PROCEEDINGS OF SCIENCE

Offline data processing and analysis at LHCb in the 2020s

Davide Fazzini[a,†,*]

[a] *Università degli studi di Milano Bicocca & INFN, Milan, Italy*

E-mail: davide.fazzini@cern.ch

The LHCb experiment is undergoing an extensive upgrade of its detector in order to meet the increased data rate expected in the LHC Run 3 with respect to the previous Runs. In particular, the offline computing and data processing has been optimized to face the new challenges posed by the data volume increase and by the related data storage resources request.The Data Processing & Analysis (DPA) project is responsible for coordinating the efforts, from both computing and data analysis experts, aimed to make the LHCb experiment able to fully exploit the new physics potential available in Run 3 data. This work represents an overview of the new LHCb software from the offline data processing to the production and storage of the final analyst-level ntuples.

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

[*] Speaker
[†] on behalf of the LHCb collaboration

PoS(ICHEP2022)230

1. Introduction

The LHCb experiment [1] is located at one of the four interaction points of the Large Hadron Collider (LHC) at CERN. LHCb is a forward arm spectrometer and was successfully operated through the first two LHC Runs. It collected data from proton-proton collisions corresponding to an instantaneous luminosity of $4 \times 10^{32}\text{cm}^{-2}\text{s}^{-1}$ and to an integrated luminosity of 9 fb^{-1}. Nowadays, the LHCb experiment is characterized by a large physics programme, including but not limited to, decays of beauty and charm hadrons, electroweak physics, exotica, exotic hadrons and heavy ion physics.

In preparation for the third Run of the LHC, LHCb has undergone a comprehensive hardware upgrade where more than 90% of its detector has been replaced [2]. Similarly, the online and offline data processing systems have been overhauled in order to face the factor five increase in instantaneous luminosity and factor 30 increase in the data volume with respect the previous Runs. The LHCb collaboration created two specific projects, the Real Time Analysis (RTA) and the Data Processing & Analysis (DPA) to face the challenges posed by the LHC Run 3 conditions. The RTA has developed a fully software-based High-Level Trigger (HLT) [3] for processing the 30 MHz input event-size rate, while the DPA is responsible for all offline data processing and analysis developments needed to transform the HLT output into data-user ntuple. This work represents a complete overview of the LHCb Upgrade offline dataflow, coordinated by the DPA, which is depicted in Figure 1. The DPA project is organised into six dedicated Work Packages (WPs):

WP1 *Sprucing*: centralized offline data processing of the data leaving the trigger;

WP2 *Analysis production*: centralized ntuple production using the DIRAC [4] transformation system with maximum automation;

WP3 *Offline analysis tools*: offline software development for a modern, thread-safe, analysis framework consistent with that of the online;

WP4 *Innovative analysis techniques*: development of novel strategies for future data analysis;

WP5 *Legacy software and data*: preservation and support of the legacy Run 1 and Run 2 data samples and software infrastructure;

WP6 *Analysis preservation and open data*: guidelines and tools for analysis preservation in LHCb and the release of LHCb datasets to the CERN Open Data portal.

2. WP1: Sprucing

Starting from Run 3, LHCb's default persistency model for physics events will be Turbo Selective Persistence (Turbo SP) [6]. Here, the trigger candidates are persisted along with a customisable set of other objects and their reconstruction; the rest of the event is discarded reducing the event size to 4-16 kB [7]. After a simple reformatting, this output stream is saved directly to disk granting an immediate access to the users. This model is expected to be used for the 70% of the LHCb analyses, however, in some particular cases access to the whole event is required. In

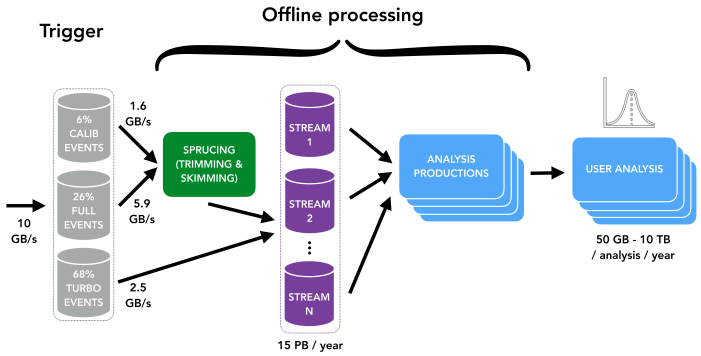

Figure 1: LHCb upgrade dataflow showing the various offline aspects coordinated by the DPA project [5].

this scenario the Full stream can be used instead, saving the whole event reconstruction to tape, with an event size of about 48-69 kB. An additional slimming and skimming step is then required in order to further compress the event-size and persist selected physics objects to disk. In Run 3 this selection step is called Sprucing, as seen in Figure 1, and it's developed within the Moore application [8], sharing the same software framework used for the trigger. By design, the Sprucing is the counterpart of the Stripping in the Run 1-2 framework [9] and it will run concurrent with the data taking and be re-run periodically whereby the data from tape must be staged.

