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 III Technical Workshop, Perugia Mag 26-29, 2025

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.

LEONARDO
CHERRY DATA
INAF ISTITUTO NAZIONALE DI ASTROFISICA

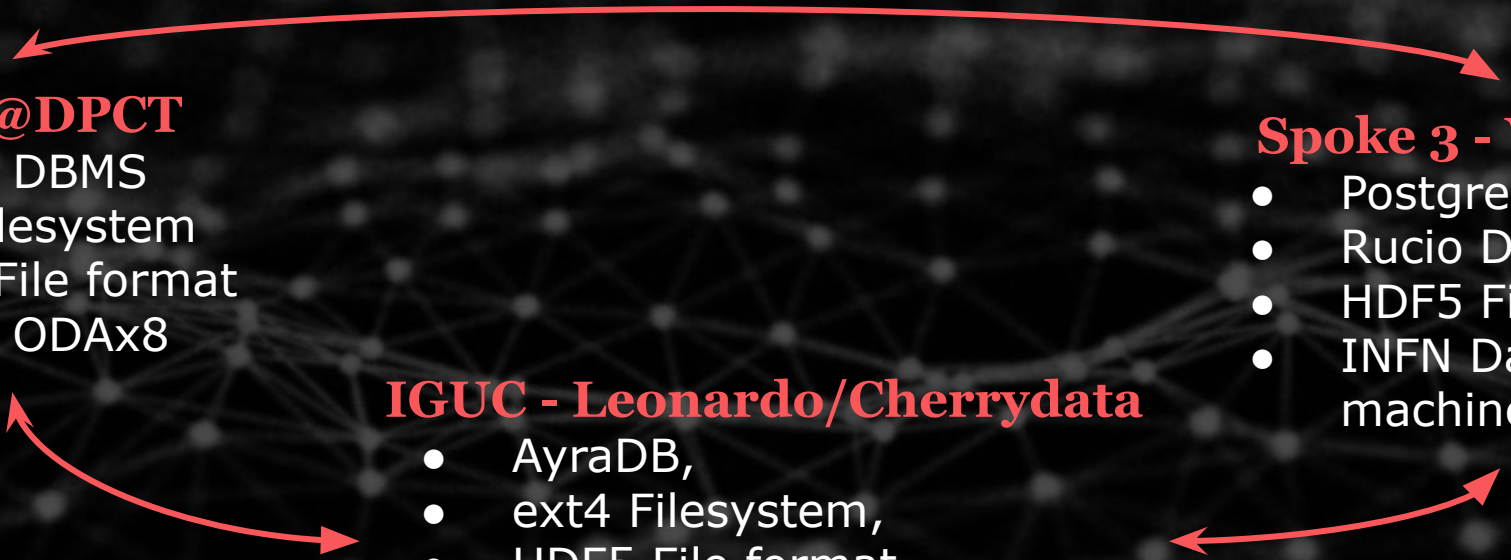
Scientific Rationale

- 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.

Sinergy between Spoke3 and IGUC Objectives

Technological testing of various DBMS and Data Management Systems, starting from the GAIA use case, with the objective of estimating and comparing the performance of the different systems, in a way that is as hardware and scale invariant as possible



OPS4@DPCT

- Oracle DBMS
- ZFS Filesystem
- HDF5 File format
- Oracle ODAx8

Spoke 3 - WP4 IDL

- Postgres DBMS,
- Rucio DMS,
- HDF5 File format,
- INFN Data Lake machines

IGUC - Leonardo/Cherrydata

- AyraDB,
- ext4 Filesystem,
- HDF5 File format,
- INAF infrastructure

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 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)

DONE

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

Project deliverables and milestones:

M13

Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (INAF) **90%**

Implementation of the Gaia PoC on the INAF HW infrastructure (Leonardo-INAF) **80%**

Database deployment, validation, and testing **80%**

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

M15 (ICSC MS10)

Final benchmark results on medium-size dataset (Leonardo-INAF)

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

M19 (**ICSC MS11)**

Final benchmark and scientific results on large-size dataset

Typical queries of the Gaia Use Case

- **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 within 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.

GAIA Test Case: Dataset 1 - Cone search + Meridian

Given a direction defined by (α,δ) and a radius ε, we have that a generic source of coordinates (α′,δ′) is inside the cone search if:

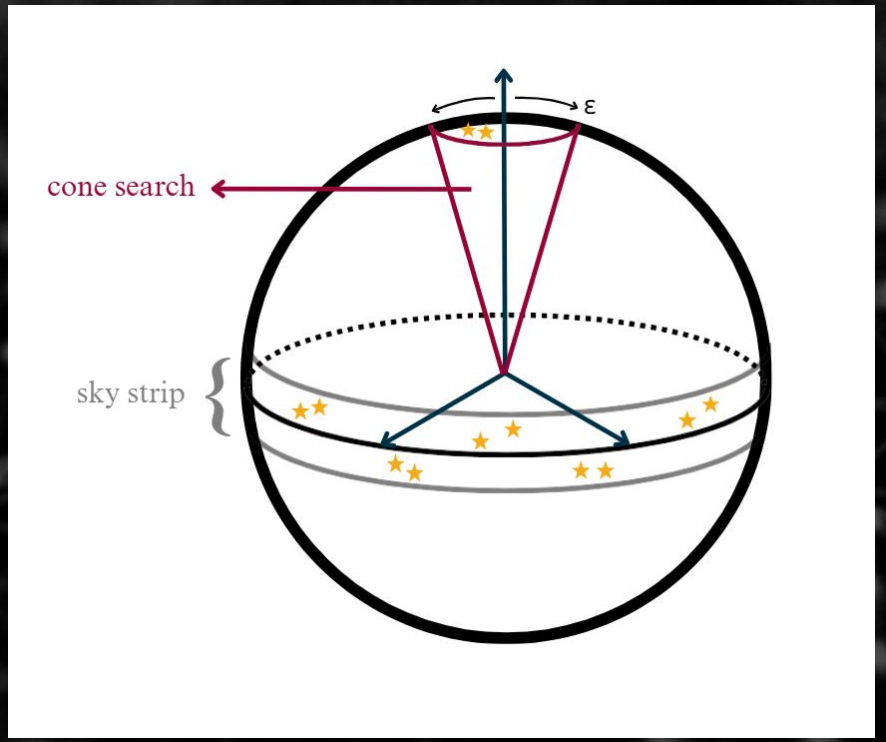
$$cos(\theta) \geq cos(\epsilon)$$

where

$$cos(\theta) = [cos(\delta)\, cos(\delta')\, cos(\alpha' - \alpha) + sin(\delta)\, sin(\delta')]$$

While it falls within the plane of semi-width ε and perpendicular to the direction (α,δ) if:

$$-\, sin(\epsilon) \leq cos(\theta) \leq sin(\epsilon)$$



GAIA Test Case: Dataset 1 - Cone search + Meridian

Select all sources and related transits for the specified regions of space and the specified timeframe

Search details:

- Cone search direction:
  - α = 0 [rad],
  - δ = PI/4 [rad] = 45 (deg)
- Cone radius & semi-width of meridian band:
  - ε = 0,002182 [rad] = 1/8 (deg)
- TransitID range:
  - Start = 6415193088000000
    Revolutions 4640.62,
    UTC 2016-12-31T23:56:32.680453840
  - End = 6586610688013107 1
    Revolutions 4764.62,
    UTC 2017-01-31T23:56:31.676574736

Search results:

- Identified ~ $4.5*10^6$ Sources

In order to perform a first test on real data we chose to limit the dataset size to ~ 1TB. To achieve this, we had to limit the timeframe for the transits to 1 month of data (over 10 years of mission)

We now have a dataset of around 1.3TB of Gbins

- ~100 GB of CompleteSource
- ~100 GB of Match
- ~1.1 TB of AstroElementary

GAIA Test Case: Dataset 2 - 20k Cone searches

Select all sources and related transits for the specified regions of space over the entire mission

Search details:

- Cone search direction:
  - α_k, δ_k = 1 of 20.000 directions equally spaced along an homogeneous spiral from celestial north pole
- Cone radius
  - $\varepsilon = 4{,}71*10^{-4}$ [rad] ~ 96,7 arcsec
- TransitID range:
  - all available mission data

Expected search results:

- Identified ~ $2.0*10^6$ Sources

This dataset is still not available, since the volume of the returned data will be well beyond the current scope of the project.

The number of gbins selected by this dataset and their total volume is overestimated due to some of the issues inherent in the gbin data format and their organization

Main Results - INAF

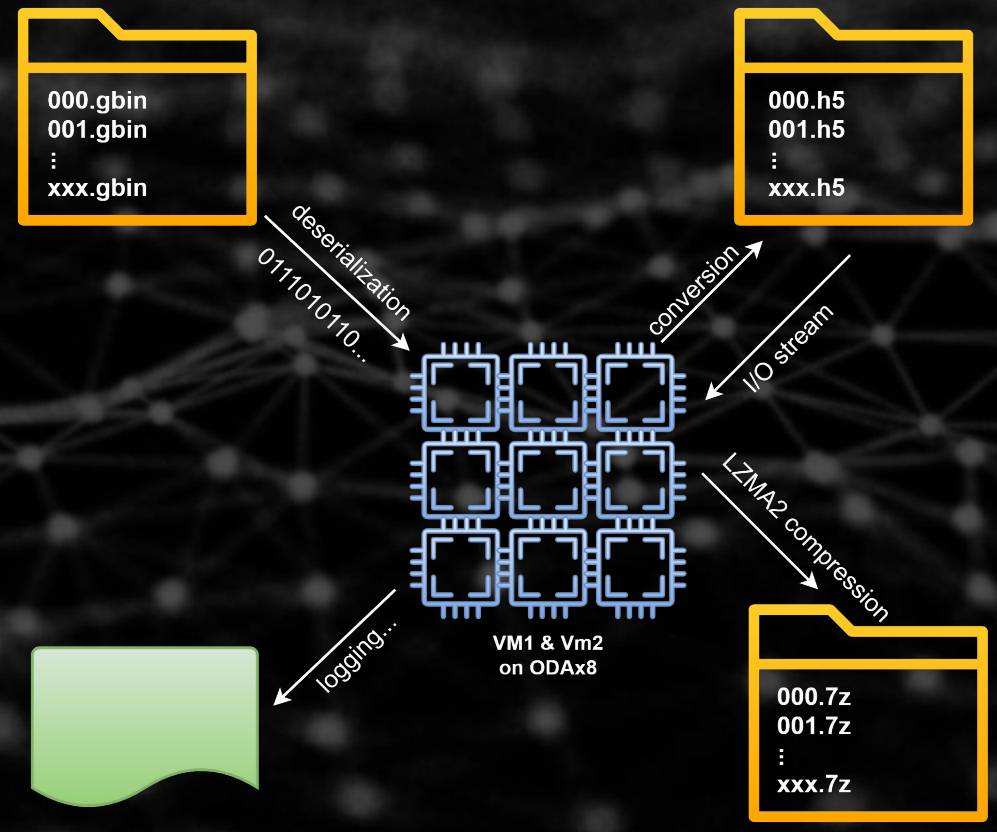
- **HDF5 converter software** for data conversion from Gaia gbin format to HDF5 in synergy with the activities carried on under Spoke3 WP4. This required a complex phase of testing and fine-tuning of the configuration parameters to achieve the required performance.
  Final version of the software is under version control on the INAF GitLab instance.
  A technical note describing the software is in preparation.

