Integration of GammaHub in ESAP

WP5 Progress Meeting, 26-27th of October 2020

J. Delgado (jordidem@pic.es), C. Nigro, L. Jouvin, A. Bruzzese, J. Rico
P. Tallada, M. Delfino, G.Merino

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Outline

- Previously on GammaHub…
- Second Use Case: Data Products
- Integration in ESAP
- Future Plans

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Previously on GammaHub… What is GammaHub?

GammaHub is intended to be an interactive science analysis platform by itself, taking previous experiences from https://cosmohub.pic.es , a platform to manage Cosmological Catalogs developed at PIC

It will provide tools to modelate large datasets on a big data environment in order to search, explore and plot billions of objects interactively using a Hadoop stack of tools

We will ingest datasets from multi-instrument astronomical gamma-ray experiments like MAGIC, HESS, VERITAS and CTA among others, but it could be applied to other science disciplines

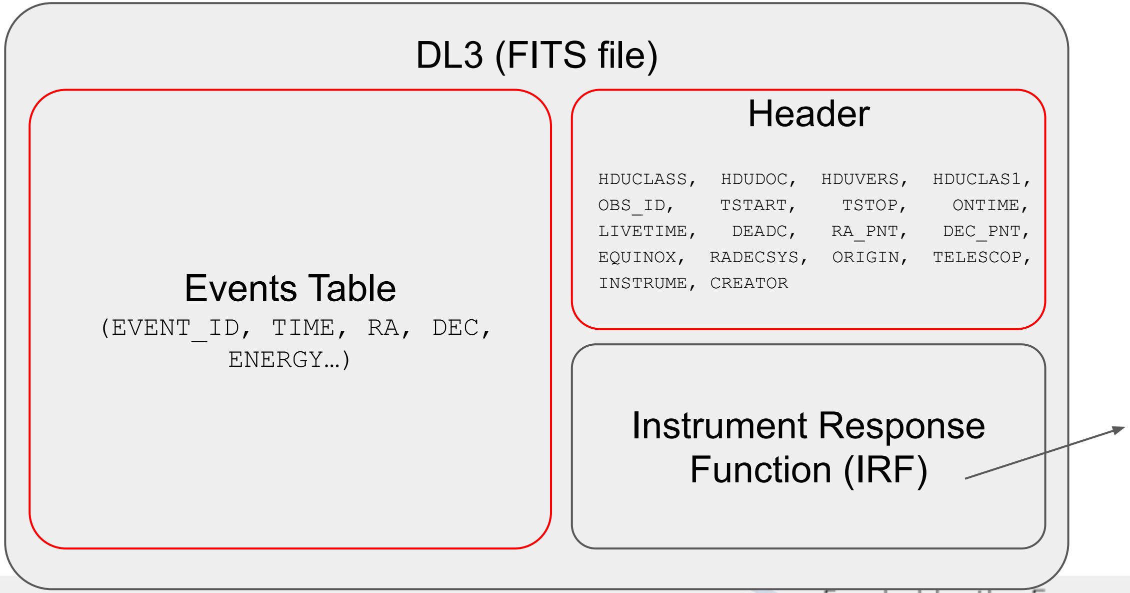
The main functionalities will be:

- Interactive data selection, exploration and plotting (scatter plots, histograms, heatmaps…)
- Automatic Data Products (Spectrum, Lightcurves…)
- Data access from Python notebooks for complex analysis

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Previously on GammaHub… The Data Model

Data Level 3 (DL3) v.0.2 (2018) contained in FITS files. The DL3 format is being promoted as a new standard in the context of building CTA



DL3 (FITS file)

Events Table

(EVENT_ID, TIME, RA, DEC, ENERGY…)

Header

HDUCLASS, HDUDOC, HDUVERS, HDUCLAS1, OBS_ID, TSTART, TSTOP, ONTIME, LIVETIME, DEADC, RA_PNT, DEC_PNT, EQUINOX, RADECSYS, ORIGIN, TELESCOP, INSTRUME, CREATOR

