A proposal for Open Access data and tools multi-user deployment using ATLAS Open Data for Education

Arturo Sánchez Pineda[1],* *Giovanni* Guerrieri[2]
on behalf of ATLAS Software and Computing

[1]LAPP, Université Savoie Mont Blanc, CNRS/IN2P3, Annecy, France
[2]INFN Gruppo Collegato di Udine, Dipartimento Politecnico di Ingegneria ed Architettura, Universita' di Udine, Udine, Italy

Abstract. The deployment of analysis pipelines has been tightly related and conditioned to the scientific facility's computer infrastructure or academic institution where it is carried on. Nowadays, Software as a Service (SaaS) and Infrastructure as a Service (IaaS) have reshaped the industry of data handling, analysis, storage, and sharing. The sector of science does not escape those changes. This situation is particularly true in multinational collaborations, where distributed resources allow researchers to deploy data analysis in diverse computational ecosystems. This project explores how the current multi-cloud (e.g., SaaS + IaaS) approach can be adapted to modest scenarios where analysis pipelines can be deployed using containers and virtual machines containing analysis tools and protocols. This approach aims to replicate sophisticated computer facilities in places with fewer resources like small universities, start-ups, and even individuals who want to learn and contribute to this and other sciences and its replicability. It is desired to explore the development of multi-cloud-compatible tools in physics analysis and operations monitoring using ATLAS experimental and simulated data, adding the Big Data component that the High Energy Physics field has by nature.

1 Introduction and Highlights

High Energy Physics (HEP) experiments' open-access data, like the multiple datasets hosted in the CERN Open Data portal [1], tend to be used as inputs in university classrooms and laboratories. Those samples help enhance curricula and teach cutting-edge research and techniques not present yet in the literature, making it a powerful tool for training and integrating new researchers and technicians into a particular scientific endeavour in this field. In the case of ATLAS [2], the usage of Open Data for education came from the use of event data by the International Particle Physics Outreach Group (IPPOG) Master Classes [3] and satellite events [4], and the dedicated project called ATLAS Open Data [5]. The latter is devoted to releasing sample datasets of data recorded by the ATLAS experiment and simulated data. The first of such releases of data recorded (1 fb^{-1}) and simulated at a centre-of-mass energy of 8 TeV was done in 2016 [6], and a more recent release of 10 fb^{-1} of real data and several 100 fb^{-1} of simulated data at a centre-of-mass energy of 13 TeV in 2020 [7] (see Figure 1).

[*] e-mail: arturos@cern.ch

The latter is the set of data samples – and tools – that this paper considers for the following sections. The ATLAS Open Data project is composed of data and Open Source tools that allow the development and execution of multiple kinds of educational and outreach activities: from a one-day workshop to a complete university course in particle physics [6]. Those tools contain, among others, a series of analysis examples that reproduce simplified versions of real public ATLAS analysis to be used in classrooms to teach experimental high-energy particle physics, as well as the software and computing techniques behind those analysis.

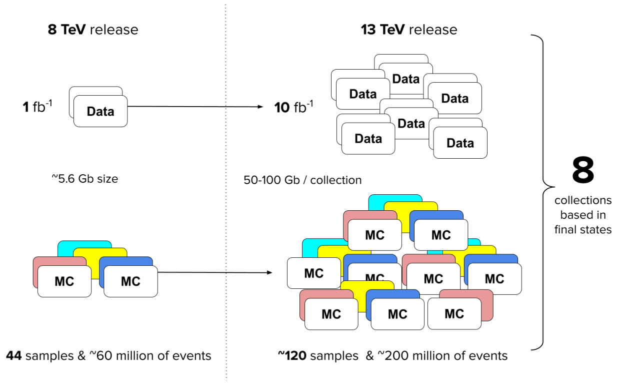


Figure 1. Summary of the composition of datasets in terms of the sizes of the last two public releases of real and simulated data for education by the ATLAS Collaboration.

A set of simulated samples suggests the physics analyses that can be performed, the so-called "signal" samples. Those contain interesting physics processes like the production of the Standard Model (SM) Higgs boson decaying in different modes, and some other Beyond the Standard Model (BSM) processes like the production of hypothetical particles like a Z' [8] or SUSY candidates [9], among others (see Figure 2).

