Access to massive catalogues in the Gaia archive: a new paradigm

Jesús Salgado

Juan González, Raúl Gutiérrez, Juan Carlos Segovia, Christophe Arviset, Sara Nieto, Bruno Merín

Gaia Archive Development Team
ESAC Science Data Center (ESDC)

Issue/Revision: 1.0
Reference: Gaia Archive
Status: Issued

gaia
European Space Agency

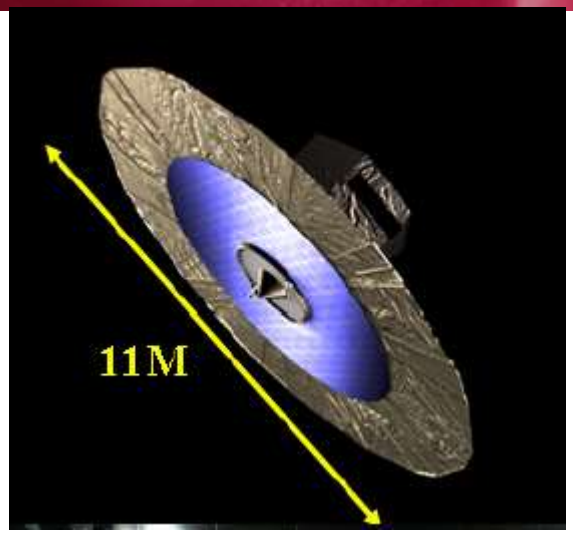
Gaia Satellite Overview

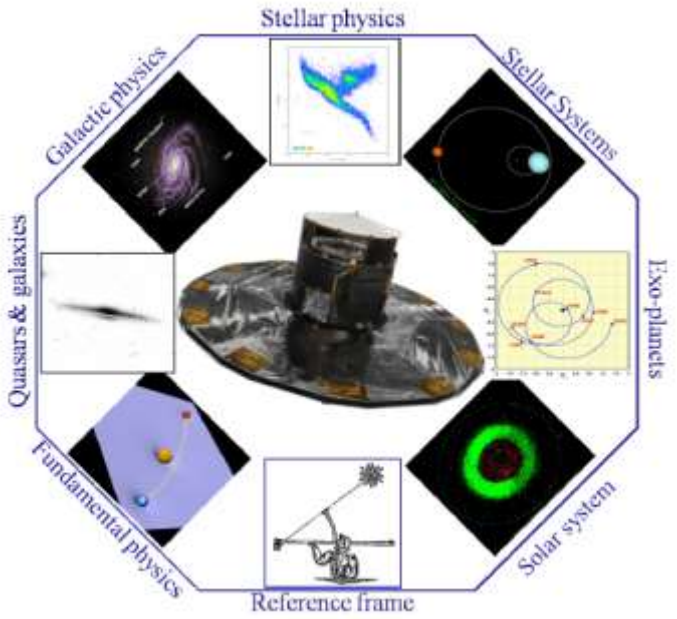
1. ESA Corner Stone 6
2. ESA provided the hardware and launch
    a. Mass: 2120 kg (payload 743 kg)
    b. Power: 1631 W (payload 815 W)

3. Launched 19th December 2013.
4. Stereoscopic Census of Galaxy over 5 years
5. Possible extension of 1 year (fuel should be ok)

