PLANES COMPLEMENTARIOS

Astrofísica y física de altas energías

PIC port d'informació científica

Multi-Messenger data hub MM - CosmoHub

Línea 8. Computación, big data e inteligencia artificial

J. Carretero on behalf of the PIC team

IFAE Institut de Física d'Altes Energies

Gobierno de España · Ministerio de Ciencia e Innovación · Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

BIST Barcelona Institute of Science and Technology

CERCA Centres de Recerca de Catalunya

UAB Universitat Autònoma de Barcelona

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

Main goals

*"Acceso eficiente a grandes conjuntos de datos astrofísicos (**CosmoHub**), **hub de astronomía multi-mensajero**. Además, fortalecer el equipo humano responsable del desarrollo de herramientas para la gestión de alertas y planificación (scheduling). Más generalmente, se **reforzarán los equipos** que trabajan en la **generación, gestión y explotación de datos masivos** de los experimentos descritos anteriormente, consolidando y expandiendo su know-how en las **herramientas** necesarias para hacer efectiva la **data science**, así como desarrollos de computación avanzada para la descripción de fenómenos astrofísicos complejos."*

Outline

- Port d'Informació Científica (PIC)
- Big Data common service
- Multi-messenger approach
- Summary

Budget allocation

- Computing & Infrastructure
  - Computing: 1024 CPUs for immersion cooling
  - Storage:
    - Spinning disk: 400 TB
    - Magnetic tape: Frame + 2 tape drives + 3PB in cartridges
  - Network: switches, optics and cabling
  - Power: adapt electrical infrastructure and other tasks
  - Other equipment: laptops
- Personnel:
  - Postdoc researcher (DevOPS) - 2.5 FTE
  - Senior technician (Maintenance & Service Development) - 2.25 FTE
  - Postdoc researcher (Data scientist) - 1.25 FTE
  - Postdoc researcher (Data scientist) - 1.5 FTE
  - Predoctoral student (Data scientist) - 0.5 FTE
  - Software engineer (Storage specialist) - 1 FTE
  - Software engineer (Web developer) - 1 FTE

- **Not started yet**
- **In progress**
- **Done**

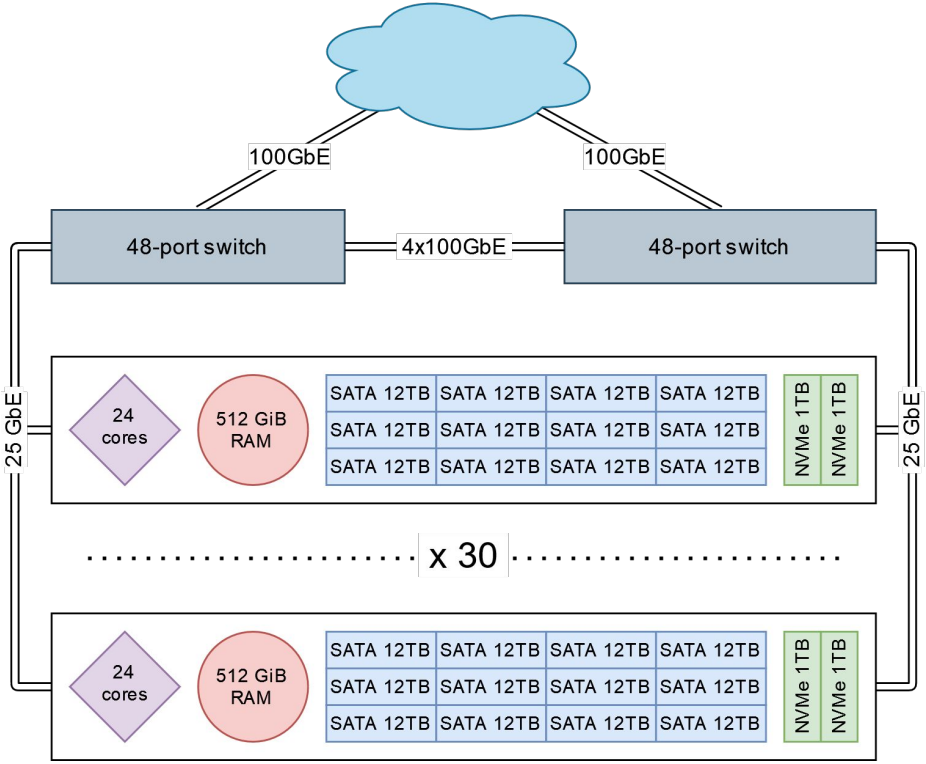
Port d'Informació Científica

- Founded in 2003: collaboration between IFAE and CIEMAT
  - **PIC joined the Spanish Supercomputing Network in 2020**
- Team of 26 people (50% scientists - 50% engineers)
  - **Agile teams that embed in scientific groups to**
    - Understand the experiment
    - Follow the evolution of data analysis requirements
    - Develop & prototype tools for data management and analysis