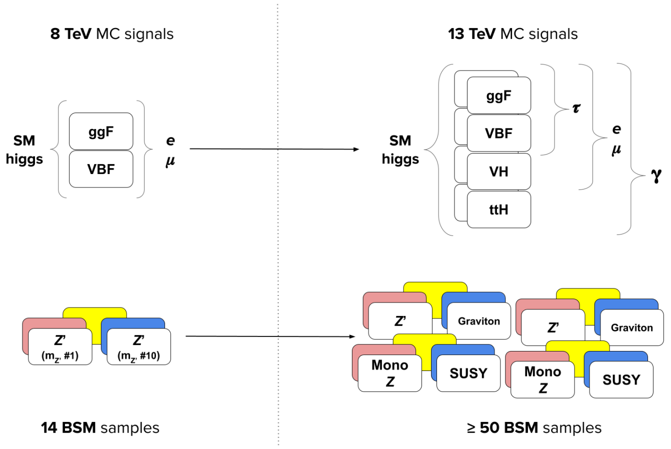


Figure 2. Overview of the composition of datasets regarding the SM and BSM simulated signal samples of the last public releases of real and simulated data for education by the ATLAS Collaboration.

2 ATLAS Open Data usage and users

2.1 Overview

The ATLAS Open Data project [5] is an effort of the ATLAS Collaboration dedicated to designing a holistic educational programme in HEP for undergraduate and master students all over the world. This effort involves the design, production, testing, validation and analysis of real and simulated Monte Carlo datasets that will be released to the public following the ATLAS Policy on Open Access with focus on educational and training objectives [10, 11]. Dedicated documentation, software tools, executable analysis examples and web-based applications are produced and delivered alongside the datasets.

2.2 Usage and Users

The ATLAS Open Data datasets and tools are already – and were for some time – used by a set of universities and dedicated academic or outreach projects [6, 12–14] with the aim to enhance their educational programmes in experimental particle physics, data analysis, computer sciences and statistics. In several cases, ATLAS Open Data has been used to develop, prepare and defend undergraduate and master theses in several universities worldwide [15–17].

The traditional way that those resources are reached by the user – like a student – is using a local computer element (e.g. a personal computer or university laboratory computer) with internet access to download the tools. Tools may include e.g. a Virtual Machine (VM), some of the analysis frameworks and a collection of Jupyter notebooks [18]. Also, the user consumes the content online, such as the documentation, video-tutorials and web-based applications. In terms of the datasets, they can be downloaded or be accessible in streaming: reading the ROOT [19] files over the internet as the analysis code runs over those files.

Figure 3 exemplifies a common user computing environment where a VM is used as an internal-and-personal server, allowing to run Jupyter notebooks, ROOT and other software directly in the host browser while accessing others resources on the ATLAS Open Data website, web-based apps and dedicated documentation [20].

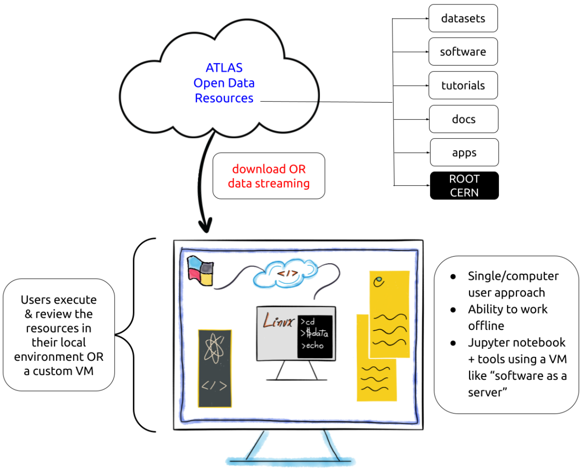

ATLAS Open Data Resources
datasets
software
tutorials
docs
apps
ROOT CERN
download OR data streaming
</>
Linux
>cd
>$data
>echo
</>
Users execute & review the resources in their local environment OR a custom VM
Single/computer user approach
Ability to work offline
Jupyter notebook + tools using a VM like "software as a server"

Figure 3. A pictorial description of the usage of the ATLAS Open Data resources in an offline environment. Tools like the VM and analysis frameworks provide a self-contained setup.

