PIC
port d'informació
científica

ACME
Astrophysics Centre for Multimessenger studies in Europe

Multi-Messenger CosmoHub

ASTRO-ORDAS

OSCARS-FIESTA

ACME-ASTRO.eu

jordidem@pic.es, *2nd Astro-Ordas workshop, 11th of December 2025*

PIC
port d'informació
científica

IFAE Institut de Física d'Altes Energies

BIST Barcelona Institute of Science and Technology

CERCA Centres de Recerca de Catalunya

UAB Universitat Autònoma de Barcelona

Gobierno de España · Ministerio de Ciencia e Innovación

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

Unió Europea
Fons Europeu
Next Generation

Gobierno de España

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

Ministerio de Ciencia e Innovación

ESCAPE

ACME

PIC
port d'informació
científica

Agenda

ACME
Astrophysics Centre for Multimessenger studies in Europe

- Origin of CosmoHub
- Big data common service
- CosmoHub recent updates
- Extensions to Multi-messenger astronomy
- To do list
- Demo

CosmoHub

The team:

- Jorge Carretero (IP)
- Pau Tallada (lead developer)
- Àlex Luelmo (web developer)
- Sílvia van Den Bogart (research software engineer)
- Elizabeth Gonzalez (postdoc researcher)
- Jordi Delgado (postdoc researcher)
- Ricard Cruz and Carles Acosta (Infrastructure)

Past members:

- Ramón Sant Feliu (predoctoral researcher)
- Agustín Bruzesse (research software engineer)

CosmoHub

As ACME Contribution:

WP4 – Provision of improved access to near-real time and archival multimessenger data (Łukasz Wyrzykowski, Andrii Neronov)

- **IFAE** (leading institute; PI: O. Blanch, co-PI: C. Nigro)
  - Research Infrastructure: MAGIC Telescopes
- **PIC as MAGIC Data Center** (J. Delgado, J. Carretero)
  - WP4: CosmoHub as multi-messenger portal
    - 1 Engineer + Operation costs

Grant ID: 101131928; Sep 2024 - Aug 2028, 14.5M€
IPs: CNRS (France); https://cordis.europa.eu/project/id/101131928

Origin of CosmoHub

Developed within the Cosmology groups at PIC and ICE (2013)

- Supporting Physics of the Accelerated Universe Survey (PAUS)
- Supporting MICE mock catalogues

Aims:

- Distribute datasets
- Perform complex queries for data selection
- Easy data exploration (quick plots)

Solution (2013): web interface + PostreSQL DB

Big data common service: Hadoop

(2016) As the size of the datasets increased, complexity of the data selection (limit of 10K rows)→ problems with performance → Apache Hadoop stack + Infrastructure

> 30 TiB of catalogued information

50×10^9 astronomical objects from a dozen different projects

Interactive visualization of datasets of thousands of millions of objects can be done in less than a minute

Customized subsets can be generated in a timescale of minutes

Tallada, P., Carretero, J., Casals, J., Acosta-Silva, C., Serrano, S., Caubet, M., … Tonello, N. (2020, July). ***CosmoHub: Interactive exploration and distribution of astronomical data on Hadoop***. Astronomy and Computing. Elsevier BV. http://doi.org/10.1016/j.ascom.2020.100391

Big data common service: Hadoop

- Software stack
  - 1th gen: Hortonworks DP 2.6.5
  - 2nd gen: Hortonworks DP 3.1.4
  - **3rd gen: Shepherd (in-house)**
- Hardware architecture
  - 1st gen: obsolete nodes
  - 2nd gen: 4 dual-twin
  - 3rd gen: 12 nodes DIY
  - **4th gen: 32 nodes**

PIC's Hadoop distribution: Shepherd

- Motivation
  - Avoid vendor lock-in (Cloudera)
  - Binaries no longer accessible
  - Outdated versions
  - Flexibility in the combination of components
- Shepherd Hadoop distribution
  - Consolidated in a **single docker image**
    - datanode and nodemanager also outside docker
  - Tested and deployed using **CI/CD** Pipeline
    - Development and preproduction setup
  - Additional RPM for client-only installation
  - Monitored using JMX protocol

Component	HDP 3.1.4	Shepherd 1.0.0
Atlas	1.1.0	2.2.0
Hadoop	3.1.1	3.2.3
HBASE	2.0.2	2.5.8
Hive	3.1.0	3.1.2
Kafka	2.0.0	2.5.0
Oozie	4.3.1	5.2.1
Ranger	1.2.0	2.4.0
Solr	7.7.0	8.11.2
Spark	2.3.2	3.4.2
Tez	0.9.1	0.10.1
Zookeeper	3.4.6	3.7.1
Kerberos	MIT KDC	FreeIPA

