Table of Contents

Tier-0 project Policies and Procedures

Project Description

Tier-0 is responsible for

- Data preservation
  - Repacking: converting the streamer files written out at Point 5 into CMSSW EDM files and organize them into datasets based on the triggers that the events passed
- Express processing
- Prompt Reconstruction

Data Processing

Tier0 is a time sensitive service that is not designed to handle permanent failures. Any new Tier-0 setup is tested in "replay" mode to make sure that it can process data without failures. If later during actual data processing a permanent failure occurs the jobs are marked failed and corresponding lumi sections won't be available in output. It is expected that such failures will be recovered during data reprocessing in the central production, which is designed to handle failures more efficiently.

Policies

Dataset Naming Conventions for Tier-0 Data

Please refer to the dataset naming policies for further information on the general conventions for dataset names for CMS data: Dataset naming policy

Current Tier0 config: You can find the current Tier0 configuration at ProdOfflineConfiguration

These are the current WMAgent based Tier-0 production system dataset naming conventions:

Express datasets

- /[ExpressDataset]/[AcquisitionEra]-Express-[ExpressProcVersion]/FEVT
- /[ExpressDataset]/[AcquisitionEra]-Express-[ExpressProcVersion]/DQM
- /[ExpressDataset]/[AcquisitionEra]-[AlcaProducer]-Express-[AlcaRawProcVersion]/ALCARECO

Repack datasets

- /[PrimaryDataset]/[AcquisitionEra]-[RawProcVersion]/RAW

PromptReco datasets

- /[PrimaryDataset]/[AcquisitionEra]-PromptReco-[RecoProcVersion]/RECO
- /[PrimaryDataset]/[AcquisitionEra]-PromptReco-[RecoProcVersion]/AOD
- /[PrimaryDataset]/[AcquisitionEra]-PromptReco-[RecoProcVersion]/DQM
- /[PrimaryDataset]/[AcquisitionEra]-[AlcaProducer]-PromptReco-[AlcaRawProcVersion]/ALCARECO

Run Coordination

Data are taken at P5 in two primary configuration of the data acquisition system: **global-run** and **mini-daq**. Mini-daq is mostly used for detector studies and is setup and configured by detector experts. Global runs are

taken by central shifters and normally include all components of the CMS detector. Data that have been acquired using global-run have higher priority for processing than the ones acquired using mini-daq. In the case of exception the request needs to be made by Run Coordination.

If Tier-0 has a problem that prevents it from processing data and Express outputs are delayed for more than 5 hours, notify P5: hn-cms-commissioning@cernNOSPAMPLEASE.ch. Single failures should be reported in a regular way.

Contact: cms-run-coordinator@cernNOSPAMPLEASE.ch

Storage Manager

The Storage Manager project is responsible for data transfer from P5 to Tier-0 storage. The data are distributed in a form of streamer files. The ownership of the data is transferred from Storage Manager to Tier-0 after Tier-0 confirms that all data were received (what is called "checked files"). Until that moment the Storage Manger team is solely responsible for the data preservation.

As a **backup** option Storage Manager tries to avoid deleting data at P5 till Tier-0 successfully repacks its copy, but it's not a requirement and data at P5 can be deleted after the ownership was transferred to Tier-0. Current deletion policy for checked but non-repacked files at P5:

- delete files older than 3 days if we reach 40% lustre occupancy
- delete files older than 2 days if we reach 50% lustre occupancy
- delete files older than 1 day if we reach 60% lustre occupancy
- delete files older than 12h if we reach 70% lustre occupancy
- delete files older than 6h if we reach 80% lustre occupancy
- delete files older than 3h if we reach 90% lustre occupancy

In addition, there is another cleanup cron which runs twice a day. This cleanup does not contact the database, but it simply removes old runs regardless of their status. We try to be extra cautious with the age of the runs that we delete using this script to try and avoid as much as possible accidental deletion of data that might still be needed. The delays are as follows:

- delete runs older than 30 days if the lustre occupancy is below 30%
- delete runs older than 15 days (or 12 days if the runs were taken with Tier0_Off flag) if the lustre occupancy is between 30% and 40%
- delete runs older than 7 days if the lustre occupancy is above 40%

https://twiki.cern.ch/twiki/bin/view/CMS/StorageManager

Contact: cms-storagemanager-alerting@cernNOSPAMPLEASE.ch

Data Handling

File Size Limitations

Tier0 processing has the following limits on the file size:

- RAW file (output of repacking) size limits:
  - Soft limit is **16GB** - only a small fraction of files can cross that limit and it's considered to be an emergency. Tier0 will notify Run and HLT coordination when it happens.
  - Hard limit is **24GB** - any output larger than that is excluded from the nominal dataset and assigned to an Error one. No further processing is done. Requires a special treatment to be used if needed.