- 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)

HDF5 Converter

The software heavily leverages reflection and recursion to explore each input objects, regardless of its type, structure or complexity, and creates a corresponding HDF5 file mapping: Java objects, arrays and primitives are mapped respectively to HDF5 Groups, Datasets, and Attributes.

- Gbin files are read from a folder and all its subfolders
- Using MDBDM (Gaia Datamodel) and HDFQL (library) an .h5 version of each gbin is created and stored in a temp folder
- Each h5 is then read through a bufferedStream, compressed using LZMA2 compression algorithm and stored into an output folder
- The hierarchical structure of the original gbin is conserved into the final HDF5



HDF5 Converter

This apparently trivial and embarrassingly parallel operation resulted being a **challenge in terms of memory, I/O, and computation time.** This lead to the selection of the ODAx8 as the execution environment

2 Available VMs on ODAx8:

- 48 CPUs each
- 180 GB RAM
- 12TB of NFS for input, temp and output

Each gbin is around 1GB of compressed data, inflating to ~10x when deserialized in RAM → impossible to read 90+ at once

This required the creation of a custom gbin reader, allowing the selection of the number of objects to deserialize (batch read)

HDF5 (x36.3 gbin size) → 7z (x2.26 gbin size)

Each HDFQL operation is analogue to the execution of an SQL statement. This required a careful consideration of the execution statements.

The refactoring of this part of the code, reduced the number of I/O operations of a factor of 10^5

Tests performed on the FS of the ODAx8 showed that a block size of 4MB was optimal for sequential read operations

A custom version of the deserializer (from GT) is in development to leverage this information

Estimated time to convert 1.3TB ~ 7gg

HDF5 Converter

We decided to create entries also for null Objects/Groups or empty Arrays/Datasets

- preserve the structure of the original DM in the exported HDFs,
- leads to a measurable increase in the final volume of the data estimated ~20%

Due to the policies on the distribution of GAIA intermediate data outside the DPAC consortium, we decided to randomize all data non strictly required for the execution of the test

- preserve the volume of the dataset to have a convincing test case
- this leads to a measurable decrease in the compression rate (data have more entropy) ~17%

Sample Configuration

```
####################
## GENERAL SETTINGS ##
####################

# input folder: all gbins found inside the specified folder
# and subfolders will be exported into HDF5 format.
# 1 HDF5 file will be created for each available gbin,
# keeping the same filename
inputFolder = E:\\CherryData\\GBins\\test_random

# folder that will contain the exported HDF5 files
# WARNING: if this folder already contains HDF files with conflicting fil
# the old HDF file will be overwritten
# WARNING: the folder must already exist
# WARNING: the .h5 files will be deleted after the creation of the compre
tempFolder = E:\\CherryData\\Temp

# folder that will contain the exported and compressed HDF5 files
# WARNING: if this folder already contains HDF files with conflicting fil
# the old HDF file will be overwritten
# WARNING: the folder must already exist
outputFolder = E:\\CherryData\\HDF5

# folder that will contain the log file keeping track of the conversion p
# if the conversion progress is interrupted can be restarted from where i
# reading this logfile.
# WARNING: delete this file if you want to restart the conversion process
# WARNING: the folder must already exist
logFolder = E:\\CherryData

# This parameter is used to filter the "get" methods returned by JAVA Ref
# it should be set to a substring of the desired fully qualified class na
# to avoid the invocation of native java methods inherited from the Objec
# default value: gaia
packageFilter = gaia

# Enables/disables the creation of empty groups in place of null objects
# WARNING: enabling this option will increase significantly the storage
# space required by the generated HDF5, but will conserve the original st
# defaults value = false
exportNullObjects = true

# Enables/disables the temp file deletion. Mainly for debug purpose
# default value = true
removeTempFiles = true

# Sets the GbinReader version
# Available values: 3, 4, 5
# default value = 4
gBinReaderVersion = 4

# enables / disables SHUFFLE and ZLIB options for datasets compression
# WARNING: depending on the type of data this option might increase the s
# default value: false
enableCompression = false

############################
## READ/WRITE OPTIMIZATION ##
############################

# Maximum number of objects to export to HDF5
# 0 means ALL available objects inside input gbin
# default value = 0
maxObjects = 0

# Maximum number of objects to be deserialized at a given time
# this property is used to limit the amount of RAM required at runtime
# default value = 10000
chunkSize = 10000

# Maximum number of objects to be stored in ram, before writing them to disk
# WARNING: higher values reduce IO activity, but heavily impact RAM
# default value = 500
flushSize = 500

# Sets the maximum number of objects to be stored insied a single h5 file to
# limit its volume on disk
# WARNING: if set to 0, removes the limit
# WARNING: hdfMaxObjects must be greater then flushSize
# default value = 10000 ~ 650MB
hdfMaxObjects = 1000


#readBlockSize = 8192
#writeBlockSize = 4096


#######################
## DATA RANDOMIZATION ##
#######################

# enables / disables data randomization.
# Each field will be replaced with a random value of the same type
# default value = false
randomizeData = true

# List of fields that will not be randomized.
# Required to specify the exact field name
# WARNING: this property is enable only when randomizeData is set to TRUE
useOriginalValuesFor = transitid sourceid alpha alphaStarError delta deltaError
```