Instrument Response Function (IRF)

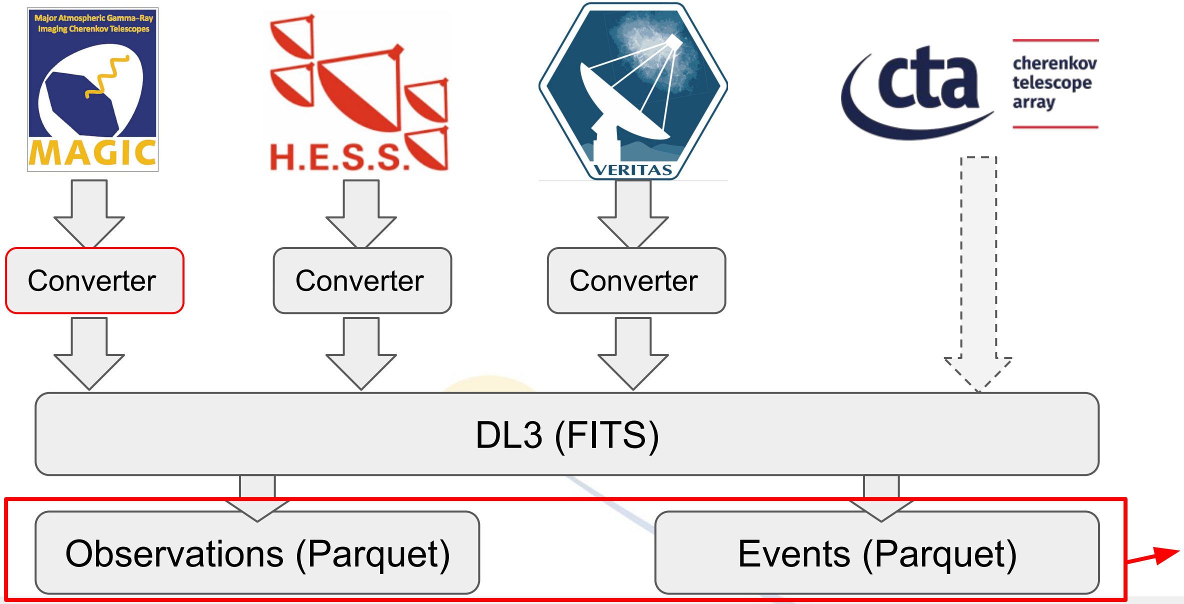
Data is modeled in the Hive DB

Not searchable, it will be accessed directly from disk

Funded by the European Union's Horizon 2020 - Grant N° 824064

Previously on GammaHub… The Data Model

Data Level 3 (DL3) v.0.2 (2018) contained in FITS files. The DL3 format is being promoted as a new standard in the context of building CTA


Major Atmospheric Gamma-Ray Imaging Cherenkov Telescopes
MAGIC
H.E.S.S.
VERITAS
cta
cherenkov telescope array

Converter

Converter

Converter

DL3 (FITS)

Observations (Parquet)

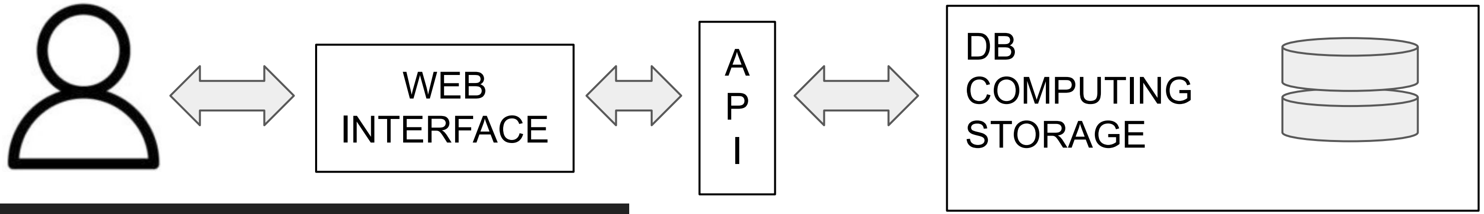
Events (Parquet)

