
CMS Internal Note

The content of this note is intended for CMS internal use and distribution only

October 1, 2004

Report of the CMS Data Management RTAG

Final Version 7

Edited by Robert M. Harris

Authors

Vincenzo Innocente, Massimo Lamanna, Emilio Meschi and Norbert Neumeister
CERN, Geneva, Switzerland

Lothar Bauerdick, Robert M. Harris, and Avi Yagil
Fermilab, Batavia, IL, USA

Lucia Silvestris
INFN, Bari, Italy

Stefano Lacaprara
INFN, Padova, Italy

Abstract

We report the recommendations of the CMS Data Management RTAG.¹⁾ This group considered the data usage patterns of CMS and explored some current models of data management in HEP (BaBar, CDF, D0, ZEUS). We found significant overlap in the data usage patterns of CMS and the Tevatron experiments in Run II. We propose a model of data organization at CMS intended for the first years of LHC operations, that is similar to the one employed by the Run II experiments. We recommend the formation of a dedicated data management project within CMS.

¹⁾ This note expresses the conclusions of the Data Management RTAG. It does not necessarily represent official CMS policy.

Contents

1	Introduction	3
2	Basic Assumptions	3
3	Recommended Data Organization Model	3
3.1	Physical Partitioning I: <i>Streams</i>	4
3.2	Physical Partitioning II: <i>Production Datasets</i>	4
3.3	Events and Data Files	5
3.4	Data Tiers	5
3.5	Skims	6
3.6	Databases	6
4	Usage Patterns within the Proposed Model	6
4.1	Re-Process Sample or <i>Production Dataset</i>	6
4.1.1	Whole Sample by Offline Group	6
4.1.2	Dataset by Physics Group	6
4.2	Validate a Dataset	6
4.3	Define an Analysis Dataset - Skim	7
4.4	Access to Data for Analysis	7
4.5	Production of a MC Sample	7
4.6	Extract DB Information at Analysis Level	7
5	Summary of Main Recommendations	7
5.1	Size of a Raw CMS Event	8
5.2	Splitting the Data into Streams and Datasets	8
5.3	Storage/Retrieval of Data	8
5.4	Data Management Project	9
A	The RTAG Process	9
A.1	Motivation	9
A.2	Mandate	9
A.3	Membership	10
A.4	Meetings	10
B	Glossary	11

1 Introduction

The CMS Data Management RTAG was charged to produce a plan for data management at CMS. We have evaluated the anticipated requirements of CMS data handling during the first years of data taking. In view of this we have studied the data usage patterns of current HEP experiments. In this report we were guided by the use cases and approaches to data management of the RunII experiments D0 and CDF at the Tevatron¹⁾.

This document is organized as follows: Section 2 discusses some basic assumptions that define the scope of the data management project. Section 3 describes the recommended model of data organization. Section 4 discusses a set of important usage patterns of CMS data and how they are achieved in the proposed model. In Section 5 we summarize our main recommendations. Details of the RTAG process, including charge, membership, meetings, minutes and supporting documents are discussed in Appendix A. Since this document uses many terms that may have different meanings to different persons based on their background, we include a glossary of some of the key terms words we use in Appendix B. These words are printed in *italics* whenever we use them in the body of the document.

Finally, it should be noted that this document is written mostly from the perspective of the physicists using the data, not strictly from the technical perspective of implementing the CMS data management project. We recommend the formation of a data management project to formulate the detailed use cases and hard requirements and to design and implement the recommended data management system.

2 Basic Assumptions

At the very beginning of our deliberations we had to decide on some “boundary conditions” relating to very basic issues concerning the expected running conditions. These impact directly the scope of the project as well as the affect the choices one may consider. We made every effort to find the “best information” available within CMS, including many informal conversations with various detector and online experts.