Main Results - INAF

- Benchmark requirements definition and benchmark implementation under Oracle DBMS on the infrastructure located at the DPCT in Altec
- metric definition to obtain an estimate of the invariant performance with respect to the execution HW platform used for the IGUC project between the Oracle environment and other DBs and the choice of the Data Lake.

ODAx8 configuration according to Project Requirements

- ODAx8 is an integrated system comprising 2 processing servers each one equipped with 2x 16-Core Intel® Xeon® Gold 5218 each, and 384 GB of memory. This machine is connected to a Storage shelf equipped with 6 x 7.68TB SSDs and 18 x 14TB HDDs for a total of 97 TB of usable space (in double mirroring).

A reconfiguration of the machine was executed to ensure security and enforce isolation protocols. ODAx8 configuration specifications were defined, encompassing IP addresses and resource allocations. The DBMS installation modalities, specifically bare-metal and DBSYSTEM, were evaluated.

- In order to optimize performance, the bare-metal installation was selected and executed at the end of March.
- Ingestion of the dataset in April before Easter.

A stop of one month has been experienced between April and May due to the huge infrastructure update of the Altec data center, mainly for which concerns the Gaia data processing center.

OPS4 @ DPCT - Oracle + ZFS

Oracle Spatial is a key feature required to efficiently identify and extract datasets like those of the GAIA use case :

- Cone Search: Distance from SDO_Geometry 2001 (point)
- Meridian Search: Distance from SDO_GEOMETRY 2002 (line) defined by 3 points:
  - 1st @ α as defined by search details
  - 2nd and 3rd @ α+180 degrees, separated by minimum TOLERANCE (this avoided computing the difference between to spherical caps)

Sample Query

```
select * from datadb_c04.completesource
where SDO_WITHIN_DISTANCE(                                  -- meridian
    COORDS,                                                 -- sdo_geometry column
    SDO_GEOMETRY(
     2002,                          -- line (Oracle code to identify a 2D line)
     20000202,                                              -- SRID (i.e., ICRS)
     null,
     SDO_ELEM_INFO_ARRAY(1,2,1),          -- Line string whose
                          vertices are connected by straight line segments
     SDO_ORDINATE_ARRAY(0,acos(-1)/4-acos(-1)/2,  acos(-1),
acos(-1)/2 - acos(-1)/4, 2*acos(-1) -0.000000000000002,
acos(-1)/4-acos(-1)/2)
                                                -- 3 vertices to define meridian
                        -- [alpha_0, delta_0-90] [alpha_0+180, 90-delta_0]
[alpha_0+360-2*TOLERANCE, delta_0-90]), 'distance =
0.00218') = 'TRUE';
```

GBIN file format

ALL GAIA SKY

- 2013 Gbins for CompleteSource
- 2170 Gbins for Match
- 71.000 Gbins for AstroElementary

Multiple AE Gbins (up to 7) insist on the same transitID range: this leads to an ambiguity on the identification of the correct gbin given a specific transitid

Impossible to select only the data required: no available off-the-shelf software to shrink the gbins and extract only the data required by the DR

GBIN features

- Compressed serialized java objects
- Requires a specific java sw with the correct DM to be able to access the data.
- Lightweight and suitable for operations
- Not suitable for scientific exploitation and dissemination

A tool based on Apache NiFi to extract the required fields from the gbins is currently under development @DPCT

Activities carried on by CherryData

Reporting period December 2024 - May 2025

- **Analysis of the amplification of disk space usage in the conversion of HDF5 files to CSV.**
- **Benchmarking of a set of machines, to select the most suitable for the deployment of the DB.**
- **Implementation of the "Cone Search + Meridian" query.**

Disk usage in the conversion from HDF5 to CSV

- The original AstroElementary, CompleteSource and CrossMatch files are in GBIN format.
- GBIN files are first converted into HDF5 format.
- HDF5 files must be converted to CSV, which is the ingestion format for the DB.
- Since the size of the CSV files can be used as an approximation of the disk usage of the DB, it is important to evaluate the amplification factor of disk usage in the conversion from HDF5 to CSV.

