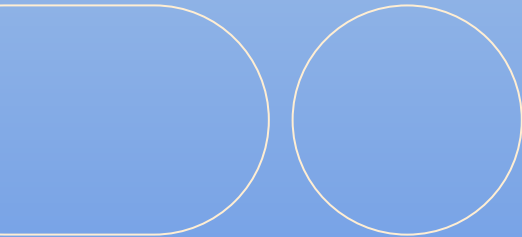


From Gaia data management solutions to an intelligent Multi-Agent system for data models and product discovery

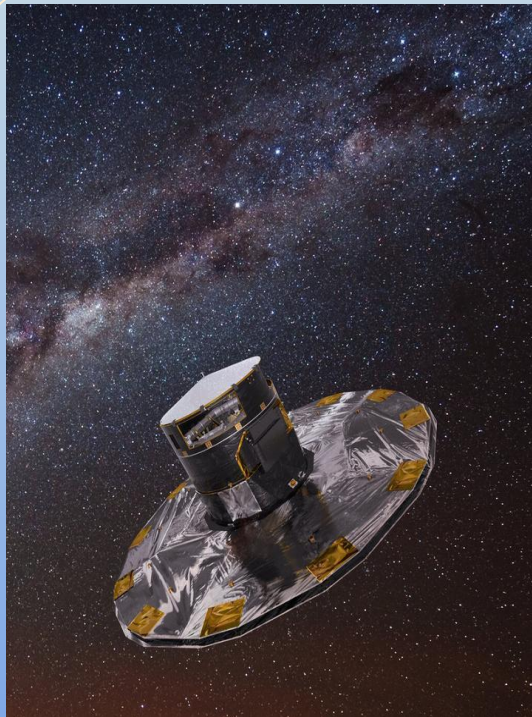


Sara Gelsumini - INAF-OATo

Main contributors: D. Busonero, E. Licata

DT-Lab KO meeting 30/09/26 Bologna

From Gaia data management solutions



Credits: ESA/DPAC

The goal is the study and implementation of a prototype **open-source platform tailored** for supporting and allowing scientific analysis on subsets of extracted Gaia data and metadata, alongside the Gaia database and data lake at DPCT, e.g., Gaia GW use case on a different platform.

To create a database and filesystem platform capable of extracting all sources within different specific areas of the sky simultaneously, and associating with each source the information regarding its transits and its calibration data.

DATA OVERVIEW

- CompleteSource: source information
- AstroElementary: transit information
- CrossMatch: association of sources and transits

From Gaia data management solutions

HDF5 converter software for data conversion from Gaia gbin format to HDF5. This required a complex phase of testing and fine-tuning of the configuration parameters to achieve the required performance. The software heavily leverages reflection and recursion to explore each input object, regardless of its type, structure, or complexity, and creates a corresponding HDF5 file mapping.

Spoke 3 - WP4 ↔ IGUC - Cherrydata

- Postgres DBMS

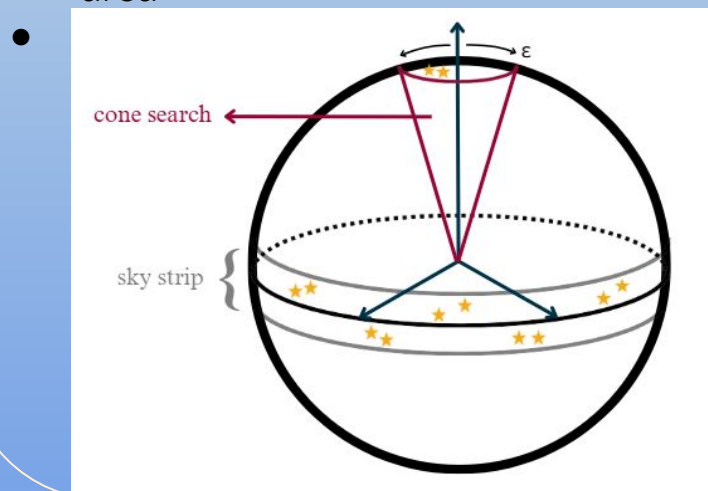
- Ayra DB

- Rucio DMS

- HDF5 file format

Some examples of types of **queries possible**:

- Find all the transit information associated with specific source
- Find all the sources and all transit information associated within a specific area



to an intelligent Multi-Agent system

Past and ongoing Science **Ground Segment projects** (like Gaia) in astrophysical missions is characterized by the increasing complexity of data volumes and processing pipelines. It is optimal to introduce an orchestration system based on **Large Language Model Agents** capable of transforming natural language requests into complex workflows, ensuring flexibility, efficiency, and full control over the computing infrastructure.

Operational Efficiency:



Reduction of the time elapsed between data acquisition and the production of scientific products.

Scalability:



The plugin-based architecture (MCP) makes the system easily adaptable to different missions or future experiments.

Lowering Technical Barriers:



Enables researchers who are not experts in specific pipelines to interact with complex systems via natural language.

Security and Costs:



The on-premise approach eliminates licensing fees and prevents the leakage of sensitive data or intellectual property to external cloud providers.

to an intelligent Multi-Agent system

The system is based on a **Multi-Agent Architecture** orchestrated via LangGraph or with an open source off-the-shelf harness like Deep Agents or TrueForge. The logical core consist of general-purpose LLMs (e.g., Qwen 3.8, Llama 3.3, or Mistral). Inference is managed locally on the institution's hardware using Ollama or similar systems.

Agents operate according to a hierarchy of specialized roles:



Coordinator Agent: Responsible for decomposing the user prompt into sub-tasks and supervising the decision graph.



Worker Agents: Specialized in specific domains (e.g., metadata querying, calibration, statistical analysis).



Validator Agent: Tasked with verifying the syntactic and logical consistency of the action plan before submission.

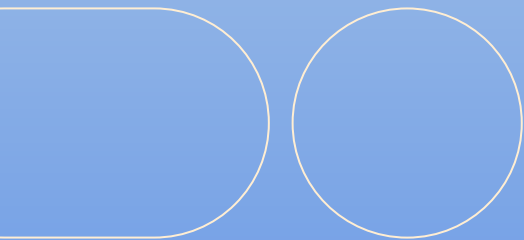
From Gaia data management solutions to an intelligent Multi-Agent system

From the knowledge of **Gaia** mission and in particular the experience in designing, developing, managing **complex data reduction pipelines**, software lifecycle, data modelling, database structuring, workflow orchestration.



To reduced replication of development effort, faster Science Ground Segment deployments for **future missions**, improved interoperability across Data Centres, and introduction of **AI-assisted data modelling workflows** that accelerate scientific exploitation.

Thanks for listening!



sara.gelsumini@inaf.it