Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Status of the IGUC Project

Deborah Busonero (INAF); Paolo Giacomazzi (CherryData)
On behalf of INAF and Leonardo/CherryData team

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024

IGUC - Interoperability Data Lake for Gaia Use Case

- IGUC borned as an additional WP (WP2_G) of the IDL project.
- The project aimed to expand the IDL project by bringing a new typology of astronomical big data offering a new challenge in big data management and recovery: the Gaia use case.
- Excellent case for testing new solutions of data management and data retrieving initially established in the contest of IDL project, stressing the performance of the technological solution.

Scientific Rationale

- The specific ***technological goal*** of IGUC project is to identify and implement additional database and data management solution to complement the traditional ones. The purpose of this activity is to support the integration and query of data coming from different sources, with a performance that enables novel application with real-time requirements, to achieve maximum effectiveness and efficiency in data provisioning and exploitation.

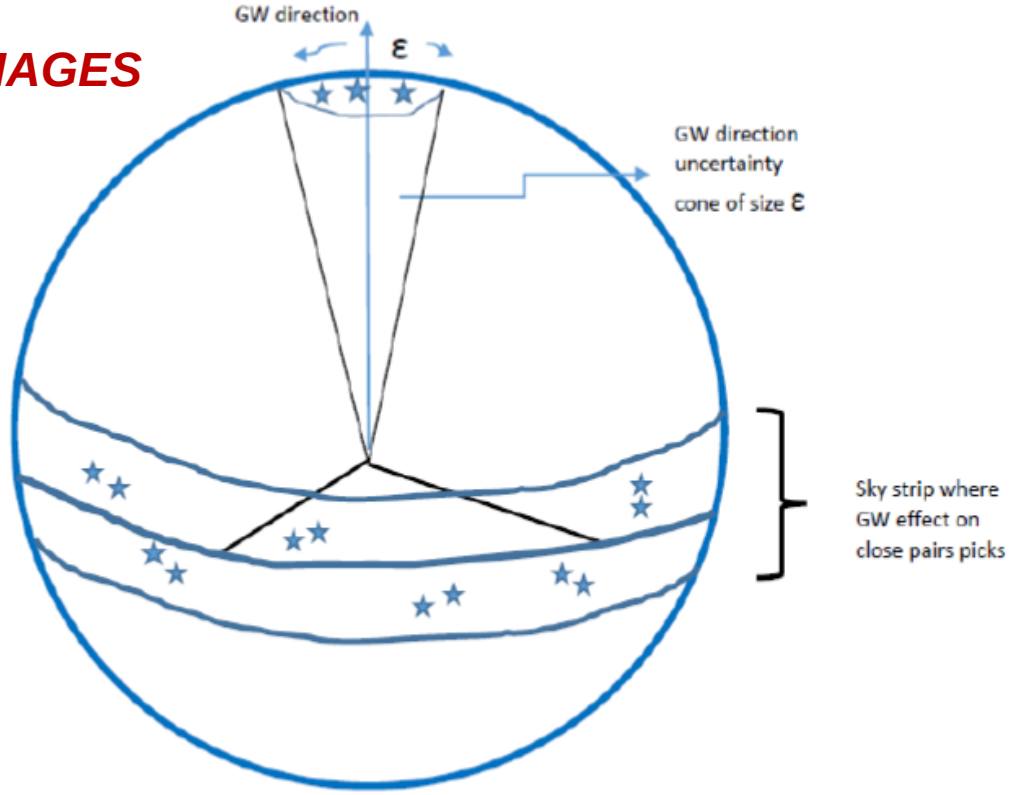
- The Gaia INAF team ***scientific goal*** is to do a further step in the implementation of the innovative platform dedicated to Gaia's legacy located at DPCT, showing the best solutions to retrieve billion of data for analyzing portions of the sky (tenth of square degrees) to identify significant variations of sources, to support science as discovery and characterization of cosmological gravitational waves or new earth-like planets.

