PROCEEDINGS OF SCIENCE

Exploiting Big Data solutions for CMS computing operations analytics

Simone Gasperini,[a,b,*] **Simone Rossi Tisbeni,**[a,b] **Daniele Bonacorsi**[a,b] **and David Lange**[c]

[a] *INFN, Bologna, Italy*
[b] *Department of Physics and Astronomy, University of Bologna, Bologna, Italy*
[c] *Princeton University, Princeton, New Jersey, USA*

E-mail: simone.gasperini@cern.ch

Computing operations at the Large Hadron Collider (LHC) at CERN rely on the Worldwide LHC Computing Grid (WLCG) infrastructure, designed to efficiently allow storage, access, and processing of data at the pre-exascale level. A close and detailed study of the exploited computing systems for the LHC physics mission represents an increasingly crucial aspect in the roadmap of High Energy Physics (HEP) towards the exascale regime.

In this context, the Compact Muon Solenoid (CMS) experiment has been collecting and storing over the last few years a large set of heterogeneous non-collision data (e.g. meta-data about replicas placement, transfer operations, and actual user access to physics datasets). All this data richness is currently residing on a distributed Hadoop cluster and is organized so that running fast and arbitrary queries using the Spark analytics framework is a viable approach for Big Data mining efforts. Using a data-driven approach oriented to the analysis of this meta-data deriving from several CMS computing services, such as DBS (Data Bookkeeping Service) and MCM (Monte Carlo Management system), we started to focus on data storage and data access over the WLCG infrastructure, and we drafted an embryonal software toolkit to investigate recurrent patterns and provide indicators about physics datasets popularity. As a long-term goal, this aims at contributing to the overall design of a predictive/adaptive system that would eventually reduce costs and complexity of the CMS computing operations, while taking into account the stringent requests by the physics analysis community.

International Symposium on Grids & Clouds 2022 (ISGC 2022)
21 - 25 March, 2022
Online, Academia Sinica Grid Computing centre (ASGC), Taipei, Taiwan

*Speaker

PoS(ISGC2022)006

1. Introduction

The current generation of High Energy Physics (HEP) experiments is approaching the exascale regime, opening several new challenges in the software and computing domain. At the Large Hadron Collider (LHC) at CERN, computing operations rely on the Worldwide LHC Computing Grid (WLCG) infrastructure, designed to provide all the services needed to guarantee the full discovery potential of each experiment [1]. This infrastructure is based on a complex and heterogeneous grid, connecting the computing centers available to the LHC experiments and providing grid-distributed services. The grid is hierarchically organized in a tiered structure including the Tier-0 data center at CERN and the Tier-1/Tier-2 sites distributed in almost 40 countries all over the world. On top of this worldwide infrastructure, the LHC experiments have developed a further set of services available for many computing operations, from data management and transfer (Phedex [2], Rucio [3]) to distributed physics analysis (CRAB [4]). Moreover, such a complex system requires a careful monitoring to be maintained in a healthy state as well as a detailed analysis to achieve the required future improvements.

In this context, Big Data derived approaches have started to be adopted to monitor computing operations [5, 6] and study resources utilisation [7]. The CERN IT department provides a set of Hadoop clusters featuring more than 30 PB of raw storage and making available a complete data analytics framework [8]. Since 2015, the Compact Muon Solenoid (CMS) experiment started collecting a large set of information (e.g. datasets access, replicas location and transfer, jobs monitoring) from many different computing services, aggregating and storing this meta-data by the Hadoop Data Files System (HDFS) on the so-called *analytix* cluster at CERN IT [9]. This new solution adopted by CMS opened the possibility to explore this data by using the Spark analytics platform, in order to investigate recurrent patterns and possibly provide useful indicators to enhance the overall efficiency of the distributed computing operations.

In this work, we use CMS meta-data about 2019-2020, aggregated combining information from several sources:

- Data Bookkeeping Service (DBS): physics data catalogue containing basic datasets information (e.g. total size, number of files);
- Physics Experiment Data Export (Phedex): data management system to handle data transfer over the grid (used up to the end of 2020, before the migration to Rucio [10]);
- Monte Carlo Management (MCM) & Production Monitor Platform (PMP): systems to manage and monitor Monte Carlo samples production [11].