Notice that after the download of all the necessary datasets and resources, the user will be able to run and further develop the exercises without an internet connection (See Figure 4).

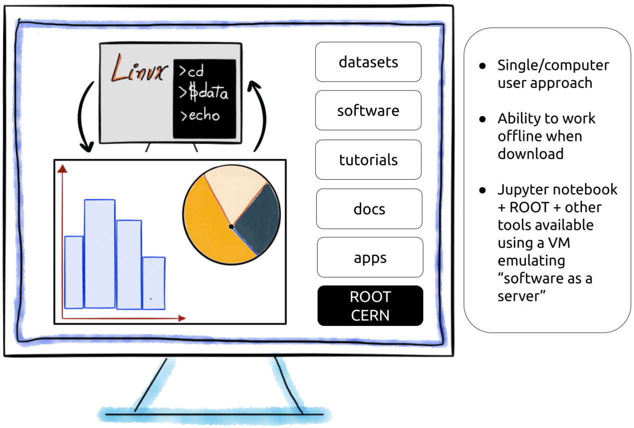

Linux
>cd
>$data
>echo
datasets
software
tutorials
docs
apps
ROOT
CERN
Single/computer user approach
Ability to work offline when download
Jupyter notebook + ROOT + other tools available using a VM emulating "software as a server"

Figure 4. A pictorial description of the usage of the ATLAS Open Data resources in an offline environment. Tools like the VM and analysis frameworks are self-contained and, together with the data located in a local volume, allow the execution of physics analysis examples – framework code or Jupyter notebooks – of multiple complexities, in term of the physics and the computing.

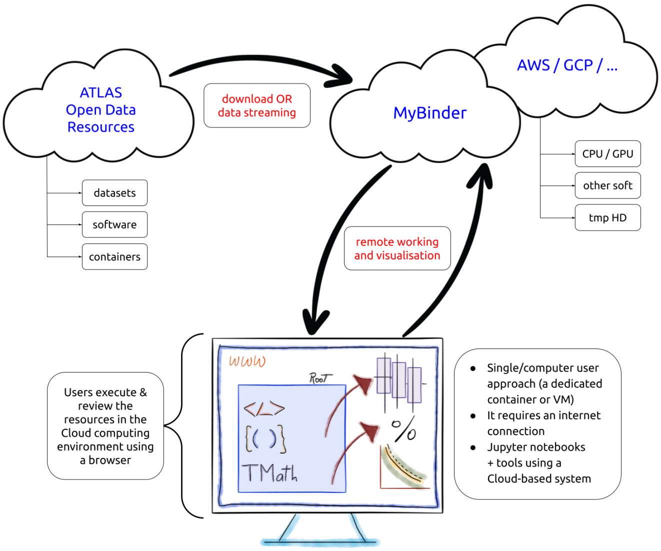
3 Institutional Infrastructure

Many (if not most) of the ATLAS Open Data users – students, professors, trainers – use such resources in institutional facilities: university computer labs, libraries and training centres equipped with the necessary infrastructure to execute a workshop, a masterclass or a complete academic period. Still, the current ATLAS Open Data tools are designed considering users that will use their computers, e.g. students using their personal laptops, at home or at school.

Over the last five years, such individual and institutional experiences helped identify ideas and technologies for designing and creating protocols to deploy and use the ATLAS Open Data resources using Infrastructure as a Service (IaaS) and Infrastructure as Code (IaC) in small and medium-size academic institutions.

For example, a popular way to use the ATLAS Open Data resources in workshops and classes is with the usage of a service called MyBinder [21]. This cloud computing service allows users to get a dedicated computer element in a cloud that – for free and anonymously – can be used to run an instance of Jupyter, and other software, defined in a Docker [22] container [23] for up to 12 hours. Figure 5 represents the way that a user uses MyBinder in combination with the ATLAS Open Data web-based resources to access, compute and modify the analysis examples. A web browser is the only application needed, something that is very attractive due to the simplicity and the minimal friction regarding local setups. A similar approach can and has been implemented using commercial cloud services. Individuals and small groups in universities have been looking for ways to implement Software as a Service (SaaS) to their trainees and students so that they can access Open Source resources and perform a lab course or a workshop.