- Input streamer file size
  - ~ **30GB** is the limit set in by the Storage Manager team to files automatically transferred to Tier0 processing. Larger files can be transferred on an explicit request from the Run Coordinators. It may cause problems in Tier0 processing, so Tier0 ops need to be notified when it happens.
    - https://github.com/cmsdaq/merger/blob/master/cmsActualMergingFiles.py#L20
  - ~ **80GB** is absolute file size limit. Files larger than that won't be repacked.

Data Distribution

Tier-0 subscribes RAW data based on the available tape space at Tier-1 sites. All other data tiers including those used by users (AOD, MINIAOD etc) are subscribed to the same site where the RAW data are subscribed. The distribution is done a few types per year for known PDs. All new PDs are subscribed to FNAL by default.

Dynamic Data Management

Tier-0 subscribes datasets on disk under AnalysisOps group. DDM takes care of replication of the data to Tier-2 sites.

Temporary data is not subscribed and Tier-0 takes care of cleaning up. Unified and DDM have mechanisms to automatically subscribe unsubscribed data. They should not touch Tier-0 data or it may cause a data loss.

As an additional protection Tier-0 publishes lock files that protect data in DataOps group.

Lock files:

- https://cmsweb.cern.ch/t0wmadatasvc/prod/dataset_locked
- https://cmsweb.cern.ch/t0wmadatasvc/replayone/dataset_locked
- https://cmsweb.cern.ch/t0wmadatasvc/replaytwo/dataset_locked

DDM policies

- https://github.com/SmartDataProjects/dynamo-policies/blob/master/detox/Physics.txt
- https://github.com/SmartDataProjects/dynamo-policies/blob/master/detox/Unsubscribed.txt - Unsubscribed data

Data Subscription

All Tier-0 subscriptions to disk are done under AnalysisOps group.

Data transfer priority

Tier-0 output has higher priority for transfers than other activities in CMS due to low latency requirements for Express and Prompt data availability.

Streamer file deletion

Summary

Streamer files are deleted in groups of run/stream (smallest run numbers first) if any of the following requirements are met:

- all files in run/stream are older than 365 days
- run/stream is successfully repacked (RAW data are produced) or Express processed

Details

We run a daily cron job that evaluates /eos/cms/store/t0streamer and removes (if required) streamer files in blocks of run/stream

```
0 5 * * * lxplus /afs/cern.ch/user/c/cmsprod/tier0_t0Streamer_cleanup_script/analyzeStreamers_pro
```

The script does not look directly into the EOS namespace, but instead makes use of the daily EOS file dump in /afs/cern.ch/user/p/phedex/public/eos_dump/eos_files.txt.gz

There are two separate policies it applies

/eos/cms/store/t0streamer/[TransferTest|TransferTestWithSafety]

- files older than 7 days will be deleted (immediately)

/eos/cms/store/t0streamer/[Data|Minidaq]

- scans for all files and calculates aggregates for all streamer files for the same run/stream: total size and modification time for newest streamer file

There is a list of runs for which we skip generating stats for streamer files in the /eos/cms/store/t0streamer/Data director, they will not be considered for deletion later.

Then the script compares a hard-coded quota (currently 75% of 1PB, so 750TB) to how much data we have in total and starts deleting until we are below quota or no more files can be deleted.

Deletions happen in run order and units of run/stream, assuming the run/stream meets these criteria:

- Tier0 Data Service run_stream_done method returns True for given run/stream to be done OR
- modification time for newest streamer file is more than 365 days ago

Alignment and Calibration

Oracle Accounts

Tier0/ProdAgent Archive Accounts

These are copies of old Tier0 production accounts. We keep them in case there are questions about how the Tier0 was processing past runs. If password is expired the owner can reset it.

