Open access for ALICE analysis based on virtualization technology

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2015 J. Phys.: Conf. Ser. 664 032008
(http://iopscience.iop.org/1742-6596/664/3/032008)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 137.138.125.164
This content was downloaded on 09/03/2016 at 08:49

Please note that terms and conditions apply.
Open access for ALICE analysis based on virtualization technology

P Buncic1, M Gheata2 and Y Schutz3 on behalf of the ALICE Collaboration

1CERN, Geneva, Switzerland
2Institute of Space Sciences, Bucharest, Romania
3Laboratoire de Physique Subatomique et des Technologies Associées

E-mail: mihaela.gheata@cern.ch

Abstract. Open access is one of the important leverages for long-term data preservation for a HEP experiment. To guarantee the usability of data analysis tools beyond the experiment lifetime it is crucial that third party users from the scientific community have access to the data and associated software. The ALICE Collaboration has developed a layer of lightweight components built on top of virtualization technology to hide the complexity and details of the experiment-specific software. Users can perform basic analysis tasks within CernVM, a lightweight generic virtual machine, paired with an ALICE specific contextualization. Once the virtual machine is launched, a graphical user interface is automatically started without any additional configuration. This interface allows downloading the base ALICE analysis software and running a set of ALICE analysis modules. Currently the available tools include fully documented tutorials for ALICE analysis, such as the measurement of strange particle production or the nuclear modification factor in Pb-Pb collisions. The interface can be easily extended to include an arbitrary number of additional analysis modules. We present the current status of the tools used by ALICE through the CERN open access portal, and the plans for future extensions of this system.

1. Introduction
The ALICE Collaboration adopts the policy of openness promoted publicly by CERN and laying the foundations of its legacy to the scientific community as well as to the general public. As stated by its data preservation policy, “the data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of investment in human and financial resources by the international community”. These data embed unique scientific information for the in-depth understanding of the profound nature and origin of matter.

Documentation, long term preservation at various levels of abstraction, data access and software availability constitute the key elements of the ALICE data preservation strategy allowing future collaborators, the wider scientific community and the general public to analyse data for educational purpose and for eventual reassessment of the published results.

The implementation of the ALICE open access policy is currently in an early stage and only allows students and interested public to have limited access to data and analysis tools for educational purposes. The main objective was to initially test the deployment of simplified tools to give the user the feel of the real tools employed by the physicists for data analysis. An important requirement is the simplicity of use, requiring no prior knowledge of ALICE software installation procedures or specific
frameworks. The use of virtualization technology that allows installing and running the software independently from the host architecture, achieved this. Applications and data are hosted by a portal used in common by all LHC experiments that collaborate with CERN IT and GS (General Infrastructure Services) divisions for its implementation, allowing for a coherent style and organization. In the following sections we briefly describe the current ALICE open access framework, its usage and the plans for extending its capabilities.

2. ALICE open access workflow

The first step to allow external users to run specific ALICE analysis programs was to deploy a dedicated open access virtual machine having the ALICE software and environment installed. The intent was to automatically enter this environment after starting the virtual machine and to provide a graphical users interface (GUI) to guide the user step by step to perform a number of pre-set analyses. Once this was possible, we redesigned the GUI to allow fetching both the analysis modules and data on demand from a remote server. The last remaining step was to set up the open data portal as entry point for the installation and contextualization of the VM, providing all necessary instructions and documentation, uploading also the data and analysis modules to a public area. The following subsections will present each of these steps in more detail.

2.1. Contextualizing with ALICE analysis environment

To make the virtual machine as lightweight as possible, we started from a microCernVM bootable machine image, which contains just a kernel and runs a CernVM-FS client [1]. We chose this approach because it minimises the amount of data to be transferred by the user and integrates in the standard VM installation prescribed by CernVM.

CernVM allows running a bootstrapping procedure called contextualization, which can make the generic virtual machine behave in a custom way. The open access ALICE contextualization script creates the default user account alice with automatic logon and customises the user start-up script to enable a specific AliRoot [2] version, set the location of the CERN open data server and initialize a few environment variables. This allows starting up automatically the analysis interface after booting the virtual machine. The usage of the contextualization procedure is explained in detail on the CernVM pages [3]; the ALICE open access context can be paired from the CernVM Market, as shown in figure1.

![Figure 1. Procedure to pair the ALICE open access context from https://cernvm-online.cern.ch/market/list](https://cernvm-online.cern.ch/market/list)

2.2. The ALICE open access graphical user interface

After starting the VM, the user is automatically logged on. The start-up script checks the presence of the package containing the GUI, which gets downloaded from the portal and unpacked. This interface is a C++ class compiled dynamically in ROOT [4]. The required versions of ROOT and of the ALICE
A software framework AliRoot are then enabled using copies stored on the CernVM-FS system, after which the GUI is automatically started. A terminal pointing to the user workspace is also opened, allowing restarting the GUI at any time.

The GUI is the single entry point for the user and allows running one of the supported analysis modules. Each module has a very simple interface that presents some documentation on the use and functionality of the given analysis, as well as a set of widgets allowing changing analysis parameters and selecting the data to be processed. An example of such interface is presented in figure 2.

A set of outreach and educational analysis exercises are based on specifically selected ALICE data organised as ROOT files, and are widely used for the particle physics master classes. These exercises highlight some of the ALICE physics. As an example, one exercise is about the search for particles containing strange quarks, based on their V0 decays; the motivation is to give an insight to how strangeness enhancement, one of the first signatures for the Quark-Gluon Plasma (QGP), is observed. Another exercise examines charged particle tracks; the aim is to calculate the nuclear modification factor RAA by comparing particle yields in the case of lead-lead and proton-proton collisions; the fact that RAA is less than one indicates suppression of charged particles due to interactions of partons with the QGP.