In section 2, we use a data-driven approach for analysing the aforementioned meta-data and exploring CMS data storage at WLCG Tier-1/Tier-2 sites in order to address some important questions related to data replicas location stored on disk. In section 3, we propose a simple supervised model able to identify popular datasets (namely physics datasets more likely to be accessed) among all CMS datasets at a given time. The model can also provide concrete hints about the most important features driving this popularity classification.

This preliminary work paves the way towards a much larger project about data management and computing operations at CMS. The long-term aim of this project is to develop an adaptive and

predictive model which can be used to evaluate the impact of possible technical upgrades in the future, or to study different scenarios and their implications, or to highlight the best choices in the software and computing domain in terms of costs and complexity.

1.1 Previous works

Most of CMS grid-distributed services are monitored through custom tools and web applications (e.g. CMS Monitoring Dashboard [12]), and logging information is scattered over several sources and typically accessible only by experts. The collected metrics are used to optimise data distribution, ensuring that the most used data are replicated and accessible on the constrained disk resources while cleaning up of unused or less used data. However, this can happen only post-factum and it requires a sufficient amount of historical data to be accumulated to trigger the replication process.

A first attempt to develop a model able to investigate the possibility to predict CMS datasets popularity dates back to 2016 [13]. In that work, a brand new tool was developed to aggregate and pre-process structured meta-data collected from several CMS data-services (e.g. DBS, Phedex, SiteDB, PopDB) and then train a Machine Learning (ML) model to make short-term popularity predictions. Additional information from CINCO (Cms INformation on COnferences [14]) was also used as a complement to datasets meta-data, in order to study the influence of up-coming conferences on the datasets popularity. That work demonstrated the possibility to make successful predictions at least on some subset of the datasets. However, the developed tool is nowadays deprecated: firstly because many changes have occurred both in CMS data-services themselves and in their monitoring systems, secondly because the computing model complexity has raised and new Big Data analytics solutions are needed to scale up with the exponentially increasing amount of meta-data.

In the last years, CMS computing fostered the adoption of common Big Data solutions based on open-source and scalable tools, such as Hadoop and Spark, available through the CERN IT infrastructure. In 2017, exploiting such Hadoop+Spark ecosystem, new ideas about complex monitoring workflows, predictive analytics, and performance studies were presented [15, 16]. In these works, the Hadoop platform is demonstrated to be a valid solution for the implementation of a CMS popularity data-service, fully replacing the previous version based on standard relational databases. Thanks to the ML library offered by the Spark analytics framework, it is also possible to efficiently make data popularity predictions, exploiting the scalability on large meta-data aggregations. These popularity predictions can play a significant role in smart data placement policies at the WLCG sites, both for newly created datasets and for holding data samples frequently accessed by CMS physicists. In conclusion, they state that the Hadoop+Spark platform is becoming a potentially crucial component in the ecosystem that enables the CMS experiment to attack Big Data analytics problems in the long run.

2. Exploratory data analytics

This section covers the initial data exploration performed on the available meta-data with the main objective of providing a point of reference for validation purposes. Physics data and its structure at the CMS experiment is described in section 2.1 to introduce the complex scenario we

are exploring. In section 2.2, we show an overview of the total amount of physics data registered on DBS, giving a general picture about CMS physics datasets. In section 2.3, we go through a more detailed analysis about data disk storage at different levels (e.g. single country or WLCG site), exploiting the monitoring data about replicas location in 2019-2020 from Phedex. Our aim is to address questions concerning the possibility to provide indicators about the effectiveness of various grid operations choices, for instance making data frequently accessed by a specific physics group (e.g. Top quark), quickly available on disk resources located in the corresponding geographical area (e.g. East Europe).

2.1 Physics data at CMS

The CMS experiment has recorded huge volumes of physics data from collisions at the LHC, and produced an even larger amount of simulated samples. Until 2020, when Phedex was the data management system used in production, the CMS data model was organized into a hierarchical structure of files, blocks, and datasets [17]. Data coming from the detector or produced by simulations was stored into files (size of few GBs), suitable for users physics analysis purposes. Files were grouped into blocks, considered as atomic units for data transferring among the sites. These blocks are replicated in multiple copies and distributed among the computing centers of the WLCG for further processing or analysis. Data blocks themselves are then logically organized into datasets which represent a processing chain of specific physics processes.

