Workflows d'anàlisi de raigs gamma basats en CosmoHub

Irene Castrillo Martínez

Resum—Aquest Treball de Fi de Grau constitueix una contribució a l'anàlisi d'observacions de raigs gamma recollides en el marc del projecte MAGIC i gestionades pel Port d'Informació Científica (PIC), amb l'objectiu de facilitar l'estudi de dades cosmològiques. El treball presenta el desenvolupament i l'optimització dels workflows SkyMap, SED i LightCurve, processos d'anàlisi habituals emprats per la comunitat científica en aquest àmbit. Amb l'objectiu d'agilitzar i simplificar el tractament de dades, aquests workflows s'han integrat dins la plataforma CosmoHub, establint una connexió directa amb la infraestructura existent i millorant-ne el rendiment mitjançant la paral·lelització amb Apache Spark. El document ofereix la contextualització necessària i descriu l'entorn informàtic que envolta la gestió d'aquestes dades, amb la finalitat de facilitar la comprensió del funcionament i la implementació del codi desenvolupat. Finalment, s'hi inclou una anàlisi comparativa del rendiment entre les versions seqüencial i paral·lela dels processos, amb l'objectiu d'avaluar els beneficis de l'execució distribuïda.

Paraules clau—raigs gamma, Gammapy, MAGIC, PIC, CosmoHub, Skymap, corba de llum, espectre de distribució d'energia (SED), Apache Spark i paral·lelització.

Abstract—This Bachelor's Thesis constitutes a contribution to the analysis of gamma-ray observations collected within the framework of the MAGIC project and managed by the Port d'Informació Científica (PIC), with the aim of facilitating the study of cosmological data. The work presents the development and optimization of the SkyMap, SED, and LightCurve workflows, which are commonly used analytical processes in this scientific domain. In order to accelerate and facilitate data processing, these workflows have been integrated into the CosmoHub platform, establishing a direct connection with the existing infrastructure and improving performance through parallelization with Apache Spark. The document provides the necessary contextual background and describes the computing environment involved in the management of these data, in order to support understanding of the implemented code and its functionality. Finally, it includes a comparative performance analysis between the sequential and parallel versions of the processes, with the aim of evaluating the benefits of distributed execution.

Index Terms—gamma rays, Gammapy, MAGIC, PIC, CosmoHub, Skymap, light curve, spectral energy distribution (SED), Apache Spark and parallelization.

1 Introducció

L'ESTUDI dels raigs gamma representa una de les àrees més fascinants i exigents de l'astrofísica moderna. Aquest tipus de radiació electromagnètica permet explorar fenòmens astrofísics de gran violència, com les explosions de supernoves, la fusió d'estrelles de neutrons o l'activitat de forats negres supermassius. No obstant això, l'anàlisi d'aquestes dades implica un alt grau de complexitat, tant per la seva naturalesa tècnica com pel gran volum d'informació generada. Per aquest motiu, s'han desenvolupat infraestructures de computació i eines específiques per facilitar-ne el tractament eficient i escalable.

En aquest context, el projecte MAGIC (Major Atmospheric Gamma Imaging Cherenkov Telescopes) és una infraestructura robusta per a la captació de la radiació Cherenkov produïda per l'impacte de rajos gamma a l'atmosfera. Amb més de 20 anys de trajectòria, ha detectat més de 200 fonts de rajos gamma i compta amb el seu lloc d'observació a l'Observatori del Roque de los Muchachos, a l'illa de La Palma, a les Illes Canàries. Dins d'aquesta col·laboració internacional, el Port d'Informació Científica (PIC) és l'únic centre de dades del projecte, encarregat de la transferència, la distribució, el reprocessament i el suport a l'anàlisi de dades.

En el marc d'una innovació constant per oferir noves i millors eines als usuaris que utilitzen el PIC per fer ciència, el centre ha desenvolupat CosmoHub, una plataforma que ofereix una interfície web per a la consulta d'aquestes, facilitant-ne l'explotació científica per part de la comunitat investigadora.

