An open source data science platform to foster innovative and production-ready machine learning systems

Romain Avouac (Insee)

1 Introduction

Following the European path towards Trusted Smart Statistics conference of 2018, the European Statistical System has adopted an ensemble of principles aiming at providing capacities to handle new data sources and statistical methods [1]. These principles involve simultaneously the need for new technical skills as well as innovative IT solutions. Not incidentally, an increasing number of public statisticians trained as data scientists have joined NSIs in recent years. However, these new profiles often find themselves isolated in national statistical systems, and their ability to deliver value is limited by several challenges.

The first challenge is related to a lack of proper IT infrastructures to tackle the new data sources that NSIs now have access to as well as the accompanying need for new statistical methods. For instance, big data sources, such as mobile phone data or receipts data, have been experimentally used to provide new statistical indicators (e.g. present population) or to refine existing ones (e.g. price indexes) [2]. However, such data requires huge storage capacities and distributed computing frameworks to be processed, which generally cannot be provided by traditional IT infrastructures. Similarly, the adoption of new statistical methods based on machine learning algorithms often require IT capacities (graphical processing units - GPUs) to massively parallelize computations [3].

Another challenge is the transition from innovative experiments to production-ready solutions. Production environments often differ from development environments, in such a way that the additional development costs needed to go from a proof of concept to a system working in production can limit the feasibility of this transition. Besides, in a production setting, a machine learning system needs both to be scaled to changing demand and to be properly monitored. Finally, it is generally the case that models need to be periodically or continuously updated, which requires proper management of their lifecycle in order to ensure reproducibility [4]. These various challenges highlight the need for both technical infrastructure and automation tools that can help statisticians and IT teams to implement the best practices advocated by the MLOps approach.

The last challenge is related to the difficulty of building proper environments for training programs on innovative statistical languages (e.g. big data frameworks such as *Apache Spark*) or tools (machine learning frameworks, databases, etc.). Since these are still rarely used in the actual production of official statistics, developing environments providing them are not available for training purposes, which in turn limit their widespread adoption. There is thus a need to provide reproducible environments in which statisticians can hone their skills by experimenting with innovative languages and tools.

Against that background, we developed the *SSP Cloud*[1], an open-innovation data science platform built upon state-of-the-art IT components to provide statisticians with scalable and reproducible environments [5]. The platform is based on three deeply structuring choices: cloud computing, object-storage and containerization, which enable to provide extensive computing resources – the benefits of a centralized infrastructure – while managing concurrency in the access to these resources and services isolation. We provide an extensive catalog of services to cover the entire lifecycle of a machine learning project : interactive services (R, Python, Julia) for the development phase and automatization tools (MLFlow to industrialize models training, argo-workflow to orchestrate parallel jobs) to develop production-ready systems.

The building principles of this platform where further refined into an open-source project : *Onyxia*[2]. As a result, public organizations can create their own internal instance of this modern data science platform and tailor it to the needs of their end users.

[1] https://datalab.sspcloud.fr

[2] https://github.com/InseeFrLab/onyxia-web

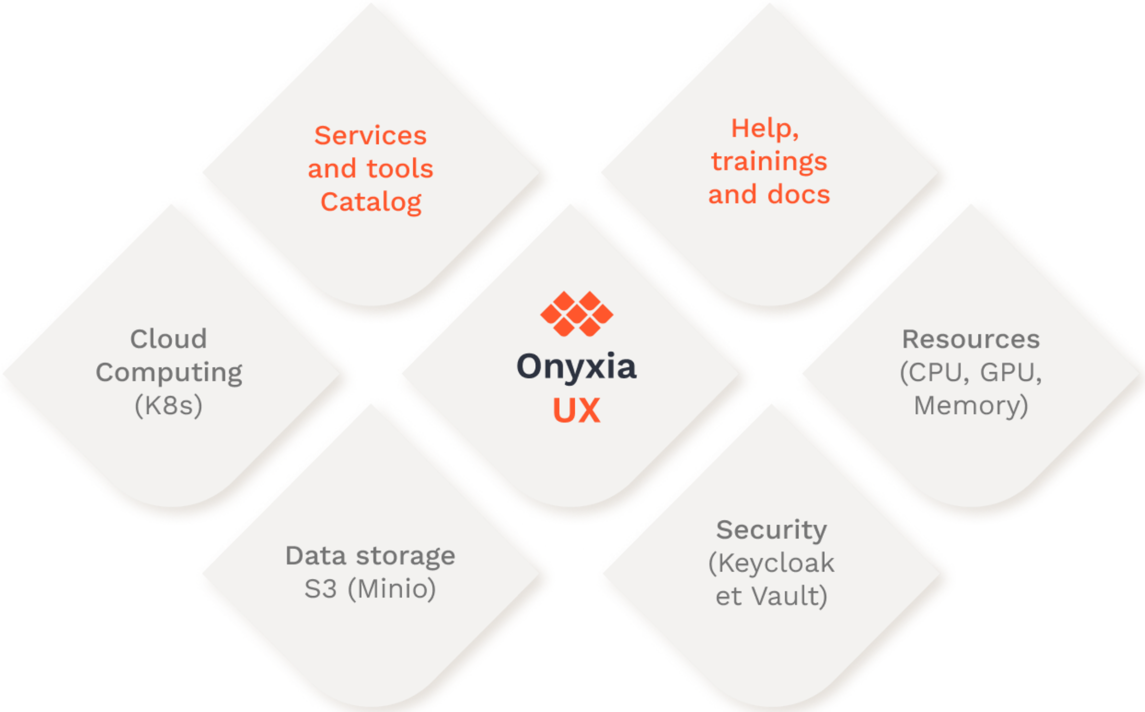
Services and tools Catalog
Help, trainings and docs
Cloud Computing (K8s)
Onyxia UX
Resources (CPU, GPU, Memory)
Data storage S3 (Minio)
Security (Keycloak et Vault)