Scientific Requirements

N.B.: GAIA RAW DATA ARE NOT IMAGES

For any sky direction of interest:

1) **Identify Sources (SourceId, S)** on the celestial sphere in cone toward direction of presumed incoming GW and in the corresponding meridian band (see figure);
2) **Pairing**, in given regions, of all of suitable sources (i.e., with angular distances > 0.2 ‘’, appropriate magnitudes and color ranges, photometric stability, etc…) (as shown in figure);
3) **Extract all of the elementary, or epoch, observations (AstroElementaries,Obs) for each pair:**
   a. ***Astrometrically calibrate angular separations***
   b. ***Build separation time series for analysis***



Analyze according to multi-direction fundamental equations (Crosta, MGL et al. 2024)

Toward the implementation of the digital Astrometric Gravitational Wave Antenna with Gaia elementary astrometry for GWB.

From MGL slide at USCVIII general meeting Oct-24

IGUC: Sizing the effort- dimensions and computational complexity

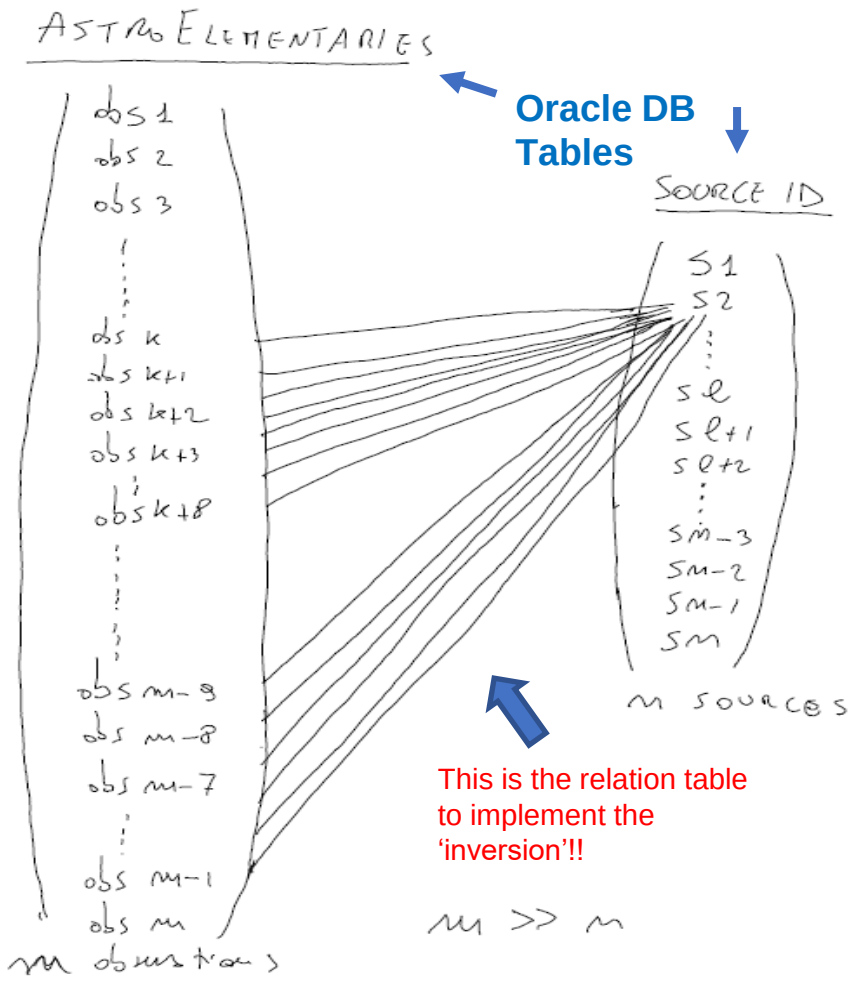
Gaia mission: observation-driven, data archiving schema (see Figure)