Database	Time period	Min Runnumber	Max Runnumber	Owner
CMS_T0AST_PROD_1@CMSARC_LB	Spring 2010 to September 2010	126940	144431	Dima
CMS_T0AST_PROD_2@CMSARC_LB	September 2010 to December 2010	144461	153551	Dima
CMS_T0AST_PROD_3@CMSARC_LB	February 2011 to May 2011	156419	164268	Dima
CMS_T0AST_PROD_4@CMSARC_LB	May 2011 until December 2011	164309	183339	Dima
CMS_T0AST_PROD_5@CMSARC_LB	2012	184147	192278	Dima
CMS_T0AST_PROD_6@CMSARC_LB	2012	192260	197315	Dima

CMS_T0AST_PROD_7@CMSARC_LB	2012	197417	203013	Dima
CMS_T0AST_PROD_8@CMSARC_LB	2012	203169	209634	Dima
CMS_T0AST_PROD_9@CMSARC_LB	2013	209642	212233	Dima

Tier0/WMAgent Archive Accounts

These are copies of old Tier0 production accounts. We keep them in case there are questions about how the Tier0 was processing past runs. Contains mostly configuration related information since most file, job, workflow, subscription related info is deleted when workflows are cleaned up. If password is expired the owner can reset it.

Database	Time period	Min Runnumber	Max Runnumber	Owner
CMS_T0AST_WMAPROD_1@CMSARC_LB	early 2013	209642	212233	Dima
CMS_T0AST_WMAPROD_2@CMSARC_LB	GRIN 2013	216099	216567	Dima
CMS_T0AST_WMAPROD_3@CMSARC_LB	AGR 2014	220647	221318	Dima
CMS_T0AST_WMAPROD_4@CMSARC_LB	Fall 2014	228456	231614	Dima
CMS_T0AST_WMAPROD_5@CMSARC_LB	Winter 2015	233789	235754	Dima
CMS_T0AST_WMAPROD_6@CMSARC_LB	Spring 2015	233998	250588	Dima
CMS_T0AST_WMAPROD_7@CMSARC_LB	Summer 2015	250593	253626	Dima
CMS_T0AST_WMAPROD_8@CMSARC_LB	Fall 2015	253620	261065	Dima
CMS_T0AST_WMAPROD_9@CMSARC_LB	HI 2015 production replay	262465	262570	Dima
CMS_T0AST_WMAPROD_10@CMSARC_LB	HI 2015	259289	263797	Dima
CMS_T0AST_WMAPROD_11@CMSARC_LB	2016	264050	273787	Dima
CMS_T0AST_WMAPROD_12@CMSARC_LB	2016	273799	277754	Dima
CMS_T0AST_WMAPROD_13@CMSARC_LB	2016	277772	282135	Dima
...	...	...	...	...
CMS_T0AST_WMAPROD_25@CMSARC_LB	2018	315252	316995	Dima
CMS_T0AST_WMAPROD_26@CMSARC_LB	2018	316998	319311	Dima
CMS_T0AST_WMAPROD_27@CMSARC_LB	2018	319313	320393	Dima
CMS_T0AST_WMAPROD_28@CMSARC_LB	2018	322680	322800	Dima
CMS_T0AST_WMAPROD_29@CMSARC_LB	2018	325112	325112	Dima
CMS_T0AST_WMAPROD_30@CMSARC_LB	2018	325799	327824	Dima
CMS_T0AST_WMAPROD_32@CMSARC_LB	2018	320413	325746	Dima

Tier0/WMAgent Replay Accounts

These accounts are used for Tier0 replays. As for March 2019, all T0 replay machines are 32cores, 60GB, CC7 VMs.

Database	Replay vobox	Owner
CMS_T0AST_REPLAY1@INT2R	vocms001	Dima
CMS_T0AST_REPLAY4@INT2R	vocms015	Dima
CMS_T0AST_REPLAY3@INT2R	vocms047	Dima
CMS_T0AST_REPLAY2@INT2R	vocms0500	Dima

Tier0/WMAgent Production Accounts

These accounts are used for Tier0 production. As for March 2019, all T0 production machines are 32cores, 60GB, CC7 VMs.

Database	Headnode	Owner
cms_t0ast_1@cmsr	vocms0314	Dima
cms_t0ast_2@cmsr	vocms0313	Dima
cms_t0ast_3@cmsr	vocms014	Dima
cms_t0ast_4@cmsr	vocms013	Dima

Tier0 Data Service Accounts

Accounts for cmsweb to report first condition safe run.

Database	URL	Owner
cms_t0datasvc_prod@cmsr	https://cmsweb.cern.ch/t0wmadatasvc/prod/firstconditionsaferun	Dima
cms_t0datasvc_replay1@int2r	https://cmsweb.cern.ch/t0wmadatasvc/replayone/firstconditionsaferun	Dima
cms_t0datasvc_replay2@int2r	https://cmsweb.cern.ch/t0wmadatasvc/replaytwo/firstconditionsaferun	Dima

-- VytautasJankauskas - 2019-03-27

This topic: CMSPublic > CompOpsTier0Policies
Topic revision: r37 - 2019-03-27 - VytautasJankauskas