Big data common service: infrastructure

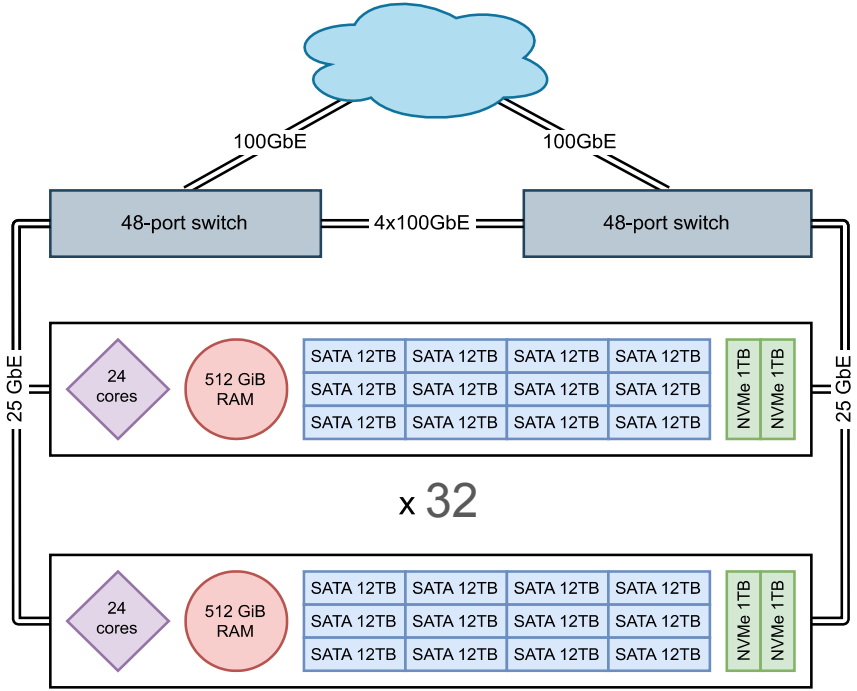
Dedicated infrastructure

32 nodes:

- 768 cores, 16 TiB RAM
- 60 TB NVMe (for cache)
- 4.3 PiB raw storage (3 PiB usable)

Main use cases:

- **CosmoHub** query processing
- Euclid **mock galaxy catalogs**
- **dCache billing** metadata analysis
- **HTCondor backfilling**



Publicación de convocatoria
Equipamiento científico – técnico
AGENCIA ESTATAL DE INVESTIGACIÓN

Convocatoria 2021 de ayudas para la adquisición de equipamiento científico-técnico
Clúster Hadoop para la gestión integral de datos científicos masivos" (EQC2021-007479-P)
Funding granted: 401k€

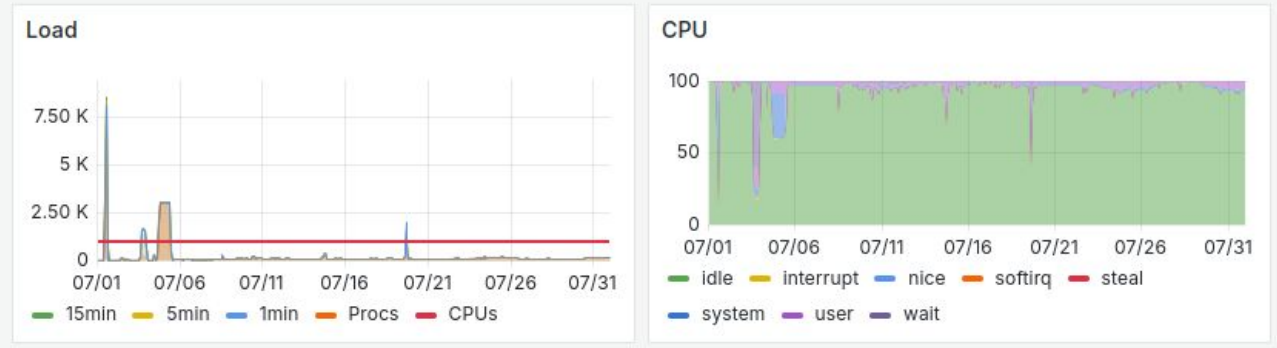
Big data common service: infrastructure

Well…not exclusively Dedicated!

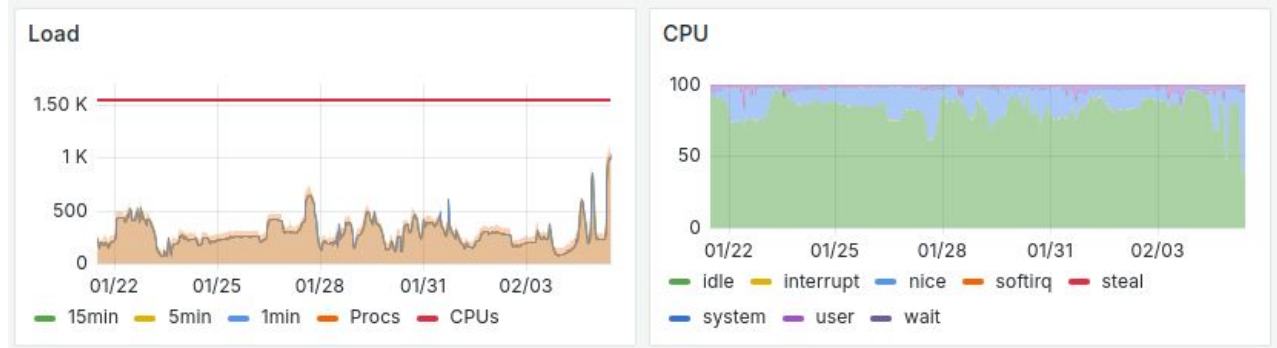
We implemented backfilling from HTCondor:

- DUNE and Virgo jobs running on the cluster
- CosmoHub performance unafected

Job submission and CPU usage of the big data cluster before backfilling



Backfilling job submission and CPU usage of the big data cluster



Current CosmoHub

- Interactive exploration (visualization)
  - **Very fast** (95% < 1')
  - **Full dataset plots** (over all rows)
    - May use sampling
    - Cone search tool
    - 1D histogram & 2D heatmap
  - **Guided process** (no SQL knowledge required)
    - Expert mode
- Distribution
  - Parquet, CSV, FITS, ASDF format
  - Email with a link to download dataset

- Data
  - 160 TiB catalogued data
  - 150 catalogs (simulated and observed)
  - **Supporting multiple projects**
    - DES, PAUS, Euclid, MICE, LST, Gaia, LSST…
- Users
  - ~2000 registered users
    - ~150 active users
  - >19K custom catalogs generated
  - >20k interactive queries
- Performance
  - >85% of custom catalogs finish in < 3'
  - Resource queues with reservation
  - Preemption to keep interactive response time

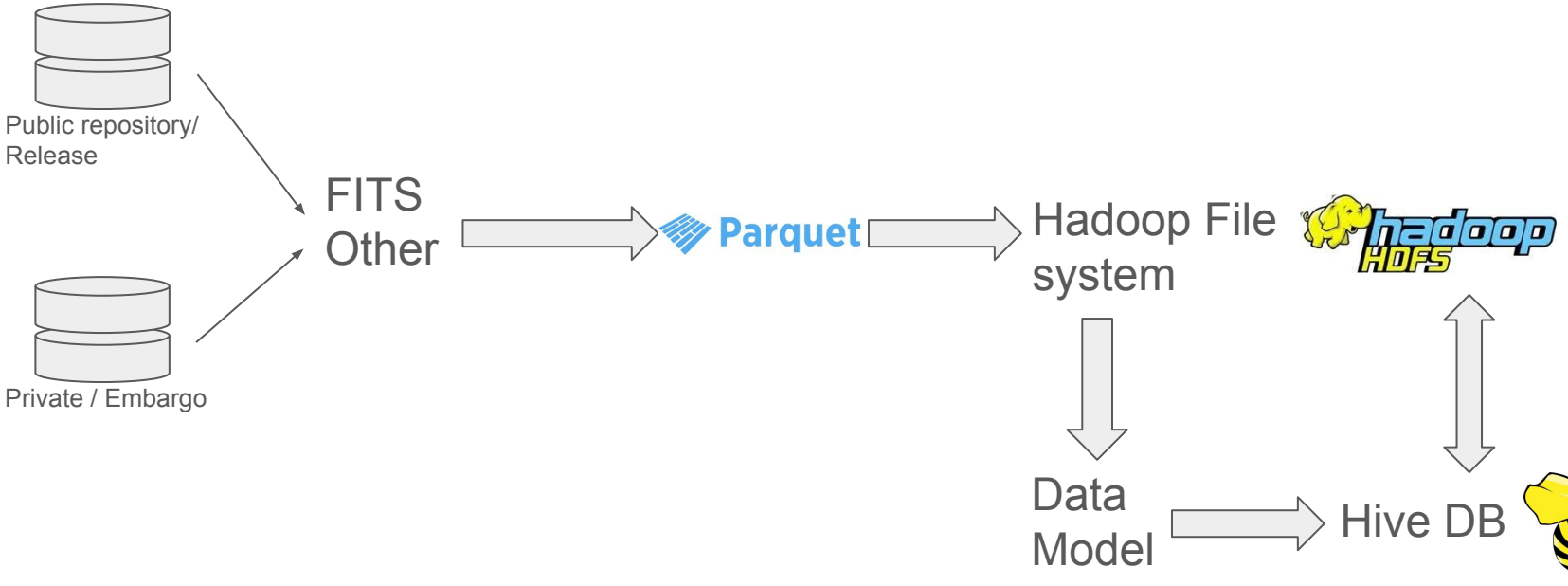
Current CosmoHub

Aligned with FAIR principles:

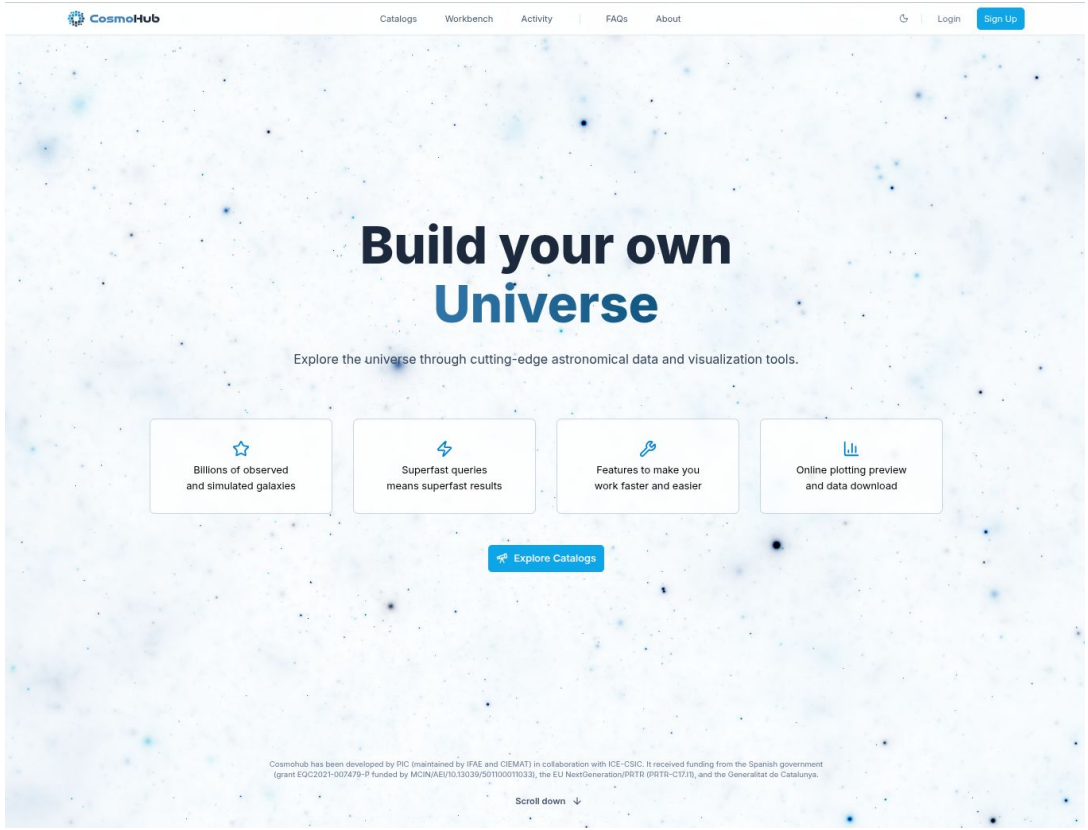
- **Findable:** It offers a centralized hub where users can easily discover a broad spectrum of cosmological datasets.
- **Accessible:** Available through a web application, CosmoHub provides a user-friendly interface designed for users with limited technical skills.
- **Interoperable:** By promoting the use of standard file formats and adopting IVOA standards (such as ADQL, ConeSearch, and TAP), CosmoHub ensures that data from various sources can be seamlessly integrated, queried, and shared across different platforms.
- **Reusable:** Every custom dataset, sample, or plot created by users is fully reproducible, ensuring transparency and supporting long-term research validity.

CosmoHub has been selected as a success use case in the **EOSC Macro-roadmap** thanks to its contributions to promoting open science, fostering interdisciplinary research, and advancing data-driven innovation in Astronomy and Cosmology (Dec 2024)