- What we do
  - R&D in methodologies and tools for **advanced data analysis**
    - Lead and participate in R&D projects. Software and Computing WPs
  - Operate services for the preservation, analysis and sharing of data
    - Run prod. services for experiments: **Euclid**, PAUS, **CTA**, MAGIC, **LIGO/Virgo**, DUNE and LHC
    - Provide data analysis services for research groups: IFAE, CIEMAT + others

Big Data common service

- Software stack
  - 1th gen: HDP 2.6.5
  - 2nd gen: HDP 3.1.4
  - 3rd gen: Shepherd
- Hardware architecture
  - 1st gen: obsolete nodes
  - 2nd gen: 4 dual-twin
  - 3rd gen: 12 nodes DIY
  - **4th gen: 20+10 nodes**



720 CPU cores, 15 TiB RAM, ~2.8PiB net storage, 60 TB NVMe cache

Big Data common service: Hadoop

- Software stack
  - 1th gen: HDP 2.6.5
  - 2nd gen: HDP 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: 20+10 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
    - pseudo-distributed, development and preproduction setup
  - Configuration tracked in a git repository
  - Additional RPM for client-only installation
  - Monitored using JMX protocol

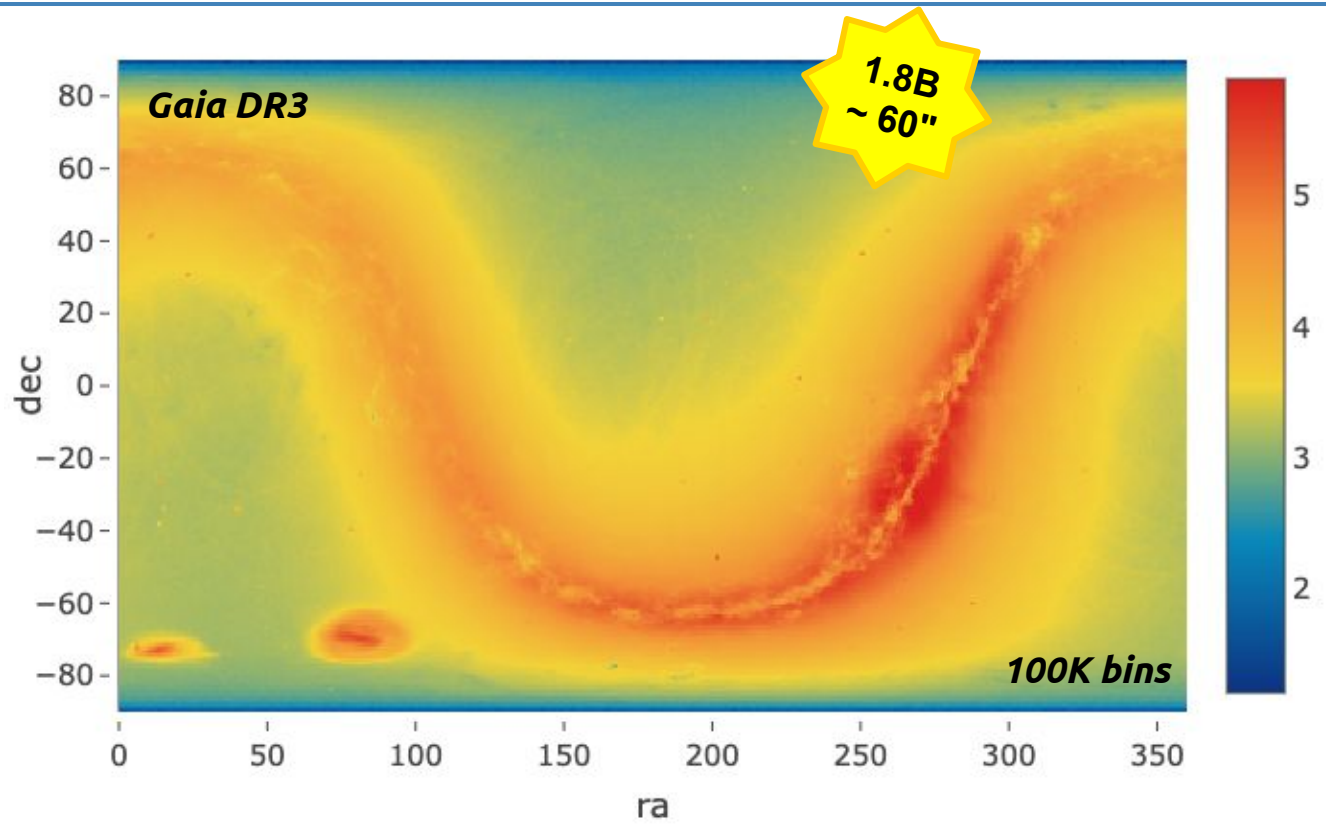
Component	HDP 3.1.4	Shepherd 1.0.0
Atlas	1.1.0	2.2.0
Hadoop	3.1.1	3.2.3
HBASE	2.0.2	2.5.8
Hive	3.1.0	3.1.2
Kafka	2.0.0	2.5.0
Oozie	4.3.1	5.2.1
Ranger	1.2.0	2.4.0
Solr	7.7.0	8.11.2
Spark	2.3.2	3.4.2
Tez	0.9.1	0.10.1
Zookeeper	3.4.6	3.7.1
Kerberos	MIT KDC	FreeIPA