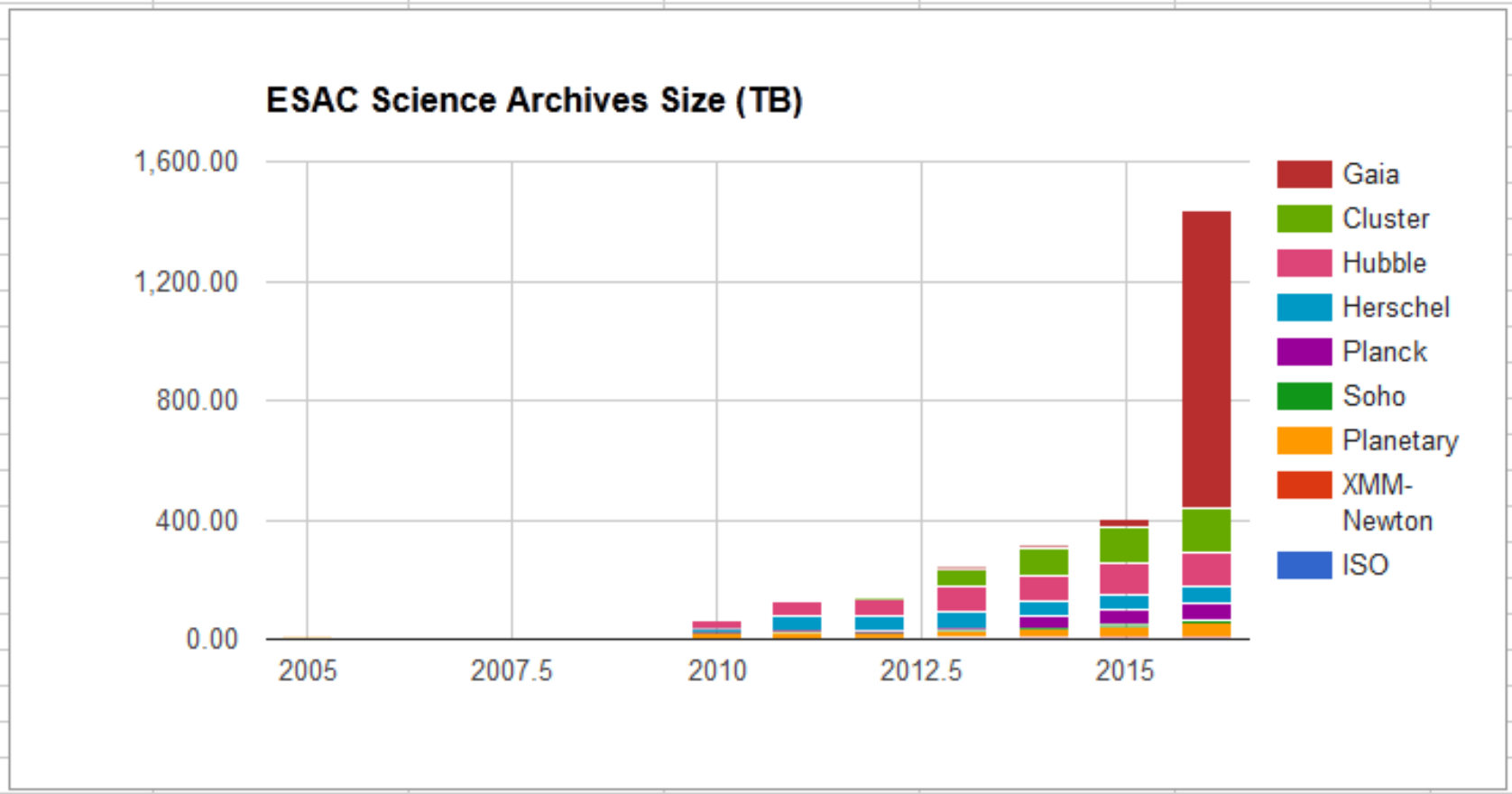
6. Astrometry G < 20 (10^9 sources)
7. 25 μarcsec at V=15
8. Radial Velocities G < 16
9. Spectro Photometry G < 20 (millimag)

10. Final catalogue ~2022





ESAC Archives Volume evolution



All data stored on hard disks and distributed through Internet

Euclid will add up to ~150 PBs by 2023

gaia
European Space Agency

Data collections stats

Statistics up to June 5

Nominal Mission Data

Type of Data	Amount	
Science telemetry	17 TB	
Astrometry transits	22.5×10^9	225×10^9 images
Photometry transits	22.5×10^9	45×10^9 images
Spectroscopy transits	1.5×10^9	4.5×10^9 spectra
Main Database	44TB	

- $\approx 30GB/day- > \approx 100TB$ total
- with products can be 1PB total data by mission end

Timo Prusti (2015)

Gaia Archive current content

Simulations

• GUMS		
• Milky Way:	2×10^9	rows
• Large Magellanic Cloud:	7.5×10^6	rows
• Small Magellanic Cloud:	1.2×10^6	rows
• Galaxies:	38×10^6	rows
• Quasars:	10^6	rows
• GOG	1.8×10^9	rows

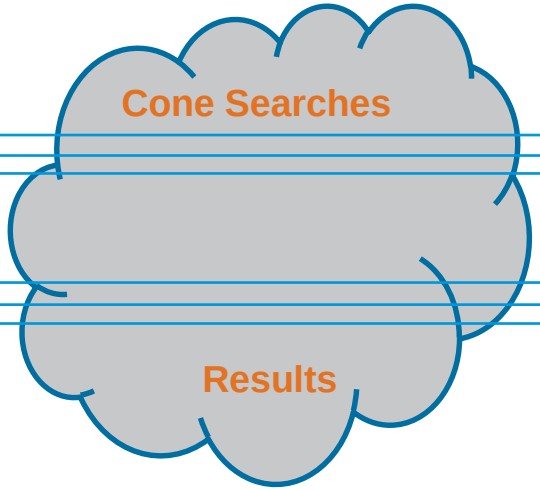
External Catalogues

• IGSL (Initial Gaia Source List)	1.2×10^9	rows
• 2MASS	9.4×10^8	rows
• Tycho2	2.5×10^6	rows
• UCAC4	1.1×10^8	rows

