Comprehensive Monitoring for Heterogeneous Geographically Distributed Storage

Ratnikova N[1], Karavakis E[2], Lammel S[1], and Wildish T[3]

[1] Fermi National Accelerator Laboratory, US
[2] CERN, CH
[3] Princeton University, US

E-mail: natalia.ratnikova@cern.ch, tony.wildish@cern.ch

Abstract.

Storage capacity at CMS Tier-1 and Tier-2 sites reached over 100 Petabytes in 2014, and will be substantially increased during Run 2 data taking. The allocation of storage for the individual users analysis data, which is not accounted as a centrally managed storage space, will be increased to up to 40%. For comprehensive tracking and monitoring of the storage utilization across all participating sites, CMS developed a space monitoring system, which provides a central view of the geographically dispersed heterogeneous storage systems. The first prototype was deployed at pilot sites in summer 2014, and has been substantially reworked since then. In this paper we discuss the functionality and our experience of system deployment and operation on the full CMS scale.

1. Introduction

CMS experiment computing [1] infrastructure spans over more than a hundred of geographically distributed sites that provide both computational and data storage resources. Storage capacity at the sites varies from hundreds terabytes to several petabytes. During the first LHC run, CMS data saturated hundred petabytes of storage. Data taking rate and total data volume will be substantially increased during the Run 2.

Storage accounting and space monitoring becomes increasingly important for data management tasks, such as efficient space utilization, fair share between users and groups, and resource planning.

Data stored at the sites are coming from a variety of sources and include i) centrally managed data sets such as real detector data and simulated data samples at various stages of reprocessing, software release validation samples; ii) files created by individual users or members of physics analysis groups stored at the associated institute sites; iii) temporary unmerged files generated as part of the data processing workflows; iiii) load test and backfill data used for availability and scalability testing [2].

CMS space monitoring system, SpaceMon [3], provides monitoring for storage space occupied by various directories in the CMS data directory tree.

Information is retrieved directly from the local site storage systems in the form of storage dumps produced by various methods or tools depending on the storage technology and local sites infrastructure, following a commonly agreed format [4].

Figure 1. World map showing locations of CMS collaborating universities and institutes

The CMS data management system [5] keeps track the centrally managed file replicas in the central data transfer management database. The SpaceMon information collected at the sites allows to account additional storage space used by the users and groups, and various other volatile data areas, not tracked in the central catalogs.

The information is aggregated locally at the sites, then uploaded into a central database at CERN.

In section 2 we discuss site information providers process and tools. Sections 3 and 4 describe SpaceMon client software and the components of the central infrastructure. In concluding sections we present project's current status and our experience of SpaceMon deployment at CMS sites.

2. Site information providers

CMS sites are asked to provide storage usage information for the space monitoring in form of storage dumps in one of the predefined formats, currently text or XML. These formats have been chosen based on WLCG recommendations described in [4]. Each line of the dump file contains XML tags or fields separated by a vertical bar with the following values: physical file name as addressed on the site storage file system, files size in bytes, optional: file creation date and checksum. The time stamp of the dump creation is provided in a separate XML tag, or encoded in the dump file name. Once created, the storage dump is processed by CMS provided SpaceMon client tool, which aggregates file sizes and creates a space usage record including directory sizes up to a certain level of depth.

CMS sites use various storage technologies for maintaining CMS data, including: Castor, dCache,DPM, EOS, Hadoop, LStore, Lustre, StoRM. There is no unique solution for retrieving the local storage information for all sites.

On the posix-like file systems, such as EOS, Lustre, Hadoop systems mounted with FUSE, or NFS-mounted dCache storage, the *find* command can do the work. However traversing the whole storage namespace by this method causes extra load on the system, or takes long time to execute. For some storage systems, e.g. dCache, direct query on namespace database is more efficient while causing less load. The impact on the production storage system can be completely avoided when storage dumps are produced on the hot standby server, onto which the namespace database is dynamically replicated for a backup. Sites that provide storage to multiple experiments can optimize by dumping only a sub-branch of the namespace under CMS-owned directory. Another possible optimization is to delegate the aggregation of the directory sizes to the database query instead of producing the whole storage dump.