ATLAS Open Data is now deployed in multiple university infrastructures [13, 14, 24] –on-premises and commercial– by the instructors and the local system administrators, or external collaborators so that the users, the students, can run exercises and conduct a practical course without the hassle of installing dedicated software locally. This brings an exciting and heterogeneous scenery where multiple users have multiple setups.



Users execute & review the resources in the Cloud computing environment using a browser

- Single/computer user approach (a dedicated container or VM)
- It requires an internet connection
- Jupyter notebooks + tools using a Cloud-based system

Figure 5. Schematic representation of how users are currently accessing, running, and manipulating ATLAS Open Data resources in the cloud. From the use of free services like MyBinder to institutional resources like SWAN [25] and ICTP [26], or commercial clouds rented for a specific time.

This SaaS and IaaS ecosystem is a fortunate sub-product of the multiple usage of Open Data for education, but it produces a "re-invention" of almost the same solution every time an individual or institution wants to deploy such resources to their students and instructors.

On the other side of the road are those individuals and small-size institutions – usually with limited SysAdmins expertise or assistance – that would like to deploy an educational analysis suite: Open Source and Open Data resources in the form of JupyterHub [27] (and its evolution as JupyterLab [28]) environments. **Here is where our next target audience resides:** educators and local resource providers who will help set up, maintain and run activities using the ATLAS (and others) Open Data and Open Source software analysis tools, profiting from a SaaS suite deployed as IaaS/IaC on-premises or commercial cloud services.

The next section explores a series of proposals to enhance the current ATLAS Open Data model, from using a single-client-based approach (Figure 6) to a multi-user approach. The plan is to develop and deliver a series of custom and reproducible instructions and scripts that allow the deployment – as easy as possible – of VMs and containers used to implement IaaS and SaaS in small and medium-size institutions.

4 Explorations and Proposals

As the usage of CERN Open Data [1] expands into different educational and training programmes [29, 30], more and more individuals and institutions want to set up a multi-user infrastructure that allows a continuous or on-demand running of experimental HEP courses and related computer and data sciences, while minimising the friction relative to setup and installation of such services by the users – students and professors – and the universities.

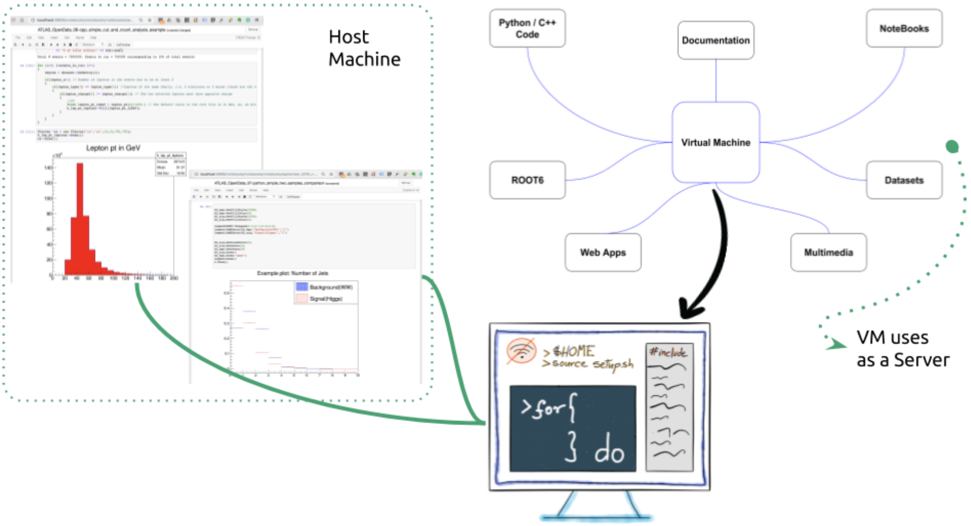


Figure 6. Schematic representation of how a VM is used to set a dedicated analysis suite for their users, allowing them to perform data analysis with Jupyter, not only to create and execute notebooks under several kernels (i.e. SaaS), but also as a simple IaaS by using the terminal provided by the Jupyter UI.

