Opportunities in the coming year

SPACIOUS challenges and science

Xavier Luri

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

GENIUS heritage

Gaia European Network for Improved User Services (GENIUS) – FP7 2013

- Predecessor of SPACIOUS, focused on Gaia so a narrower scope "*The objective of the GENIUS proposal is to contribute to the design and implementation of the Gaia archive*"
- **Architecture:** Developed the Gaia Data Mining and Big Data prototype (GDAF), based on Hadoop, YARN, and Spark, deployed on bare-metal infrastructure.
- **Scope:** Primarily aimed at enabling access and validation tools for the Gaia archive, with some early exploration of data mining and visualization.
- **Limitations:** The prototype was not scalable for widespread community use. Deployment was tied to specific hardware environments, making it difficult to maintain, expand, or adapt to new missions. Lacked flexibility for cloud integration and broader interoperability.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

Science Platform Cloud Infrastructure for Outsize Usage Scenarios - SPACIOUS (Horizon Europe)

The massive growth of ESA mission datasets (Gaia DR4 ~0.5 PB, Euclid forthcoming) made the GENIUS approach insufficient. A more flexible, scalable solution was needed to deploy, maintain, and expand a viable service for Big Data astrophysics.

- Core Innovation: Builds on GENIUS/GDAF experience but introduces the Big Data Analytics Framework (BDAF):
  - Cloud-native and virtualized design, deployable on supercomputing centers (BSC), OpenStack, and commercial clouds (Azure, AWS, Google Cloud).
  - Compliant with European Open Science Cloud (EOSC) standards for interoperability and sustainability.

- Objectives:
  - Provide a portable, open platform for Big Data and Data Mining, enabling "code-to-data" analysis rather than downloading massive datasets.
  - Support Gaia and Euclid exploitation through internal and external scientific challenges.
  - Leave a legacy system with templates and virtualization for future missions.

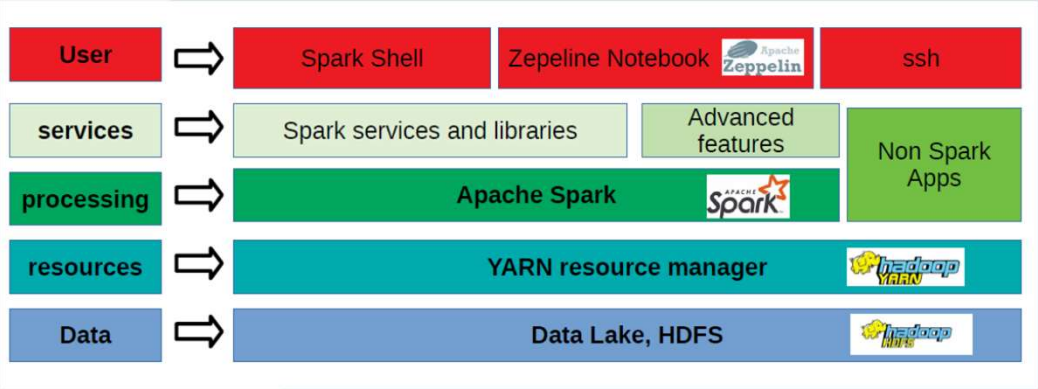
- Impact: Aims to democratize access to Big Data tools, train new professional profiles, and influence future ESA data exploitation paradigms.

