


Onyxia: An Open Source Cloud Native Data Science Platform

Frédéric Comte

Institut National de la Statistique
et des Études Économiques
INSEE

frederic.comte@insee.fr

Romain Lesur 

Institut National de la Statistique
et des Études Économiques
INSEE

romain.lesur@insee.fr

Abstract Onyxia is a project developed by the French public service that aims at providing a state-of-the-art working environment for data teams. It makes it possible to build a data science oriented cloud service using cloud technologies in an agnostic way. It avoids the use of cloud providers proprietary services and prevents its user from any vendor lockins from cloud providers.

1. Introduction

In recent years, big data and data science have become increasingly important and have an impact on many domains, including official statistics (Schweinfest and Jansen 2021). Big data offers new opportunities but also poses challenges for official statistics, such as big data methodologies, privacy concerns, and processing issues (Struijs, Braaksma, and Daas 2014). National statistical institutes need to address these issues at both strategic and operational levels (Kitchin 2015) including computational aspects (Van Der Loo 2021). This paper examines the challenges of building computational environments for data science and the solution proposed by the Onyxia project, an open source initiative of the French National Institute for Statistics and Economic Studies (INSEE).

2. Challenges and Solutions for Data Science Environments

In the rapidly evolving of a data-centric world, national statistical institutes must embrace big data and data science, which pose numerous operational challenges. These include the high cost of computing resources, the complexity of managing diverse software applications, and the need for reproducible working environments.

A key obstacle in data science is the need for robust computing resources to process and analyse large datasets, and train machine learning models, which often require graphics processing units (GPUs). Personal computers are not suitable for these tasks and data science environments need to be built. Traditional models, such as Hadoop clusters, tightly coupled data with computational capacity, which introduced limitations on scalability and portability, often leading to inefficiencies (Sobha Rani and Lakshmi 2018). The new paradigm, however, decouples data from computing power, allowing for dynamic scalability and improved cost efficiency. Leveraging cloud technologies, such as containerisation, container orchestrators, and object storage — whether through commercial cloud services or on-premises solutions — offers ways to allocate resources more efficiently and improve the efficiency of the computational environment (Barua 2021).

The adoption of cloud technologies also makes it possible to respond to other challenges, such as the diversity of applications and reproducibility.

In data science operations, the integration of multiple software applications is complex. Each stage of the data science process, from data collection and storage to processing, modelling, and visualisation, requires different tools. Ensuring seamless interoperability and management of these tools is a challenging task. Within national statistical institutes, the IT focus has traditionally been on transactional informatics, primarily for data collection and dissemination. While these operations remain crucial, there has been relatively little emphasis on developing analytical and data science platforms. Moreover, statisticians and data scientists often lack autonomy in choosing and implementing their tools, making it difficult to adopt and benefit from innovative solutions. It is important to identify and address these gaps and ensure that analytical efforts receive as much attention as traditional data collection and dissemination processes. Containerisation offers a particularly interesting technical solution for managing this diversity (Pahl 2015). In short, containers

can be managed in the same way from a technical point of view and offer the possibility to contain a very wide range of applications. This is a very practical way to manage the diversity of tools required for data science.

Reproducibility has become a fundamental aspect of scientific research and, increasingly, in official statistics (Panel on Transparency and Reproducibility of Federal Statistics for the National Center for Science and Engineering Statistics et al. 2022). This poses a number of challenges, as reproducibility requires in particular the archiving of code, data and the computing environment. Here again, containerisation, which allows a set of applications to be virtualised, offers a particularly an interesting technical solution: containers can be stored in the form of images, which in turn allow identical containers to be recreated. Containerisation is therefore a technology easing the creation of reproducible computing environments. Last but not least, it is this ability to replicate computing environments that makes containers such a useful technology for distributed computing (Bentaleb et al. 2022).

3. Challenges and Considerations in Using Public Cloud Services for Data Science

To build a cloud environment for data science, the most natural choice at first glance might seem to be to use public cloud offerings. The public cloud offers several benefits to organisations, including scalability and convenient access. However, there are risks associated with these benefits.

A predominant concern is the potential leakage of confidential information (Sun et al. 2014). Storing sensitive data with third-party public cloud providers poses risks such as unauthorised access and potential data breaches, concerns that can be magnified in shared multi-tenant environments. Despite the existence of preventative measures, the risk still remains (Cheng, Liu, and Yao 2017). In addition, several countries have enacted privacy regulations that limit the ability of public administrations, such as national statistical institutes, to use public cloud services. In order to mitigate these risks, national statistical institutes might choose to retain sensitive data within their private cloud or private data centers (on-premises).