Starting from raw data produced by the detector online system or by the simulation software, successive degrees of processing (event reconstruction) refine this data, apply calibrations and create higher-level physics objects. CMS uses a number of data formats (e.g. RAW, RECO, GEN, SIM, DIGI) with varying degrees of details and refinement. Event information from each step in the reconstruction chain is logically grouped into the so-called data tiers (e.g. RAW, RECO, AODSIM but also RAW-RECO, GEN-SIM-DIGI) [18]. A data tier may contain multiple data formats, for instance a given GEN-SIM-DIGI dataset includes the physics process generation (GEN), the detector simulation (SIM), and electronics digitization step (DIGI). The most important data tiers for the physics analysts community are AOD* (AOD or AODSIM), containing real or simulated data actually used in the analysis, and their derivatives MINIAOD* (MINIAOD or MINIAODSIM) and NANOAOD* (NANOAOD or NANOAODSIM), produced to grant only the essential physics information to reduce the size.

2.2 CMS datasets overview

As mentioned in section 1, DBS can be seen as a huge catalogue of all the existing datasets at CMS. It records general datasets information as their total size, number of files, number of events, data tier, etc.

Figure 1 shows an overview of the total amount of CMS data (almost 0.5 EB including both real and simulated samples) registered on DBS up to the end of 2020. For detector collisions data ($\sim$ 215 PB), the most relevant data tiers are: RAW, particles collisions raw data collected directly by the detector electronics; RECO, particle collision data produced by prompt reconstruction at Tier-0 or Tier-1 sites; AOD (Analysis Object Data), processed collisions data containing higher-level information about the reconstructed physics objects.

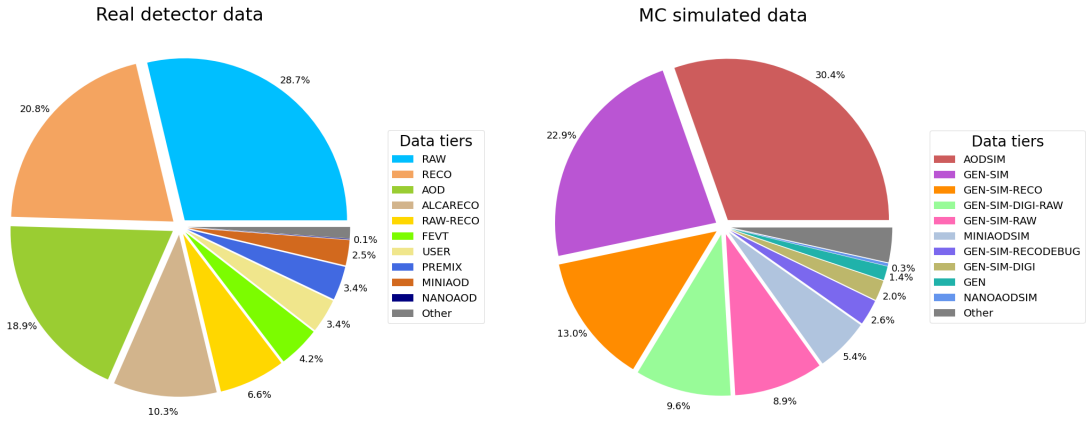

Figure 1: Overview of the CMS physics data collected/produced up to 2020. On the left, real detector data are shown (total size ~ 215 PB); on the right, Monte Carlo simulated data are shown (total size ~ 259 PB).

The prevalent data tiers for simulated data (~ 259 PB), instead, are: AODSIM, analysis objects data deriving from simulations; GEN-SIM, data combining Monte Carlo generated physics events with their energy deposits in the simulated detector; GEN-SIM-RECO, including also reconstructed hits/tracks/clusters of physics objects.

2.3 CMS Tier-1/2 disk storage

In 2019-2020, the CMS experiment included 7 Tier-1 and 55 Tier-2 active sites over the WLCG infrastructure. The total amount of data stored on the available disk was beyond 100 PB over all the period (~ 60% of real detector data, ~ 40% of MC simulated data) with peaks of about 125 PB.