HDF5 to CSV conversion: treatment of Blob fields

- In the HDF5 files, there are objects that represent a collection of multiple fields (Blobs).
- In the HDF5→CSV conversion, Blobs are converted to explicit scalar fields.
- The column structure of the tables in the DB accounts for all the individual scalar fields, including those in the Blobs, and the Blobs are eliminated.

HDF5 to CSV conversion: measurements

The benchmarking of the HDF5→CSV conversion has been performed on a sample of files. The results are shown in the table.

Table	Number of HDF5 files	Size of HDF5 files (byte)	Size of CSV files (byte)	Amplification factor
AstroElementary	85	7,258,710,928	1,817,069,849	0.25
CompleteSource	6	3,482,403,160	179,090,338	0.05
CrossMatch	43	458,092,152	69,795,170	0.15

HDF5 to CSV conversion: results

- It is interesting to observe that the amplification factor is smaller than 1.

- This means that the CSV version of a file occupies less disk than the original HDF5 format.

- This is an important metric to be accounted for in the dimensioning of the DB.

Benchmarking of candidate machines for the DB

- The adopted database is AyraDB, which is disk-based, that is, the tables are stored on the disks (only the indices are in memory).

- Therefore, the performance of the disks impacts directly on the overall performance of the DB.

- We have carried out a series of disk benchmarks on a selection of candidate machines, to identify the best configuration for the deployment of the DB.

Disk benchmarking: the candidate machines

- The following machines (all of INAF) have been examined:

  A) Virtual machine with a 3-Tbyte disk, in non-VMWARE configuration.

  B) Virtual machine with a 3-Tbyte disk, in VMWARE configuration.

  C) Virtual machine with a 100-Tbyte disk, in VMWARE configuration.

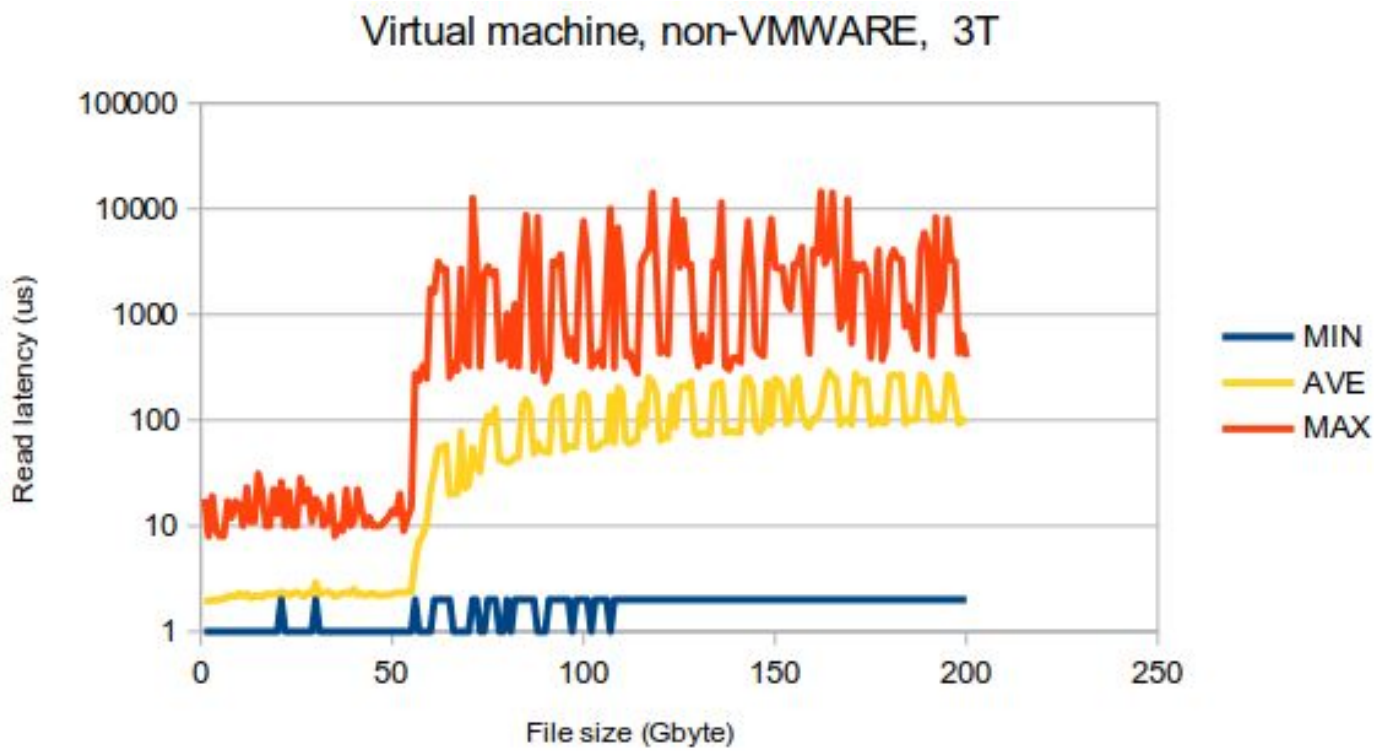
  D) Physical machine with a 4-Tbyte local disk.