Data ingestion



CosmoHub recent updates

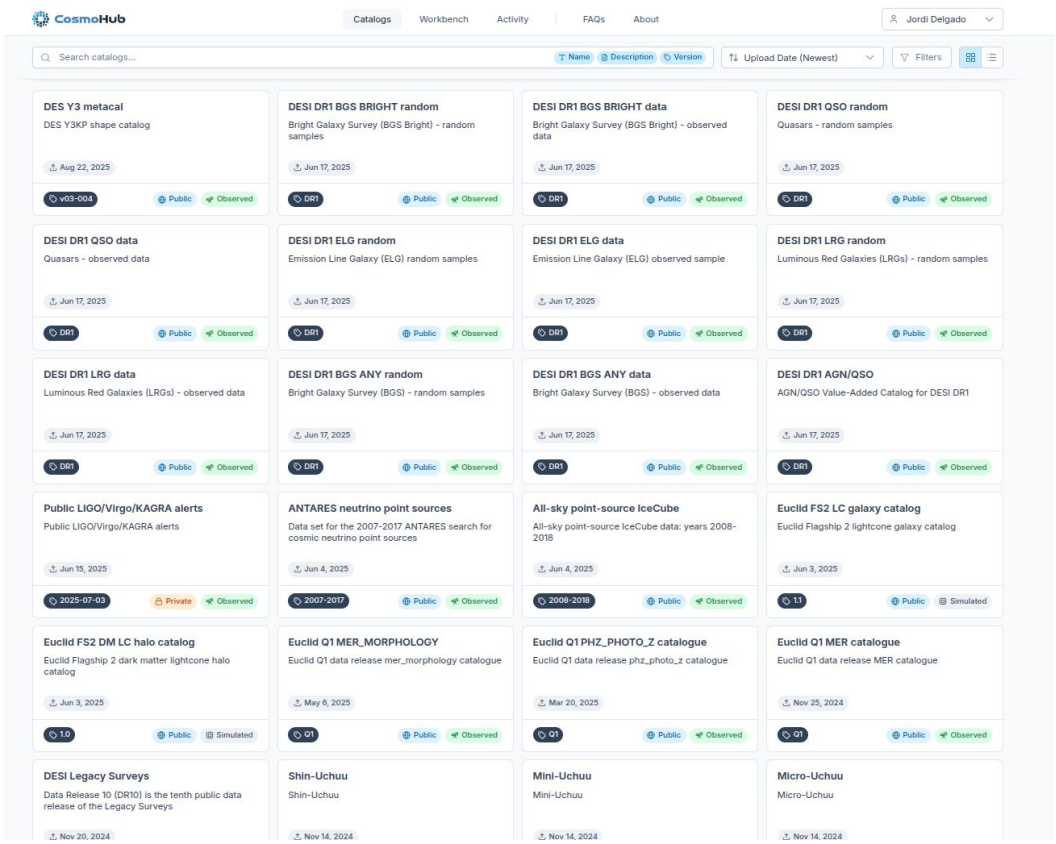


New release: 2025.11.25

New features:

- Modernized Look and feel
- Filters
- Better browsing experience
- Data visualization improvements

CosmoHub recent updates

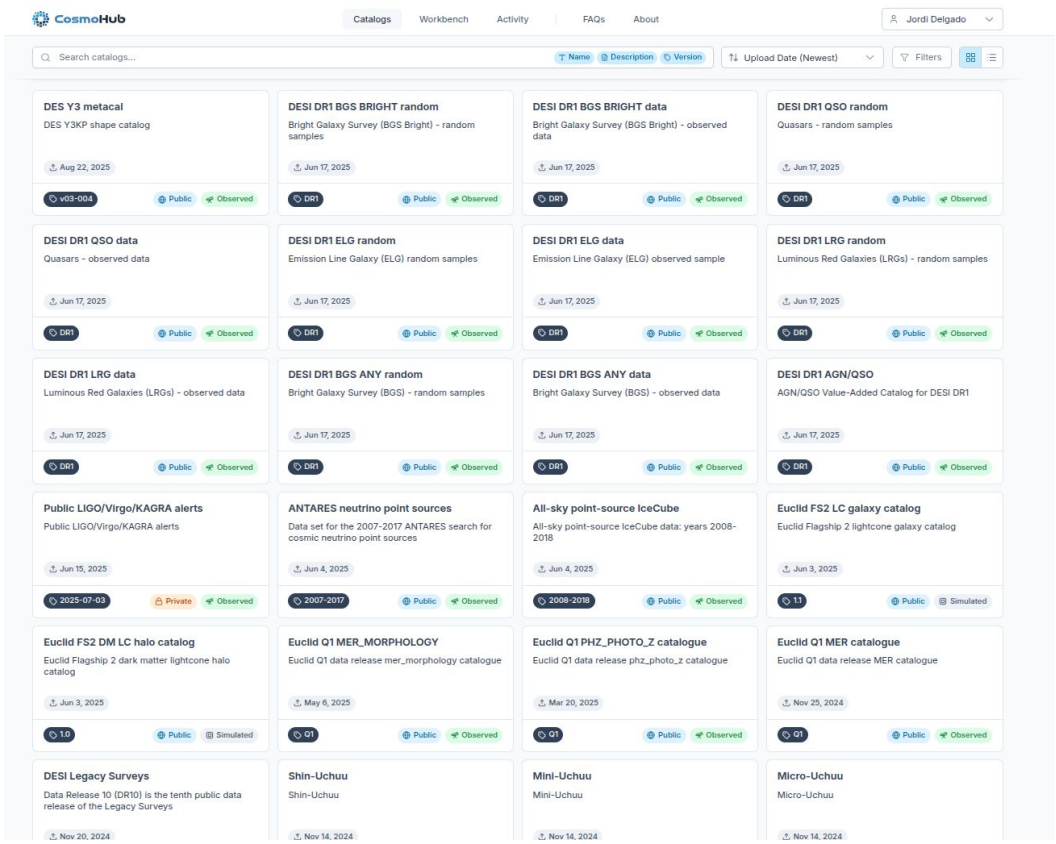


New release: 2025.11.25

New features:

- Modernized Look and feel
- Filters
- Better browsing experience
- Data visualization improvements