The choice of the information provider tools remains with the site admins. In many cases the successful recipe can be reused with minimal configuration adjustments. CMS site admins exchange tools and solutions for creating storage dumps via an open source repository on github. Fig. 2 shows statistics of commits to the repository for the last year.

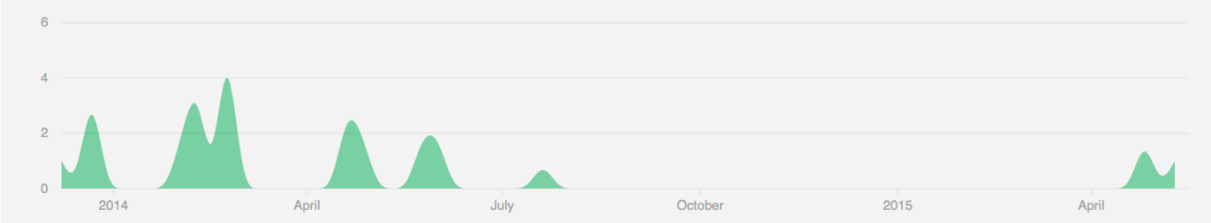

Figure 2. Commits contributed by CMS site admins to SiteInfoProvider tools repository

3. Site client software

SpaceMon client tools are installed at the sites for processing storage dumps and uploading aggregated space usage information into the central data store. Deployment campaign at the sites started with the prototype of the client tool developed within PhEDEx software framework [6] and presented in [3].

New SpaceMon client software includes the following main components:

- *Record* - holds aggregated node space usage info
- *StorageDump* - machine representation of the storage dump info
- *Format (XML, TXT)* - rules for parsing storage dump in corresponding format
- *RecordIO* - various input/output operations, such as read or write Record to file, upload or download from the central database via data service, compare two records
- *Aggregate* - algorithm to convert StorageDump into Record

Components-based design allows for easy extentions of the functionality and separates the workflow from the implementation details. Additional work has been done to eliminate dependency on PhEDEx for independent packaging and release management. Particularly, PhEDEx Namespace framework originally used for supporting various storage technologies is replaced by Format framework, which implements similar idea of plugins used for various storage dump formats.

4. Central infrastructure

CMS space monitoring central infrastructure is integrated into CMS central computing services supported by CMS computing operations and CERN IT division. Historically, SpaceMon project

stemmed from PhEDEx project for CMS data transfer infrastructure [7] and inherited many of PhEDEx developed solutions: Data Service interface to oracle database [9] and security model. Authenticated write access to the information store for a given site is based on the individual's certificate and role registered in the CMS SiteDB [8] for that site.

Any CMS user with valid certificate can retrieve space usage information via Data Service *storageusage* API. The data directory paths are organized by level of depth and are available in perl. xml, or json formats.

Another component of the SpaceMon central infrastructure currently in development is visualization of space monitoring data. The goal of visualization is to present information in a convenient form, enabling users to check space usage across the sites, explore historical views or drill down into a particular directory in the CMS storage namespace. While this functionality is intented primarily for CMS computing management and the central data operations, it will likely become handy also for the site administrators and individual storage users. After trying several prototypes, we opted for WLCG Experiment Dashboard [10] framework for the space monitoring visualization. CMS dashboard, based on this framework, already provides visualization for monitoring job processing, data transfers and site/service usability.

Recently started work with WLCG dashboard developers resulted in the architecture proposal shown in Figure 2.

