Implementing data placement strategies for the CMS experiment based on a popularity model

Fernando Barreiro Megino, Maria Girone, Nicolo Magini, Valentina Mancinelli, Daniele Spiga
Domenico Giordano,
(CERN IT-ES)

Computing in High Energy and Nuclear Physics 2012
24/05/2012
Outline

- Motivations
- Concept and architecture of data popularity service
- The CMS experience based on
  - distributed analysis tools (CRAB)
  - Xrootd monitoring
- A popularity-based site cleaning agent
- Conclusions
The CMS Experiment makes a considerable usage of the distributed grid resources for the storage and offline analysis of the collected data

- Several hundreds of users submit daily up to 500,000 jobs accessing distributed data

It is a challenge to fully profit of the available network and storage resources and to facilitate daily computing operations

- Resource monitoring plays a key role to take strategic choices and promptly react to unexpected effects
Many data in many places

Official datasets needed for analysis (real data, sim data) are distributed among the ~50 CMS T2 sites

- ~14.5 PB of data are resident @ T2 site
- ~4.5 PB added just in the last year of LHC operation
  - NB: copies of a datasets are stored in different places

O(20PB) : volume of data transferred to the CMS T2 sites in the last year

- Transfers ~ 5 x Resident

Need to know how we are using our data

- Data used / reused / unused by physics groups in different sites
Automate the operations

Given the scale of the CMS Grid infrastructure, to control and optimize the usage of the storage is a complex problem

- CMS analysis teams coordinated in 20 physics groups
  - Each group has allocated space in several CMS T2 sites
  - Overall: 124 group-site associations

Data transferred systematically and/or on request to several T2 sites

Significant human effort needed to handle and free the storage space

- Automation of the procedures plays a key role to reduce this workload
  - First step is to discover data most (un)used

For this purpose a popularity data service has been provided to CMS

- inspired by a similar Atlas experience
  - Profit of common needs and shared experience to provide common solutions

More info:
[110] The “Common Solutions” Strategy of the Experiment Support group at CERN for the LHC Experiments.
CMS Data Popularity service tracks over time the official data accessed by the users (mainly) on WLCG

The purpose of a Data Popularity service is to provide

- Usage statistics **along time** about all the levels of data aggregation (files / blocks / datasets) accessed by the users
  - Information in terms of number of accesses, file access success/failure, CPU hours of processing, dataset name, number of users
  - Traces the evolution in time of the popularity metrics

- data service for further applications
  - eg. a Site Cleaning agent, dynamic data placement agent
Architecture

Approach:
- keep the overall design simple and optimized for data mining

Components
- Plugins to interface various data sources
- Popularity database (Oracle)
  - Allows collection of the file based data extracted from data sources
  - Aggregate at different level of collection granularity (block/datasets)
  - Correlate with other attributes (site, user, #users, # CPU hours)
- Lightweight web layer
  - implements multiple interfaces: json API, plots, tables
“Distributed” Data Popularity

Data Popularity Service based on distributed analysis tools

- based on CMS CRAB + Dashboard
- Allows to define the Data Popularity based on the user activity on WLCG
  - Distinct per site and datasets
  - Correlate with the job exit status

Restriction: collects ONLY statistics from grid jobs

- No hint about activity on local batch systems or interactive jobs
The CRAB-based Popularity workflow is steadily collecting data since ~12 months

- Amount of data uploaded:
  - $O(90k)$ jobs/day, $O(360k)$ files/day

- Speed of the procedure
  - Raw-Table upload: $O(1\ h)/day$, MV update: $O(1.5h)/day$

- Size of the Table
  - Raw-Table: $O(120\ GB)$, MVs: $O(2.6\ GB)$
CMS is deploying Xrootd front-end servers in several Tiers, allowing remote access of the data

The Xrootd detailed monitoring contains information suitable for a data popularity

Advantages of a Xrootd-based popularity:

- Monitor data popularity also for batch/interactive job submissions
- Can help in managing the user space on a site:
  - Providing feedback about not only the popularity of the official datasets, but also of the user data

First deployed use-case: monitor the file usage @ CERN from the Xrootd based CMS-EOS DataSvc

More info: [381] Using Xrootd to Federate Regional Storage
ES

Xrootd-based popularity Architecture

Components

- Collector of the Xrootd detailed monitoring stream
  - Based on UDP packets

- Messaging System for Grid (MSG)
  - Adopted Publish-subscribe model
    - Provides a temporary cache for the messages to be uploaded to the popDB
    - Several consumers can access the MSG Broker

- Popularity DB, Lightweight web layer

More info: [233] Xrootd Monitoring for the CMS experiment
Identification of corrupted files

When based on the distributed analysis tool (CRAB), the service allows the identification of corrupted files in a given site:

- Account for ~3% of the CRAB job failures
  - Cause the users move away from submitting jobs on a site, black listing it
- The prompt identification allows to retransfer the corrupted file from another site
Usage Monitoring

Data usage statistics (CPU time, # accesses, # users)

- Several level of aggregation (blocks, datasets, processing campaigns, data tiers)
- Absolute and relative (%) metrics
- Configurable time window aggregation
- Views per single site or aggregated
Most used data format