Structure of the bechmarks

- The benchmarks have been structured to simulate the typical disk activity of the DB.
- In each benchmark run, a file that represents a tablet is written progressively, adding 1 Gbyte at each step.
- For each file size, 1000 read operations are performed, where each operation reads a block of 1000 bytes, in a random position in the file.
- For each file size, the following statistics are recorded:
  - minimum read time,
  - maximum read time,
  - average read time,
  - standard deviation of read time.

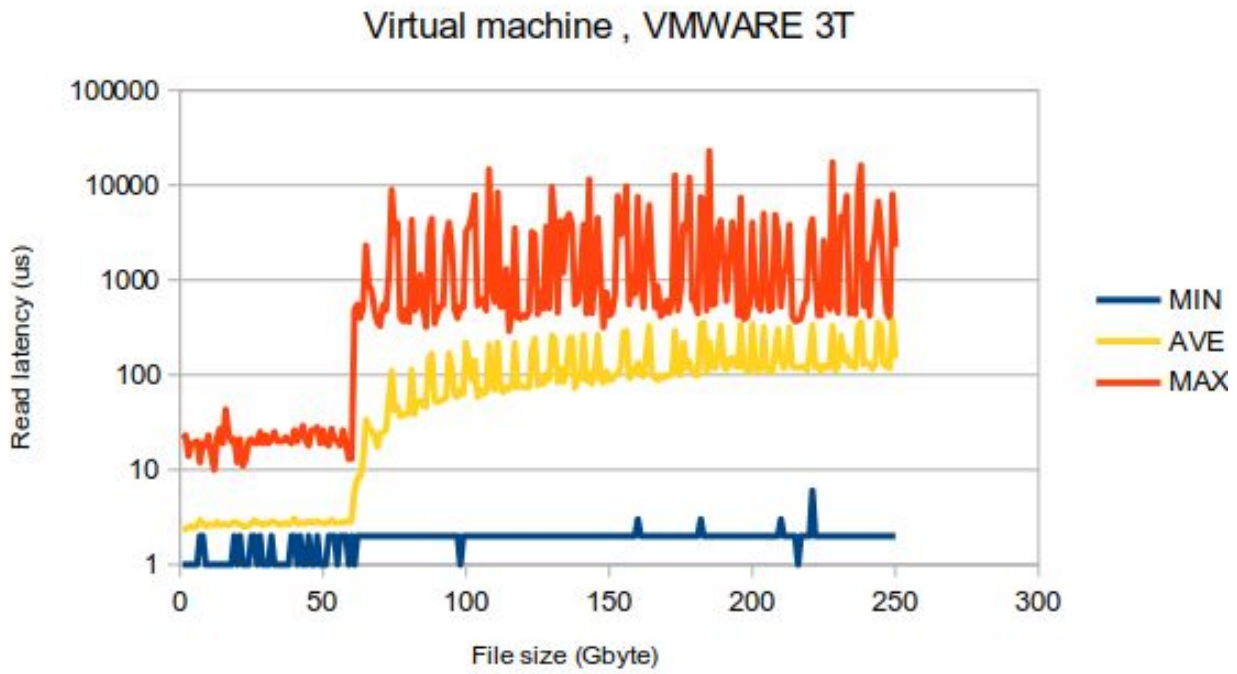
A) Virtual machine, non-VMWARE 3-Tbyte disk

-We register a sharp increase of latency when the size of the file exceeds 50 Gbyte.
-The average latency stabilizes at about 100 us.
-There are very high peaks of latency (even 10000 us).