Funded by the European Union

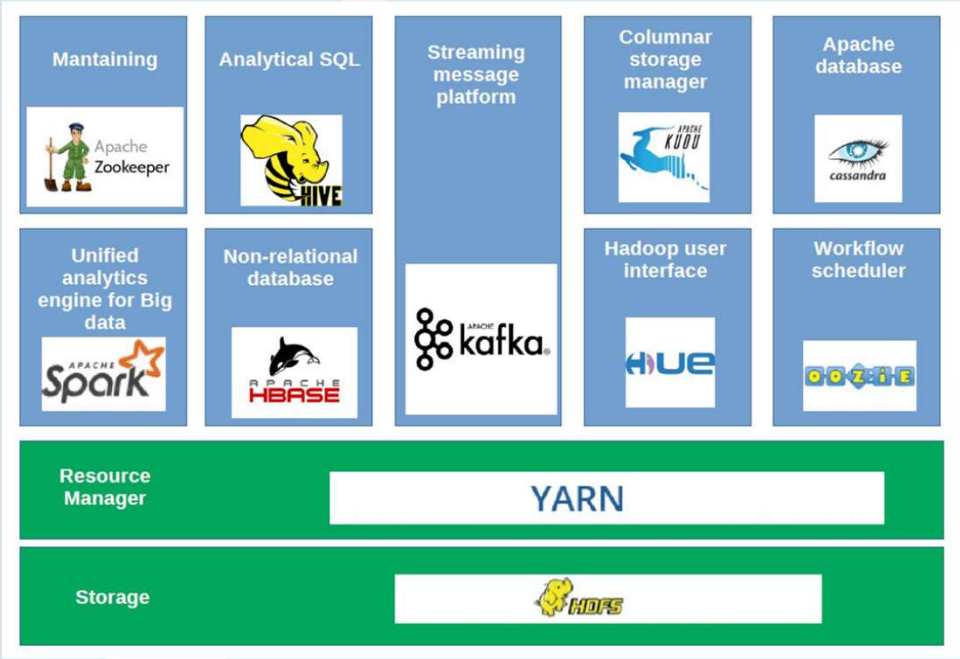
SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

GENIUS



SPACIOUS



GDAF has been a successful prototype. However, it can be only used by a limited number of users. The Big Data Analytics Framework (BDAF) that we will develop under SPACIOUS will use GDAF as a starting point. BDAF will rely on similar technologies, but its concept is wider as it will be designed to be compatible with different environments, such as on-premise facilities or cloud services. We aim to build a virtualized BDAF platform capable of being deployed in different computational environments. The architecture of BDAF will be initially different for each computational environment with the ultimate goal of generating as a final product a virtual BDAF based on Kubernetes compatible with as many infrastructures as possible.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

SPACIOUS goals

SPACIOUS objectives are both scientific and technological:

1. The Scientific objective: **To increase the number of scientific products and publications** using data from ESA archives by addressing key astrophysical research problems (challenges) that require Big Data and Data Mining technologies for the present and upcoming Gaia and Euclid data releases.

2. The Technological objectives:
   - **To increase the number of institutions and researchers with the capability of performing an advanced processing of data** by **developing** a Data Mining and Big Data framework (BDAF) including ESA data combined with other archives **enabling** the analysis of large data sets. This platform is essential to achieve our scientific objective.

To increase the number of scientific teams (users) exploiting data from ESA archives by **opening** the BDAF framework to the community, minimising the human resources required to set up and effectively use a Big Data infrastructure.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

SPACIOUS platforms

Project resources:

- Barcelona Supercomputing center: 162 VCPUs - 500GB RAM - 11 VM Instances – up to 1PB of disk
- Google Cloud: flexible assignation of resources (~100k€)

Additional resources:

- West Cambridge Data Centre (WCDC): 800 vCPU, up to 0.9 PB of storage
- Edinburgh Parallel Computing Centre (EPCC): 800 vCPU, up to 0.9 PB of storage

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

Internal challenges

- Challenge 1: The star formation history and the stellar initial mass function of the Milky Way disc
- Challenge 2: Completing and characterising the open cluster census of the Galaxy
- Challenge 3: The cosmological challenge: large-scale power-spectrum analysis
- Challenge 4: The Gaia enhanced data products

Ongoing, development to be completed in 2026, but application to DR4 (early Dec. 2026) very difficult. Ask for SPACIOUS extension?

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

External challenges

Open call for proposals just closed

- Received five proposals
- Evaluation of technical feasibility ongoing
- The Resource Allocation Committee will evaluate their scientific merits
- Execution planned during 2026
- May have enough resources for more challenges (TBD, 2nd call in 2026?)