- Interactive exploration (visualization)
  - **Very fast** (85% < 30s)
  - **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
  - 90 TiB catalogued data
  - >130 catalogs (simulated and observed)
  - **Supporting multiple projects**
    - DES, PAUS, Euclid, MICE, LST, Gaia, LSST…
- Users
  - >1750 usuarios registrados
    - ~150 active users
  - >17K custom catalogs generated
  - >20k interactive queries
- Performance
  - >75% of all queries finish in <3 min
  - Resource queues with reservation
  - Preemption to keep interactive response time

SQL:

```
SELECT `ra`, `dec`
FROM gaia_dr3_source
```

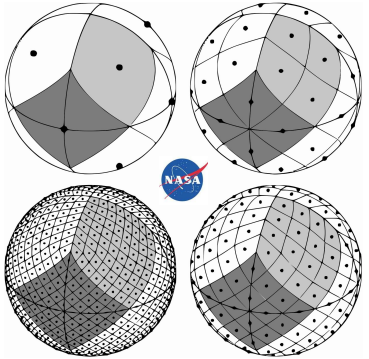


Gaia affiliate data center

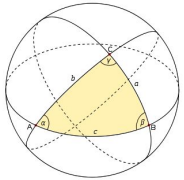
MM-CosmoHub

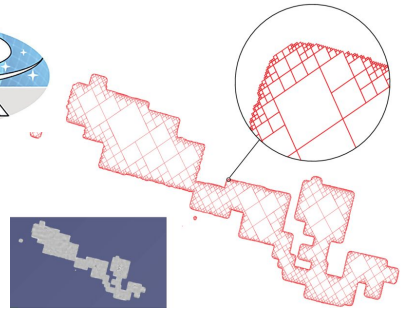
- Multi-instrument, -frequency, -messenger
  - Electromagnetic
    - Optical (Euclid ✔, Gaia ✔, LSST⚙)
    - Gammas (MAGIC ✔, CTAO/LST⚙)
  - GWs (LIGO/Virgo) ⚙
  - Neutrinos (ANTARES, KM3NeT)
- Massive transfers:
  - Rucio + FTS ✔
  - Rclone ✔
- Standardize & facilitate access:
  - Federated Authentication ⚙
    - Tokens
  - Virtual observatory standards ⚙
    - VOTable, ADQL, TAP, UWSd
  - Persistent identifiers (DOIs)
- Advanced features
  - Users can upload catalogs ⚙
  - User Defined Functions (UDFs) ✔
  - Enhanced plotting
    - HEALPix maps
    - Improved density
- Jupyterlab integration ✔
  - Notebooks over PIC's HTC cluster
  - SSH terminal, VNC desktop, VS Code IDE
- Massive algorithms (Dask ✔ / Spark ✔)
  - Spatial cross-match ⚙
  - Light curves & SED ⚙
  - Synthetic galaxy catalog generation ✔
- Infrastructure expansion ⚙
  - Computing / storage
    - 1024 cores, 0.4 PB disk, 3PB (tapes)