Taking in consideration such tendency, the proposal of this paper is the design, development and proof-of-concept of ready-to-use recipes and publicly release reproducible instructions and Open Source IaC software tools to easily, safely and predictably deploy, change, and improve SaaS in their on-premises facilities, or rented-commercial clouds.
Such a proposal takes advantage of the usage of collaborative tools, Continuous Integration (CI) and Continuous Delivery (CD) – or CI/CD [31] – and current ATLAS know-how in containerisation and software distribution to implement a complete pipeline between the developers and the users in the universities.

Figure 7 illustrates this overall idea by presenting a schematic view of how that deployed IaC, plus the ATLAS Open Data resources (hosted at CERN or elsewhere) are used in a multi-user environment. Such an environment does not need to be in a single physical place, e.g. a lab or classroom, but can also support users dispersed in different locations. This distributed approach is a popular and needed feature in the current tendency to remote learning and asynchronous learning.

4.1 ATLAS Open Data VMs and containers

The design of possible solutions has started, and a first approach has been implemented at CERN: the use of IaC under the Open Source software Terraform [32] and the CERN OpenStack [33] internal infrastructure during a winter student project in 2020. In this case, a series of Terraform configuration files are able to automatically create and configure an OpenStack instance, or VM, with all the needed elements to execute the ATLAS Open Data resources – a Ubuntu OS-based VM, plus the sequential installation and configuration of Python3, ROOT, JupyterHub and other Linux-related libraries.

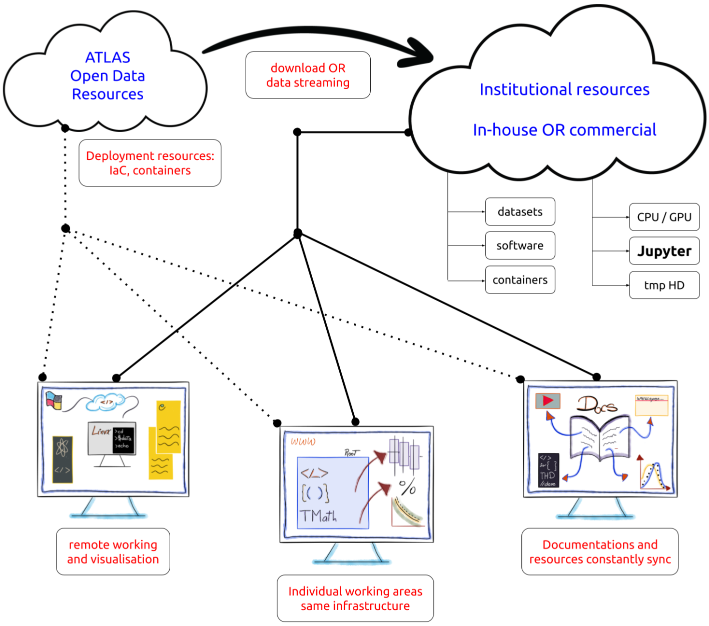


Figure 7. Schematic representation of a proposal on how users can access, run, and manipulate ATLAS Open Data resources in the cloud. From the use of university or institutional resources out of the HEP community (i.e. without previous access or knowledge to HEP software) and limited SysAdmins, or commercial clouds rented for a specific time by such universities or individuals running short-term workshops. A multi-user approach is implemented by design with as minimal setup as possible.

The final setup looks very similar to the one described in Figure 7, where a VM is used to deploy SaaS, making it possible to run a complete workshop or training sessions for users at CERN. In this particular proof-of-concept, the CERN cloud computing facility was used, accessible from inside the CERN network. The next steps in this approach are the design and proof-of-concept of dedicated Docker [22] containers that can be delivered and deployed in a similar fashion. This approach will allow to use more up-to-date techniques (CI/CD) and to make the process modular: multiple containers –or numerous different recipes to build a custom-made container– can be designed and be deployed into a cloud instance, instead of a monolithic VM which is less flexible to continuous updates. The publication of those resources will be in the form of Git repositories and container registries. It can also be a method that allows a single development for multiple target audiences: the same container that is deployed in a cloud instance can be installed locally in a single-user computer element, like a laptop, making the overall project more sustainable in the future.