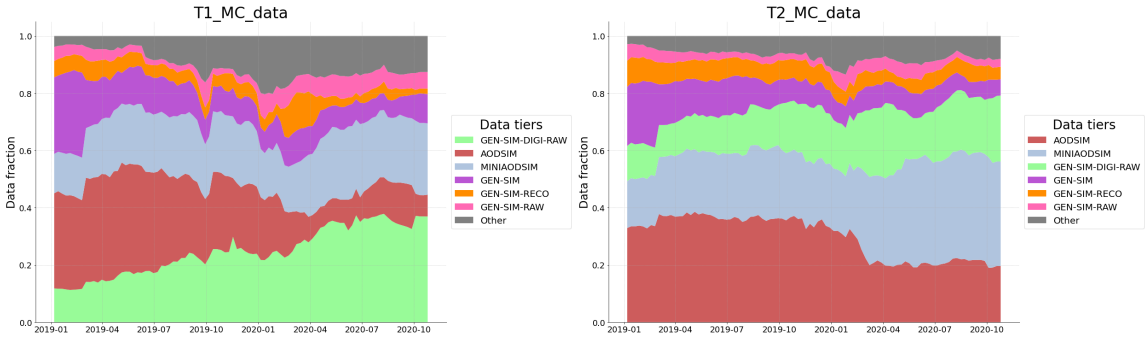

Figure 2: Weekly time trend of CMS Monte Carlo simulated data in 2019-2020. On the left, Tier-1 disk storage is considered (7 WLCG sites); on the right, Tier-2 disk storage is considered (55 WLCG sites).

In Figure 2, the trend of Monte Carlo data stored on Tier-1 and Tier-2 disk is shown, grouping by data tier. In Tier-1 sites, GEN-SIM-DIGI-RAW fraction linearly increases over time and it takes up a large part of the total amount of data in the end of October 2020. The other more significant data tiers are AODSIM and MINIAODSIM. While the former seems to decrease over time, the latter increases both in Tier-1 and Tier-2 sites, demonstrating that lighter data formats are more and more appropriate for most of the physics analysis.

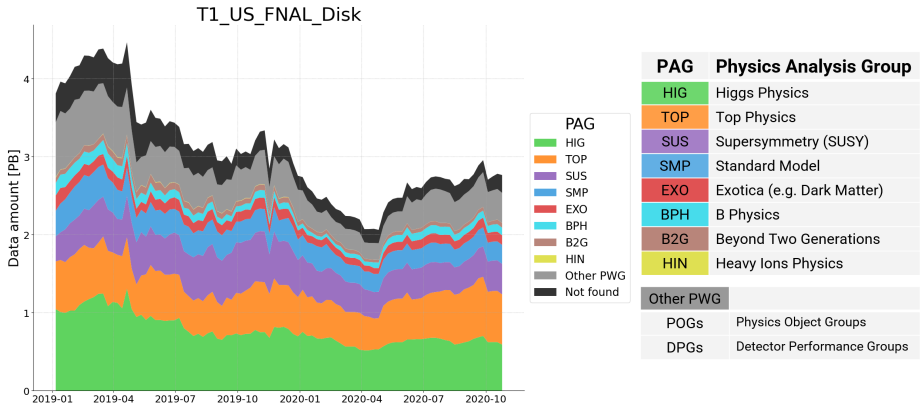

Figure 3: Weekly time trend of *AODSIM data at FNAL Tier-1 site (USA) in 2019-2020, grouped by PAG. Datasets assigned to others PWGs are in grey, while datasets not found on the MC management system are in black. On the right, the table maps each PAG code to the corresponding full name.

Finding new solutions for the targeted removal of large AOD* replicas in favour of the corresponding MINIAOD*, could be a central factor in reducing disk storage needs. The measured average size per event for AODSIM datasets is $\langle S \rangle = 388 \pm 97$ KB (nominal size is 400-480 KB), while for MINIAODSIM we have $\langle S \rangle = 51 \pm 14$ KB (nominal size is 35-60 KB), showing a reduction of almost one order of magnitude [19]. Further size reduction to 1-2 KB per event is achieved by the NANOAOD format, started to be adopted very recently [20].

