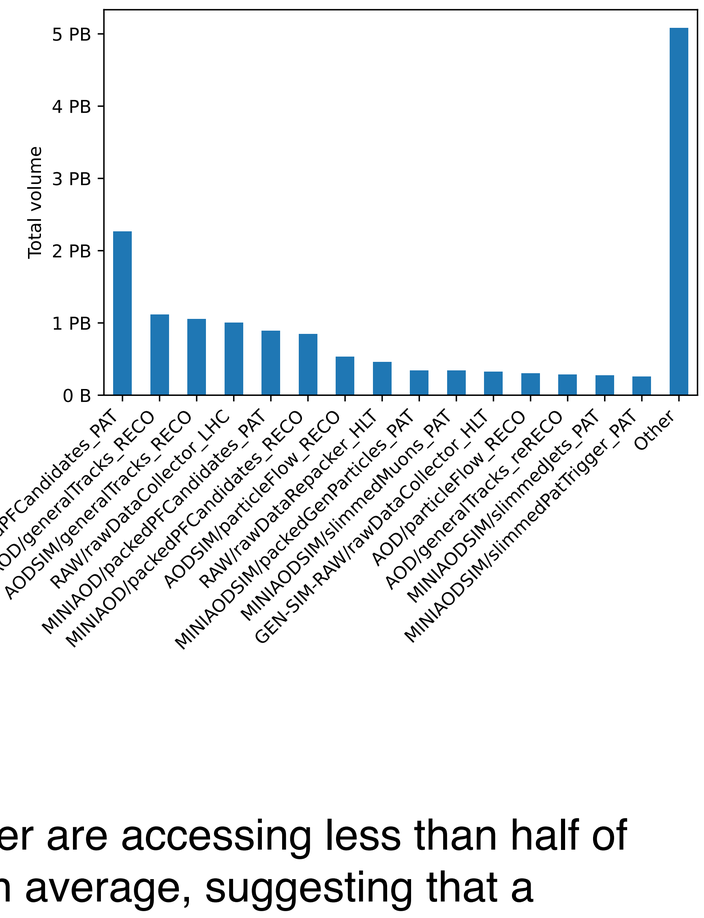
Since February 2024, CRAB records a list of branches accessed per workflow, allowing a sub-dataset popularity analysis of event data for user workflows. This data source was implemented by Dario Mapelli (CERN) and Stefano Belforte (INFN Trieste).

Popularity analysis

The compressed size per branch is extracted from a representative file for each dataset. Then, for each dataset, the union of branches accessed by any user task is found. The size of this set is then summed over all datasets, grouped by data tier, to estimate the minimum volume necessary to serve analysis needs, shown in orange. This is compared to to the volume of all branches per accessed dataset, summed over datasets. This volume, shown in blue, is what must be held on disk in the current data model.



To right, the total unique volume of data accessed by user analysis, grouped by data tier and data product branch, is shown. The largest volume accessed by CMS users is the packedPFCandidates data product found in MiniAOD, which stores compressed kinematic information and limited particle ID information for all particle flow candidates in an event.



Conclusions

Users of the AOD data tier are accessing less than half of the data stored in files on average, suggesting that a refinement of the data tier definition or a data management model that stores each data product separately could yield significant savings in disk usage. For MiniAOD, users read 80% of the data, so potential savings are less significant.

Ref. 1: EPJ Web of Conferences 295, 01003 (2024)
This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.