3. WP2: Analysis Production

The LHCb experiment has improved the system creating analyst-level tuples. In Runs 1 and 2 production of tuples was performed by the analysts through a dedicated application, named Ganga [10]. Ganga allows analysts to submit jobs to the LHCbDirac [11] transformation system, accessing the Turbo/Stripping output files and selecting the physics quantities of the candidates of interest for a specific analysis.

In order to make tuple-production more efficient in Run 3 the features of the LHCbDirac system have been extended, in so-called Analysis Productions. The idea behind Analysis Productions is to create an automated centralized procedure for the processing of the LHCb data and simulation. In particular, the key points of this new application are a simpler user job configuration, that now can be declared via YAML files, and better file and job failure handling, that is now fully automatic. In order to wast computing resources, extensive tests are run on the GitLab Continuous Integration Platform validating every new production before the submission. Meta-data of the output files is saved in the LHCbDirac bookkeeping system, along with their provenance and job configuration options, ensuring a better long-term analysis preservation. Finally, the result and details of final production are collected in a dedicate web page, granting a more transparent overview.

4. WP3: Offline analysis tools

For the LHCb upgrade, the offline analysis framework developed in the DaVinci application [12], has been completely revised, with respect to the one used in the Run 1-2 runs. This

has been guided by the need to have a more modern software, able to exploit the latest computing technologies and features. Differently from the offline software used in Run 1 and Run 2, the new framework is fully consistent with the online one used for developing the trigger and sprucing selections. In this way code and effort duplication is avoided and the whole software stack becomes more efficient and is more easily maintained. Similarly, the main building blocks of physics selections (*functors*), implemented in the Moore ThOr [13] and LoKi [14] framework, are also shared with the online system ensuring a one-to-one correspondence between the physics quantities evaluated and used in the online and offline systems [15].

As a consequence of this code sharing, the DaVinci software is now only focused on creation of the output data and simulation tuples. The new tupling algorithm for the LHCb upgrade and beyond, namely FunTuple, has been designed in such a way to exploit the Moore functors to fill the final tuples with a specific and restricted set of information selected by the analyst. With respect to the Run 1-2 software, the new tupling software provides full control over the choice of the variables to be persisted in the final tuple, resulting in a generally more light-weight output file.

5. WP4: Innovative analysis techniques

The WP4 is focused on the development of innovative analysis techniques and new facilities that may be exploited in future LHCb analyses. The first stage of the Upgrade HLT consists of a fully software system based on GPUs [16]. Therefore, LHCb can exploit the GPU farm to run computationally detector simulation and computationally intensive analysis tasks.

A very new technology, the quantum computing, has been tested in LHCb. A quantum machine learning algorithm has been setup to identify jet originated by b quark. [17]. The new algorithm exploits the intrinsic properties of the quantum computation to train a Variational Quantum Classifier with the aim to identify hadronic jets, originated by a b or $\bar{b}$ quark at the moment of production. The performance of the algorithm has been evaluated on a quantum simulator and recently on the real hardware and compared with a classical deep neural network method.

6. WP5: Legacy software and data

LHCb analysts will continue to perform studies on the data collected during the Run 1 and Run 2 (legacy) data taking. For this reason the DPA project created a dedicated work-package to coordinate all the activities related to the maintenance of these datasets and the tools required to analyze them. In the legacy data flow model the full event information is always saved and this allow to rerun periodically a reprocessing of the various datasets, the so-called "*re-stripping campaigns*". This allows to re-analyze these datasets with the same software used originally but introducing improvements in the reconstruction and selection criteria, allowing to the analysts to run their analysis on a more updated ntuple.

The WP5 is also collaborating with the WP6 providing metadata and information related to the legacy datasets for the Open Data releases.

7. WP6: Analysis preservation and open data

At the end of 2017 the LHCb Collaboration adopted a minimal set of mandatory Analysis preservation (AP) practices as part of the internal analysis review process. Following these rules the source code and meta-data related to any analysis are preserved in dedicated Gitlab repositories at gitlab.cern.ch along with the instructions to run the code and install the runtime environment. As a further step, these analysis code bases will also be deployed via REANA [18]. REANA is a reproducible analysis platform allowing full access to all the libraries used in an analysis such as RooFit [19] and Scikit-HEP project packages [20] that are available through CVMFS [21].

