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

Gaia use case

Sara Gelsumini, Deborah Busonero

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

SCIENTIFIC RATIONALE - the goal beyond the ICSC

- Generation of a deep and complete sky, on 4π sterad, as a reference tool and therefore interoperable for the integration of multiband data (from radio to high energies) and multimessenger data (e.g. sources of gravitational waves, neutrinos, ...) for efficient data mining aimed at fast multidimensional scientific data exploitation;
- Capacity for ad hoc recalibrations of astrometric and photometric data for the reclassification and redetermination of the fundamental properties (motions and magnitudes) of classes of objects of particular astrophysical interest;
- Interoperability and integration of metadata from non-astronomical databases, i.e. engineering and orbital data, data from service modules or payloads, or data coming, e.g., from Space Weather and/or surveillance of space debris (space debris surveillance);
- Operations of telescopes from Earth and space and support for studying new missions/projects.

OBJECTIVES

Overall goal: an infrastructure for archiving, management, processing, visualization, reprocessing, and analysis of Gaia data from raw data to processed data, not only for astrophysical exploitation but also for space science technological exploitation to enable large-scale reprocessing (see Busonero talk's 1° Tech Meeting 10/10/23)

In particular: study and implement 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

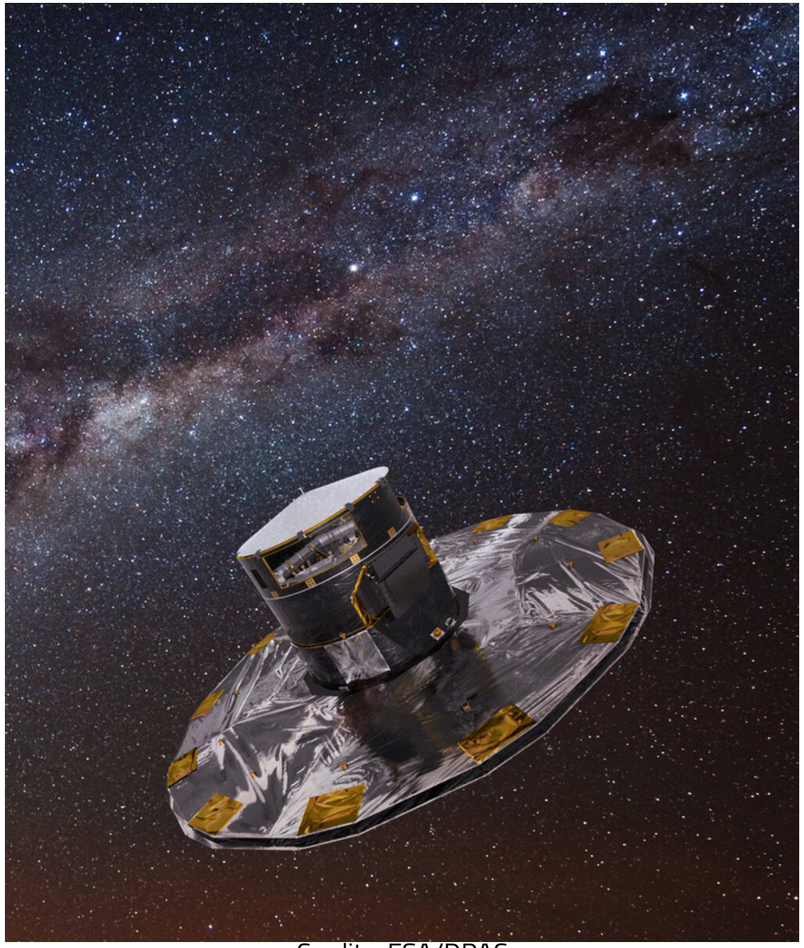
To create a database and filesystem platform capable of extracting all sources within different specific areas of the sky simultaneously and associating with each source the information regarding its transits and its calibration data

OBJECTIVES

We need fast queries and analysis of data from different perspectives:

- Run queries at billions of rows (sources) per second
- Switching between a source-oriented search by row (space) to a columnar search by transit (time) leveraging both indexing methods without the need to duplicate the DB volume;
- We also need to pre-aggregate and pre-calculate the information in the database before delivering it to the users.