The developed analytics toolkit is also able to provide more precise storage insights at the level of a single country or single WLCG site. As an example, Figure 3 shows the amount of *AODSIM (AODSIM, MINIAODSIM, or NANOAODSIM) data stored at the Fermi National Accelerator Laboratory (FNAL) facility in the USA, grouped by HEP analysis community. In CMS, Physics Working Groups (PWGs) are sub-grouped into Physics Analysis Groups (PAGs), Physics Object Groups (POGs), and Detector Performance Groups (DPGs) [21]. Using the MC data management system, it is possible to retrieve the information about which PWG submitted the dataset production request, mapping each *AODSIM sample to the corresponding group of interest. In this case, the plot shows that more than half of the *AODSIM datasets stored on FNAL disk actually contains data interesting in physics studies about Higgs boson, Top quark, or SUSY theory.

Such analysis could be useful to investigate new ideas for dynamic replicas placement over the WLCG architecture, making the most popular kind of data efficiently available to local HEP analysis groups. Unfortunately, at this stage, it is still difficult to make precise statements about the effectiveness of this range of solutions because no clear patterns emerge from the available data.

3. Popularity analysis

In this section, we adopt a targeted approach to discover more specific correlations in the meta-data previously explored, introducing the concept of data popularity. This approach could allow to study CMS data placement policies based on the analysis of replicas storage across WLCG sites and their access patterns. In particular, our purpose is to develop a supervised model able to classify AODSIM datasets as *popular* or *not_popular* in a given ΔT (e.g. in month M) and to

provide some indicators about the most important features driving the classification. The training features extracted for each dataset and used in this binary classification are the following:

- *tot_size*: total size;
- *avg_file_size*: average size per file;
- *avg_event_size*: average size per event;
- *num_replicas*: number of replicas on T1/T2 disk (in month M);
- *fract_replicas*: fraction of replicas on T1/T2 disk (in month M);
- *pag*: Physics Analysis Group that submitted the Monte Carlo production request (e.g. *HIG*);
- *campaign*: Monte Carlo production campaign (e.g. *RunII-Summer16*);
- *generator*: Monte Carlo event generator (e.g. *sherpa*).

The target label of the classifier is the popularity class P_M, defined as:

$$P_M = \begin{cases} popular & \text{if } N_M^{\text{access}} > T \\ not\_popular & \text{if } N_M^{\text{access}} \leq T \end{cases} \tag{1}$$

where N_M^{access} is the total number of dataset accesses in month M and T is the threshold used to discriminate popular datasets. For the following analysis, we set this threshold T to a specific fixed value, computed by running a simple 1D clustering algorithm (K-means clustering or Natural Breaks Optimisation) on the distribution of the average number of datasets monthly accesses in the whole period 2019-2020.

3.1 Random-forest classification

For the dataset popularity classification, a random-forest binary classifier is trained and tested for each month from January 2019 to October 2020. As shown in Figure 4, the classification accuracy is always greater than 85%, which is quite promising at this preliminary stage.

However, to have an unbiased measure of the ability of the model to classify a dataset as *popular* only if it is actually *popular*, we need to consider other metrics as well. In particular, we are dealing with an imbalanced binary classification problem ($\sim$ 15% of *popular* and $\sim$ 85% of *not_popular* datasets in each month) and this can lead to the tendency of the classifier to assign datasets to the majority class. For this reason, we also computed precision and recall metrics for the minority class of *popular* datasets:

$$\text{Precision} = \frac{TP}{TP + FP} \simeq 0.63 \qquad \text{Recall} = \frac{TP}{TP + FN} \simeq 0.82 \tag{2}$$

where TP = True Positives, FP = False Positives, FN = False Negatives.

The result could be further improved adding other carefully engineered features for the classifier (e.g. replicas time creation/deletion, number of replicas transfers requests) but this requires more detailed monitoring information about data placement, transfer, and access not yet available.