CosmoHub recent updates

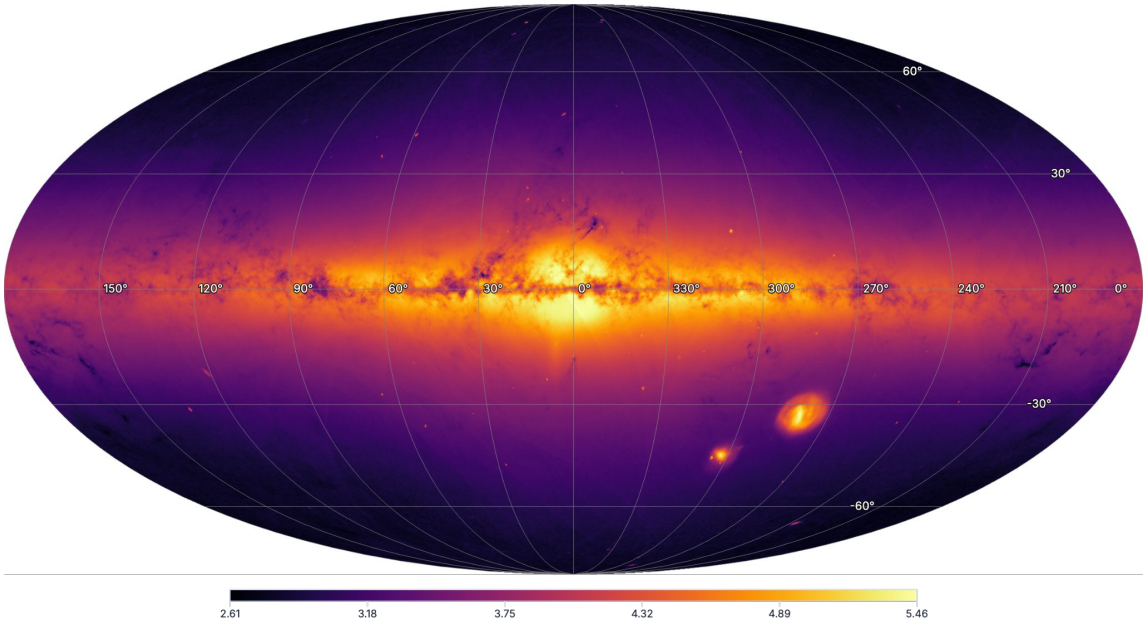


New release: 2025.11.25

New features:

- Modernized Look and feel
- Filters
- Better browsing experience
- Data visualization improvements

CosmoHub recent updates



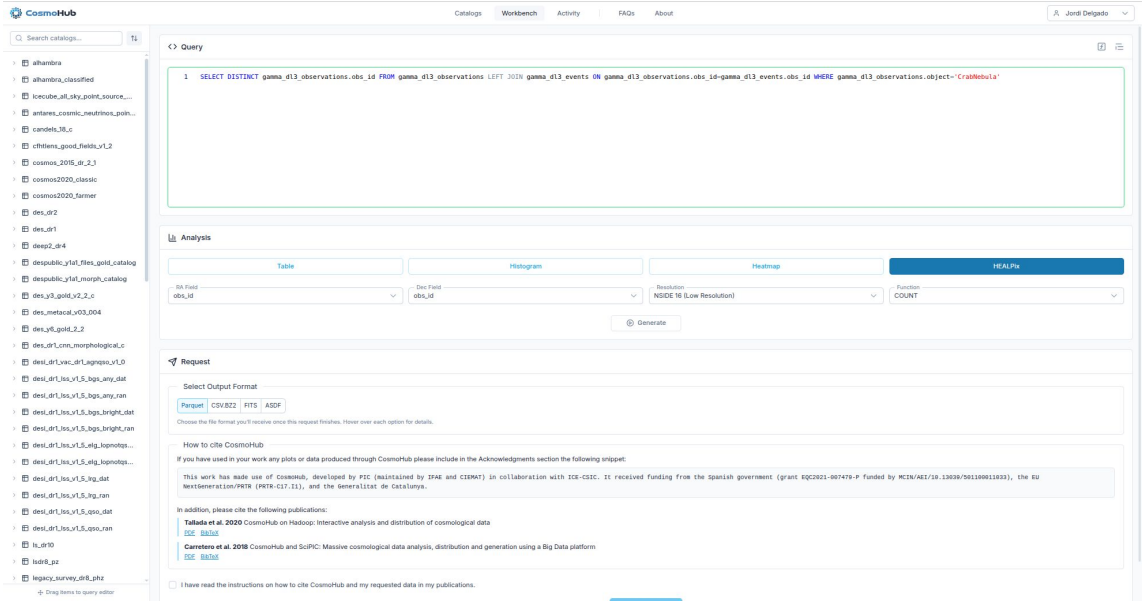
Gaia DR3 objects plotted from CosmoHub

New release: 2025.11.25

New features:

- Including enhanced plotting tools
- Support for HEALPix maps
- Plotting of array-based data.