- *E-mail de contacte: 1637742@uab.cat*
- *Treball tutoritzat per: Eduardo Cesar Galobardes (Àrea d'Arquitectura i de Tecnologia de Computadors)*
- *Curs 2024/25*

Aquest treball de fi de grau s'insereix dins aquesta línia d'innovació, amb l'objectiu d'aportar millores en el rendiment i l'escalabilitat dels workflows d'anàlisi de dades de raigs gamma mitjançant tècniques que permetin la combinació de dades massives d'un o més instruments, amb l'ús de tècniques de paral·lelització com és Apache Spark. Concretament, el projecte s'ha centrat en tres fluxos de treball: la generació de mapes celestes (Skymap), l'obtenció de la distribució espectral d'energia (Spectral Energy Distribution, SED) i el càlcul de la corba de llum (LightCurve).

L'objectiu principal ha estat implementar i optimitzar aquests workflows dins l'entorn del PIC, aprofitant tant la infraestructura de computació com la base de dades en Hive que alimenta CosmoHub. El projecte persegueix tres objectius principals: comprendre l'entorn i les eines implicades, implementar i millorar els workflows existents, integrant-los a CosmoHub, i dur a terme una comparativa de rendiment entre les implementacions basades en Dask i Spark.

Aquesta tasca suposa la continuació i ampliació del treball previ «Paral·lelització de workflows d'anàlisi de dades d'astronomia de raigs gamma», en el qual es van implementar aquestes funcions mitjançant Dask. El present projecte pretén anar un pas més enllà, aprofitant les capacitats de Spark i millorant la integració amb CosmoHub, de manera que es pugui oferir noves eines i aplicacions a través de la plataforma.

Cal destacar que, durant el desenvolupament, han sorgit diversos reptes tècnics, com la configuració del clúster Spark dins del PIC i la compatibilitat entre els formats emmagatzemats al catàleg SQL i les classes adients de GammaPy. Aquestes dificultats, tot i haver-les solucionat amb èxit, han provocat que ens quedéssim sense temps per realitzar l'últim objectiu, una comparativa amb la paral·lelització realitzada anteriorment amb Dask.

Així doncs, es busca no només optimitzar el rendiment dels workflows, sinó també assentar les bases per a una millor integració entre tecnologies de còmput distribuït i eines d'anàlisi científica, contribuint a fer més àgil, accessible i eficient la investigació en l'àmbit de l'astrofísica i facilitar la feina als analistes.

El treball s'estructura en una part teòrica i una part pràctica. La Secció 2 presenta el marc contextual del projecte, incloent-hi la tecnologia d'observació de raigs gamma, el Port d'Informació Científica (PIC) i la plataforma CosmoHub. La Secció 3 aborda el tractament de les dades, tot descrivint els formats emprats i els processos de filtratge aplicats. La Secció 4 detalla els workflows implementats —SkyMap, SED i Light Curve. La Secció 5 analitza els resultats obtinguts, comparant l'execució seqüencial amb l'execució paral·lela mitjançant Spark. Finalment, la Secció 6 recull les conclusions i planteja possibles línies futures de treball.

2 Marc contextual

Per entendre les motivacions, eines i metodologies d'aquest projecte, és essencial situar-lo dins el seu context científic i tecnològic. En aquesta secció s'ofereix una visió general de l'escenari informàtic implicat en l'estudi dels raigs gamma, així com de les infraestructures computacionals que en permeten el processament.

2.1 Tecnologia Raigs Gamma

Un raig gamma és una forma de radiació electromagnètica amb una energia extremadament alta, sent la més intensa dins de l'espectre electromagnètic. Aquesta radiació es genera a través de processos astrofísics i nuclears d'alta energia, com les explosions de supernoves i les fusions d'estrelles de neutrons. Quan els raigs gamma i les partícules que les componen arriben a l'atmosfera terrestre, es produeix la radiació Cherenkov, un fenomen que pot ser detectat mitjançant telescopis especialitzats, equipats amb càmeres fotosensibles que capten aquesta radiació[2].