User Defined Functions



- 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





MM-CosmoHub: Gamma rays

- Datasets from different astronomical gamma-ray experiments
  - This work continues the efforts of the project ESCAPE
- Data model based on VODF (Very-High-Energy Open Data Format)
  - Data Level 3 (DL3) science rady file format
    - Supported by Gammas and Neutrinos
    - Deploy DL3 model into Hive DB
- CosmoHub web interface
  - Cone search
- Implementing DL3 analysis
  - Based on Gammapy (core library of CTA science tools)
    - Interface to Hive
    - Dask parallelism, on top of the jupyter.pic.es + HTCondor cluster
    - Spark parallelism, on top of the jupyter.pic.es + Hadoop cluster

Summary

- Expand computing and infrastructure
  - Computing: 1024 CPUs
  - Storage: spinning disk (400 TB) & magnetic tapes (3 PB)
  - Network & power installation
- Develop advanced features
  - Data from different sources (optical, gamma-ray, GW, neutrinos)
  - Federated Authentication
  - Interoperability (VO standards)
  - User Defined Functions (HEALPix, array aggregation, spherical geometry)
  - Additional file formats (e.g. Parquet, DL3)
- Personnel with crucial knowledge about big data management
  - Shepherd: Integration of a custom Hadoop distribution
  - Rucio / Rclone for massive data transfer

Thanks for your attention!

Credits to: E. Acción, V. Acín, C. Acosta, A. Alou, A. Bruzzese, L. Cabayol, J. Carretero, J. Casals, R. Cruz, M. Delfino, J. Delgado, M. Eriksen, D. Graña, J. Flix, E. Johana, G. Merino, C. Neissner, A. Pacheco, A. Pérez-Calero, E. Planas, M.C. Porto, J. Priego, P. Tallada, F. Torradeflot

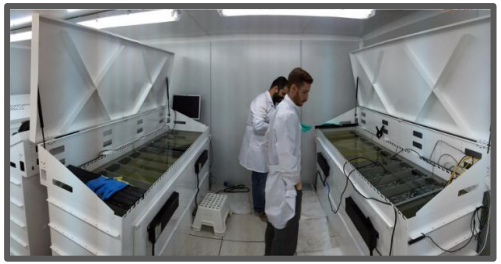
www.pic.es

PIC data center

- Connectivity
  - 2x100 Gbps to Academic Network
  - Largest data mover in Spanish academic network: 100 PB in+out per year

- Data processing services
  - Disk - dCache: 30 PB
  - Tape - Enstore: 70 PB
  - Computing - HTCondor: 13000 cores, 18 GPUs
  - Computing - Hadoop: 720 cores, 2.8 PB disk

- Facilities, ~120 kW IT
  - ~80 kW in 150 m^2 air-cooled room
    - high efficiency, PUE 1.44
  - ~40 kW in 25 m^2 liquid immersion cooling system
    - very high efficiency, PUE 1.1



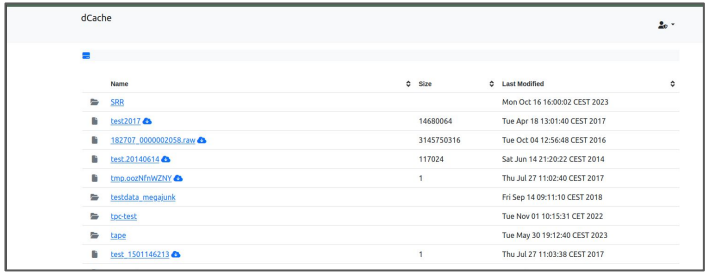
Massive data transfer: Rucio + FTS

- Rucio: comprehensive solution for managing, organizing, and distributing data
- Led an effort to **integrate a Rucio deployment in a single helm chart**
  - Includes a PostgreSQL instance, the Rucio server and the daemons.
  - Monitoring using Grafana and ElasticSearch
  - Token support (in progress)
- Deployed several instances to manage different project's transfer needs
  - MAGIC
  - CTAO/LST (in progress)
  - ICFO (in progress)
  - InCaem (in progress)
  - LSST lite-IDAC (evaluating)