Another challenge associated with using public cloud services is cost management. While cloud services may appear cheaper at first glance, the ease of scalability can result in high costs. This is particularly true for compute-intensive data science workloads. Controlling cloud costs is now a major issue and the way to manage this risk is to create a FinOps team whose role is to

monitor, control and limit the risks of over-consumption. Initially motivated by the search for flexibility and adaptability, organisations find themselves implementing administrative processes that make operations more cumbersome.

Another significant risk with commercial cloud services is vendor lock-in (Opara-Martins, Sahandi, and Tian 2016). Organisations may become increasingly integrated with the tools and services of one cloud provider, potentially making transitions to other platforms difficult and resource-intensive. This dependency can restrict adaptability and might increase operational costs over time. Recognizing these challenges, there is a trend towards endorsing cloud-neutral strategies (Opara-Martins, Sahandi, and Tian 2017). These methods aim to engage with cloud services in a way that reduces reliance on a single vendor’s specific solutions.

The risks associated with using public cloud services also explain why some organisations have adopted a hybrid cloud strategy or have started to repatriate their data from the public cloud to their own infrastructure (Jewargi 2023).

4. Onyxia: An Autonomous Data Science Environment Solution

Building robust and efficient data science environments presents many challenges. From the need for scalability, to the complexity of managing heterogeneous software applications, to the imperatives of reproducibility and sovereignty, the data science landscape demands innovative solutions.

The Onyxia project, initiated by the French Public Service¹ and available at onyxia.sh, is an open source project aimed at creating self-sufficient data science environments in the cloud or on-premises. This project can be seen as a “Platform as a Package” (PaaS) solution for organisations wishing to create a data science environment based on cloud technologies.

It is designed to provide a helpful working environment for data teams, and specifically aims to reduce the complexity of setting up a data science oriented cloud platform.

This initiative makes it easier to set up a modern data science working environment with a focus on accessibility, especially for data analysts or data scientists who are not well versed in using cloud technologies. Onyxia addresses this

¹French National Institute for Statistics and Economic Studies (INSEE), supported by the Interministerial Digital Directorate (DINUM, CodeGouvFr)

issue by providing an interconnected environment designed to streamline the management of multiple software applications. It aims to act as a bridge between multiple open source backend technologies, and provide a conducive working environment for its users.

Onyxia also enables organisations to adopt DevOps and MLOps practices by providing automated environment provisioning, version control and infrastructure as code capabilities. Through containerisation, it ensures consistent and reproducible environments, fostering collaboration between data analysts, data scientists, data engineers, machine learning engineers and operations teams.

Onyxia can be installed either on public cloud platforms or on-premises, embodying a cloud-neutral perspective to avoid vendor lock-in concerns. Beyond this cloud flexibility, Onyxia is committed to user empowerment. While its interface streamlines tasks, it also educates users on how to replicate their actions directly from the command line. This ensures that users are not only never locked into a particular cloud provider, but are also equipped with the skills to operate autonomously outside of the Onyxia environment.

Finally, by providing access to advanced technological tools, Onyxia aims to make data science more accessible to a broader audience within organisations. The design of the platform allows for user-driven customisation, providing flexibility to meet specific needs.

5. SSPCloud: Fostering Open Innovation and Collaboration in Europe

The SSPCloud² is an open infrastructure dedicated to open innovation (Comte, Degorre, and Lesur 2022). Run by the French National Institute for Statistics and Economic Studies (INSEE), it is offered to the European Statistical System and many European universities. The initial goal of the SSPCloud is to support open innovation and collaboration. Leveraging the advances offered by Onyxia, the SSPCloud is now more than just an open infrastructure; a community has been built around this infrastructure to share knowledge and collaborate in data science across Europe.

The community of the SSPCloud users now includes several hundred data scientists, professors and researchers in data science. Recognising the diverse needs of these professionals, the SSPCloud provides a range of data science

²<https://datalab.sspcloud.fr>

tools and compute resources ready to support a wide range of research and analysis activities. But beyond the service offered, the main aim of the SSP-Cloud is to support open innovation. The platform encourages its users not only to use its services, but also to share their achievements and courses, thereby enriching the community's collective knowledge base.

This culture of knowledge sharing is reinforced by the Onyxia principles that underpin SSPCloud. Promoting user autonomy and flexibility, SSPCloud advocates an agnostic approach, ensuring that users are not tied to the Onyxia environment. It also actively promotes the idea that every tool and service should be as accessible outside the platform as it is inside. This fosters an environment in which users are not just passive consumers, but active contributors, freely innovating.