According to the AP rules, the LHCb datasets are produced centrally, preserved in the distributed computing infrastructure and made available to the LHCb collaboration via a dedicated bookkeeping system. In accordance with the CERN Open Data policy [22, 23], the LHCb experiment agreed to make the output of the Turbo/Sprucing steps available to the general public through the CERN Open Data portal. To facilitate access to the datasets, LHCb has developed a dedicated web interface, named "*nTuple Wizard*" [24], with the aim to provide secure access to the data replicas stored on the grid, working at the same time as a firewall against any injection of nefarious code into the production system. The key point of this new interface is to allow the users to reproduce the LHCb data ntuples without any *a-priori* knowledge of the LHCb software. This is possible thanks to a fully automatic process that, starting from a few simple user inputs, is able to recreate the required job options and run the Analysis Production jobs.

8. Summary

The Data Processing & Analysis (DPA) project at LHCb coordinates the offline system development efforts to allow the full exploitation of the Physics potential available in Run 3 and beyond despite the unprecedented challenges posed by the increased data rate. With the contribution of a variety of experts, the DPA builds a fully centralized system for the skimming and trimming of the data leaving the software trigger and a centralized analysis production system for creating ntuples ready to be used in physics analyses.

References

[1] A. A. Alves Jr. et al. *The LHCb detector at the LHC*. In: JINST 3 (2008), S08005. doi: 10.1088/1748-0221/3/08/S08005.

[2] LHCb collaboration. *Framework TDR for the LHCb Upgrade: Technical Design Report*. Tech. rep. CERN-LHCC-2012-007. Geneva CERN, 2012.

[3] LHCb collaboration. *LHCb Trigger and Online Upgrade Technical Design Report*. Tech. rep. CERN-LHCC-2014-016. Geneva: CERN, 2014.

[4] A. Tsaregorodtsev et al. "*DIRAC3: The new generation of the LHCb grid software*". In: J. Phys. Conf. Ser. 219 (2010), p. 062029. doi: 10.1088/1742-6596/219/6/062029.

[5] LHCb Collaboration. *RTA and DPA dataflow diagrams for Run 1, Run 2, and the upgraded LHCb detector*. Tech. rep. LHCb-FIGURE-2020-016. Sept. 2020.

[6] R. Aaij et al. *Tesla : an application for real-time data analysis in High Energy Physics*. In: Comput. Phys. Commun. 208 (Apr. 2016). arXiv: 1604.05596.

[7] Alejandro Alfonso Albero et al. *Upgrade trigger selection studies*. Tech. rep. CERN-LHCb-PUB-2019-013 Geneva: CERN, Sept. 2019.

[8] LHCb Collaboration. *Moore application* GitLab repository.

[9] LHCb Collaboration. *LHCb computing: Technical Design Report*. Tech. rep. CERN-LHCC-2005-019, LHCb-TDR-11. Geneva, 2005.

[10] *Ganga* documentation.

[11] *LHCbDIRAC* documentation.

[12] LHCb Collaboration. *DaVinci application*, GitLab repository.

[13] LHCb Collaboration. *ThOr* functors.

[14] I Belyaev et al. *PYTHON-based Physics Analysis Environment for LHCb*. In: LHCb- PROC-2004-021 (Oct. 2004), 4 p.

[15] LHCb Collaboration. *Computing Model of the Upgrade LHCb experiment*. Tech. rep. CERN-LHCC-2018-014, LHCb-TDR-018. Geneva: CERN, May 2018.

[16] R. Aaij et al. *Allen: A High-Level Trigger on GPUs for LHCb*. In: Computing and Software for Big Science 4.1 (Apr. 2020). issn: 2510-2044. doi: 10.1007/s41781-020-00039-7.

[17] A. Gianelle et al., *Quantum Machine Learning for b-jet charge identification*. Journal of High Energy Physic (2022). arXiv:2202.13943. doi: 10.1007/jhep08(2022)014

[18] *REANA* documentation.

[19] Wouter Verkerke and David P. Kirkby. *The RooFit toolkit for data modeling*. In: eConf C0303241 (2003). Ed. by L. Lyons and Muge Karagoz, MOLT007. arXiv:physics/ 0306116.

[20] Eduardo Rodrigues et al. *The Scikit-HEP Project overview and prospects*. In: EPJ Web of Conferences 245 (2020). doi: 10.1051/epjconf/202024506028.

[21] *CernVM-FS* documentation.

[22] *CERN Open Data Policy for the LHC Experiments*. Tech. rep. https://cds.cern.ch/record/2745133. Geneva: CERN, Nov. 2020.

[23] CERN Scientific Information Policy Board. *CERN Open Data Policy for LHC Experiments: implementation plan*. In: (Nov. 2020). url: https://cds.cern.ch/record/2745081.

[24] O'Neil, Ryunosuke. *The NTuple Wizard An NTuple production service for accessing LHCb Open Data*. url: https://cds.cern.ch/record/2815814