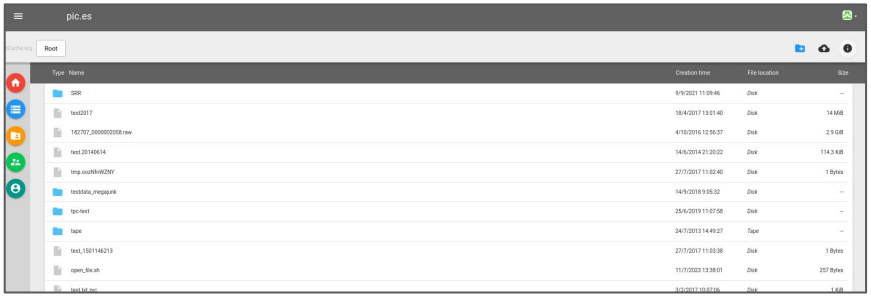
Storage (https access)

External access to mass storage (https)

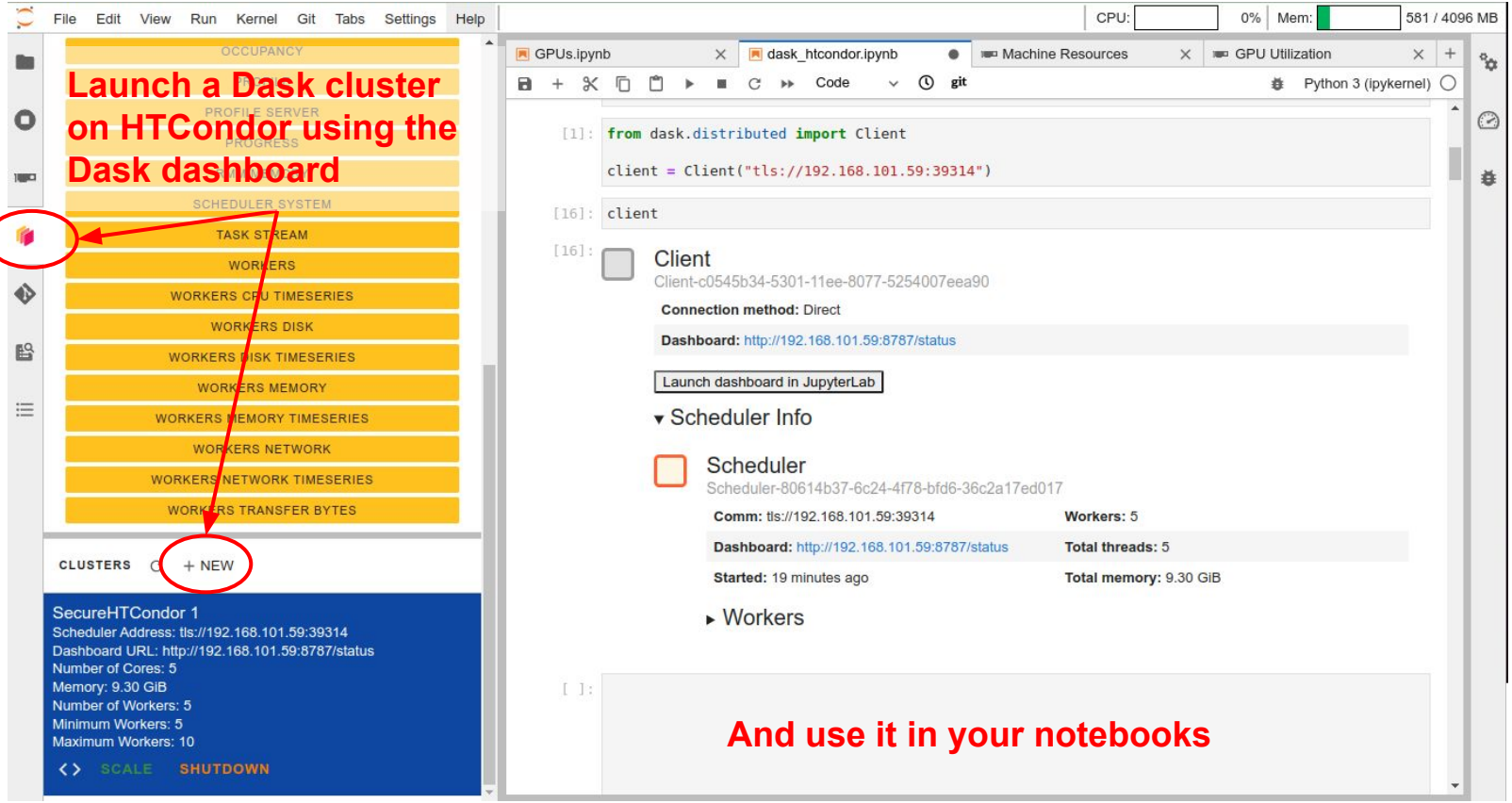
- Webdav door
  - http protocol to read / download files
  - Command line access (upload included)
  - Work with external tools (Rclone)