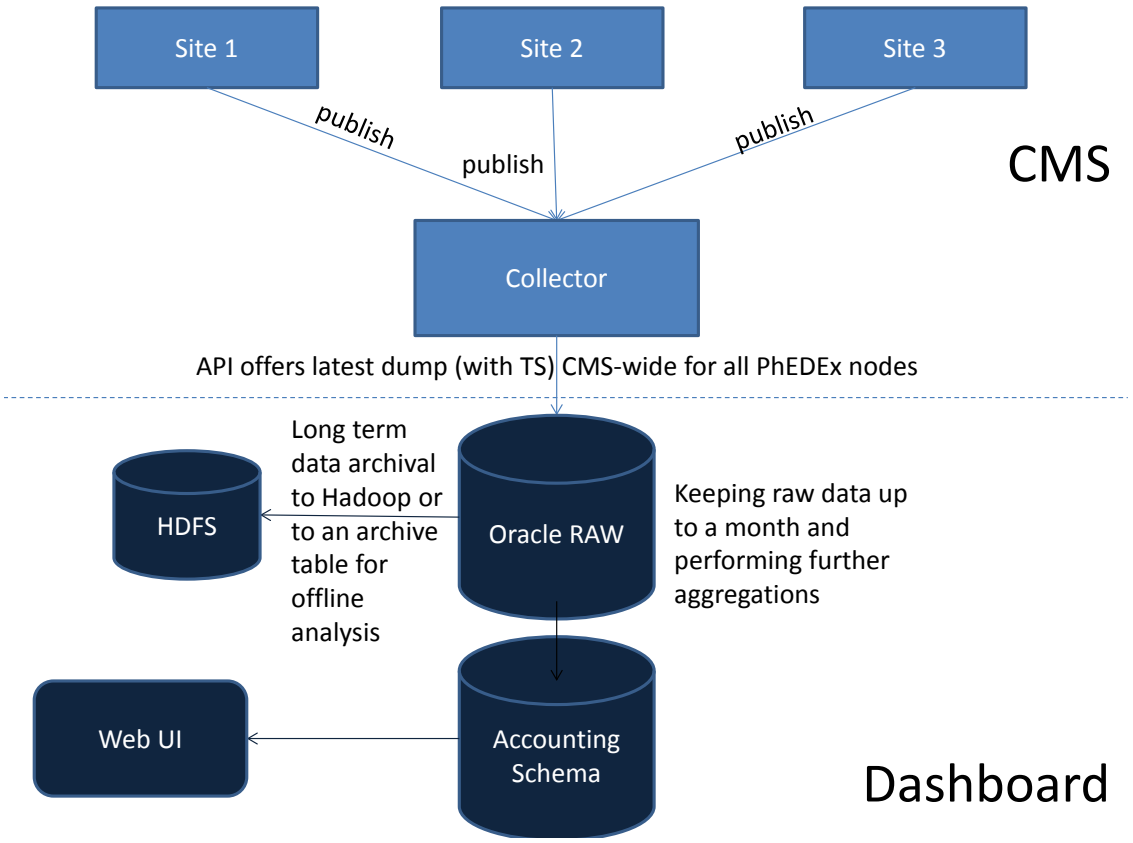

Figure 3. Proposed architecture for storage accounting and visualization in CMS Dashboard based on CMS Space Monitoring information

Similar design was developed and deployed for ATLAS storage accounting [11] based on storage summaries from the central data management catalogs. The challenge in CMS case is to represent uniformly the monitoring data asynchronously pushed by the sites.

5. Deployment

The deployment of the first prototype started in late spring of 2014. A few pilot sites with various storage technologies were contacted, and voluntiered to participate in the initial deployment tests. The goal of this campaign was to collect and upload to the central data service the storage information from a few sites. This allowed to independently verify the provided client software tools and that the documentation is complete and well structured. It also provided some first data that could be used to design and test the visualization tools. Administrators from several sites contributed with feedback and improvements to the site information provider tools. The pilot deployment led to enhancements in the central data service and to updates of the documentation.