1. **Running Period:** We consider the first one or two years of LHC running. That is, we do not consider the first weeks during the early commissioning of CMS, nor the out years of high luminosity stable running. This document addresses the initial low luminosity, steady running period aimed at obtaining the first physics results and possible discoveries. The mode of operation may be very different both during the turbulent first weeks of commissioning or during stable mature running after the experience gained in the first few years.
2. **Event Size:** The projected event raw data size is approximately 0.3 MB/event during the running period considered. We note that this estimate is consistent with that given in the DAQ TDR, but it differs from numbers used in various later presentations that seem to prefer a number of 1 MB/event. We also note that the event raw data size may depend on the running period, and is expected to be larger in the commissioning phase, decreasing in early stable operations, and increasing again with increasing luminosity in the out years.
3. **Event Rate:** CMS expects to write to tape events at a rate of 100 Hz. We note that a projected filter farm bandwidth of 100 MB/s, with 300 KB events, could allow for a higher event rate. We expect the extra rate above 100 Hz would populate event samples at lower p_t . In the proposed data organization scheme, this would not have a strong impact on first physics and discovery at event samples of high p_t .

In case that these assumptions would change in significant ways, the recommendations in this report would necessarily have to be revisited.

3 Recommended Data Organization Model

In this section we describe the proposed organization of the CMS data; discuss how the complete *data sample* is split into functional and manageable *streams* and *datasets* and stored in *files*. Mention (some of the) expected *event data formats* *data tiers* as well as *skims*.

¹⁾ We recognized that for example CDF, in its current run, is writing events to tape at a sustained rate of 70 Hz. CDF has a raw event data size of 200 KB, not so different from the expected CMS raw event size of 300 KB. CDF is preparing to significantly increase trigger rates, to an event output rate of up to 300 Hz, well before the start of CMS data taking in 2007. Taking into account the expected increases in computing capacity through Moore’s law, we believe that the RunII data management systems provide a valuable and almost-to-scale data point.

3.1 Physical Partitioning I: *Streams*

A *stream* is the largest unit of data organization and can be thought of as a collection of HLT triggers. Here we are discussing physically separate *streams*, written to different physical *files*, as opposed to logically separate streams in the same physical *files*. A *stream* is a selection of the output event data from the CMS detector which is physically grouped together to optimize data access for production processing or analysis. It is natural to define it as a set of HLT triggers that “go together” for analysis purposes.

As an example, in table 1 we decompose the CMS HLT trigger table from the DAQ TDR [2] into five *streams*: Electron, Photon, Muon, JetMET, and Calib. Here the electron stream contains only two trigger types, the inclusive electron and di-electrons, while the JetMET stream contains eight trigger types. In this example there are only 5 *streams* and only 16 triggers (di-photon 14 and 25 are separate triggers and 1-jet, 3-jet and 4-jet are separate triggers).

Table 1 is merely an example from the DAQ TDR: it is not a realistic trigger table for CMS, or even the most current trigger table. We expect the number of *streams* in the actual experiment to be closer to 10 *streams* based on at least 50 HLT triggers. For example, CDF currently has 200 HLT triggers, and we do not expect CMS to have less.

Stream	Trigger/Dataset	Threshold (GeV)	Rate (Hz)	Cumulative Rate (Hz)
Electron	Inclusive electron	29	33	33
	Di-electrons	17	1	34
Photon	Inclusive photons	80	4	38
	Di-photons	40, 25	5	43
Muon	Inclusive muon	19	25	68
	Di-muons	7	4	72
JetMET	Inclusive τ -jets	86	3	75
	Di- τ -jets	59	1	76
	1-jet * ET miss	180 * 123	5	81
	1-jet OR 3-jets OR 4-jets	657, 247, 113	9	89
	Electron * Jet	19 * 45	2	90
	Inclusive b-jets	237	5	95
Calib	Calibration and other events (10%)		10	105

Table 1: For illustrative purposes, we show a HLT trigger table, taken from the DAQ TDR for low luminosity, and an example of a decomposition into streams. Each stream would be written to physically separate output files by the filter farm. Each HLT trigger could correspond to a *production dataset* written to separate files out of the production engine.

There is some duplication of events between two different *streams*, but as long as this duplication is kept small (a design goal would be $\sim 10\%$) it can be worth the convenience of storage and access.

Managing event reconstruction in the production engine benefits from *streams* as well. If physical *streams* are identified before processing begins, production has the option of easily processing data one *stream* at a time, providing additional flexibility in processing and reprocessing.