- Frontend dCacheView
  - User-friendly frontend to upload/download files
  - Share a URL with a temporal token to download files (up to 1 week)

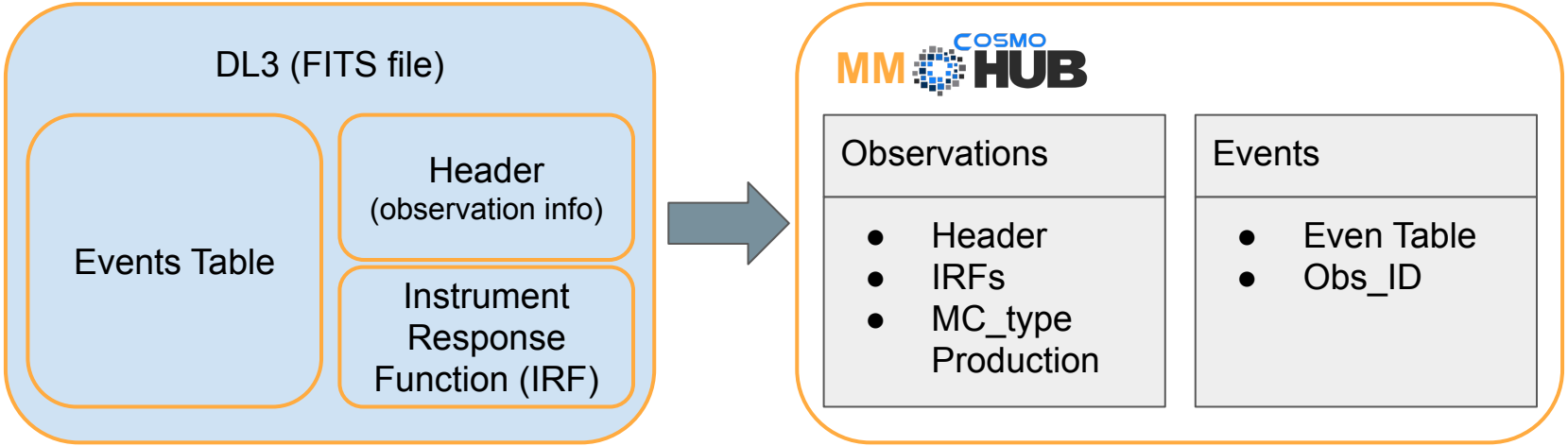


Jupyter: Dask



Gamma ray data model

- DL3 modeled as tables in Hive DB:



- Parallel processing framework
  - 3 compatible APIs
    - SQL
    - Dataframes
    - RDD
  - Interfaces with Hive/CosmoHub tables
  - Can also access massive storage (PNFS/Ceph/NFS)
  - Dual execution: notebook and batch

- Big Data algorithms:
  - SciPIC: virtual galaxy catalogs
  - Deepz: photometric redshift
  - SparkTreecorr (in development)

```
df = spark.sql("""
  SELECT id, ra, dec
  FROM cosmohub.micecatv1_0_hpix
  LIMIT 100
""")
df
```

```
DataFrame[id: int, ra: double, dec: double]
```

```
df.show(5)
```

```
+---------+---------+---------+
|       id|       ra|      dec|
+---------+---------+---------+
|191225057|18.523232|79.887398|
| 49810401|59.949303|20.816753|
|  9887201| 22.78075|46.971172|
| 11503841|51.193577| 17.27203|
| 43089377| 8.418952|16.733221|
+---------+---------+---------+
only showing top 5 rows
```