The GAIA operations DM is not suitable for technical/scientific exploitation



Credits: ESA/DPAC

Accomplished Work - DATA OVERVIEW

- **CompleteSource:**

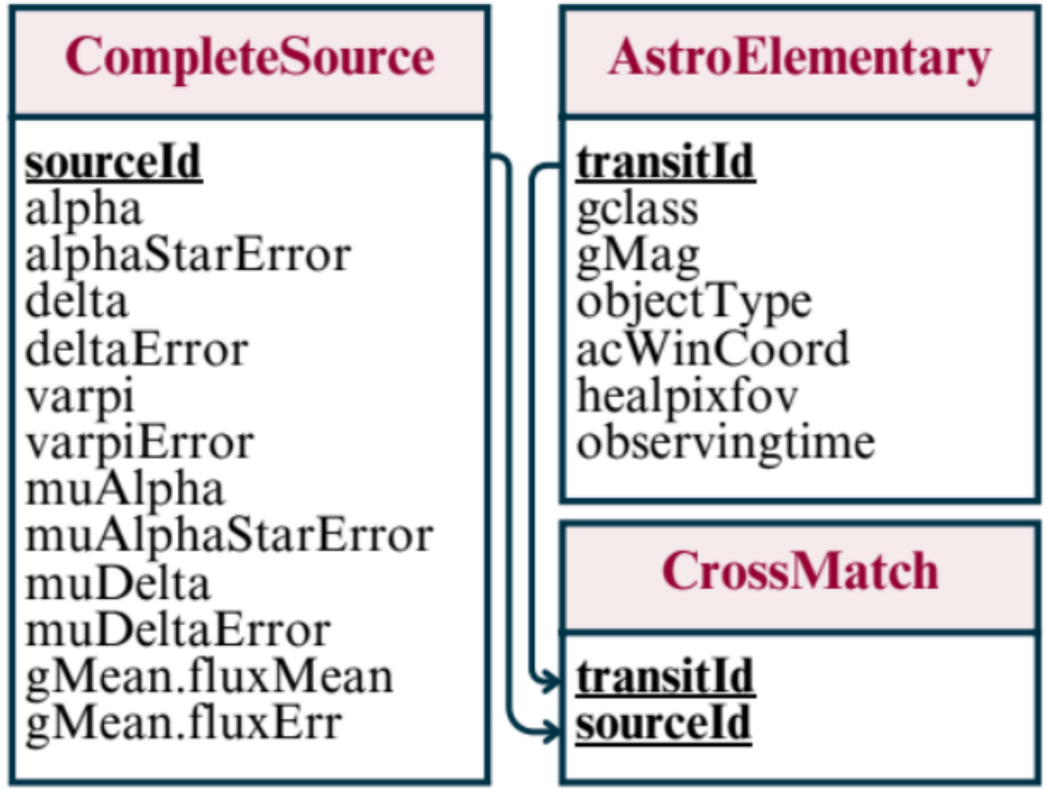
source information (180 attributes), ~ 4.8 TB with 2.793*10^9 elements.

- **AstroElementary:**

transit information (33 attributes), ~ 41 TB with 99.9*10^9 elements.

- **CrossMatch:**

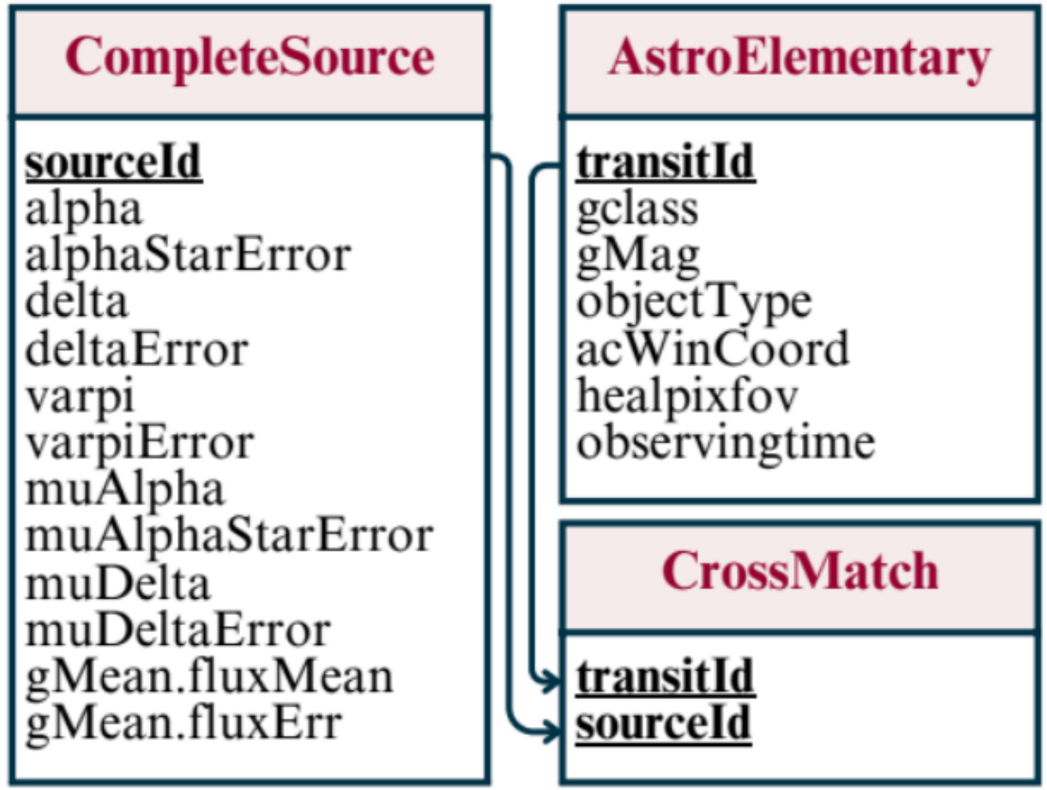
association of sources and transits (8 attributes), ~ 1.4 TB with 88.997*10^9 elements.