Fig.1 - Components of the Onyxia project

2 Methods

2.1 An open platform scaled for innovative data science experiments

The platform we develop is best described as a *Datalab*: it aims at providing statisticians with both the physical and logical resources necessary to properly prototype and test a data science pipeline from end-to-end. Contrary to conventional IT infrastructures, access to these resources is immediate: there is no need for instance to ask beforehand to provision a storage space or an execution environment. It is open on the internet and can thus be accessed from everywhere, independently of the NSI infrastructures.

On the physical side, the platform is a private cloud based on a cluster of about 20 servers, for a total capacity of 10 TB of RAM, 1100 CPUs, 34 GPUs and 150 TB of storage. These resources enable the platform to scale to most data science experiments. Big data sources can be handled using frameworks such as *Spark* to distribute computations over the servers. Furthermore, the availability of graphical processing units (GPUs) allow for the training and use of large deep learning models, which are becoming prevalent in many machine-learning applications.

2.2 Architectural choices aimed at fostering scalability, autonomy and reproducibility

The platform is based on three deeply structuring choices: cloud computing, containerization and object-storage. These technologies have become the new standards in modern data science infrastructures.

In a cloud environment, the computer of the user becomes a simple access point to perform computations on a central infrastructure. This enables both ubiquitous access to and scalability of the services, as it is easier to scale a central infrastructure — usually horizontally, i.e. by adding more servers. This rationale has also influenced the choice of the data storage technology. The platform uses *MinIO*, an open-source and S3 compatible object storage framework. In this model, users can store data as “objects” (data and metadata) in their own “buckets” (data store). This storage model is optimized for scalability and intensive computations. Data access is also quite easy and cloud-native as buckets can be queried through a REST API.

In line with the cloud approach, the cluster is running on the open-source framework *Kubernetes* to deploy and manage containerized services. The choice of containerization is fundamental as it tackles the two main issues pertaining to data processing environments: managing concurrency in access to processing resources (RAM, CPUs, GPUs..) while properly isolating the running services from one another. As a result, users can both freely tailor services to their needs (programming language, system libraries, packages and their versions, etc.) while scaling their applications to the computing power and storage capacities it demands, e.g. by distributing computations over several containers.

Beside autonomy and scalability, these architectural choices also foster reproducibility of statistical computations. Contrary to traditional IT infrastructures — either a personal computer or a shared infrastructure with remote desktop access — the user must learn to deal with resources which are by nature ephemeral, since they only exist at the time of their actual mobilization. This fosters the adoption of development best practices, notably the separation of the code — put on an internal or open-source forge such as GitLab or GitHub — the data — persisted on a specific storage solution, such as MinIO — and the computing environment. The projects developed in that manner are usually more reproducible and portable — they can work seamlessly on different computing environments — and thus also more readily shareable with peers.

2.3 A catalog of services which covers the entire lifecycle of a data science project

The aim of the platform is to provide statisticians an environment to prototype their data science projects end-to-end. To do so, it offers a wide range of services which cover the entire lifecycle of a data science project:

- Data services: as aforementioned, the platform offers an object-storage service, but also various databases (PostgreSQL, MongoDB, ElasticSearch..)
- Execution environments : RStudio for R processing, Jupyter and VSCode for Python processing, distributed computation engines (Spark, Dask, Trino)
- Automatization tools : a batch deployment service (argo-workflow), a GitOps deployment service (argo-cd) and a MLOps service (MLFlow)
- Dissemination tools : visualization/BI services (Redash, Superset) and an API management platform (Gravitee)