I. Requirements 1 and 2 (in previous slide) use SourceId table which is sky referenced, indexed, via HEALPix : estimated **10^6 suitable pairs** to G < 17 for each GW direction;

II. Req. 3 is the current challenge: Extract from AstroElementary Table all of the observations of each pair formed in steps 1 and 2. This means that the relation in Figure must be inverted. Given the dimensions of the AstroElementary table this operation requires proper partitioning of the CrossMatch table (shown in Figure as connections from obs-i to s-j)

Unrealistic processing time to analyse GW sky at a resolution of, say, 1 square degree . HTC breakthrough required, and\or resort to mitigation strategies (reduce no. of pairs)!!

From MGL slide at USCVIII general meeting Oct-24



IGUC - Interoperability Data Lake for Gaia Use Case

Partners involved: ONLY Leonardo/Cherry Data and INAF/Gaia team

- The Data are covered by an **NDA - NO PUBLIC DATA**

Database deployment in ICSC infrastructure **BUT IGUC experiment will carry on a dedicated INAF infrastructure due to the data policy.**

The agreement provides that once the project is concluded the data will be eliminated.

The work plan of [IDL] is organized in [1] work package (WP)

WP1 – Data Models and metadata definition, data archiving and database for Gaia Use Case

Planned Tasks:

- T1 [Gaia Use case requirements, data model and metadata definition]
- T2 [Benchmark requirements definition]
- T3 [Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT]
- T4 [Database deployment, validation, and testing]
- T5 [Implementation of the Gaia PoC on the INAF infrastructure]
    - Final benchmark results

Project deliverables and milestones:

M1-M2

Project Kick-off **2024 June 6th**

Gaia Use case requirements, data model and metadata definition (INAF)

M4 (ICSC MS9)

Benchmark requirements definition: definition of the metric to obtain an estimate of the performances invariant with respect to the execution platform (INAF-Leonardo)

M4 deliverable: Technical reports

M13

Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (INAF)

Implementation of the Gaia PoC on the INAF HW infrastructure (Leonardo-INAF)

Database deployment, validation, and testing

M13 deliverable: Report including hw and sw architecture description and verification tests (Leonardo-INAF)

M15 (ICSC MS10)

Final benchmark results (Leonardo-INAF)

M15 deliverable: Report on final results (Leonardo-INAF)

Project deliverables and milestones:

M1-M2

Project Kick-off **2024 June 6th**

Gaia Use case requirements, data model and metadata definition (INAF)

M4 (ICSC MS9)

Benchmark requirements definition: definition of the metric to obtain an estimate of the performances invariant with respect to the execution platform (INAF-Leonardo)

M4 deliverable: Technical report (Leonardo-INAF)

DONE

M13

Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (INAF)

Implementation of the Gaia PoC on the INAF HW infrastructure (Leonardo-INAF)

Database deployment, validation, and testing

M13 deliverable: Report including hw and sw architecture description and verification tests (Leonardo-INAF)

M15 (ICSC MS11)

Final benchmark results (Leonardo-INAF)

M15 deliverable: Report on final results (Leonardo-INAF)

Main Results - INAF

Reporting period: [July – December 2024]

- **Specification of current data model of the Gaia Use Case** (technical report TN-IGUC-INAF-EL-001-01 and appendix) https://owncloud.ia2.inaf.it/remote.php/webdav/PNRR-IG-IGUC/Gaia_dataset/TN-IGUC-INAF-EL-001-01.pdf
- **Development of the software for data conversion from Gaia gbin format to HDF5** that are used to allow access to the Gaia data in synergy with the activities carried on under Spoke3 WP4. **First version of the software is under version control on the INAF GitLab instance.**

- Original GAIA Data Format
  - Gbin data format is a zipped serialized java object
  - very limited interoperability
  - very lightweight (structure not included)
- HDF5
  - greatly improved interoperability
  - increased storage requirements (by a factor of 2)