5 Current and future activities

Right now, the gathering of information regarding the creation and support of containers is ongoing. First, this *learning* process is done in the context of the ATLAS Collaboration, i.e. internal resources and know-how so that we can profit from the expertise available in-house. The composition of the workforce to develop this project during 2021 is a researcher and a PhD student, besides the parallel collaborations and help between the group and partners at CERN, LAPP, and other universities in Europe and the Americas. The next figures give a overview of the different institutional scenarios, or multi-user architectures that we would like to support under this IaC project:

Figure 8 shows a multi-user architecture where the users' computing elements are connected to an institutional on-premises infrastructure, usually – but not mandatory – to an internal network. The cloud "Institutional resources" provide the SaaS, computing power, permanent storage and authentication methods so the students can use the mentioned educational resources. In such a model, it is expected that the datasets are also stored on-premises for faster access, and also to reduce the consumption of bandwidth by individual users, and avoid possible security threats coming from the exposure to the internet.

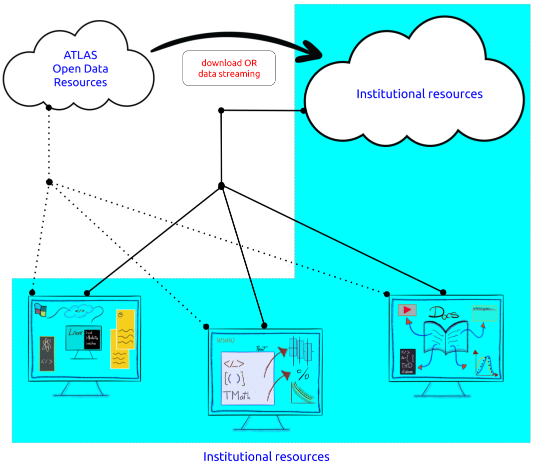


Figure 8. A multi-user architecture where the users' computing elements are connected to an institutional on-premises infrastructure. The cloud "Institutional resources" provide the SaaS, the educational resources and datasets, computing power, permanent storage and authentication methods. Internet access allows users to consult the web-based documentation and applications.

Figure 9 shows a similar multi-user architecture as in Figure 8, but, the users' computing elements, like the computers in a classroom, are connected to an external (e.g. commercial) cloud computing infrastructure. The cloud provides the SaaS, the educational resources and datasets, computing power, permanent storage and authentication methods. In this architecture, the institution or individuals use the flexibility to rent the computational resources only when needed. Because of this IaC approach, the deployment of the SaaS and educational resources should be relatively fast (e.g. up to one hour for a single-VM JupyterHub deployment), allowing to contract some commercial services (or book some dedicated external resources) by a predefined amount of hours or days, and a precise amount of computational resources, that will be needed for the activity.

Finally, Figure 10 shows a scenario where an institution provides SaaS to external users. This is a popular approach in terms of delivering computational and educational resources for a limited amount of time, like in an online workshop, an enhanced IPPOG-kind masterclass or thematic schools where students are following the activities remotely. The relevance also resides in the increase in the number of possible students who can have access to such activities traditionally done on-site with selected groups.

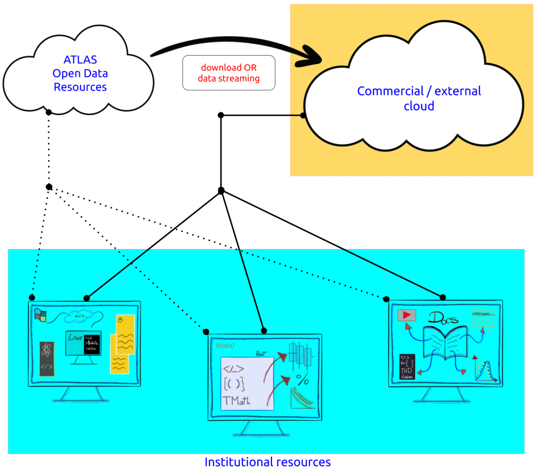


Figure 9. A multi-user architecture where the users' computing elements are connected to an external (e.g. commercial) cloud computing infrastructure. In this architecture, the institution or individuals use the flexibility to rent the computational resources only when needed.