La captació i anàlisi d'aquests raigs són essencials per a l'astrofísica, ja que permeten l'estudi de fenòmens astrofísics extremadament violents. Per tal de dur a terme aquesta tasca, s'han desenvolupat diversos instruments com els telescopis terrestres MAGIC[3] (Major Atmospheric Gamma Imaging Cherenkov Telescopes), HESS[4] (High Energy Stereoscopic System) i VERITAS[5] (Very Energetic Radiation Imaging Telescope Array System), o el satèl·lit FERMI de la Nasa, els quals se centren en la detecció d'aquesta radiació. Actualment, MAGIC continua operatiu, mentre que el projecte CTAO[6] (Cherenkov Telescope Array Observatory) està desenvolupant una nova generació d'instruments amb una sensibilitat millorada que substituirà els actuals MAGIC, HESS i VERITAS.

Aquesta situació genera un conjunt de dades antigues, actuals i futures, amb una resolució cada cop més alta. Per facilitar la comparació i la interoperabilitat de les dades entre aquests diferents conjunts de dades, s'ha creat el format DL3 (Data Level 3)[7], amb l'objectiu d'estandarditzar la informació i permetre'n l'anàlisi de manera més eficaç i coherent. Aquestes dades es comparteixen amb la comunitat científica de manera oberta i estandarditzada mitjançant el format VODF[8] (Very-High-Energy Open Data Format).

A més, l'anàlisi de dades DL3 es pot realitzar mitjançant Gammapy[9], una biblioteca de codi obert en Python dissenyada per optimitzar l'anàlisi i el processament de dades de raigs gamma. Entre les eines que ofereix, destaquen opcions de visualització com els Skymaps, que permeten una representació gràfica detallada i eficient de les dades observacionals. També s'inclouen les corbes de llum (light curves) i l'anàlisi de l'espectre energètic (Spectral Energy Distribution, SED),

que permeten caracteritzar els esdeveniments observats a partir de les dades generades pels telescopis.

Dins d'aquesta gestió informàtica, col·labora el PIC[10] actuant com a infraestructura central per al processament i l'anàlisi de les dades obtingudes pel telescopi MAGIC i també dels telescopis CTAO.

Certs nivells de dades recollides, com les dades DL3, es poden convertir i explotar mitjançant CosmoHub[11], una plataforma dissenyada específicament per a l'emmagatzematge i la gestió de grans volums de dades en cosmologia. Aquesta està dirigida a la comunitat científica, facilitant el processament i l'anàlisi de dades observacionals, simulades o sintètiques a gran escala.

Un dels objectius de l'equip investigador del PIC és expandir l'ús d'aquesta plataforma a altres camps científics, com oferir eines costoses de processar per l'anàlisis de raigs gamma, entre d'altres. Per això, es treballa en el desenvolupament de models de dades i eines de processament integrades amb Gammapy.

Així doncs, aquest projecte constitueix una petita contribució, en col·laboració amb el PIC, a la millora del rendiment de les funcions de visualització de raigs gamma. Es considera una continuïtat del TFG titulat *"Paral·lelització de Workflows d'Anàlisi de Dades d'Astronomia de Raigs Gamma"*[14], realitzat per Yasmin L'Harrak, on es paral·lelitzaven les funcions utilitzant Dask. En aquest cas, s'explora la possibilitat de millorar-les mitjançant l'ús de Spark[12] i incorporar-les dins de CosmoHub.

De manera similar a Dask, Apache Spark[13] és un motor de dades distribuït que permet processar grans volums d'informació de manera eficient, repartint les tasques entre múltiples nodes d'un clúster. Ofereix el patró MapReduce de Hadoop per paral·lelitzar l'execució de codi. A més, és plenament compatible amb la infraestructura de CosmoHub, a diferència de Dask, la qual cosa suposa una millora substancial, ja que aquesta plataforma opera sobre un clúster dedicat i homogeni, en contrast amb la infraestructura genèrica del PIC utilitzada anteriorment. Aquesta compatibilitat evita els problemes derivats de la heterogeneïtat i de les limitacions de xarxa detectats durant les proves amb Dask, i permet obtenir resultats més fiables i comparables.