CosmoHub recent updates



New release: 2025.11.25

New features:

- Workbench

Extensions to multi-messenger astronomy

Towards a Multi-Messenger data hub:

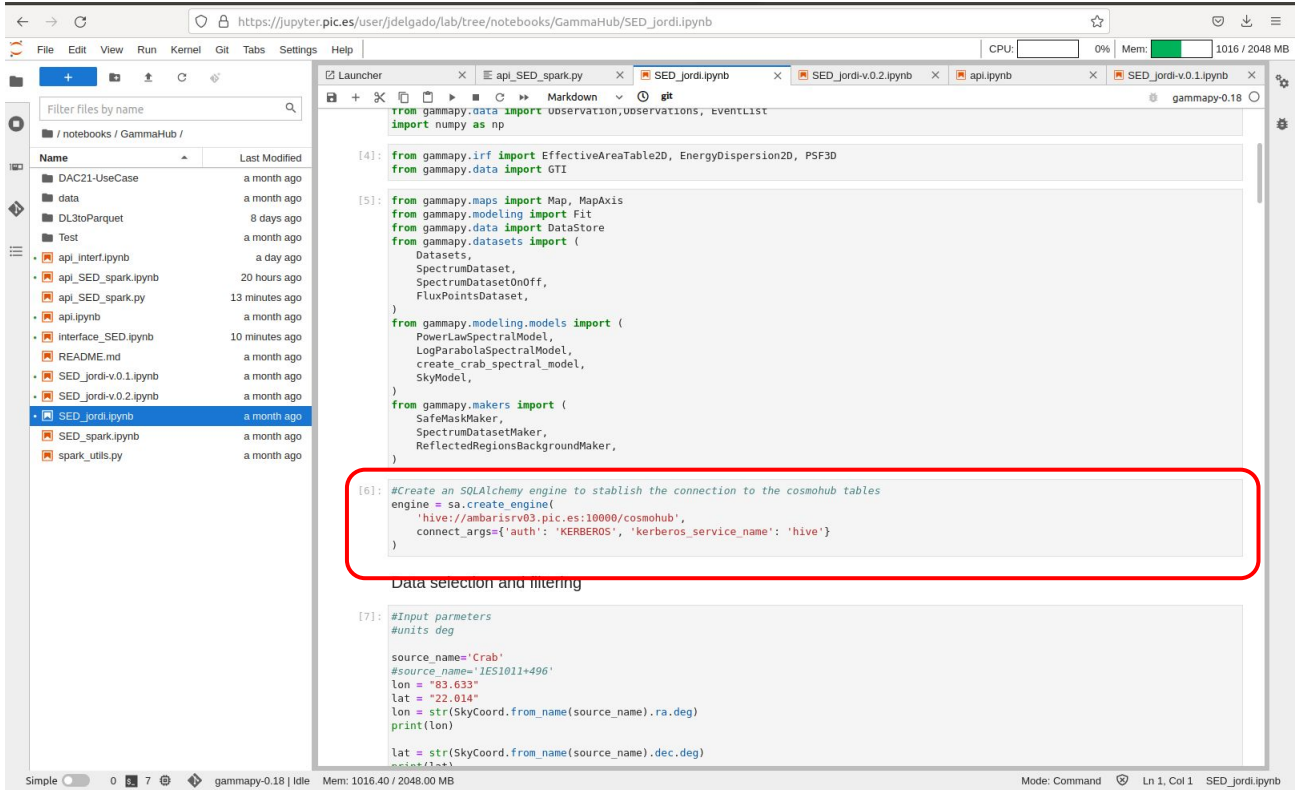
- Multi -instrument, -messenger, -wavelength
  - Electromagnetic:
    - Optical (Euclid ✔, Gaia ✔, LSST⚙)
    - Gammas: (MAGIC ✔, CTAO/LST⚙ , FERMI⚙)
  - GWs (LIGO/VIRGO) events ✔
  - Neutrinos (ANTARES ✔, KM3NeT⚙)

Extensions to multi-messenger astronomy

Towards a Multi-Messenger data hub:

- Gamma-ray modeling

CosmoHub extensions for Gamma-ray astronomy (II)



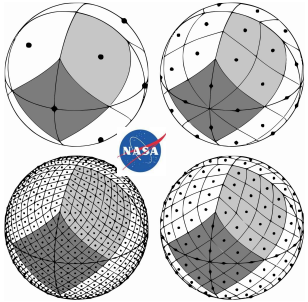
CosmoHub DB

- Accessible through Jupyter (jupyter.pic.es)
- Support Gammapy
- Spark parallelization

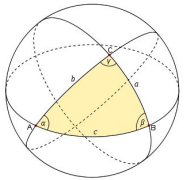
Dask parallelization

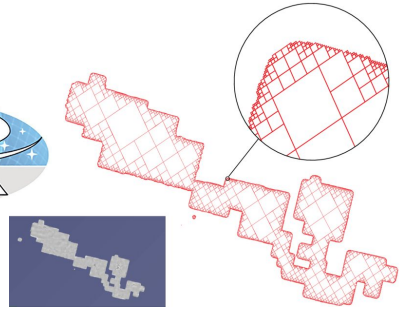
Extensions for Multi-Messenger astronomy: User Defined Functions

API extensions:



- HEALPix (enable spherical analysis)
  - Conversion: ang2pix, pix2ang, ang2vec, vec2ang, vec2pix, pix2vec
  - Ordering: nest2ring, ring2nest
  - Sizing: nside2npix, npix2nside, order2npix
  - Other: neighbours, maxpixrad, nside2order
- Array aggregation
  - Summary: array_min, array_max, array_count, array_sum
  - Dispersion: array_avg, array_stddev_pop, array_stddev_samp array_var_pop, array_var_samp
- Spherical geometry
  - Types: point, box, circle, polygon
  - Constructors: from coordinates / pixels
  - Simple: area, centroid, coord1, coord2, distance
  - Spatial relationships: contains, intersects
  - Region extension (MOC): complement, intersection, union





To Do list

A list of missing functionalities and *ongoing* developments:

- Multi-messenger supports
  - Scientific data products
- New backend API
- A&A federation
  - Implementation of automatic federation using Keycloak plugin
  - Full integration with PIC systems
- Webdav exportation of the HDFS
  - Development of a client extension
- Exploring alternatives to Hadoop and Hive
  - Starrocks

Demo

Part 1: Browsing through CosmoHub

Part 2: producing plots using CosmoHub

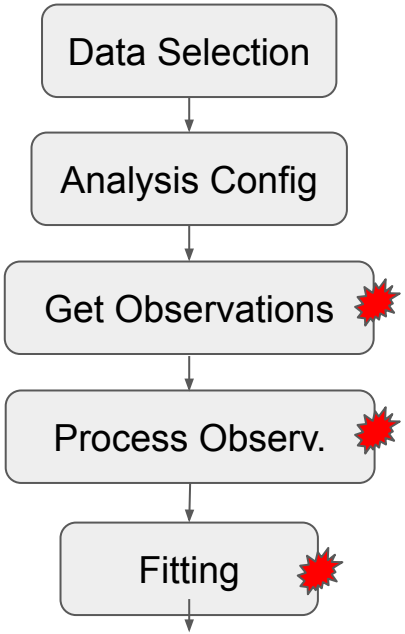
Part 3: jupyter notebooks interfacing CosmoHub and Spark nodes

Example notebooks at: https://gitlab.pic.es/jdelgado/acme-wp4-cosmohub-demo.git

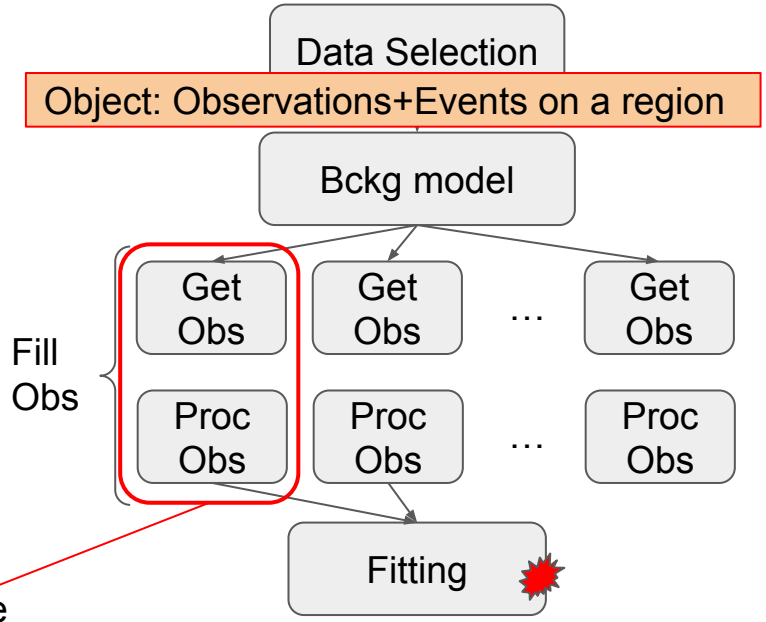
Demo

Jupyter data products: introducing Spark parallelism

Gammapy High level SED workflow



Gammapy low level SED workflow
+ data parallelism



map-reduce
mapper function call per Observation

ACKs

Thanks!

Acknowledgements: "This study was supported by MICIN with funding from European Union NextGenerationEU(PRTR-C17.I1) and by Generalitat de Catalunya".
"This work has been financed by the grant EQC2021-007479-P funded by MCIN/AEI/ 10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR".

PIC
port d'informació
científica

ACME
Astrophysics Centre for Multimessenger studies in Europe

Multi-Messenger CosmoHub

ASTRO-ORDAS

OSCARS-FIESTA

ACME-ASTRO.eu

jordidem@pic.es, *2nd Astro-Ordas workshop, 11th of December 2025*

PIC
port d'informació
científica

IFAE Institut de Física d'Altes Energies

BIST Barcelona Institute of Science and Technology

CERCA Centres de Recerca de Catalunya

UAB Universitat Autònoma de Barcelona

Gobierno de España · Ministerio de Ciencia e Innovación

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

Unió Europea
Fons Europeu
Next Generation

Gobierno de España

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

Ministerio de Ciencia e Innovación

ESCAPE

ACME