Figure 2. The GUI allows running few ALICE analysis modules. Each module can be selected from a separate tab and allows selecting the input datasets or setting some specific parameters.

Figure 3. The particle calculator used to identify the V0 decaying particles.
At this stage we have not released modules performing analysis of ALICE primary datasets and reproducing published results. We have made however available the standard $p_T$ analysis tutorial presented to ALICE newcomers, which gives an introduction by examples to the use of the ALICE analysis framework and allows plotting inclusive transverse momentum and pseudorapidity distributions for reconstructed particles. Since the analysis is compiled dynamically as a ROOT macro, the users can run a “living” analysis by changing input parameters or adding new output histograms in a hands-on session.

2.3. Accessing ALICE primary datasets
The primary datasets already available on the open data portal are intended for demonstration purposes only and are not statistically representative. From the $p_T$ analysis tutorial interface, the user can select and download one of the ALICE event summary data files made available from the portal. These data can be also separately downloaded from the portal.

Following its data preservation policy ALICE aims to make publicly available 10% of the data after 5 years, and has currently approved the release of about 10 $TBytes$ of 2010 p-p and Pb-Pb reconstructed datasets. These will become available from the open access portal by the summer 2015. At the moment this paper was written, the data were already copied on the final EOS storage at CERN and have only to be indexed and made available in the open access portal. Each dataset has associated metadata kept in a JSON schema, embedding the year, collision type, centre of mass energy and file-related information, as shown in Figure 4.

```
Dataset": {  
  "name": "LHC2010h_PbPb_ESD_138275",  
  "description": "Pb-Pb ESD data sample at 3.5 TeV",  
  "path": "/eos/opendata/alice/2010/…",  
  "files": 100,  
  "file_details": [  
    {  
      "file_path": "/eos/opendata/alice/…",  
      "file_size": 200762886,  
      "file_checksum": "41b466d6eb9b…",  
      "file_timestamp": "2014-09-24 12:32:31"  
    }, …
  ]
```

In future, ALICE plans to release more compact datasets in a suitable analysis format, including filtered data for specific analysis. This will require extending the JSON schema to include new tags such as pointers to the filtering procedure or statistics information relevant for the sample.

2.4. Analysis software versioning
The versioning of the analysis software made available for open access is using CernVM-FS in the same way as for the ALICE production software. Both analysis software and the object database (OADB) required by different analysis modules are regularly tagged, built and deployed within the ALICE distributed analysis infrastructure. On the public open access CernVM-FS enabled virtual machine any tag can be enabled on demand and become immediately available for one of the available analyses. This allows easy extension of the system in future for embedding new modules, even if requiring different versions of AliRoot.

3. ALICE information and tools on the open access portal
The CERN open access portal [6] is meant as an entry point to the information, tools and data provided by all LHC experiments [7]. It is developed by the CERN IT and GS groups in collaboration with researchers from the HEP community, and contains a growing range of data produced through the
research performed at CERN, together with analysis software and demonstration tools intended for the general public.

The portal uses INVENIO [8] for the digital library technology, CernVM as virtual software appliance allowing for portable and easy to configure user environment to develop and run locally LHC data analysis, and EOS as disk based low latency storage system allowing XROOT [9] based access to data. Among the four types of data categorized by the DPHEP study group (2009) [10], the portal focuses on the release of data from levels 2 and 3, namely simplified data formats for analysis in outreach and training exercises, and reconstructed data and simulations as well as the analysis level software to allow for full scientific analysis.

ALICE has released so far data from the two categories meant only for education purposes, sustained by a thorough documentation and usage instructions for the software. The available datasets contain minimum bias reconstructed events in the event summary data format (ESD), usable directly with scientific ALICE analysis tools. In the near future ALICE will release a larger set of such ESD datasets representing 5 to 10% of the data taken in 2010, together with datasets in a more compact analysis format (AOD) [2] and simulated samples anchored on the acquisition conditions for the released data.

The software suite available so far on the portal contains educational tools and a demonstrator for the usage of the ALICE analysis framework, as described in the previous section. This corresponds to the education section of the open data portal, and will be further extended with more realistic analysis examples. Our intention is to extend the scope of our open access initiative to research, exposing real data analysis examples that were already published by the ALICE collaboration and that are reproducible to some extent using a limited statistics, or filtered datasets. This will include event characteristics observables as well as rare signals relevant to the scope of ALICE physics research.

4. References

[1] Blomer J and Fuhrmann T 2010 *Proc. ICNCN* doi: 10.1109/ICCCN.2010.5560054
[2] http://aliweb.cern.ch/secure/Offline/sites/aliweb.cern.ch.Offline/files/uploads/OfflineBible.pdf
[3] Buncic P et al. 2010 *J. Phys.: Conf. Ser.* 219 042003
[4] http://root.cern.ch
[5] Peters A and Jansyt L 2011 *J. Phys.: Conf. Series* 331 052015
[6] http://opendata.cern.ch
[7] J Cowton et al. 2015 Open data and data analysis preservation services for LHC experiments Current proceedings
[8] N Robinson, J-Y Le Meur and Simko T 2007 *ERN-IT-Note-2007-008* 27p
[9] Dorigo A, Elmer P, Furano F and Hanushevsky A 2005 *WSEAS Transactions on Computers*
[10] DPHEP Study Group 2009 arXiv:0912.0255 [hep-ex]