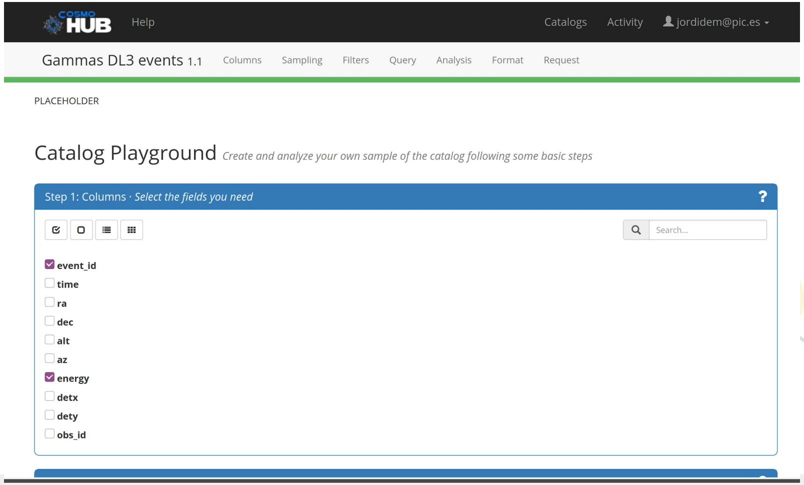
Parquet files will be ingested in the platform

Funded by the European Union's Horizon 2020 - Grant N° 824064

Previously on GammaHub… First use case

1. Web interface for interactive/on-line exploration and visualization of the selected datasets.





Gammas DL3 events 1.1

PLACEHOLDER

Catalog Playground *Create and analyze your own sample of the catalog following some basic steps*

Step 1: Columns · *Select the fields you need*

☑ event_id
□ time
□ ra
□ dec
□ alt
□ az
☑ energy
□ detx
□ dety
□ obs_id

✓ Data Searching (single and multi-instrument exploration)
✓ Data Plotting (Scatter, Histogram 1D, Heatmaps 2D…)
✓ Batch custom subsets

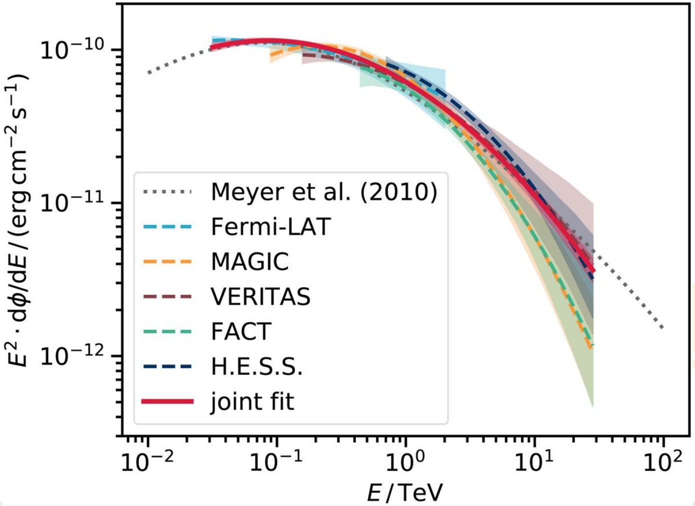
✓ User guidance / User friendly look and feel
✓ No SQL knowledge needed (but Expert mode included)