Gaia (not public data)

• TGAS	2.2×10^6	rows
• MDB	$\sim 3 \times 10^9$	rows
• Foreseen, first Gaia catalogue	$> 2 \times 10^9$	rows

Crossmatch for my objects



User Disk

Crossmatch for my objects (II): TAP approach

White Dwarfs

User Disk

TAP
Upload & Query

Results

gaia archive

Repeat operation n-times

gaia
European Space Agency

Crossmatch for my objects (III): TAP+ approach & VOSpace+

White Dwarfs

User Disk

TAP Upload

n TAP Queries

Results

gaia archive

Results

gaia

European Space Agency

Crossmatch for my objects (III): TAP+ approach & VOSpace+ & SaaS

White Dwarfs

User Disk

TAP Upload

n TAP Queries

Results

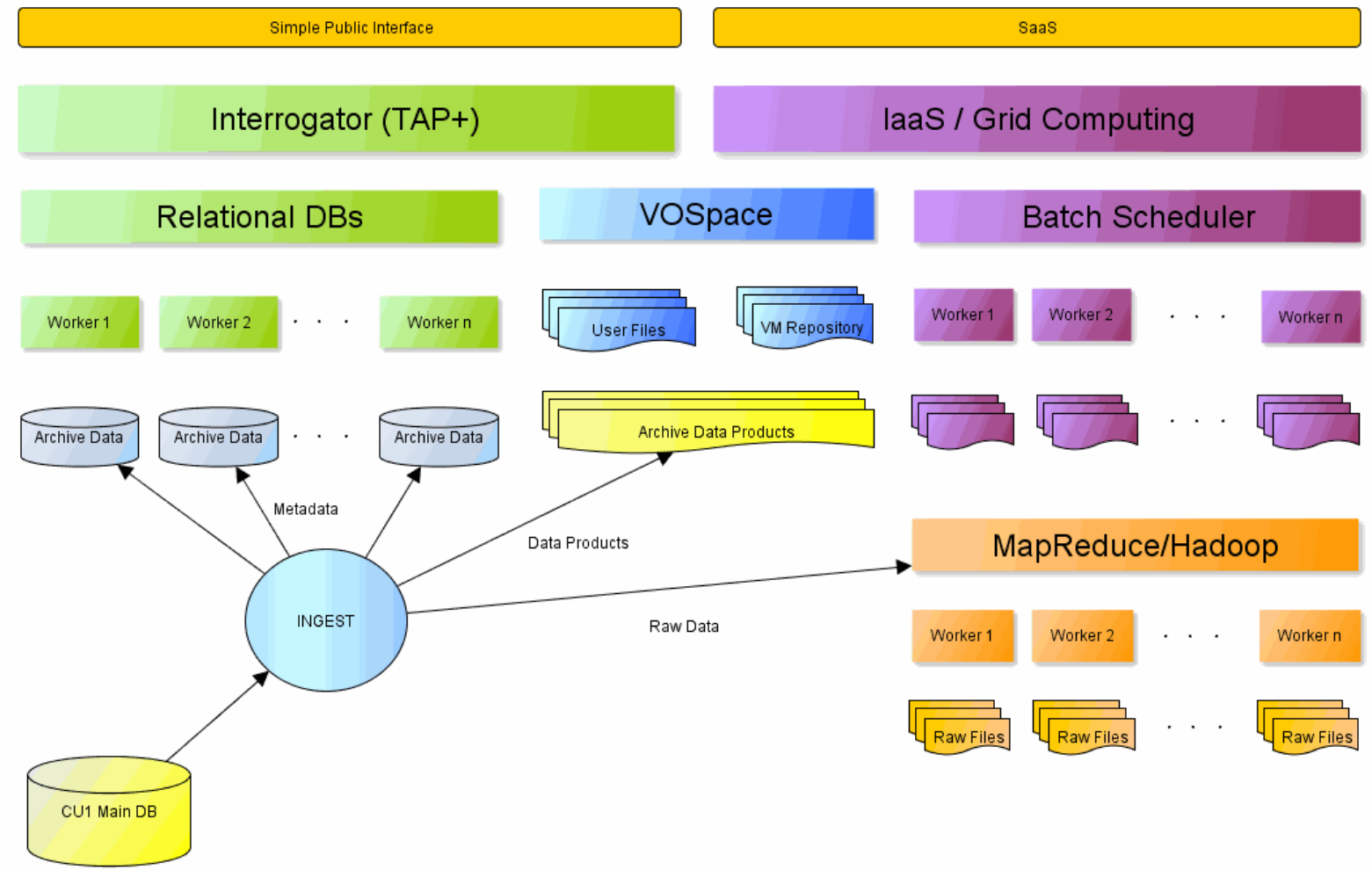
gaia archive

VOSpace

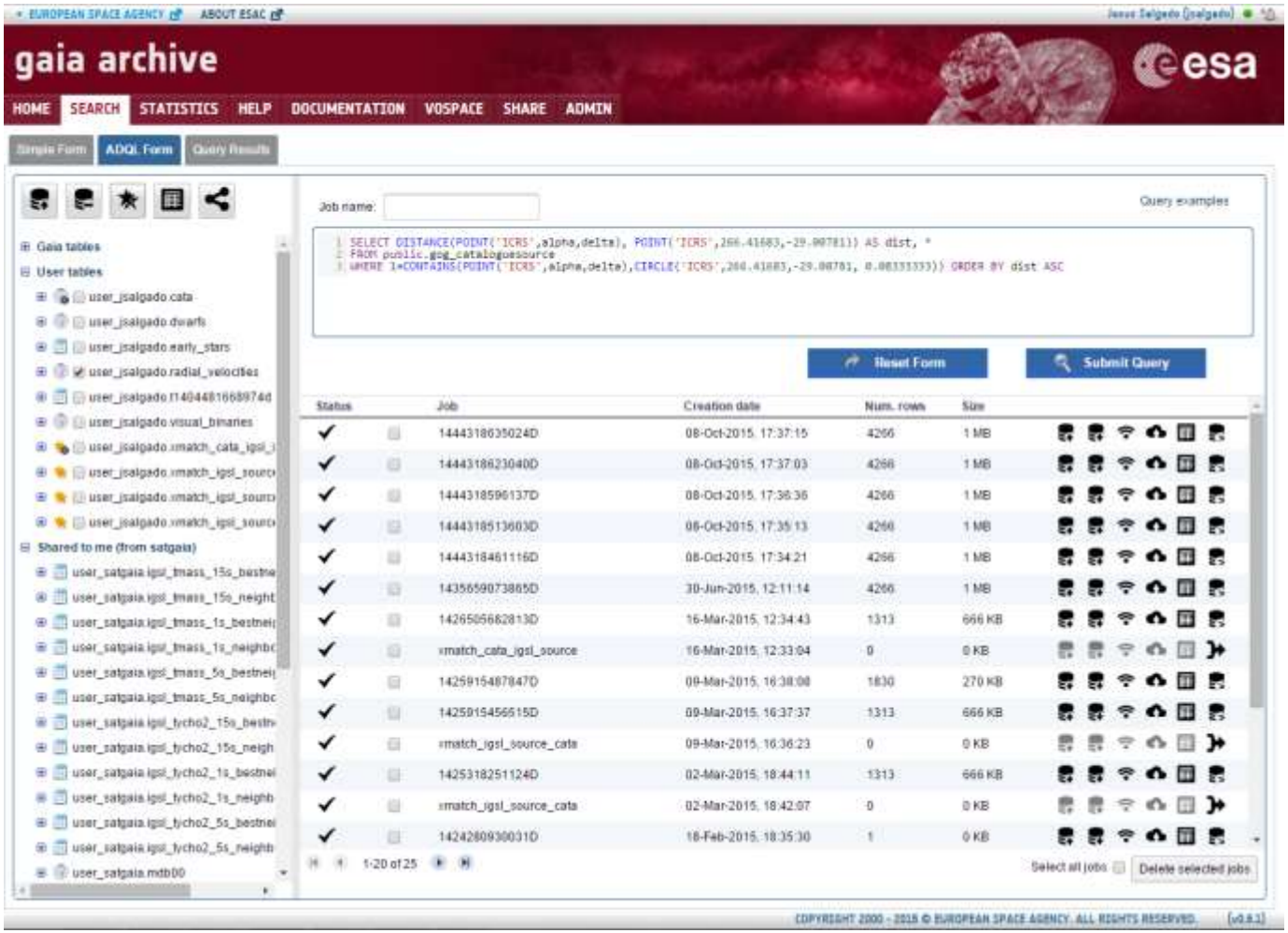
Results

gaia European Space Agency

Gaia Archive Architecture