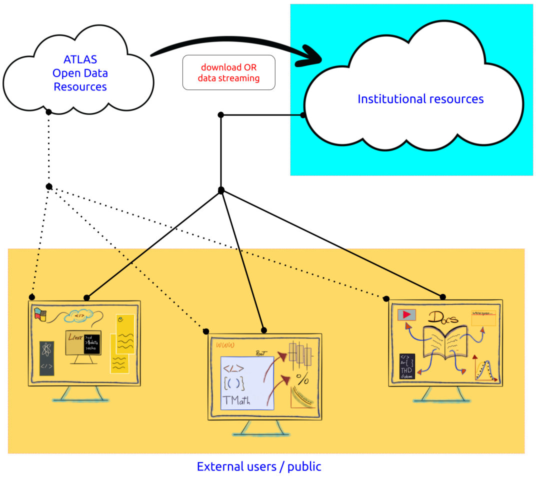


Figure 10. A multi-user architecture with external users. An institution delivers specific resources for a limited period to activities where students work remotely, and sometimes asynchronously. Internet access is needed, also to access to the ATLAS Open Data web-based documentation and applications.

It is important to mention this is approach does not mean an anonymous-user service, like MyBinder [21], but a case where temporary or long-term accounts are created and distributed among the participants that, beforehand, could pass a registration procedure. The usage of tools like JupyperHub [27] ensures minimal privileges for the users. Also, the SaaS/IaaS proposals will be constrained to live in a container (that can be, at the same time, hosted inside a VM), strengthening the security of the system in use.

6 Summary

Thanks to several Open Data and Open Source projects, multiple physics and computer sciences educational programmes worldwide have been implementing – or enhancing – their curricula to include hands-on sessions and long-term projects in experimental data analysis.

This paper describes how resources, like the ATLAS Open Data project, can be deployed in a multi-user approach, taking advantage of current and widely available software and computing tools. This proposal intends to deliver a series of IaC recipes and containers that can be used by educators and IT service providers in small and medium-size institutions to deploy IaaS and SaaS on-premises or on commercial clouds. A first proof-of-concept was performed using Terraform and the CERN OpenStack internal cloud service. The current activities aim to design, create, test and adequately deliver IaC and IaaS/SaaS recipes using VMs, containers, and ATLAS know-how to cover several multi-user configuration scenarios. Those recipes and containers will be public on the ATLAS Open Data website [5].

References

[1] CERN, ***The CERN Open Data portal***, https://opendata.cern.ch/, accessed: 2021-05-30

[2] ATLAS Collaboration, ***The ATLAS Experiment at the CERN LHC***. 2008 *JINST* **3** S08003.

[3] ***The International Particle Physics Outreach Group (IPPOG)***. http://ippog.org/

[4] ICTP. ***The CODATA-RDA Research Data Science Applied workshops*** http://indico.ictp.it/event/8170/. (2017), accessed: 2021-05-30

[5] ***The ATLAS Open Data project for Education***, http://opendata.atlas.cern/about/ (2021), accessed: 2021-05-30

[6] ATLAS Collaboration, ***Review of ATLAS Open Data 8 TeV datasets, tools and activities***. ATL-OREACH-PUB-2018-001 https://cds.cern.ch/record/2624572/

[7] ATLAS Collaboration, ***Proposal for an ATLAS endorsed 13 TeV data set for Outreach Purposes***. ATL-OREACH-PUB-2020-001 https://cds.cern.ch/record/2707171

[8] ATLAS Collaboration, ***Search for heavy particles decaying into top-quark pairs using lepton-plus jets events in proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector***, Eur. Phys. J. C 78 (2018) 565, arXiv: 1804.10823 [hep-ex]

[9] ATLAS Collaboration, ***Search for electroweak production of supersymmetric particles in final states with two or three leptons at $\sqrt{s}$ = 13 TeV with the ATLAS detector***, Eur. Phys. J. C 78 (2018) 995, arXiv: 1803.02762 [hep-ex]