The mission of the SSPCloud thus goes beyond the mere provision of a service and is positioned at the intersection of technology and collaborative practices. By supporting open innovation and knowledge sharing, the SSPCloud offers an ambitious approach to promoting the adoption of data science within the official statistics community.

6. Perspectives

As an open source initiative, Onyxia attracts a growing community of contributors from official statistics, research centres and even industry. This collaborative approach will shape the direction of the project in the coming years.

Data scientists, developers, and organisations are invited to contribute and improve the project on an ongoing basis. This open model ensures that Onyxia stays up-to-date with the evolving needs of the data science community.

The Onyxia community is not limited to code; it is also a hub for sharing knowledge and best practices. Collaborative efforts also focus on creating a repository of insights, tutorials, and resources that benefit the entire data science community.

Bibliography

Barua, Hrishav Bakul. 2021. 'Data Science and Machine Learning in the Clouds: A Perspective for the Future'. <https://doi.org/10.48550/ARXIV.2109.01661>.

Bentaleb, Ouafa, Adam S. Z. Belloum, Abderrazak Sebaa, and Aouaouche El-Maouhab. 2022. 'Containerization Technologies: Taxonomies, Applications and Challenges'. *The Journal of Supercomputing* 78 (1): 1144-81. <https://doi.org/10.1007/s11227-021-03914-1>.

Cheng, Long, Fang Liu, and Danfeng (Daphne) Yao. 2017. 'Enterprise Data Breach: Causes, Challenges, Prevention, and Future Directions'. *WIREs Data Mining and Knowledge Discovery* 7 (5): e1211. <https://doi.org/10.1002/widm.1211>.

Comte, Frédéric, Arnaud Degorre, and Romain Lesur. 2022. 'Le SSPCloud : Une Fabrique Créative Pour Accompagner Les Expérimentations Des Statisticiens Publics'. *Courrier Des Statistiques*, no. 7 (January): 68-85. <https://www.insee.fr/fr/information/6035940?sommaire=6035950>.

Jewargi, Kiran. 2023. 'Public Cloud to Cloud Repatriation Trend'. *Scholars Journal of Engineering and Technology* 11 (1): 1-3. <https://doi.org/10.36347/sjet.2023.v11i01.001>.

Kitchin, Rob. 2015. 'Big Data and Official Statistics: Opportunities, Challenges and Risks'. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2595075>.

Opara-Martins, Justice, Reza Sahandi, and Feng Tian. 2016. 'Critical Analysis of Vendor Lock-in and Its Impact on Cloud Computing Migration: A Business Perspective'. *Journal of Cloud Computing* 5 (1): 4. <https://doi.org/10.1186/s13677-016-0054-z>.

———. 2017. 'A Holistic Decision Framework to Avoid Vendor Lock-in for Cloud SaaS Migration'. *Computer and Information Science* 10 (3): 29. <https://doi.org/10.5539/cis.v10n3p29>.

Pahl, Claus. 2015. 'Containerization and the PaaS Cloud'. *IEEE Cloud Computing* 2 (3): 24-31. <https://doi.org/10.1109/MCC.2015.51>.

Panel on Transparency and Reproducibility of Federal Statistics for the National Center for Science and Engineering Statistics, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, and National Academies of Sciences, Engineering, and Medicine. 2022. *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies*. Washington, D.C.: National Academies Press. <https://doi.org/10.17226/26360>.

Schweinfest, Stefan, and Ronald Jansen. 2021. 'Data Science for Official Statistics: Views of the United Nations Statistics Division'. *Harvard Data Science Review* 3 (4). <https://doi.org/10.1162/99608f92.c1237762>.

Sobha Rani, Neelam, and Neelam Venugopal Muthu Lakshmi. 2018. 'Major Challenges with Hadoop Distributed Framework: An Overview'. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3200427>.

Struijs, Peter, Barteld Braaksma, and Piet Jh Daas. 2014. 'Official Statistics and Big Data'. *Big Data & Society* 1 (1): 205395171453841. <https://doi.org/10.1177/2053951714538417>.

Sun, Yunchuan, Junsheng Zhang, Yongping Xiong, and Guangyu Zhu. 2014. 'Data Security and Privacy in Cloud Computing'. *International Journal of Distributed Sensor Networks* 10 (7): 190903. <https://doi.org/10.1155/2014/190903>.

Van Der Loo, Mark P. J. 2021. 'Computing in the Statistical Office'. *Statistical Journal of the IAOS* 37 (3): 1023-36. <https://doi.org/10.3233/SJI-210862>.