A physical *stream* is conceived of as a sequential access container: raw data *files* belonging to that *stream* can for example be stored together on a slow-access storage medium like a set of tapes in a mass storage system. This storage organization of *streams* allows that data processing of single *streams* can be managed independently, according to physics priorities, needs of the experiment and the availability and readiness of reconstruction software releases.

We recommend that CMS have roughly ten physical *streams* out of the filter farm. They should be organized keeping in mind the subsequent processing by the production engine and distribution across Tier-1 sites.

3.2 Physical Partitioning II: *Production Datasets*

The production engine takes as input one of the aforementioned *streams*, runs the reconstruction executable and splits the output into *production datasets*. A *production dataset* corresponds to one or more HLT triggers, and is identified by those triggers. For example, in Table 1, the HLT trigger table from the DAQ TDR was decomposed

into 5 filter farm *streams*, that are fed separately into production. At the output of production, the electron stream would then be split into the large inclusive electron dataset and the small di-electron stream, greatly speeding up data access for those interested in di-electrons. In Table 1 we have simplistically indicated each trigger as a separate *production dataset*, written to separate *files* out of production.

The *production dataset* is a most important concept in data management because it is the unit of data that a physicist must access to analyze a given trigger. Data discovery is largely *production dataset* discovery, and data delivery for analysis is largely *production dataset* delivery. We recommend that CMS have roughly 50 *production datasets*.

3.3 Events and Data Files

All *event* data is stored in *files*. A *production dataset* physically consists of *files* of *events*. A *file* is a container of event records that supports sequential and random access. *Files* may have a richer sub-structure (like being a zip-archive of sub-files or support random access to sub-sections of events), but storage systems (disks, tape libraries, hierarchical storage managers), the network and the grid middleware will handle files "natively" as the smallest granularity unit of objects (data containers).

A data management system at its simplest level is a *file* management system: the finest granularity of delivered data is generally a single *file*. Many mass storage systems that write to tape assume that data is in *files*, and some of their performance characteristics for storage and retrieval are directly related to *file* size. The choice for how the data is organized into files has immediate ramifications for the performance and integrity of data storage and retrieval.

We recommend keeping the *event* together in a single *file*, while maintaining logical partitioning within an event. This flexibility will enable CMS to benefit in I/O performance as well as allow for future physical splitting into multiple *files* if desired and needed.

Since the production engine has many compute nodes and many output *production datasets*, the natural file size out of a single compute node is much smaller than the file input to the engine. The data management system will need to control *file* sizes to be efficient, requiring merging of *files*: a process we call concatenation. This procedure is essential and needs to be planned for carefully as it can be logistically daunting.

3.4 Data Tiers

The basic definitions of *data tiers* we use in this section are:

Raw The raw event data written by the filter farm. The digis.

DST All reconstructed objects coming from the production engine.

ESD We define the ESD to be the raw data plus the reconstructed objects. We recommend that this *data tier* be the physical output of the production engine.

For the period under consideration by this RTAG, the use of raw, reconstructed and ESD data tiers is expected to dominate physics analysis needs. Moreover, a useful definition of highly compressed data formats (sometime lossy) or usage of a subset of the reconstruction information is expected to evolve along with the understanding of the detector and the reconstruction software, alignment, calibration etc.

With the ESD data tier the user must be able to do everything that is possible with either the Raw or DST data tier, including scanning events with an event display.

For high level *data tiers* to be useful independently of lower level *data tiers*, there may be some objects from the low level *data tiers* that need to be duplicated at the higher levels. In order for *data tiers* to be a useful method of data organization, the contents of the *data tiers* needs to be defined in close consultation with the PRS groups, who are the principle users.

For MonteCarlo events, we expect the ESD tier to contain sufficient MC truth information. This RTAG did not study the event sizes needed to fulfil this requirement for MC data. This needs to be studied and the physical layout of MC datasets needs to be defined.