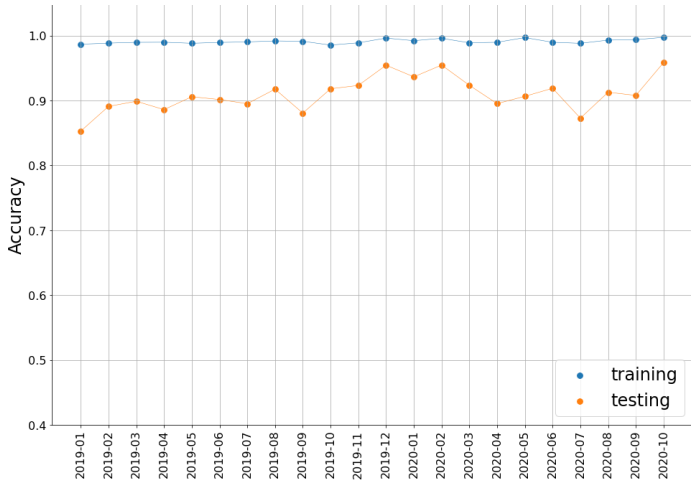

Figure 4: Accuracy of the random-forest classifier on the training and testing set in each month.

3.2 Feature importance

Finally, in Figure 5, we show the feature importance results comparing the two years 2019 and 2020: at the current status of our work, *num_replicas* (number of replicas on disk) and *tot_size* (total size of the dataset) are the most important features to drive the classification. Although this preliminary result may seem trivial and does not provide very insightful indicators in the scope of replicas placement, it is nevertheless useful in the validation process of what has been done so far and in the evaluation of what the next steps may be.

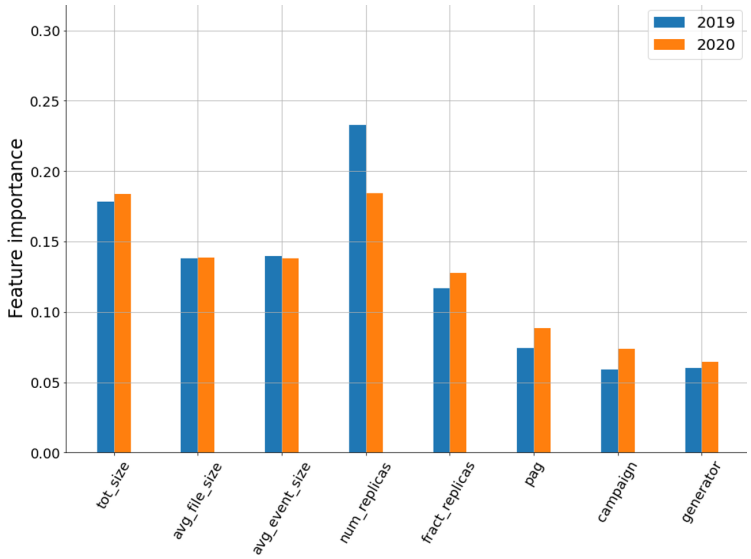

Figure 5: Features importance of the trained random-forest classifier for 2019 and 2020.

A possibility that needs further investigation in this context, is to exploit this kind of popularity analysis to understand if there is a trace of the so-called seasonality effect, namely a raising of the popularity of some datasets based on the approaching of HEP conferences during the year. This

idea can be explored having more detailed data available and trying different adjustments in the classification timescale (e.g. shrinking the time window to a week).

4. Future developments

To improve the classification performance of the dataset popularity model and look for interesting correlations among the training features, more monitoring data have to be collected from the different CMS services, exploiting the full set of information aggregated on the CERN IT *analytix* cluster. In particular, the analytics toolkit we drafted and tested in this work could be re-designed in order to directly extract the meta-data from HDFS, running even complex queries through an effective and handy interface implemented on top of Pyspark (Spark Python API). This would also allow us to perform data pre-processing and custom analysis in a more streamlined manner and for a wider set of use-cases.

The following step will be to complete the development of a meta-data analytics platform for CMS computing operations, providing tools to further investigate the popularity problem (e.g. searching for seasonality effects) exploiting the Big Data analytics framework. In this direction, the purpose of our long-term project is to build a robust and effective model able to make predictions about datasets popularity in the future, to simulate and evaluate different replicas placement and data caching strategies, and to explore various possibilities to reduce costs and complexity of the CMS computing model.

5. Conclusions

In this work, we developed a preliminary analytics toolkit to analyse CMS computing operations meta-data, collected by the Apache Spark platform available on the dedicated Hadoop *analytix* cluster at CERN IT department.