Gaia Archive: TAP+ Interface



gaia
European Space Agency

Gaia Archive: TAP+ Interface

- Gaia tables
- User tables
  - user_jsalgado.cata
  - user_jsalgado.dwarfs
  - user_jsalgado.early_stars
  - user_jsalgado.radial_velocities
  - user_jsalgado.t1404481668974d
  - user_jsalgado.visual_binaries
  - user_jsalgado.xmatch_cata_igsl_source
  - user_jsalgado.xmatch_igsl_source_dwarfs
  - user_jsalgado.xmatch_igsl_source_radial_velocities
  - user_jsalgado.xmatch_igsl_source_visual_binaries_2
- Shared to me (from satgaia)
  - user_satgaia.igsl_tmass_15s_bestneighbour
  - user_satgaia.igsl_tmass_15s_neighbourhood
  - user_satgaia.igsl_tmass_1s_bestneighbour
  - user_satgaia.igsl_tmass_1s_neighbourhood
  - user_satgaia.igsl_tmass_5s_bestneighbour
  - user_satgaia.igsl_tmass_5s_neighbourhood
  - user_satgaia.igsl_tycho2_15s_bestneighbour
  - user_satgaia.igsl_tycho2_15s_neighbourhood
  - user_satgaia.igsl_tycho2_1s_bestneighbour
  - user_satgaia.igsl_tycho2_1s_neighbourhood
  - user_satgaia.igsl_tycho2_5s_bestneighbour
  - user_satgaia.igsl_tycho2_5s_neighbourhood
  - user_satgaia.mdb00

Persistent Upload

Server Crossmatch

Table sharing

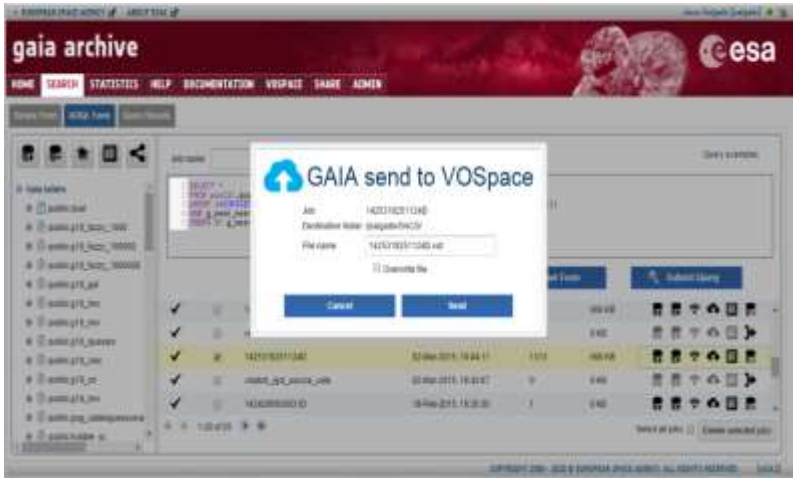
gaia
European Space Agency

Gaia Archive Crossmatch Examples (20 threads)

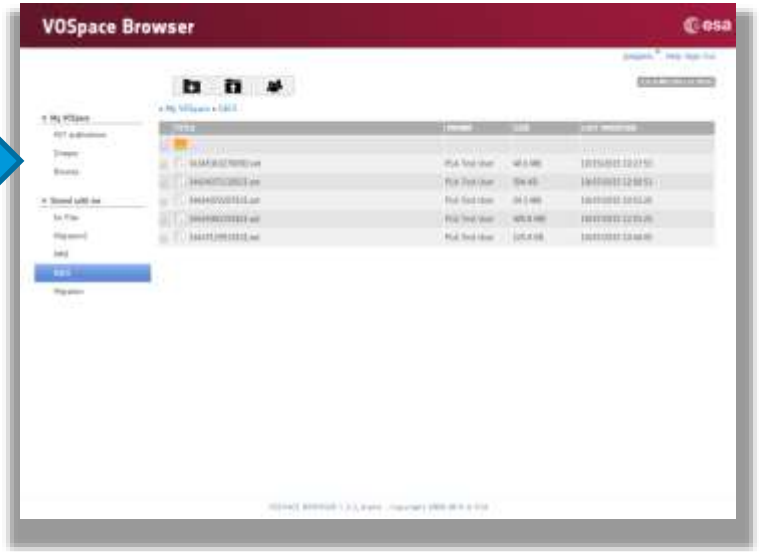
Catalogue 1	Catalogue 2	Radius (arcsec)	# results	Time
Tycho2 2.5×10^6 rows	2MASS PSC 4.7×10^8 rows	1″	2,495,304	49s
Tycho2 2.5×10^6 rows	2MASS PSC 4.7×10^8 rows	5″	2,614,163	116s
Tycho2 2.5×10^6 rows	IGSL 1.2×10^9 rows	1″	2,600,542	46s
Tycho2 2.5×10^6 rows	IGSL 1.2×10^9 rows	5″	2,829,401	55s