2.2 PIC

El Port d'Informació Científica (PIC)[10], és un consorci entre el Centre d'Investigacions Energètiques, Mediambientals i Tecnològiques (CIEMAT) i l'Institut de Física d'Altes Energies (IFAE), és un centre de dades i computació d'altes prestacions especialitzat en ciència de grans volums. La seva missió és integrar, emmagatzemar i distribuir petabytes d'informació procedents de col·laboracions internacionals, tot oferint serveis robustos i segurs a la comunitat científica, principalment en els àmbits de la física de partícules, l'astrofísica i la cosmologia, la física de materials, entre d'altres. Paral·lelament, el PIC impulsa la seva pròpia recerca i el desenvolupament d'eines i plataformes Big Data que maximitzen l'eficiència del processament i de l'anàlisi, adoptant aquestes a arquitectures distribuïdes i optimitzant fluxos de treball amb paral·lelització massiva del codi. D'aquesta manera, redueix les barreres tècniques de l'accés a dades a gran escala i accelera el cicle de descoberta científica.

Per les característiques d'aquest centre de dades, s'adopta una filosofia de paral·lelització orientada al *throughput*, fet que es té en compte en el desenvolupament d'aquest treball. El *throughput* mesura el rendiment global d'un sistema de computació, la quantitat de treball que pot completar per unitat de temps. A diferència de la latència, que indica el temps que necessita cada tasca individual, el *throughput* prioritza maximitzar la producció sostinguda del sistema. L'objectiu és mantenir totes les unitats de càlcul ocupades de manera contínua, encara que això comporti un augment lleu del temps de resposta per tasca.

2.3 CosmoHub

CosmoHub[12] és una plataforma web d'exploració interactiva i de distribució centralitzada de catàlegs cosmològics de gran escala, desenvolupada pel PIC i suportada sobre l'arquitectura de còmput distribuït Apache Hadoop/Hive. En la seva versió actual allotja més de 60 TiB d'informació catalogada, prop de 5×10^{10} objectes astronòmics, procedent de col·laboracions com Euclid, DES, PAUS, MICE i observacions de raigs gamma com MAGIC, entre d'altres.

La plataforma[11] proporciona eines avançades que permeten als investigadors consultar, filtrar i descarregar subconjunts de dades de manera ràpida i intuïtiva, simplificant així la complexitat tècnica associada al tractament de dades massives. Per a dur a terme tota aquesta transformació, la interfície web s'executa sobre una arquitectura dedicada, la qual es descriu més endavant. Actualment, els usuaris poden adquirir les dades en diversos formats, com taules, histogrames, diagrames de dispersió (scatter) i mapes de calor (heatmaps) o també en dataframes a través de Jupyter notebooks.

En la primera capa de la infraestructura es troba CosmoHub[14], la plataforma web que actua com a interfície entre l'usuari i el sistema. Des d'aquesta interfície, l'usuari pot seleccionar els conjunts de dades d'interès, així com definir els criteris que aquestes han de complir. A partir d'aquesta selecció, es genera una sol·licitud que es transmet a una segona capa del sistema, on es troba l'API de CosmoHub, que en facilita la integració, així com quaderns d'execució (notebooks).

Per tal de permetre aquesta funcionalitat, s'ha implementat una API específica que recull els paràmetres de filtratge definits per l'usuari i s'encarrega de gestionar la sol·licitud, retornant el resultat corresponent un cop finalitzat el processament.

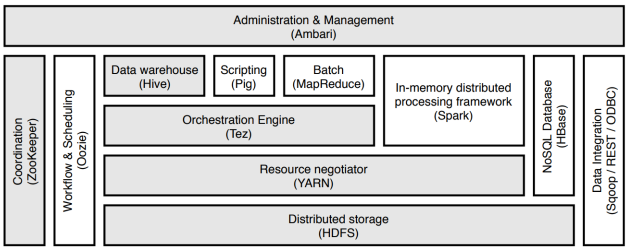


Figura 1[1]: Il·lustració de les capes d'infraestructura que sustenten CosmoHub

En el marc d'aquest sistema, il·lustrat a la Figura 1, la lectura de les dades es realitza mitjançant una consulta SQL generada dinàmicament. Aquesta és interpretada per Apache Hive[20] (*High-level Interface for Virtual Execution*), que actua com a motor de base de dades SQL-like. Hive permet realitzar consultes sobre conjunts de dades massives mitjançant una sintaxi SQL familiar, però amb l'avantatge que aquestes consultes es poden executar en paral·lel sobre múltiples nodes dedicats.