We started to test the current implementation against monitoring data aggregations from 2019 and 2020, exploring some ideas about how to exploit this meta-data to investigate data storage on disk at WLCG Tier-1 and Tier-2 sites. We also designed, trained, and tested a *random-forest classifier* to identify popular physics datasets in each considered month. The overall classification accuracy turns out to be around 90% and the most important features driving the popularity class assignment for dataset X in month M are the total size of X and its number of replicas stored on disk in M. A larger set of more detailed monitoring data and further optimisation in the classifier design are needed to provide robust and more effective indicators that could be eventually used to improve replicas placement strategies.

Exploiting Big Data Analytics solutions adopted by the CMS experiment in the recent years by using a data-driven approach, this work laid the foundation to investigate many problems in the computing model scenario of the experiment: in this direction, the long-term aim of our project is to develop an adaptive and predictive model which can be used to simulate CMS data management and CMS computing operations in order to design, test, and evaluate future improvements in the software services domain.

References

[1] Worldwide LHC Computing Grid. https://wlcg.web.cern.ch/.

[2] Physics Experiment Data Export - CMS TWiki. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhEDEx.

[3] Rucio for CMS experiment - CMS TWiki. https://twiki.cern.ch/twiki/bin/viewauth/CMS/Rucio.

[4] Software Guide on CRAB - CMS TWiki. https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCrab.

[5] Ariza-Porras, Christian, Kuznetsov, Valentin and Legger, Federica, *Big data solutions for CMS computing monitoring and analytics*, *EPJ Web Conf.* **245** (2020) 03022.

[6] Ariza-Porras, Christian, Kuznetsov, Valentin, Legger, Federica, Indra, Rahul, Tuckus, Nikodemas, Uzunoglu, Ceyhun et al., *The evolution of the CMS monitoring infrastructure*, *EPJ Web Conf.* **251** (2021) 02004.

[7] Lange, David, Bloom, Kenneth, Boccali, Tommaso, Gutsche, Oliver and Vaandering, Eric, *CMS Computing Resources: Meeting the demands of the high-luminosity LHC physics program*, *EPJ Web Conf.* **214** (2019) 03055.

[8] CERN IT Hadoop Service - User Guide. https://hadoop-user-guide.web.cern.ch/.

[9] CERN IT Hadoop Cluster - CMS TWiki. https://twiki.cern.ch/twiki/bin/view/CMS/CMSComputingAnalyticsDatasets.

[10] Vaandering, Eric, *Transitioning CMS to Rucio Data Managment*, *EPJ Web Conf.* **245** (2020) .

[11] CMS Monte Carlo Management (MCM) system. https://cms-pdmv.gitbook.io/project/.

[12] CMS Monitoring Project - Grafana Dashboard. https://monit-grafana.cern.ch/d/000000530/cms-monitoring-project?orgId=11.

[13] V. Kuznetsov, T. Li, L. Giommi, D. Bonacorsi and T. Wildish, *Predicting dataset popularity for the CMS experiment*, *Journal of Physics: Conference Series* **762** (2016) .

[14] Cms INformation on Conferences (CINCO). https://cms-mgt-conferences.web.cern.ch/conferences/conf_listing.aspx.

[15] M. Meoni, T. Boccali, N. Magini, L. Menichetti and D.G. and, *XRootD popularity on hadoop clusters*, *Journal of Physics: Conference Series* **898** (2017) 072027.

[16] M. Meoni, V. Kuznetsov, L. Menichetti, J. Rumševičius, T. Boccali and D. Bonacorsi, *Exploiting Apache Spark platform for CMS computing analytics*, *J. Phys.: Conf. Ser.* **1085** (2017) 032055. 9 p.

[17] D. Bonacorsi, V. Kuznetsov, N. Magini, A. Repečka and E. Vaandering, *Exploiting analytics techniques in CMS computing monitoring*, *Journal of Physics: Conference Series* **898** (2017) 092030.

[18] Data Formats and Data Tiers - CMS TWiki. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDataFormats.

[19] G. Petrucciani, A. Rizzi and C. Vuosalo, *Mini-AOD: A New Analysis Data Format for CMS*, *Journal of Physics: Conference Series* **664** (2015) 072052.

[20] Rizzi, Andrea, Petrucciani, Giovanni and Peruzzi, Marco, *A further reduction in CMS event data for analysis: the NANOAOD format*, *EPJ Web Conf.* **214** (2019) 06021.

[21] Physics Working Groups - CMS TWiki. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsConveners#Physics_Analysis_Groups_PAG.