Funded by the European Union's
Horizon 2020 - Grant N° 824064

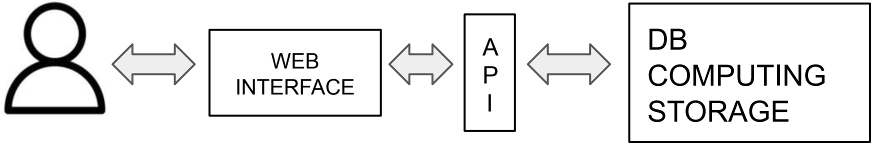
Second Use Case: Data Products

1. Data Products: implement interactive common data analysis workflows using Gamma-ray data (spectrum, lightcurves…)

Example: *Spectrum calculation*



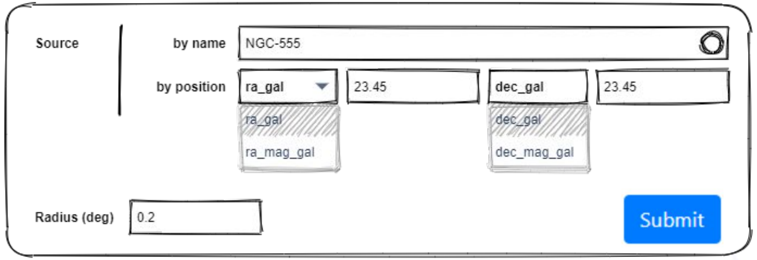
← Crab Nebula SED.
C.Nigro et Al (2019)