There will likely be many attempts to define and use higher level *data tiers* for analysis, but in the first few years of LHC running they will not replace the need to frequently access the ESD. Making the ESD available to a wide com-

munity of physicists, and optimizing access to the ESD sample through the data organization methods described above will facilitate a single event data format and promote the use of standardized software environments.²⁾

3.5 Skims

A *skim* is a method by which physics groups form selected and condensed *analysis datasets* from *production datasets*. The physics groups then typically do analysis on the more manageable *analysis datasets*. Skims make it easier to fit the "hot" data on disk, either cache disk or semi-static disk, and also make it easier to access the data if it continues to reside primarily on tape.

3.6 Databases

Databases will be required to manage all non-event, run-dependent information: calibrations, detector conditions, alignment, luminosities, etc. There is a significant overlap and interplay between the databases that manage this information and the databases that manage event level metadata and we recommend these issues be considered in the same context as the rest of the Data Management Project.

4 Usage Patterns within the Proposed Model

Here we go over some of the foreseen CMS usage patterns and try to demonstrate how they may look in the context of the proposed model. These examples are more general than detailed use-cases and are not meant to replace them.

4.1 Re-Process Sample or *Production Dataset*

These are two distinct cases, they differ in scope, who manages the process, who is responsible for validation, how the results are stored and distributed, etc.

4.1.1 Whole Sample by Offline Group

This activity may happen when a major software upgrade is available that justifies a full scale reconstruction of the complete *data sample*. It requires a validated new reconstruction release, spans all the physics groups in the experiment. All *streams* out of the filter farm are reprocessed. The name of the *production datasets* are changed to reflect the new version of the offline code, and the resulting datasets are added to the catalog (list of datasets). The old *production datasets* remain available for a period of time until they are declared obsolete by the collaboration and then they are removed from the data management system.

4.1.2 Dataset by Physics Group

This is smaller in scope, may happen more frequently, and can be triggered by an individual analysis group. It typically is caused by a more modest improvement in calibration/alignment/correction scheme etc. that benefits a particular analysis significantly. The physical division of the data into *streams* and split *production datasets*, each containing the relevant raw data, enables this reprocessing by a physics group when required.

4.2 Validate a Dataset

This is a crucial part of the process of turning the output of the detector and the offline reconstruction into a usable dataset for analysis. This process relies on expertise that resides in two tiers:

1. **Expert level:** Usually an offline working group (e.g. tracking, electron, jet/met, etc). Typically low level variables are studied (e.g. impact parameter resolution, material distribution, hit usage on tracks, calorimeter occupancy, etc.)

²⁾ CMS will need to verify, track and optimize the resources required by this approach. A good test-bed would be the production and analysis systems used over for the next years (2005-2006). We recommend that APROM, DM, Production and Deployment tasks in CCS should be involved and collaborate on this.

2. **Physics groups:** Interested parties looking at higher level variables and try to make sense of things (e.g. j/ψ mass and width, electron E/P Vs. ϕ , jet energy calibration, etc.)

In the end, the final responsibility is on the physics groups and they sign off on an offline release used to define a *production dataset* (and of course the em data sample). This takes time and many peoples effort. The result is a recipe that ensures stability of physics analysis using the *production dataset*.

4.3 Define an Analysis Dataset - Skim

A *skim* is done from a given *production dataset* to produce an *analysis dataset*. An *analysis dataset* usually selects out a single well defined *trigger path* from the production dataset. In addition to selecting a single *trigger path*, the physics group may choose to:

- Impose additional cuts
- Add high level objects to the event record (e.g. b_{tag} , V_0 ...)
- Drop objects from event record (e.g. raw data, some objects)
- Store the info in a different format (e.g. root tree, ntuple)

The *analysis datasets* are the basic "property" of the relevant physics group(s) not usually generated by production.

4.4 Access to Data for Analysis

Typically users performing analysis will access data in "units" of *production datasets* or *analysis datasets*. This is expected to be the main use-case, and the data management system must clearly support it. Interactive analysis or the use of an event display for visualization adds another use-case: a user has selected a list of events during her analysis and wants to visualize just those events using the event display. The data management system, in coordination with the framework, should supply a mechanism to deliver to the user individually specified events.

4.5 Production of a MC Sample

Made for a given physics group. Managed in a similar way to a "real" dataset, with the additional steps of specification, generation and simulation leading to descriptions unique from data that complicate the data discovery process. An additional complication is the (potential) need for realistic simulation - trying to match the MC data to data taken under dramatically varying running conditions, which requires extensive DB access.