CHALLENGES

- DM and metadata definition to be queried in an efficient way
- Blob attributes as links to other tables
- The data are covered by an NDA - NO PUBLIC DATA

CompleteSource	~ 4.8 TB	2.793*10^9 elements
AstroElementary	~ 41 TB	99.9*10^9 elements
CrossMatch	~ 1.4 TB	88.997*10^9 elements



GBIN

- Requires specialized tools for interpretation
- Converted gbin to another format for easier usage
- GBIN files can contain multiple CS/XM/AE entries

```
'sourceId': 3376960784291370112,
'alpha': 1.6257970447712626,
'alphaStarError': 415.1820205532908,
'delta': 0.389743536501208,
'deltaError': 304.3747106045399,
'linDecompNormalsParamSolved': 31,
'muAlphaStar': None,
'muAlphaStarError': None,
'muDelta': None,
'muDeltaError': None,
'radialVelocity': None,
'radialVelocityError': None,
'varpi': None,
'varpiError': None,
'linDecompNormals': [0.17978651002121898,
0.21658186521428793, 0.021359902368825224,
0.15672888162006487, -0.007329793929945749,
0.1774723087349154, -0.12139129917298573,
0.09194578488838766, -5.103146522638524e-05,
0.01918613887682385, -0.1353116037702648,
0.09556740757878916, -5.341107177382422e-05,
0.0028955756289015286, 0.019095000561812434,
4.645244283992821e-18, -3.3773274466869448e-18,
1.875151880209317e-21, -1.0141508864111856e-19,
-9.097583078950226e-20, 0.0010000000474974513],
'refEpoch': '<javaobj:gaia.cu1.tools.time.GaiaTime>',
'colConstLevel': None,
'f2': 1.2288812398910522,
'noiseFlag': 8,
'solutionId': 1636042515805110273,
'bpMean': None,
'fieldOriginators': '<javaobj:java.util.EnumMap>',
'gMean':
'<javaobj:gaia.cu1.mdb.cu5.photpipe.phot.dmimpl.MeanPhotImpl>'
'rpMean': None,
'Gof': 0.0,
'assumedModelOrigin': 0,
'assumedPhysicalMultiple': False,
'assumedVariableCombSpec': False,
'astrometricDuplicateSourceId': 0,
'astrometricPseudoColor': None,
'astrometricPseudoColorError': None,
'astrometryFromEarlierCycle': False,
'bpIntegratedSpectrum': None,
'converged': True,
'deltaQ': None,
'emissionLinesCombined': False,
'epoch': None,
'excessNoise': 9.363175726773374,
'excessNoiseSig': 16.60973007898822,
'expectedSigToNoise': None,
'gRvs': None,
'gRvsConstancyProbability': None,
'gRvsError': None,
'hasRadVelSpeBarSys': False,
'inPencilBeam': False,
'inverseConditionNumber': 2.485626464476809e-05,
'ipdFracHighGof': 11,
'ipdFracMultiPeak': 0,
'ipdFracOddWin': 0,
'ipdGofHarmonicAmplitude': 194292.4375,
'ipdGofHarmonicPhase': 39.968650817871094,
'isGrvsValid': False,
'isPhotometricOutlier': False,
'isRadVelVariable': False,
'isSB2': False,
'isWeakClassification': False,
'matchedObservations': 2,
'matchedObservationsUsedByAgis': 2,
'meanFluxExcess': None,
'meanOnBoardGMag': 20.7109375,
'meanVarpiFactorAc': 0.7114633321762085,
'meanVarpiFactorAl': -0.5583772659301758,
```