Tycho2 vs IGSL crossmatches are even faster than the ones with 2MASS as IGSL is located in the fastest local storage (PCIe), even when IGSL (similar to the final Gaia catalogue) is around 3 times bigger than 2MASS

VOSpace :
Virtual storage for collaboration



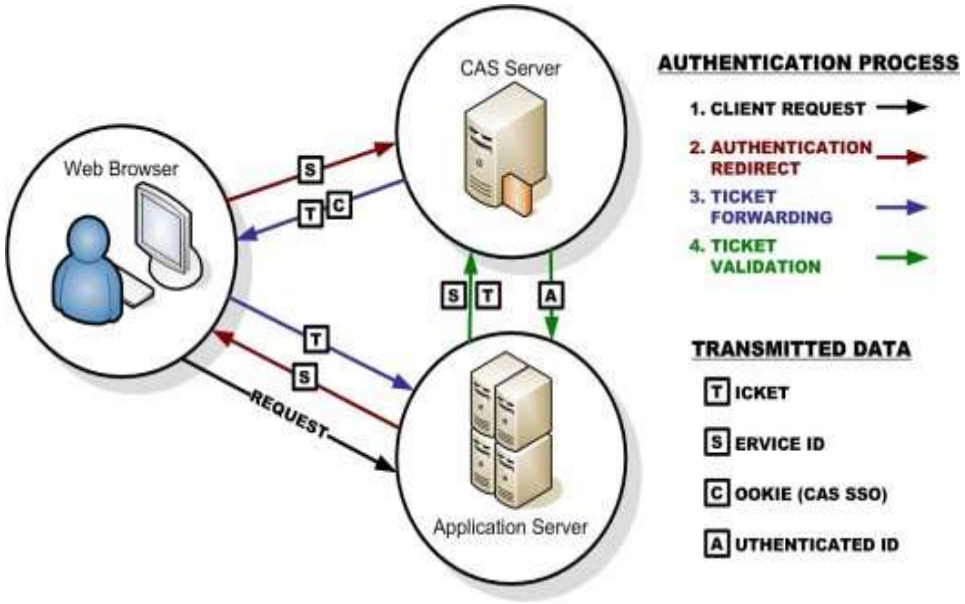
1. Astronomy DropBox
2. VO protocol
3. InterOperable
4. Accessible by VO Applications
5. Accessible by REST
6. Close to data
7. Repository VOSpace IDs?



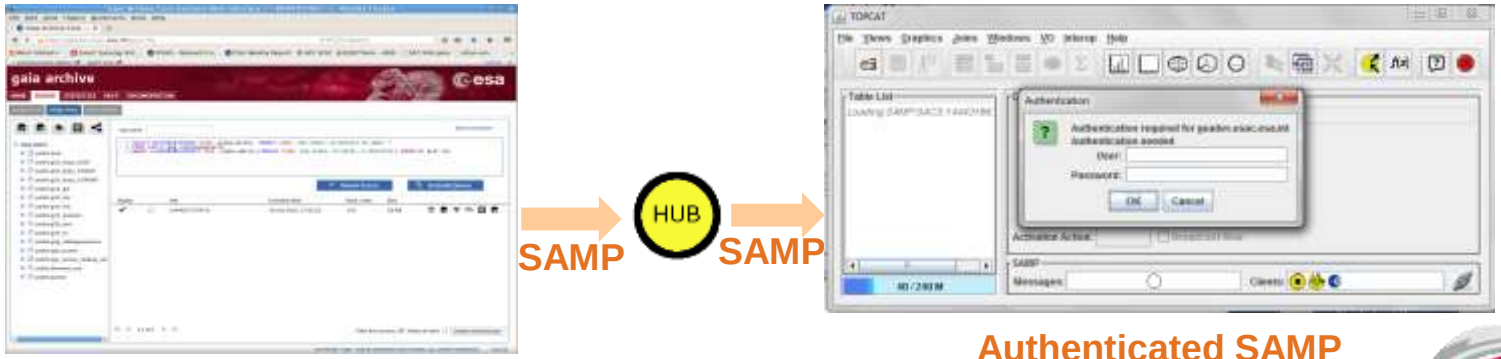
P082 Sara Nieto

gaia
European Space Agency

Gaia Archive Security



1. Single Sign-on through CAS server (both GACS and VOSpace)
2. Authorization through CAS proxy tickets
3. SAMP simple authentication
4. Full support on SAMP https for Web profile still pending at protocol level