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

Contribution to Gaia DR4 papers

SPACIOUS resources offered to DPAC for the data processing of Gaia DR4 papers

Will contribute to two papers in 2026, to be published with DR4:

- The Bottom of the Main Sequence
- Star clusters in Gaia DR4: The interplay of multiplicity, variability, and dynamics

Will also contribute to the computation of the Gaia DR4 selection function (supporting the GaiaUnlimited team)

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

Google cloud evaluation

One of the goals of the project is to evaluate the use of commercial cloud services (specifically Google cloud) for massive data processing.

Based on previous projects, the use of a commercial cloud can be a cost effective option for science data exploitation:

- Good (and large) scalability
- Resources on-demand, no cost while not using them
- Many hardware and software options

The question is how they compare *in practice* with the use of scientific data centers.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

We have been experimenting with Google cloud for more than a year. Some personal conclusions:

- Many, many options, sometimes it can be confusing. Also, technology and the resources available can change quickly.
- Efficient configuration and use requires specialized personnel (the same applies to an institute computing centre, but here we face it more directly, for the moment).
- Errors cost money (they always do, but in this case the cost is more direct)

- Allows access to the best suited hardware for a given task (e.g. GPUs, SSDs). But cost optimization takes time and resources.
- Great flexibility, can quickly set-up a tailor-sized resources (cluster, disk, etc.) quickly.

Personally, I think SPACIOUS can be a good tool to hide much of the complexity while allowing a transparent access to commercial cloud resources. But it will still need a specialized technician to manage it.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

In short, I think the (intensive) use of the commercial clouds (for scientific data processing) would require the support of a specialized team to manage its complexity and take advantage of its flexibility. Relying on this team tools like SPACIOUS can provide a simple, familiar work environment for scientists that can then benefit of its scalability, variety of resources and cost effectiveness.

> ***To increase the number of scientific teams (users) exploiting data from ESA archives*** *by* ***opening*** *the BDAF framework to the community, minimising the human resources required to set up and effectively use a Big Data infrastructure.*

Not unlike the current supercomputer centres; specialized teams manage the complexities of the hardware, maintenance and software deployment. Scientist then "log-in and calculate" using familiar tools (MPI, slurm, etc.).

Some centres (e.g. BSC, WCDC, EPCC) are moving in this direction (private clouds, data mining environments), from HPC to HTC.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

In 2026 we will continue experimenting with Google cloud, but in this case with full-sized scientific challenges. Some examples of already available findings:

Storage: *as expected, local SSD is considerably faster at both writing and reading data (to a maximun of 700 MB/s) at the cost of a higher budget, but proving useful for specific tasks like storing program files and offering low latencies when needed. The buckets are still a powerful service when storing large files due to its massive storage capability and the fact that multiple services can access their data simultaneously. As to the second point, buckets have proven to be scalable and well-behaved when serving multiple data requests. Its transfer speed is limited to around 180 MB/s but does not seem to be affected by the number of VMs reading its data, making it the best service to store the massive datasets involved in the SPACIOUS project.*

Virtual machines: high-end, expensive virtual machines can be more cost efficient that cheaper ones for a given task. The task is completed so more quickly, and the total cost is lower. More expensive virtual machines with GPUs can be very cost efficient if the SW can properly use them.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

SPACIOUS legacy

Our aim is that SPACIOUS developments are used beyond the end of the project

- Continued usage at the data centers participating in SPACIOUS. **Potential extended collaborations. Submission of new projects?**

- Can be adopted by individual teams:
  - Installation in local clouds/clusters
  - Commercial cloud deployment templates
  - **To be promoted**

- Interest by ESA for their own developments of data mining environments (next slide).

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

SPACIOUS and ESA DataLabs

ESA has developed DataLabs, a data-science / e-science platform under the paradigm “bring code to the data” to enhance the exploitation of ESA’s mission data.

The SPACIOUS features could complement the existing ones in DataLabs by integrating in it.

ESA is also exploring the possibility of deploying DataLabs beyond its data centers, to national institutes or research centers. SPACIOUS could benefit/join in this effort.

Funded by the European Union

SPACIOUS is a Horizon Europe HORIZON-CL4-2023-SPACE-01-71 project funded under grant agreement no. 101135205

SPACIOUS

Thank you!

SPACIOUS

Website.com
info@spacious.com