In February of 2015 CMS computing management decided to proceed with the roll out of the space monitoring. CMS site support setup a metric in the CMS dashboard that monitors the space information uploaded by sites. When this was ready, sites were contacted in small batches and ask to setup regular uploading of storage information to the central data service. About ten site administrators were contacted via email each week. This allowed to sort out any remaining hidden issues, fine tune instructions, and provide prompt response to questions and in case of problems. The last sites were contacted beginning of May. The issues encountered during this first stage of this deployment can be categorized into three groups: 1) Questions from sites about why they need to provide storage usage information and at what level of detail; 2) Authentication problems uploading the information to the central data service; 3) The long time it takes to take a dump for some storage systems. Sites were informed about how CMS plans to use the storage information. Site were reminded about the three main reasons for the space monitoring project: i) consistency checks between site storage and central PhEDEx catalogue, ii) monitoring the inventory of areas that are not managed by PhEDEx, and iii) to apply user and group quotas fairly at the experiment lavel, i.e. across sites. For level of detail on the storage information guidance was provided and the documentation clarified. The by far largest issues were related to the authentication to upload the information. Instructions and also a script to troubleshoot the authentication problems has been provided. In the process we found the perl-based upload client to not function correctly with the one of the perl-SSL libraries and a problem with the curl command in the current Scientific Linux distributions. On some large storage systems a dump can take several days to complete and this needs to be accounted for. As of middle May 2015, about half of the sites are uploading space information regularly. Several sites are in the process of setting it up. There are no outstanding technical issues preventing sites from providing space information. The plan is to follow the first stage with a ticketing campaign for any remaining sites. We anticipate the roll out to be completed in June.

With plans to enhance the schema of the central data service, sites have been ask to keep the storage dumps they currently take. Given the upload issues, we are also looking into at other secure upload solutions.

6. Summary

CMS space monitoring system is designed to provide a central view of space occupied by all CMS data stored in heterogeneous storage systems deployed at CMS sites. Storage dumps produced locally by the sites include user areas and CMS production data not accounted in the CMS central data catalogues. Central infrastructure has been deployed as CMS web service at CERN and started collecting data uploaded by the sites. In the course of the deployment

campaign all system components benefited from the improvements based on feedback from the sites. Remaining steps are: complete deployment at CMS sites and provide the visualization.

References

[1] Bonacorsi D *et al.* 2007 The CMS computing model *Nucl. Phys.* B *(Proc. Suppl.)* **172** 53-56

[2] Belforte S *et al.* 2010 Bringing the CMS distributed computing system into scalable operations *J. Phys.: Conf. Ser.* **219** 062015

[3] Ratnikova N *et al.* 2014 CMS Space Monitoring *J. Phys.: Conf. Ser.* **513** 042036

[4] Huang C-H *et al.* 2012 Data Storage Accounting and Verification at LHC experiments *J. Phys.: Conf. Ser.* **396** 032090

[5] Giffels M *et al.* 2014 The CMS Data Management System *J. Phys.: Conf. Ser.* **513** 042052

[6] Sanchez-Hernandez A *et al.* 2012 From toolkit to framework - the past and future evolution of PhEDEx *J. Phys.: Conf. Ser.* **396** 032118

[7] Egeland R, Wildish T and Metson S 2008 Data transfer infrastructure for CMS data taking *XII Advanced Computing and Analysis Techniques in Physics Research (Erice, Italy: Proceedings of Science)*

[8] Metson S *et al.* 2010 SiteDB: marshalling people and resources available to CMS *J. Phys.: Conf. Ser.* **219** 072044

[9] Egeland R *et al.* 2010 PhEDEx Data Service *J. Phys.: Conf. Ser.* **219** 062010

[10] Andreeva J *et al.* 2012 Experiment Dashboard - a generic, scalable solution for monitoring of the LHC computing activities, distributed sites and services *J. Phys. Conf. Ser.* **396** 032093

[11] Karavakis E *et al.* 2014 Common Accounting System for Monitoring the ATLAS Distributed Computing Resources *J. Phys.: Conf. Ser.* **513** 062024