DataTier time evolution

Accomplished in the last year transition to the lighter data format (AOD, AODSIM)

- Allows better usage of the storage resources

D. Giordano
Dataset time evolution

Allow to monitor the trend of usage of the same kind of data (same MC process, same data acquisition period) along the different processing campaigns

- New reconstruction release, new detector conditions, new simulated scenarios, etc

Users follow the processing campaigns

At some point becomes useful to remove old datasets

![Time evolution of W+jet datasets](chart.png)
Site Cleaning Agent

Scan Tier2 sites reaching their space quota and suggest obsolete, unused data that can be safely deleted

- Adopt Popularity information in conjunction with PhEDEx (the CMS data placement and file transfer system)
- Acts on group space and central space

Cleaning inspection is triggered when a group reaches 90% of its pledged resources on a site.

- Tries to free up space to 70% (configurable thresholds)
- Only old not-custodial copies unused since 30 days are removed
Benefits of the Cleaning Agent

- Optimal handling of the disk quotas
- views (plot & tables) of the available and used space per site and physics group

Monitoring of the evolution in time of the used and available disk space

* views of the Site Cleaning web interface
Benefits of the Cleaning Agent

- Optimal handling of the disk quotas
- views (plot & tables) of the available and used space per site and physics group

- Monitoring of the evolution in time of the used and available disk space

* views of the Site Cleaning web interface
Deletion Request

Sets of data to clean for groups that exceed the pledge are published on web interface

- Possibility for group responsible to select entries and submit the request to the dedicated service (CMS PhEDEx)

| Dataset name                                      | Replica creation dates   | Size (GB) | # file accesses | CPU time | Selected blocks |
|---------------------------------------------------|--------------------------|-----------|-----------------|----------|-----------------|
| /MinimumBias/Commissioning10-May19ReReco-v1/RECO  | 2011-06-17 to 2011-06-17 | 242.82    | 0               | 0        | 5               |
| /ZeroBias/Commissioning10-May19ReReco-v1/RECO    | 2011-06-17 to 2011-06-24 | 9140.22   | 0               | 0        | 46              |
| /AllPhysics2760/Run2011A-16Jul2011-v1/RECO       | 2011-07-27 to 2011-07-27 | 33.04     | 0               | 0        | 1               |
| /MinimumBias/Commissioning10-07JunReReco_900GeV/RECO | 2011-08-29 to 2011-08-29 | 1863.37   | 0               | 0        | 11              |
| /MinimumBias/Commissioning10-May19ReReco-v1/RECO                  | 2011-08-09                | 242.82    | 0               | 0        | 5               |
| /MinimumBias/Commissioning10-May19ReReco-v1/RECO                  | 2011-06-24                | 9140.22   | 0               | 0        | 46              |
| /AllPhysics2760/Run2011A-16Jul2011-v1/RECO                                                                                     | 33.04     | 0               | 0        | 1               |
| /MinimumBias/Commissioning10-07JunReReco_900GeV/RECO                  | 2011-08-09                | 1863.37   | 0               | 0        | 11              |
| /MinimumBias/Commissioning10-May19ReReco-v1/RECO                  | 2011-08-09                | 242.82    | 0               | 0        | 5               |
| /ZeroBias/Commissioning10-May19ReReco-v1/RECO                  | 2011-06-24                | 9140.22   | 0               | 0        | 46              |
| /AllPhysics2760/Run2011A-16Jul2011-v1/RECO                                                                                     | 33.04     | 0               | 0        | 1               |
| /MinimumBias/Commissioning10-07JunReReco_900GeV/RECO                  | 2011-08-09                | 1863.37   | 0               | 0        | 11              |
| /MinimumBias/Commissioning10-May19ReReco-v1/RECO                  | 2011-08-09                | 242.82    | 0               | 0        | 5               |
| /ZeroBias/Commissioning10-May19ReReco-v1/RECO                  | 2011-06-24                | 9140.22   | 0               | 0        | 46              |
| /AllPhysics2760/Run2011A-16Jul2011-v1/RECO                                                                                     | 33.04     | 0               | 0        | 1               |
| /MinimumBias/Commissioning10-07JunReReco_900GeV/RECO                  | 2011-08-09                | 1863.37   | 0               | 0        | 11              |
CMS is investigating dynamic data placement strategies

- **Benefits**
  - Reduce at minimum the pre-placed replicas
  - Optimize the usage of storage resources

The Popularity service will play a key role in this sense

- Promptly determine hot data more often requested by users and trigger replication
- Deletion of unpopular replicas triggered by Site Cleaning agent
Conclusions

Common challenges for all experiments

‣ Automate daily operations
‣ Optimize the usage of the storage and computing resources
  • Evolving computing models
  • Improving data placement strategies

The Popularity service provides a direct monitoring along time of data (un)needed by users

‣ Several sources of information
  • distributed analysis tool; Xrootd front-end monitoring

A popularity-based Site Cleaning agent has been deployed for CMS

‣ implements a strategy to free up space at Tier2 sites