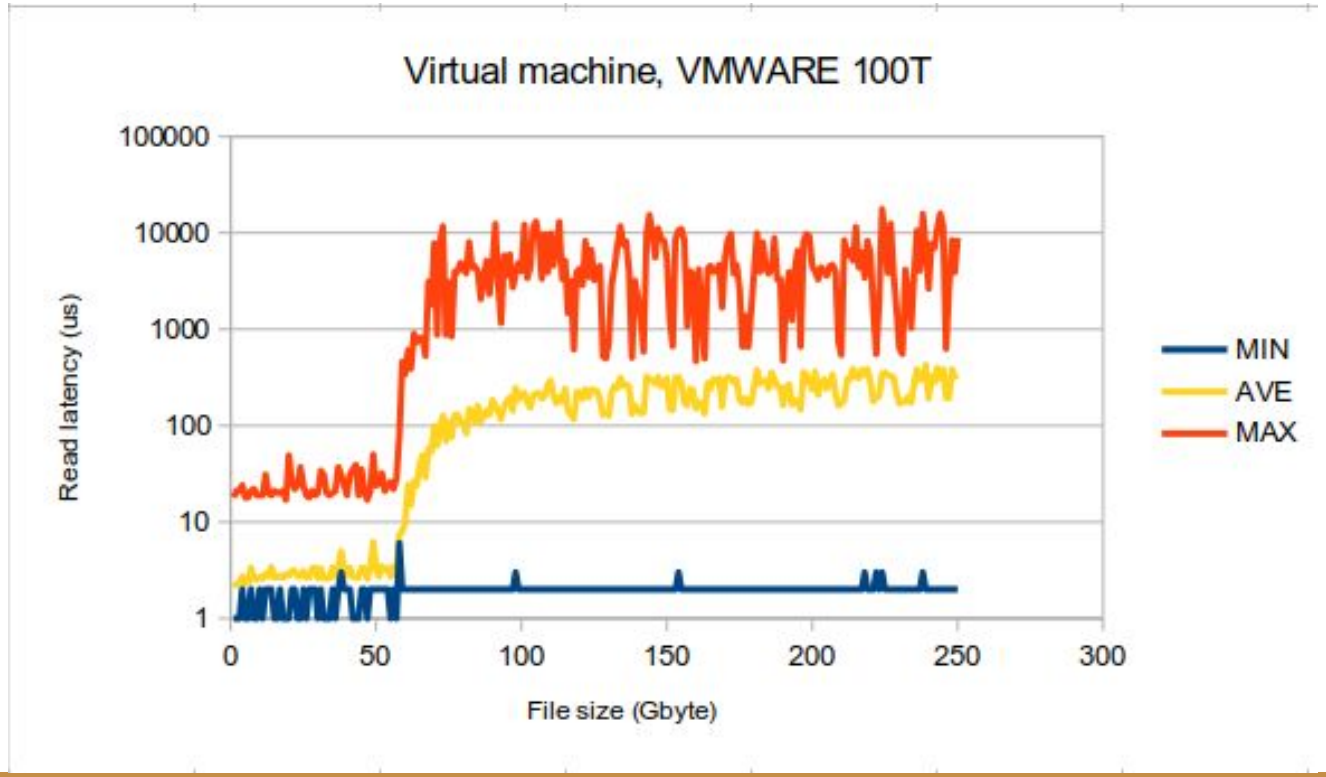
B) Virtual machine, VMWARE 3-Tbyte disk

-The results are similar to those of the previous configuration.
-We register a sharp increase of latency when the size of the file exceeds 50 Gbyte.
-The average latency stabilizes at about 100 us.
-There are very high peaks of latency (even 10000 us).



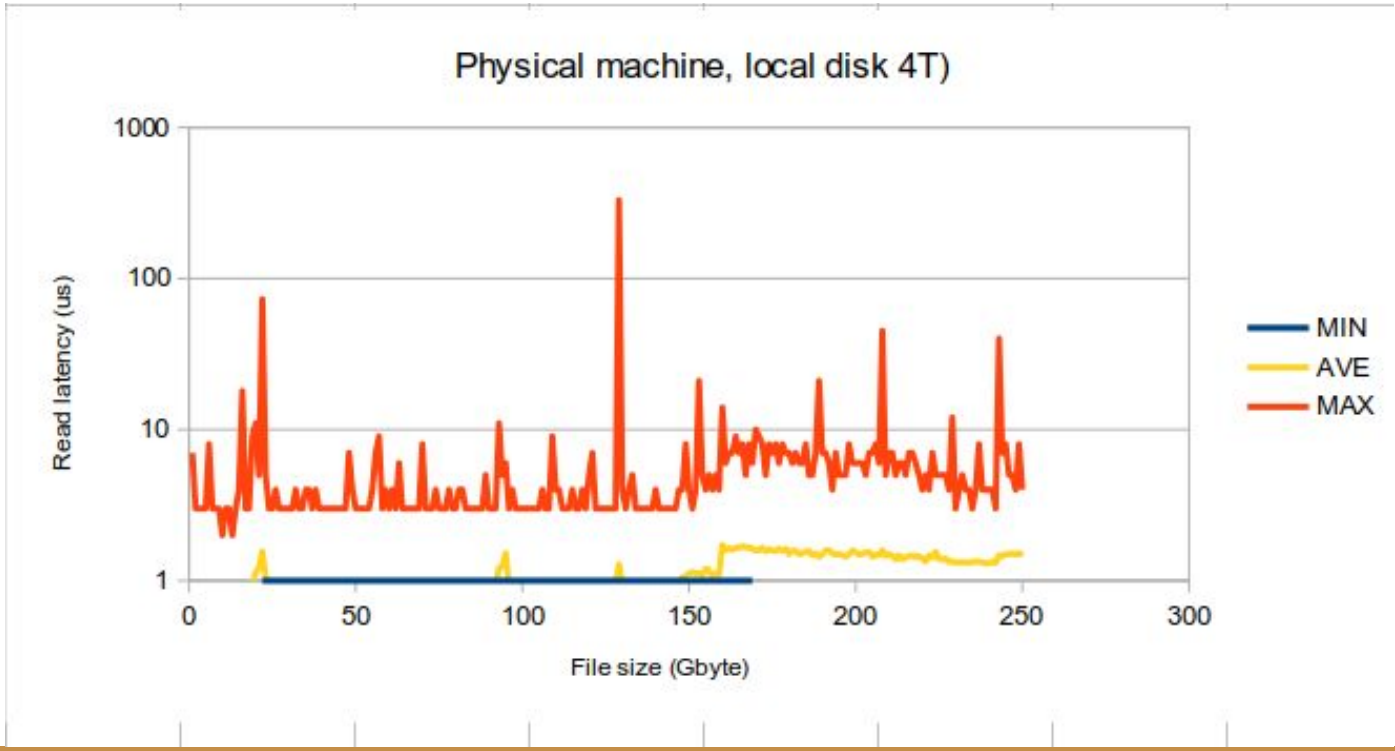
C) Virtual machine, VMWARE 100-Tbyte disk

-The average and maximum latency is significantly larger than that of the configurations with a 3 Tbyte disk.
-The same sensitivity to the disk size above 50 Gbyte is present.