Authenticated SAMP

gaia
European Space Agency

Gaia Added Value Interfaces Portal

Portal Systems

- First point of access
- Manage accounts
- Authentication/authorisation
- Web-interface to system
  - Web/API

AVI container

- **AVI**
- Accessed by web-interface
- Runs AVI framework
- Runs astronomy pipelines
  - Your code

Supporting Systems

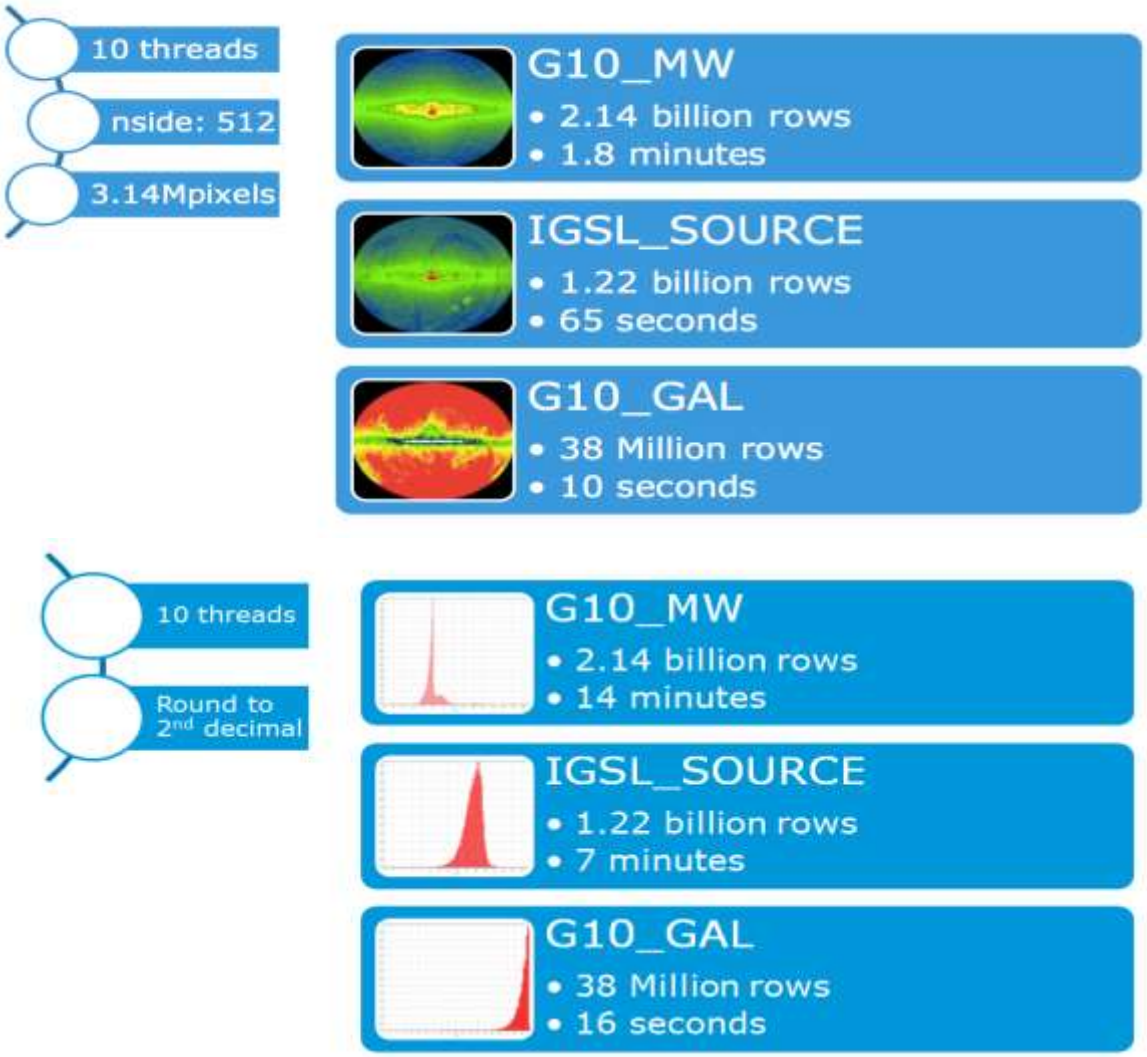
- Monitor system
- Host message bus
  - Several queues
  - Distributed