WEB INTERFACE

API

DB COMPUTING STORAGE



Source

by name NGC-555

by position ra_gal 23.45 dec_gal 23.45

ra_gal
ra_mag_gal

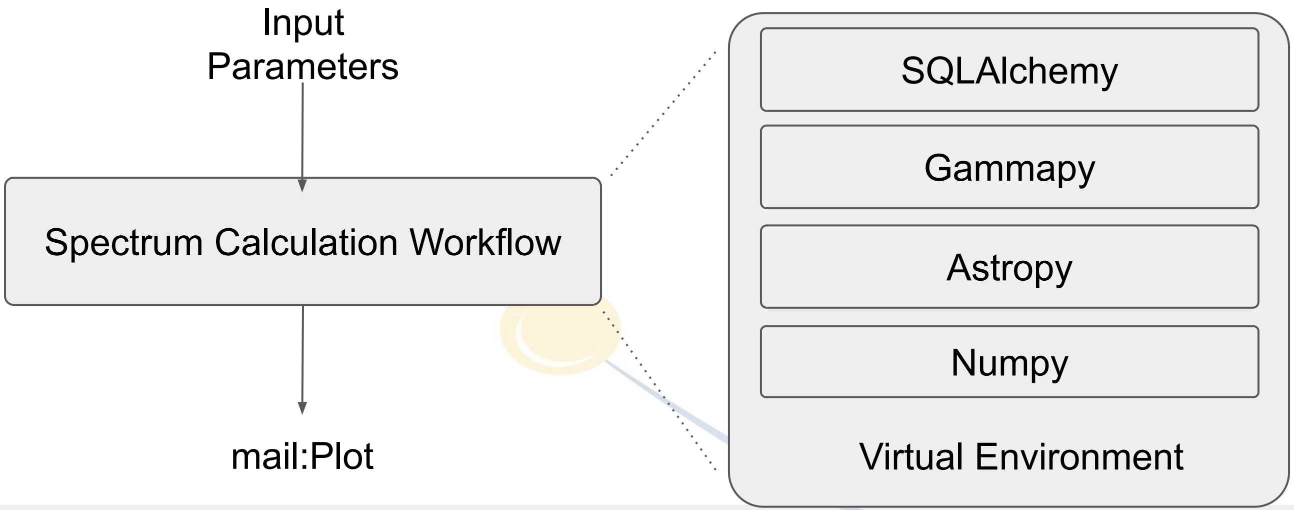
dec_gal
dec_mag_gal

Radius (deg) 0.2

Second Use Case: Data Products

1. Data Products: implement interactive common data analysis workflows using Gamma-ray data (spectrum, lightcurves…)

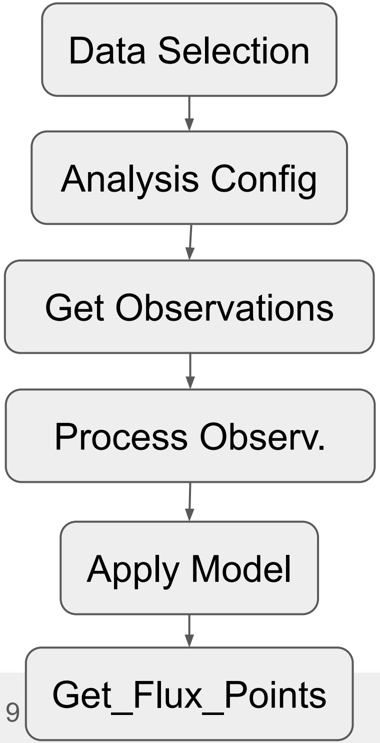
Example: *Spectrum calculation*



Funded by the European Union's Horizon 2020 - Grant N° 824064

Second Use Case: Data Products

Spectrum calculation workflow using Gammapy

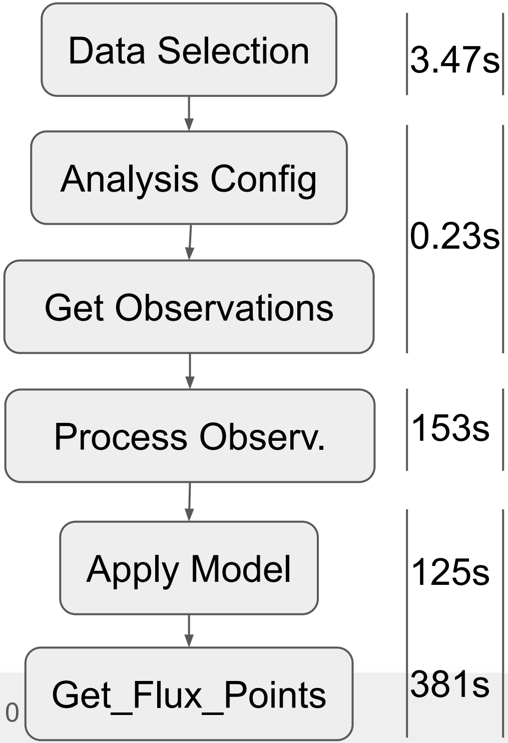


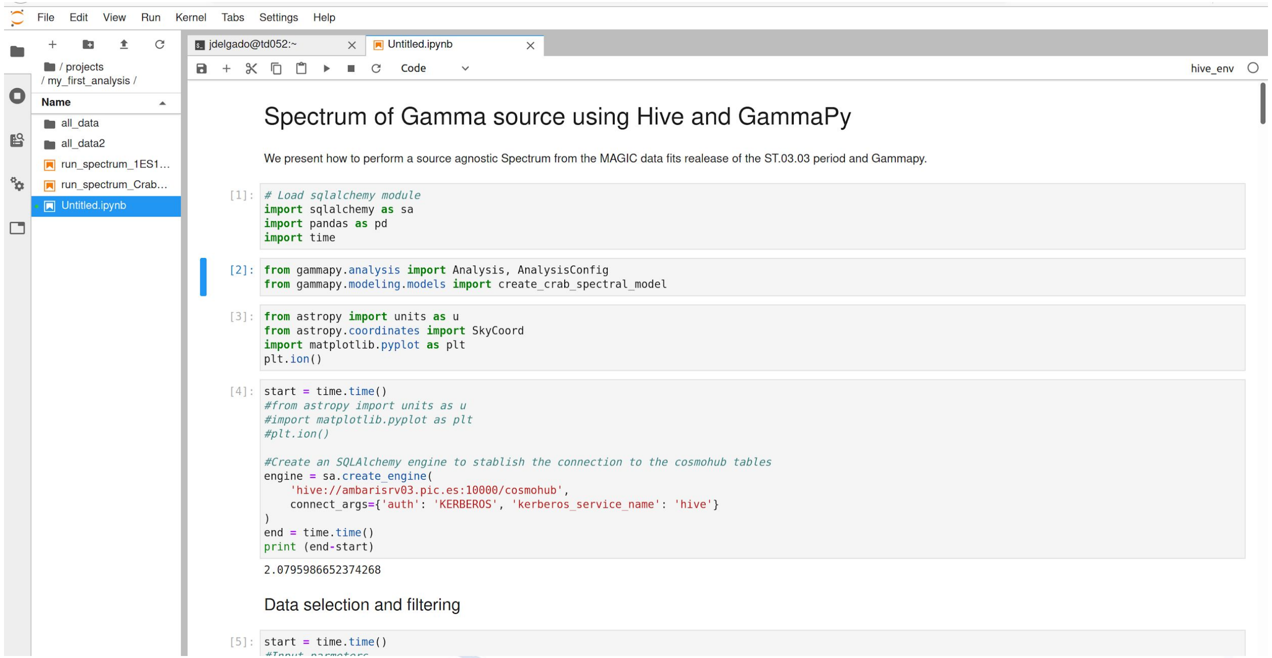
Data Selection	Interactive: HIVE DB - Cone Search	List of observations and events
Analysis Config	Backend: define analysis parameterization	
Get Observations	Backend: Stage FITS files	RUCIO data staging
Process Observ.	Batch: parallel processing (scatter - gather)	Spark?
Apply Model	Batch: processing	
Get_Flux_Points		