Seguidament, el gestor YARN[21] (*Yet Another Resource Negotiator*) entra en acció com l'orquestrador principal de la plataforma. És responsable de coordinar l'execució de tots els processos que es llancen dins l'ecosistema Hadoop, distribuint els recursos computacionals entre aplicacions i assegurant que cada tasca tingui accés als recursos necessaris.

Les dades sobre les quals es fan les consultes es troben emmagatzemades en el sistema HDFS[22] (*Hadoop Distributed File System*), un sistema de fitxers distribuït. HDFS reparteix automàticament les dades entre diferents nodes de l'arquitectura, creant còpies redundants per garantir la tolerància a fallades i l'alta disponibilitat. Cada node del clúster conté tant capacitat d'emmagatzematge (disc) com de processament (CPU), la qual cosa permet accedir i processar les dades de forma paral·lela i escalable. Aquest enfocament assegura un rendiment òptim fins i tot amb volums massius d'informació.

Els fitxers utilitzats en aquest procés es troben emmagatzemats en format Parquet. Un dels avantatges principals d'aquest format és la seva capacitat d'incorporar metadades, les quals contenen informació sobre els valors dels camps que s'utilitzen en el filtratge. Aquesta característica permet realitzar consultes eficients, ja que no cal llegir tot el contingut del fitxer per determinar si conté la informació rellevant per a la informació sol·licitada.

Aquesta infraestructura també és accessible des de l'entorn del PIC mitjançant notebooks Jupyter i la funcionalitat enableHiveSupport com a interfície de connexió amb el metastore que conté les taules dels catàlegs de CosmoHub. Les dades es consulten amb sentències SQL, tècnica que s'ha implementat en el codi del treball.

3 Tractament de dades

Per tal de comprendre el funcionament del codi desenvolupat en aquest projecte, és necessari entendre els processos generals de tractament de dades que s'apliquen a les observacions de raigs gamma. Aquesta secció descriu com s'organitzen i es transformen les dades al llarg del flux d'anàlisi, així com els identificadors i mecanismes de filtratge emprats per seleccionar i estructurar la informació.

3.1 Formats i transformacions

En la cadena de reducció dels telescopis CTAO, el nivell de dades 3 (DL3)[7] marca el primer punt en què la informació deixa de dependre del programari intern de l'instrument i esdevé reutilitzable per qualsevol investigador. Per aconseguir-ho, cada observació s'empaqueta en un o diversos fitxers FITS que segueixen una convenció oberta mantinguda pel projecte gamma-astro-data-formats. Dins d'aquests FITS hi ha tres taules essencials, els múltiples esdeveniments reconstruïts (temps, energia, coordenades, etc), les funcions de resposta i caracterització de l'instrument (àrea efectiva, dispersió d'energia, etc), i una tercera que actua com a índex, amb les metadades que descriuen tota l'observació i que serveixen per identificar i relacionar les dues taules anteriors.

El pas de DL3 a DL4 consisteix una agrupació els esdeveniments bruts en productes científics concrets. Es llegeixen els FITS DL3, s'apliquen talls de qualitat, se selecciona un rang d'energies o un període temporal i es projecten les dades sobre clases especifiques de Gammapy espacials i d'energia. El resultat són mapes de comptes, exposicions, fons i PSF combinats, d'entre altres.

Finalment, quan s'agrupen molts productes DL4 provinents d'observacions diferents o, fins i tot, d'instruments diversos, es construeixen els DL5, catàlegs homogenis de fonts. Aquests catàlegs, que integren informació de diferents origens, permeten no només una visió global, sinó també la seva comparació sistemàtica amb models teòrics, facilitant així l'estudi de la seva naturalesa i evolució.