Typically, MC samples are stored by each physics group in the data management system, on tape, for common use. This requires the physics groups to be able to write into the data management system and register their MC samples under their physics group name.

4.6 Extract DB Information at Analysis Level

User analysis code may access the DB for various reasons. This may result in a very large load on the servers as well as other resources when running on highly selective datasets, where there are only a few events per run in the *analysis dataset*. Examples of quantities extracted from DB are the beam-line, calorimeter tower constants, and the like. Additional cases where user code may be pounding the DB are trigger info for selection of events, hot/dead strips in b-tagging applications and the like. Database information is relatively easy to access, but it can be costly if not thought about beforehand.

5 Summary of Main Recommendations

The following is a brief summary of our main observations and recommendations in the context of the data organization model described above.

5.1 Size of a Raw CMS Event

Here we do not refer to what the front end writes out, but to an experiment-wide optimized and well-defined CMS raw data format as used in reconstruction and analysis.

1. We observe that the event is not yet fully defined in CMS, although subdetectors have worked out their local schemes.
2. We recommend that the PRS groups centralize and broaden their process to define, prototype and implement an event format for storing raw data as output from the production process (as distinct from the raw data written by the filter farm). MC data formats and eventually compression schemes (lossless, lossy) should also be studied.
3. We adopt 300 KB/evt as a baseline "best estimate" event size for the period of steady running at low luminosity.

5.2 Splitting the Data into Streams and Datasets

We recommend that CMS raw data be split into a number of physical *streams* based on trigger information.

1. We recommend of order 10 *streams* out of the filter farm separately written to tape.
2. We recommend of order 50 *production datasets* out of the production engine separately written to tape.
3. We recommend that *streams* as well as *production datasets* are defined by unique *trigger paths*.
4. For this to work, the HLT farm must tag all events, and therefore process the events it accepts through its full algorithm set.
5. We note that the overlap between *streams* is tunable and is expected to be between 5-20%.

5.3 Storage/Retrieval of Data

We conclude the following

1. During the first year(s), we expect that access to raw data will be required as well as to reconstructed information.
2. We foresee a reconstruction output containing the event raw data as well as the reconstructed data stored together. We call this the ESD: Event Summary Data.
3. CMS must plan for a system that functions with tapes as well as disk, most likely utilizing caching to disk from tape as well as the capability to pin selected datasets to disk.
4. Freight-train approach to analysis is not desirable.
5. CMS should not by default split a single event between multiple files, but should maintain the flexibility to do so if desired.
6. We recommend that Storage/Retrieval of Run/Time dependent non-event data such as calibrations, alignment and configuration must be considered in the same context as the rest of the data management issues.
7. CMS should plan that a user running her job on an ESD or DST, will use the information stored in the event record as the default setting. Reconstruction-on-demand is an option that should be reserved for well controlled circumstances and require explicit-invocation.

5.4 Data Management Project

We recommend the formation of a data management project to formulate the detailed requirements and to design and implement the data management system. The system should operate within the model we have described and implement our recommendations.

Appendix

A The RTAG Process

The CMS data management RTAG was initiated on June 23, 2004 by the CMS CPT Joint Technical Board with the following introduction, motivation and mandate. The requirements for the CMS Data Management System (DMS) are defined by the needs of the PRS groups and the use cases for performing data analysis, event simulation and data handling by physicists, the CMS offline group and the regional center staff in the LHC distributed computing environment. The DMS should be build on top of the emerging grid tools and grid services and interface to the CMS COBRA framework and the (yet to be defined) environment for Distributed Analysis (DA). A blueprint for the CMS DMS needs to be developed in time for the CCS Computing TDR.

A.1 Motivation

1. To agree on a set of use cases to be addressed by the CMS data management system
2. To agree on requirements which will allow CCS to provide a focus of effort
3. To provide guidance to CCS on development directions and interfacing work to match the requirements, defining the scope of the CMS data management task.
4. To identify the roles and responsibilities of the components/layers/services in CMS data management, distributed analysis and LHC computing grid services
5. To give guidance to the community on the expected division of work between the experiments, the LCG and the external projects.