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Second Use Case: Data Products

Spectrum calculation workflow using Gammapy





Spectrum of Gamma source using Hive and GammaPy

We present how to perform a source agnostic Spectrum from the MAGIC data fits realease of the ST.03.03 period and Gammapy.

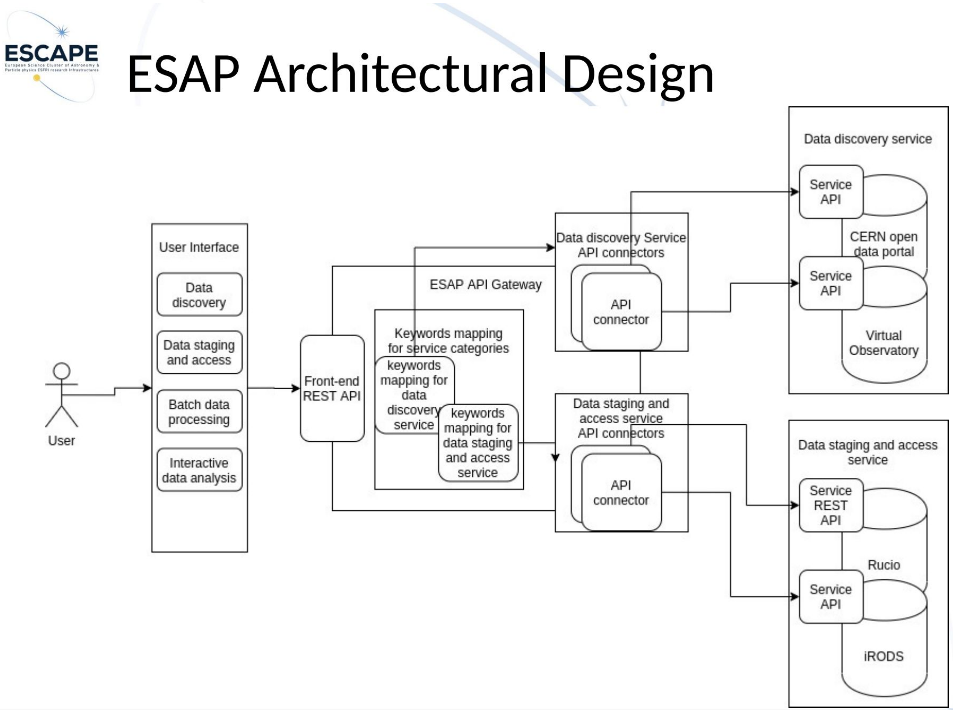
```
[1]: # Load sqlalchemy module
     import sqlalchemy as sa
     import pandas as pd
     import time

[2]: from gammapy.analysis import Analysis, AnalysisConfig
     from gammapy.modeling.models import create_crab_spectral_model

[3]: from astropy import units as u
     from astropy.coordinates import SkyCoord
     import matplotlib.pyplot as plt
     plt.ion()

[4]: start = time.time()
     #from astropy import units as u
     #import matplotlib.pyplot as plt
     #plt.ion()

     #Create an SQLAlchemy engine to stablish the connection to the cosmohub tables
     engine = sa.create_engine(
         'hive://ambarisrv03.pic.es:10000/cosmohub',
         connect_args={'auth': 'KERBEROS', 'kerberos_service_name': 'hive'}
     )
     end = time.time()
     print (end-start)

     2.0795986652374268
```