AVI Infrastructure

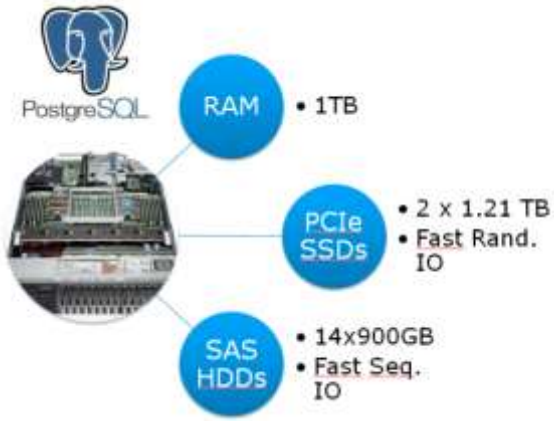
- Manages AVI containers

O1.4 William O'Mullane
O11.3 Christophe Arviset
P080 Vicente Navarro Ferreruela

Vizualization

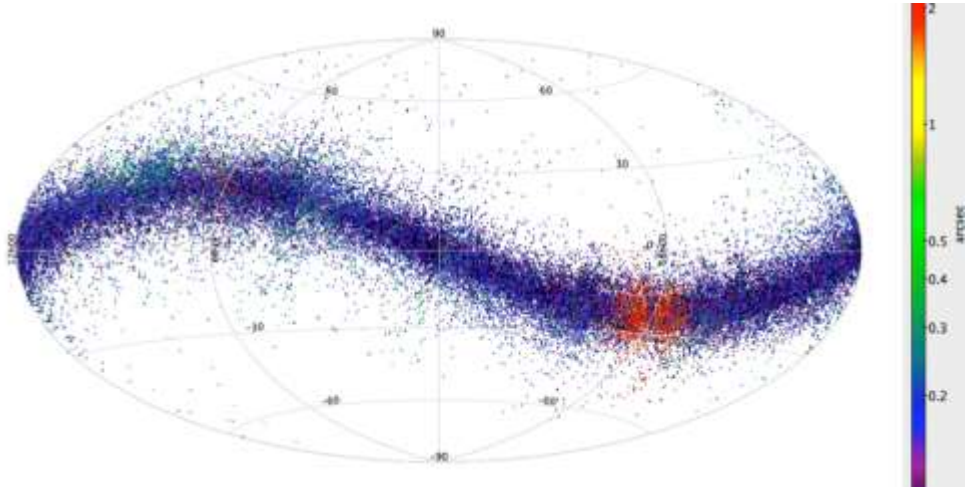


1. Visualization is a need
2. Statistics provide holistic views. Big data techniques
3. Specific visualization packages within Gaia consortium

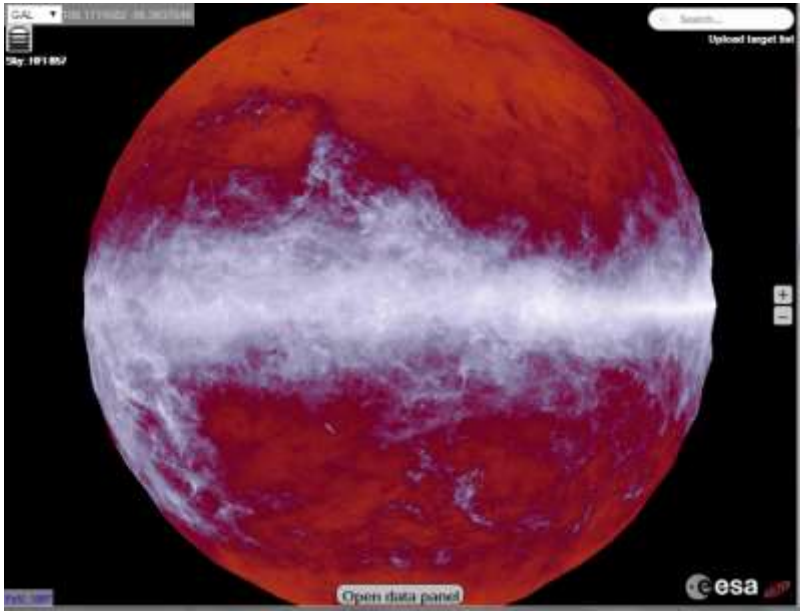


gaia
European Space Agency

Asteroids



"Asteroids position knowledge will be increased on a factor of 50 to 100 due to Gaia mission"



gaia
European Space Agency

Conclusions

1. New Astronomical missions are facing the same challenges about big data of other communities. That implies:
    a. A new way to offer the data
    b. A new way to work for scientists
2. Techniques can be reused like:
    a. Asynchronous jobs, visualization, Map/Reduce
    b. VLDBs (Very Large Databases)
3. Resources for the community
    a. User work space *inside* the archive
    b. Analysis work is done where the data is
4. Security is now an issue

5. First public version coming soon: Summer 2016