EUROPEAN
STATISTICAL
SYSTEM

The Eurostat Data Science Lab

Tools and infrastructure for sharing, testing, developing, deploying and integrating data sources

© Shutterstock 2024

The project

The EC Data Platform is an environment developed by the central IT services of the European Commission (DG DIGIT) to offer the tools and infrastructure necessary for sharing, testing, developing, deploying and integrating a wide range of data sources. The platform also provides data analytic solutions to analyse, process, manipulate and explore data. Eurostat in cooperation with DG DIGIT deployed a version of this platform that is adjusted to the needs of statistics.

In line with the aims of the European Statistical System (ESS) Innovation Agenda, the Data Science Lab is a flexible and ready-to-use laboratory for data science, designed to innovate and experiment with data, using pre-installed advanced analytics tools (e.g. R and Python).

Through these innovative techniques, the Data Science Lab aims to provide services to share and reuse data, experiment, prototype and develop applications.

The motivation

Eurostat created the Data Science Lab to enable users to explore and analyse data from an accessible platform. The goal is to set up capabilities that can serve researchers, data scientists, professional statisticians and policy analysts working with official statistics and micro datasets.

By offering tools to adapt, create and innovate new statistical products and insights, the Data Science Lab encourages the development of new methodologies, or improving existing ones, and applications in data analysis. As such, the Data Science Lab seeks to innovate and experiment with data through use cases such as the Web Intelligence Hub (WIH), a key initiative of Eurostat and the ESS that aims to produce high-quality statistics based on web content.

The Data Science Lab also leverages existing open source tools, deploying them in diverse technological contexts. This ensures that the Data Science Lab remains adaptable and relevant across different technological environments.

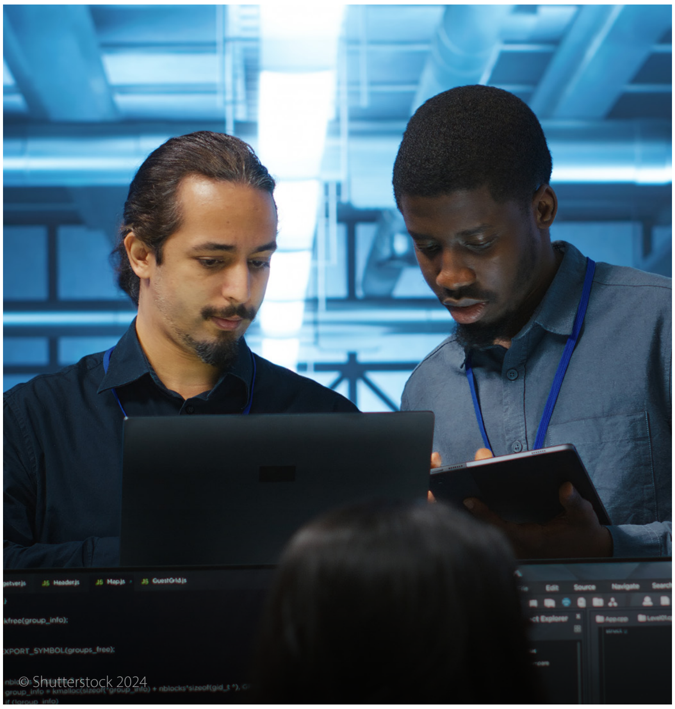
The methodology

Eurostat conducted an evaluation of various open source solutions, which was used for defining the requirements of the cloud-agnostic Data Science Lab.

The robustness of the technical infrastructure was tested during the 2023 EU Big Data Hackathon. This event provided a dynamic environment to assess the performance and scalability of the solutions.

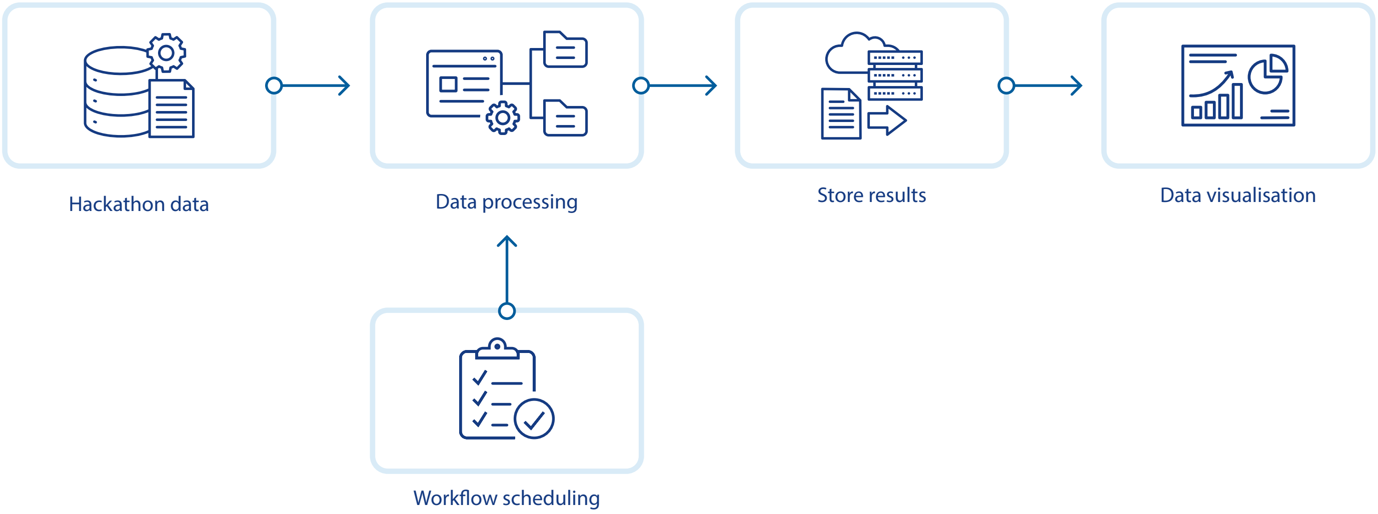
Following the Hackathon, an updated version of the infrastructure was deployed to support the onboarding of initial projects at Eurostat. These projects include:

- experimental indicators for traffic and mobility;
- trainings on R and Python;
- analysis of financial transaction data;
- testing Large Language Models (LLM);
- processing data on online job advertisements (OJA).



© Shutterstock 2024

Technologies involved: Cloud services, Kubernetes cluster with Helm charts, EU login authentication, open source data science tools.



The team

- **Project owner:** The ESS Directors group for methodology and IT (DIME-ITDG)
- **Solution provider:** EC DG DIGIT
- **Service manager:** Eurostat

The timeline

- **Project initiation:** 2023
- **The first phase 2023–2024:** Deployment of the Data Science Lab by Eurostat for use by its staff.
- **The second phase 2025 onwards:** Making the Data Science Lab accessible to staff of ESS members to create a community of data scientists who can collaborate around data sources available at Eurostat. This will allow them to co-experiment and develop innovative work in this area.