Data selection and filtering

```
[5]: start = time.time()
     #Input parameters
```

Funded by the European Union's Horizon 2020 - Grant N° 824064

Integration in ESAP



ESAP Architectural Design

User
User Interface
Data discovery
Data staging and access
Batch data processing
Interactive data analysis
Front-end REST API
ESAP API Gateway
Keywords mapping for service categories
keywords mapping for data discovery service
keywords mapping for data staging and access service
Data discovery Service API connectors
API connector
Data staging and access service API connectors
API connector
Data discovery service
Service API
CERN open data portal
Service API
Virtual Observatory
Data staging and access service
Service REST API
Rucio
Service API
iRODS

Open questions to solve:

- Data selection:
  - is HIVE/HADOOP compatible with ESAP?
  - How to use IVOA tools to publish and get data?
- Data staging:
  - We will upload DL3 files in the Data Lake
  - Install Rucio Client (Rob’s demo)
- Computing: how we manage the parallelism?

Funded by the European Union’s Horizon 2020 - Grant N° 824064

Future plans

Next steps in the developing of GammaHub

- Continue the data ingestion from other instruments, other MAGIC periods
- Upload DL3 data in the Data Lake and test the Rucio Client to stage and read data
- Implement parallelism using Spark (from jupyter notebooks to standalone)
  - Perform a multi-instrument spectral analysis of the Crab Nebula with Gammapy (reproduce C.Nigro et al *Towards open and reproducible multi-instrument analysis in gamma-ray astronomy*)
- Open to contribute to the resource federation, workflows publication...
- Implement a Data Product to calculate Light Curves

Funded by the European Union's Horizon 2020 - Grant N° 824064

Thank you!

WP5 Progress Meeting, 27th of October 200

J. Delgado (jordidem@pic.es), C. Nigro, L. Jouvin, A. Bruzzese, J. Rico,
P. Tallada, M. Delfino, G.Merino

Funded by the European Union's
Horizon 2020 - Grant N° 824064

GammaHub components: Infrastructure

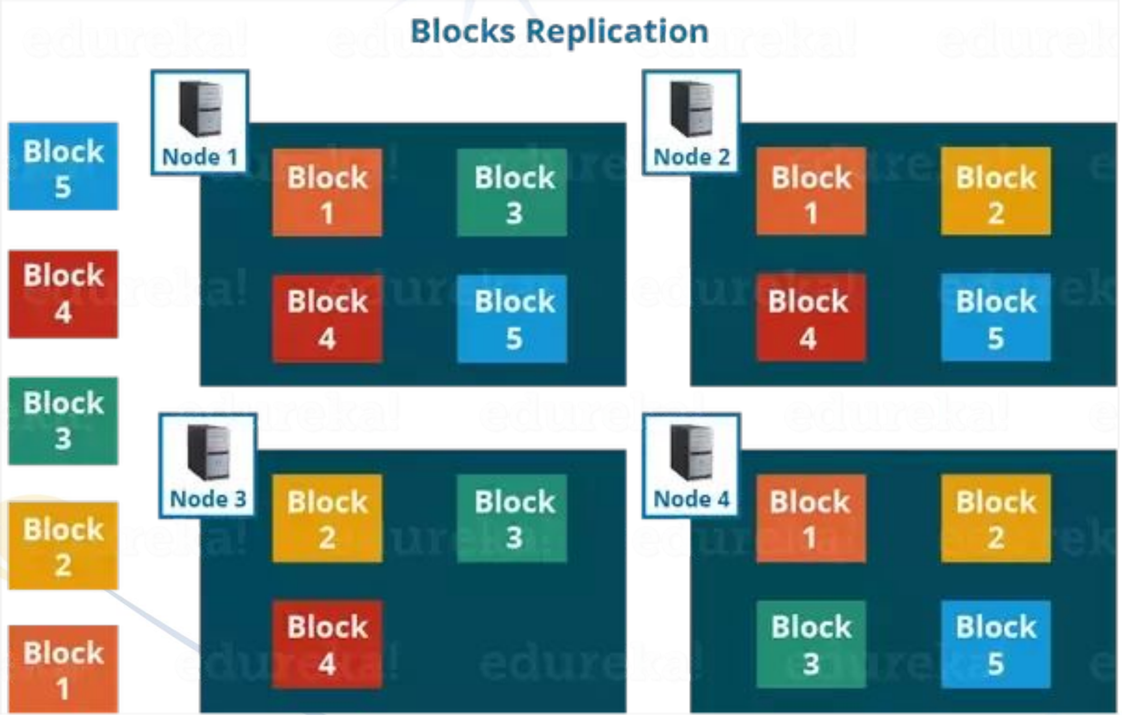
Data replication on HDFS:

- File is divided in blocks of data
- Default replication factor x3

File of 128MB -> 384MB stored in the platform

- Hadoop nodes includes access

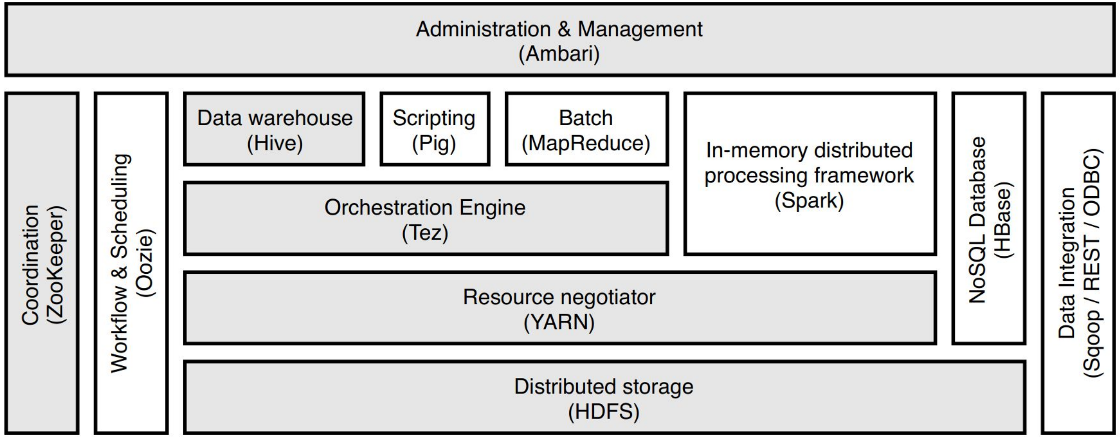
- Rack awareness algorithm, to distribute data on different data nodes to improve network performance and fault tolerance



GammaHub components: Infrastructure

Apache Hadoop:

- Distributed processing
  - Map-reduce model
- Dedicated computing cluster
  - Each node is a computing and data node
- Easy to scale from single server to thousand machines, each one offering local computation and storage
- Fault tolerance



Administration & Management (Ambari)

Coordination (ZooKeeper)

Workflow & Scheduling (Oozie)

Data warehouse (Hive)

Scripting (Pig)

Batch (MapReduce)

In-memory distributed processing framework (Spark)

Orchestration Engine (Tez)

NoSQL Database (HBase)

Data Integration (Sqoop / REST / ODBC)

Resource negotiator (YARN)

Distributed storage (HDFS)

Hadoop Layered ecosystem applied to CosmoHub (P.Tallada et. al. 2020 arxiv)

Funded by the European Union's Horizon 2020 - Grant N° 824064