Main Results - INAF

Reporting period: July 2024 – December 2024

- **Specification of typical queries of the Gaia Use Case** (technical report TN-IGUC-INAF-EL-001-01 and TN-IGUC-INAF-EL-002-01) https://owncloud.ia2.inaf.it/remote.php/webdav/PNRR-IG-IGUC/Gaia_dataset/TN-IGUC-INAF-EL-001-01.pdf, https://owncloud.ia2.inaf.it/remote.php/webdav/PNRR-IG-IGUC/Gaia_dataset/TN-IGUC-INAF-EL-002-01.pdf

  - **Multiple Cone search:** Given a direction identified by alpha (right ascension) and delta (declination) and a circle arc of amplitude ε, all the sources in the cone (solid angle) must be identified
  - **Cone search + meridian:** Given the results of the Cone Search, all the sources in a band of amplitude ε around the meridian orthogonal to the selected direction must be identified. Then, all the couples of sources withing a given angular separation must be identified, both for the sources in the error cone around the GW direction, and for the sources in the band around the big orthogonal circle. When the sources have been identified, all the associated transits, in a given time interval, must be retrieved through the CrossMatch table.

- Started on December 1° the technical support to INAF for the benchmark requirements definition and benchmark implementation under Oracle DBMS on the infrastructure located at the DPCT in Altec, i.e. metric definition to obtain an estimate of the invariant performance with respect to the execution HW platform used for the IGUC (Interoperability data lake for Gaia Use Case) project, in order to guarantee interoperability between the Oracle environment and other DBs and the choice of the Data Lake.

Main Results - CherryData

Reporting period: [July 2024 – December 2024]

1. Analysis of samples of datasets from the three tables of the data model of the Gaia Use Case:
   - AstroElementary
   - CompleteSource
   - CrossMatch
2. Development of the software for data extraction from the HDF5 files that are used for batch data interexchange.
3. Definition of a data model for the AyraDB database that reflects the data model of the Gaia Use Case.
4. Deployment of a test AyraDB cluster on the INAF infrastructure@ OATs
5. Test of batch ingestion procedures on a small sample dataset on the test AyraDB cluster.
6. Definition of an SQL query for AyraDB to implement the cone search query.
7. Test of the cone search query on the test AyraDB cluster.

The Data Model

The Data Model (DM) is structured on three main tables:

- **CompleteSource**
  - **Indexed by the SourceId, describes the sources.**
- **AstroElementary:**
  - **Indexed by the TransitId, describes the observations (for example flux, magnitude, color, and "quality" of the observation).**
- **CrossMatch:**
  - **It is the table that links AstroElementary and CompleteSource.**

The typical queries

We are working on two queries:

- **Multiple Cone Search:**
  - Given a direction identified by alpha (right ascension) and delta (declination) and a circle arc of amplitude ε, all the sources in the cone (solid angle) must be identified
  - When the sources identifiers (SourceID) have been identified, through the CrossMatch table, all the transits (AstroElementary) of each source must be identified and retrieved, in a selected time interval.
- **Cone Search and meridian:**
  - Given the results of the Cone Search, all the sources in a band of amplitude ε around the meridian orthogonal to the selected direction must be identified.
  - Then, all the couples of sources withing a given angular separation must be identified, both for the sources in the error cone around the GW direction, and for the sources in the band around the big orthogonal circle.
  - When the sources have been identified, all the associated transits, in a given time interval, must be retrieved through the CrossMatch table.

The DataBase choice

- **The system has been deployed (at the moment with a sample dataset of small size) with AyraDB, a database developed by Cherrydata, with a Key-Value core, and allowing SQL operations.**
- In order to enhance the performance of SQL queries, **two additional tables have been created**:
  - **AstroElementaryDigest:**
    - It contains an extract of the AstroElementary table, and contains exclusively the fields strictly necessary for the execution of the SQL queries (identification of sources and transits).
  - **CompleteSourceDigest:**
    - It contains an extract of the CompleteSource table, and contains exclusively the fields strictly necessary for the execution of the SQL queries (identification of sources and transits).