A.2 Mandate

1. Review, reconcile and define the use cases and CMS requirements for Data Management and capture them in a consistent way, taking into account the agreed HEP CAL use cases and the existing and expected sets of middlewares, experience from other experiments, the CMS application environment and application services like COBRA, POOL, the CMS production environment and the CMS user's potential work environments etc.
2. Consider the interfaces of a CMS DMS to Cobra/POOL, the distributed analysis environment, grid middleware and LHC computing grid services and experiment-specific services.
3. Consider the functionality of existing CMS packages, state of advancement and role in the experiment and identify functionalities and components that could be integrated in the CMS DMS.
4. Develop a roadmap specifying wherever possible the architecture, the components and potential sources of deliverables to guide the medium term work of CCS and the DMS and DA planning in the experiment

A.3 Membership

The CPT JTB has determined that the CMS RTAG would be composed of

- Members from PRS groups
- Members from CCS
- Editor for the RTAG document
- Liaison with the LCG

Members were understood to both contribute their own expertise and to represent the interest of their group and facilitating communication with their constituencies. The RTAG would co-opt or invite representatives from DM projects at RunII and non-LHC running experiments with DM experience, if that was deemed useful. The RTAG members were

Emilio Meschi (TriDAS Online and PRS member)

Norbert Neumeister (PRS member and Deputy RPROM)

Lucia Silvestris (PRS member, CMS-DAPROM)

Stefano Lacaprara (PRS Member, CCS Contributor)

Avi Yagil (PRS member, RUN II Experience)

Vincenzo Innocente (CCS Architect)

Lucas Taylor (CCS DPM)

Massimo Lamanna (LCG-Liaison)

Lothar Bauerdick (CCS Member, Chair)

Robert Harris (Editor)

The PRS and the CCS project manager were invited to attend the meeting, but formally were not members of the RTAG.

A.4 Meetings

The RTAG met 7 times from July through September. Minutes and the 22 documents that were collected and discussed are available at <http://www.uscms.org/s&c/cms/dm-rtag/dm.html>. The RTAG began by considering a seed document of draft requirements [1] which focused us on the issues necessary to be addressed.

B Glossary

Data sample The complete mix of triggers (e.g. calibration, min-bias, jet, incl leptons, di-leptons, etc.)

Trigger Path The set of L1 and HLT triggers that an event has to pass. The primitive used to define a stream or a production dataset.

Stream The output of the filter farm consisting of a collection of HLT triggers physically grouped together in a file. Usually based on common utility (e.g. jet triggers, incl e+di-electro+photon, etc.). Number/size of streams is optimized based on ease of access, utility.

Production Dataset Datasets output from the production engine (including MonteCarlo production) containing all the events from a single HLT trigger, or possibly from two or more logically related trigger paths.

Analysis Dataset Smallest "unit" for an analysis. Defined by its utility (e.g. a given trigger path, calibration, a family of physics analysis, etc.). Obviously, a subset of the data sample.

Skim A skim is done from a given production dataset to produce an analysis dataset.

Data Tier Composition of object stored on an event record in a given dataset. Examples are raw, reconstructed, combinations of raw and reconstructed, ntuples, etc. We have defined the data tiers raw (digis), reconstructed (DST), and raw+reconstructed (ESD).

File A container for the data related to a number of events, typically from a given dataset, that allows sequential or direct access. Size determined by convenience/cost for data handling.

Run A collection of events during which detector conditions are sufficiently stable (e.g. so that a single calibration set can be applied).

Event Smallest unit of physics analysis. Consists of raw data, trigger info, reconstructed objects. May be kept together or broken-up. Has various levels of compression/reduction.

References

- [1] **CD-doc-481-v0**, Edited by Robert M. Harris, *Draft Requirements for a CMS Data Management System*, Draft Version 4, June 24, 2004. Available at <http://www.uscms.org/s&c/cms/dm-rtag/docs/DM-Requirements-v4.pdf>
- [2] **CERN /LHCC 2002-26, CMS TDR 6.2** *The Trigger and Data Acquisition project, Volume II, Data Acquisition & High-Level Trigger, Technical Design Report*, 15 December, 2002.