Disclaimer: example with fake data

GBIN

- Requires specialized tools for interpretation
- Converted gbin to another format for easier usage
- GBIN files can contain multiple CS/XM/AE entries

```
'sourceId': 3376960784291370112,
'alpha': 1.6257970447712626,
'alphaStarError': 415.1820205532908,
'delta': 0.389743536501208,
'deltaError': 304.3747106045399,
'linDecompNormalsParamSolved': 31,
'muAlphaStar': None,
'muAlphaStarError': None,
'muDelta': None,
'muDeltaError': None,
'radialVelocity': None,
'radialVelocityError': None,
'varpi': None,
'varpiError': None,
```

Disclaimer: example with fake data

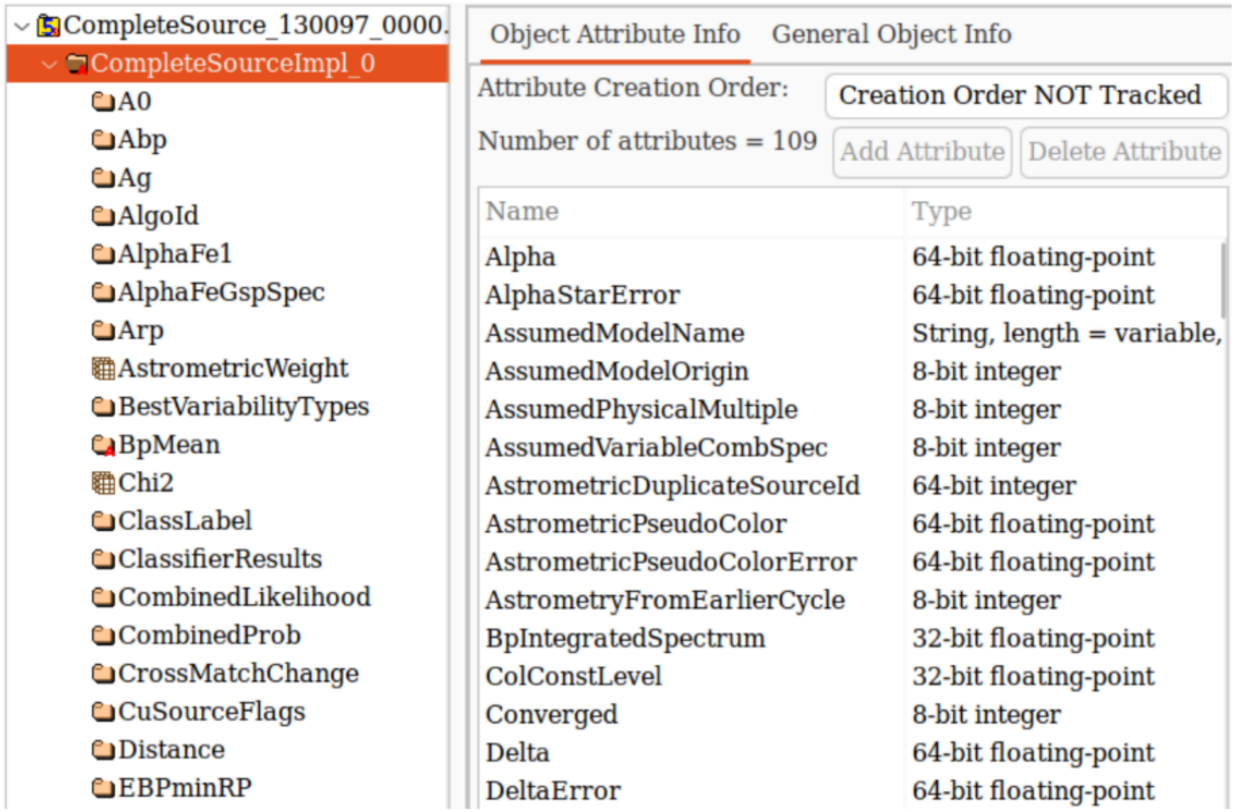
FITS

- Initially considered FITS format but faced challenges
- Created FITS files for each gbin, defining expected structures.
- Challenges with FITS rigid structure for dynamic needs.

HDF5

- Considered HDF5 for a more flexible structure
- Intuitive search
- Better blob integration

Java Objects ⇒ HDF5 Groups
Java arrays ⇒ HDF5 Datasets
Java primitives ⇒ HDF5 Attributes



QUERIES

Typical query:

- Given a specific area of the sky
- Identify all the sources within
- Pair the source information with the transit data

Another example of a query:

- Given a specific time
- Identify all the transit data
- Pair the source information with the transit data

Each of the billion astronomical objects is observed on average 200 times over the 10 years of the mission's duration

Thank you for your attention!