The Client

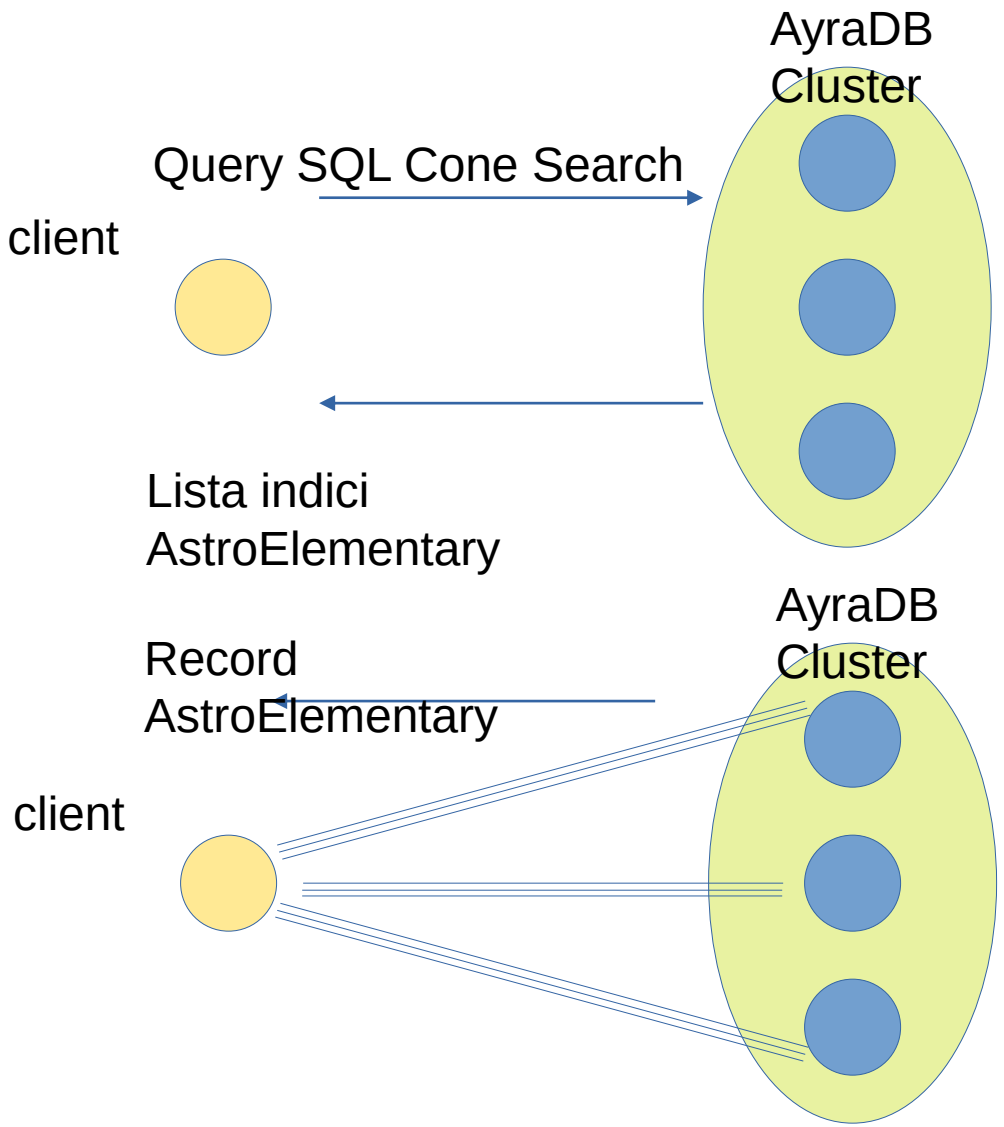
- We have developed an ad-hoc Python client that exposes simple APIs implementing the queries of the described Use Cases.
- The implementation of the Cone Search Use Case is complete.
- The use case Cone Search and Meridian is under development.