D) Physical machine, local 3-Tbyte disk

-The performance of this configuration is significantly superior.
-The average latency is stable at few us.
-Most peaks are below 10 us, with a few peaks above 10 us.



The Cone Search plus Meridian query

-In this query, at first we identify the sources in a given angular distance from a given direction.

-Then, we search the sources that are in a given angular distance from the meridian orthogonal to the same direction.

-Finally, we identify among these sources the couple of sources with a mutual angular distance smaller than a given target.

SQL phase of the query

```
SELECT
   astroelementarydigest.TransitId,
   filtered_triplets.SourceId,
   astroelementarydigest.AyraDBMainTablePackedRecordIndex
FROM ayradb.astroelementarydigest AS astroelementarydigest
JOIN (
   WITH {cos_delta} * digest.CosDelta * cos(digest.Alpha - {alpha}) + {sin_delta} * digest.SinDelta AS computed_expr
   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 (
         (computed_expr > {cos_epsilon})
         OR (
            (computed_expr >= {-sin_epsilon})
            AND (computed_expr <= {sin_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);
```

Angular distance

Temporal filter

Direction filter 1

Direction filter 2

Then, the query proceeds as follows

- **Phase 2:**
  - **We retrieve in key-value fashion the complete data of the transits, from the table AstroElementary.**
- **Phase 3:**
  - **We retrieve, in key-value fashion, the [SourceId, Alpha, Delta] of the identified sources, from the table CompleteSource**
- **Phase 4:**
  - **We identify the pairs of sources within a given angular distance.**
  - **Phase 4 is executed with an efficient ad-hoc software written in C language.**

Phase 4

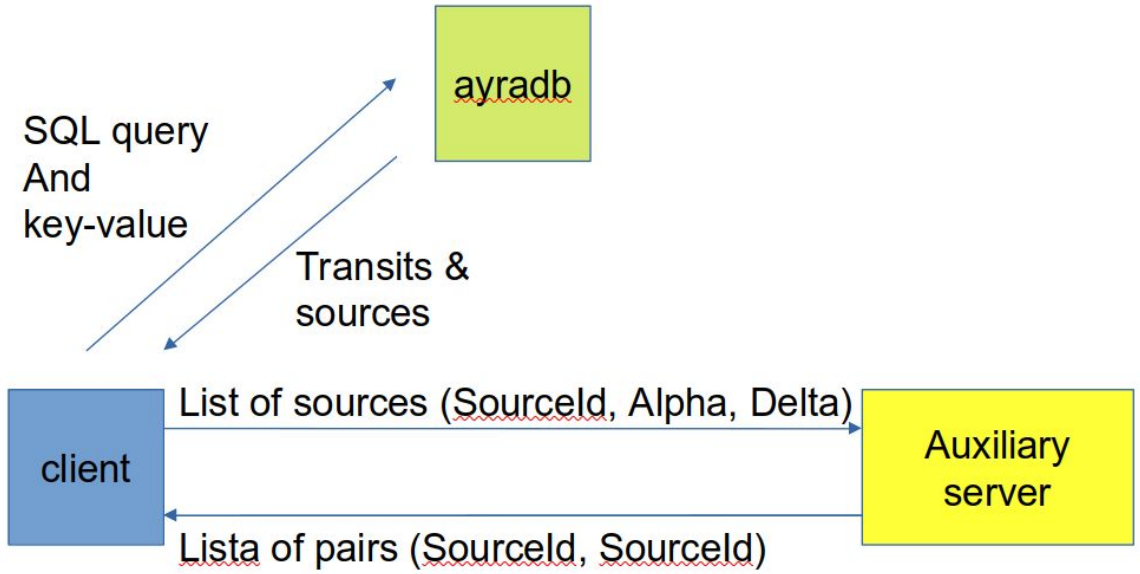
- The angular space is reprojected on a sphere, in a (x, y, z) cartesian space.
- In this way, sources can be indexed with a kdtree of order 3.
- In the indexed reprojected space, the search for the pairs has complexity Nlog(N), where N is the numbe of sources.

System architecture

-The angular space is reprojected on a sphere, in a (x, y, z) cartesian space.

-In this way, sources can be indexed with a kdtree of order 3.

-In the indexed reprojected space, the search for the pairs has complexity Nlog(N), where N is the numbe of sources.



Milestone 10: **[Nov 2024 – Aug 2025]**

Final steps

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

Expected Targets

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

KPI _ Report including HW and SW architecture description: **90%** → **100%.**

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

KPI _ Report including HW and SW architecture description: **80%** → **100%**

Database deployment, validation, and testing

KPI _ Report including HW and SW architecture description: **80%** → **100%.**

Milestone 11: **[Sept – Dec 2025]**

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.


Thank you for your attention