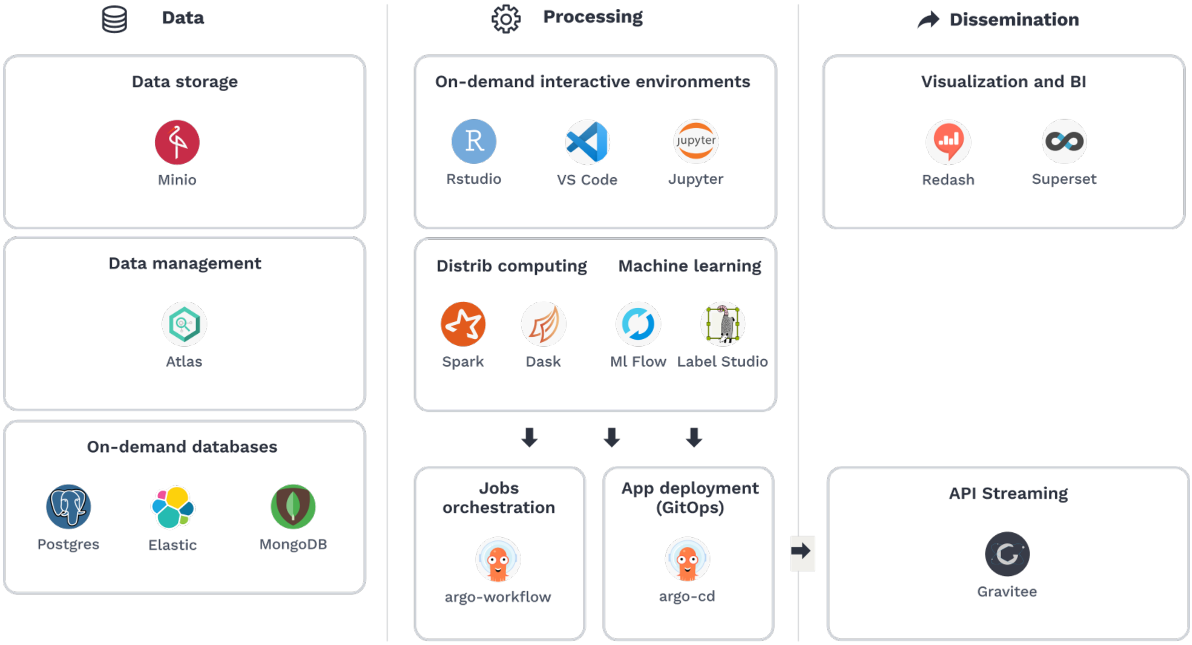

Fig.2 - Catalog of services of Onyxia

2.4 Providing reproducible environments for training programs

The fact that deployed services are simply containers running on a centralized infrastructure make it very easy to provide reproducible environments for training programs. Teachers can provide trainees with environments tailored to the specific needs of their training program, e.g. by pre-downloading necessary data, packages, etc. Besides, these environments can be deployed directly in the training catalog of the platform, so that trainees can launch them using a simple URL. This ensures a seamless training experience for both teachers and trainees.

2.5 A fully open-source project aimed at fostering reusability

In order to improve reusability, an open-source project was developed in order to make the deployment of similar state-of-the-art data science platforms possible in other organizations. The full code-source is available on GitHub and an accompanying documentation website thoroughly details the steps needed to instantiate a platform.

3 Results

The platform is now widely used in the national statistical system and even beyond, with about 800 unique users per month. These users form a dynamic community which, through the use of a centralized discussion canal, help improve the experience by reporting bugs and suggesting new features or even directly contributing to the codebase. It is also used in several data science schools and universities to host courses on statistical languages and frameworks. Finally, it is used to host innovative events such as hackathons, both at the national and international level.

The open source project has also been getting a lot of attention. Multiple organizations already have instantiated a platform or plan to do so, both at the national and international level, and also in the private sector. Some of these organizations have also started to contribute to the code base.

4 Conclusion

We developed a modern data science platform which aims at making data scientists in NSIs more autonomous by providing them with scalable computing resources and a wide range of modern data science services to prototype their projects from end to end. In order to encourage reusability, an open-source project is also developed to facilitate the instantiation of similar data science platforms in other organizations.

5 References

- [1] EUROSTAT, Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics) (2018). https://ec.europa.eu/eurostat/fr/web/european-statistical-system/-/dgins2018-bucharest-memorandum-adopted
- [2] UNECE, Big data and modernization of statistical systems, Report of the Secretary-General, 45th Statistical Commission (2013). https://unstats.un.org/unsd/statcom/doc14/2014-11-bigdata-e.pdf

- [3] UNECE, HLG-MOS Machine Learning Project (2021). https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project
- [4] S. Luhmann, J. Grazzini, F. Ricciato, M. Meszaros, K. Giannakouris, J.M. Museux, & M. Hahn, Promoting reproducibility-by-design in statistical offices, Proceedings for New Techniques and Technologies for Statistics (NTTS) (2019).
- [5] F. Comte, A. Degorre & R. Lesur, Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics, Courrier des statistiques (2022).