[10] ATLAS Collaboration, ***ATLAS Data Access Policy***. ATL-CB-PUB-2015-001 https://cds.cern.ch/record/2002139/

[11] CERN, ***CERN Open Data Policy for the LHC Experiments***. CERN-OPEN-2020-013 https://cds.cern.ch/record/2745133/

[12] Sánchez Pineda, Arturo. ***The CEVALE2VE case***. ATL-OREACH-PROC-2017-001 https://cds.cern.ch/record/2241903/

[13] Camacho Toro, Reina, ***Outreaching particle physics to Latin America: CEVALE2VE and the use of ATLAS open data***. ATL-OREACH-PROC-2017-003 https://cds.cern.ch/record/2286585/

[14] Doglioni, Caterina, ***The ATLAS Open Data project***. ATL-OREACH-PROC-2018-001 https://cds.cern.ch/record/2637284/

[15] Domenico Franco, Maria. ***Reconstruction of the invariant masses of bosons of the Standard Model using public data from ATLAS Open Data***. CERN-THESIS-2017-239. https://cds.cern.ch/record/2293251/

[16] Garcia, Iskya. ***Perspectives and Evaluation of Dark Matter production in association with a light quark, a heavy quark (b-quark) or an electroweak boson in particle colliders at a centre-of-mass energy of 8 TeV***. CERN-THESIS-2017-217. https://cds.cern.ch/record/2291838/

[17] Evans, Meirin. ***Enabling Open Science with the ATLAS Open Data project at CERN***. CERN-THESIS-2018-099. https://cds.cern.ch/record/2630961/

[18] Kluyver, Thomas. *et al.* ***Jupyter Notebooks – a publishing format for reproducible computational workflows***. Positioning and Power in Academic Publishing: Players, Agents and Agendas. http://ebooks.iospress.nl/publication/42900/

[19] Brun, R and Rademakers, F. ***ROOT - An Object Oriented Data Analysis Framework*** *Nucl. Inst. & Meth. in Phys. Res. A* **389** pp 81-86, 1996.

[20] ***ATLAS Open Data 13 TeV docs***. http://opendata.atlas.cern/release/2020/documentation/ (2020), accessed: 2021-05-30

[21] Jupyter et al., ***"Binder 2.0 - Reproducible, Interactive, Sharable Environments for Science at Scale."*** Proceedings of the 17th Python in Science Conference. 2018. doi://10.25080/Majora-4af1f417-011

[22] Merkel, Dirk. *et al.* ***Docker: lightweight Linux containers for consistent development and deployment***. Linux Journal. 2014. https://dl.acm.org/doi/10.5555/2600239.2600241

[23] IEEE. ***Cloud Container Technologies: A State-of-the-Art Review***. 2019 DOI: 10.1109/TCC.2017.2702586

[24] Sánchez Pineda, Arturo and Mehlhase, Sascha. ***ATLAS Outreach: on the dissemination of High Energy Physics and Computer Sciences***. ATL-OREACH-PROC-2019-006 https://cds.cern.ch/record/2699514/

[25] CERN. ***The SWAN Service***. https://swan.web.cern.ch/swan/, accessed: 2021-05-30

[26] ICTP. ***The International Center for Theoretical Physics***. http://ictp.it/ (2021), accessed: 2021-05-30

[27] ***JupyterHub*** https://jupyterhub.readthedocs.io/ (2021), accessed: 2021-05-30

[28] ***JupyterLab*** https://jupyterlab.readthedocs.io/ (2021), accessed: 2021-05-30

[29] Serkin, Leonid, ***The release of the 13 TeV ATLAS Open Data: using open education resources effectively***. ATL-OREACH-PROC-2020-002 https://cds.cern.ch/record/2710384/

[30] Wunsch, Stefan, ***Using CMS Open Data for education, outreach and software development***. https://cds.cern.ch/record/2753429/

[31] ***Continuous Integration / Continuous Delivery*** https://en.wikipedia.org/wiki/CI/CD (2021), accessed: 2021-05-30

[32] ***Terraform*** https://www.terraform.io/docs/ (2021), accessed: 2021-05-30

[33] ***OpenStack*** https://docs.openstack.org/ (2021), accessed: 2021-05-30