- For example, the Cone Search API has the following prototype:

```
cone_search(\
      servers: List,\
      credentials: Dict[str, str],\
      alpha: float,\
      delta: float,\
      epsilon: float,\
      str_date_start: str,\
      str_date_end: str,\
      target_astroelementary_column_labels: List)
```

Cone Search: an SQL + Key-Value query

- **To optimize the query, we have divided it into two phases:**
  - **PHASE 1 (SQL): operates only on the digest tables, to retrieve the indexes of the AstroElementary records (and also the associated SourceId) matching the requirements of the query.**
  - **PHASE 2 (Key-Value): the actual AstroElementary records are retrieved in key-value fashion, with a pool of parallel connections to all the servers of the AyraDB cluster. This guarantees a high throughput and a short execution time.**



Cone Search: an SQL + Key-Value query

```
SELECT astroelementarydigest.TransitId, filtered_triplets.SourceId,
astroelementarydigest.AyraDBMainTablePackedRecordIndex FROM ayradb.astroelementarydigest AS
astroelementarydigest JOIN (SELECT DISTINCT crossmatch.SourceId, crossmatch.TransitId FROM
ayradb.crossmatch AS crossmatch JOIN ayradb.completesourcedigest AS digest ON
crossmatch.SourceId = digest.SourceId WHERE crossmatch.TransitIdTimeDt64 >=
toDateTime64('{str_date_start}', 6) AND crossmatch.TransitIdTimeDt64 <=
toDateTime64('{str_date_end}', 6) AND ({cos_delta} * digest.CosDelta * cos(digest.Alpha - {alpha}) +
{sin_delta} * digest.SinDelta > {cos_epsilon})) AS filtered_triplets ON filtered_triplets.TransitId =
astroelementarydigest.TransitId WHERE astroelementarydigest.TransitIdTimeDt64 >=
toDateTime64('{str_date_start}', 6) AND astroelementarydigest.TransitIdTimeDt64 <=
toDateTime64('{str_date_end}', 6);
```

Results summary and next steps

- We have implemented and tested the Cone Search query on a small sample dataset (few millions of records).
- We have verified that the query is functionally correct.
- Given the small size of the dataset, it has not been yet possible to asses the performance of the query.

- We are in the process of deploying a realistic test on a dataset of about one TByte, on which we will also test the Cone Search and Meridian query.
- The dataset for the final test will have a size of about 10 TByte.

Final steps

Milestone 10: [Nov 2024 – Aug 2025] **Milestone 11: [Sept – Dec 2025]**

[M10 – November 2024 – March 2025]

WP1 – Data Models and metadata definition, data archiving and database for Gaia Use Case

Expected Targets

TAR1.2 Benchmark requirements definition (1 TB and 10 TB dataset)

KPI:

- Update benchmark requirements for larger dataset.
- Dataset identification report: **100%** completed.

TAR1.3 Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (1 TB dataset)

KPI _ Report including HW and SW architecture description: **70%.**

TAR1.4 Database deployment, validation, and testing (1 TB dataset)

KPI _ Report including HW and SW architecture description: **50%.**

Final steps

Milestone 10: [Nov 2024 – Aug 2025] **Milestone 11: [Sept – Dec 2025]**

[M10 – March – August 2025]

WP1 – Data Models and metadata definition, data archiving and database for Gaia Use Case

Expected Targets

TAR1.1 Database deployment, validation, and testing (1 TB dataset)

KPI _ Report including HW and SW architecture description: **100%.**

TAR1.2 Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (10 TB dataset)

KPI _ Report including HW and SW architecture description: **100%.**

[M11 - Sept – Dec 2025]

TAR1.1 Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (10 TB dataset) and on INAF infrastructure

KPI _ Final report including also the scientific validation on the two systems: **100%.**