En conjunt, les transformacions de DL3 a DL5[16], il·lustrat a la figura 2, constitueixen tres graus d'especificació. Primer, es garanteix la preservació a llarg termini de cada observació (DL3), després, es generen els productes específics necessaris per abordar preguntes

[1] Figura 2 del article *CosmoHub: Interactive Exploration and Distribution of Astronomical Data on Hadoop*[19]

científiques concretes (DL4) i finalment, el coneixement s'integra en bases de dades estructurades que poden ser explotades per altres investigadors (DL5). Així, la comunitat científica obté una major interoperabilitat, cohesió i un accés molt més àgil a l'anàlisi de l'astronomia de raigs gamma, ja que aquest patró es segueix en cada modelització de dades.

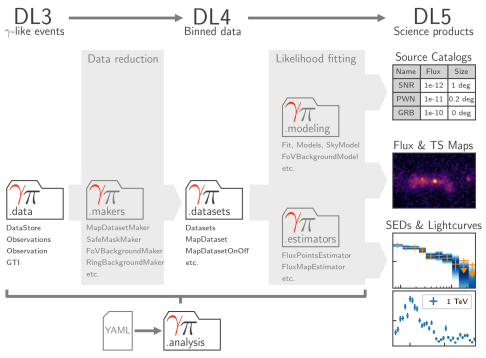


Figura 2[2]: Flux general de transformacions amb observacions

3.2 Filtratge

Per filtrar i ubicar les observacions a CosmoHub s'utilitza el mecanisme Cone Search[18], el mateix principi adoptat per la llibreria Gammapy. El procediment consisteix a fixar una posició al cel mitjançant coordenades equatorials, Right Ascension (RA) i Declination (Dec), equivalents a longitud i latitud, i especificar un radi al voltant d'aquesta posició. Així doncs, es seleccionen tots els esdeveniments observats dins del con definit. L'ús d'aquest sistema de referència facilita l'anàlisi focalitzada de regions concretes del firmament, fet que millora significativament la detecció i caracterització de fonts d'energia. Alhora, esdevé una eina clau per a la modelització precisa dels fenòmens astrofísics detectats.

La fórmula general del Cone Search on es defineix el diametre del con; on α és l'ascensió recta (RA), δ la declinació (Dec) i θ la separació angular entre el dos punts; és la següent:

$$\cos(\theta) = \sin(\delta_1)\sin(\delta_2) + \cos(\delta_1)\cos(\delta_2)\cos(\alpha_1 - \alpha_2)$$

Per tant, a partir d'aquesta fórmula, s'adapta l'expressió per identificar els punts que es troben dins d'aquest con:

$$\cos^{-1}(\sin(\delta_1)\cdot\sin(\delta_2)+\cos(\delta_1)\cdot\cos(\delta_2)\cdot\cos(|\alpha_1-\alpha_2|))\leq\theta$$

Aquesta inequació és la que s'aplica a les consultes SQL executades a CosmoHub, garantint una implementació òptima, ja que les dades de RA i Dec de cada observació es troben emmagatzemades com a metadades dels fitxers Parquet, evitant així la necessitat de llegir tot l'arxiu per descartar o validar si es troba dins del rang d'anàlisi desitjat.

4 Workflows implementats

En aquesta secció es presenten els workflows desenvolupats durant el projecte, els quals s'han escollit perquè són anàlisis generals i habituals entre la comunitat d'astrofísics. A continuació, es detallen les característiques principals de cadascun d'aquests workflows, emprant com a exemple la font astronòmica Crab Nebula.

4.1 SkyMap

Existeixen diverses metodologies per modelitzar i representar la informació obtinguda a partir de les observacions astronòmiques. Una de les més rellevants és la visualització espacial mitjançant la representació de la distribució de l'energia dels raigs gamma al cel, coneguda com a SkyMap. En el cas d'aquesta implementació, aquests SkyMaps es generen emprant el tipus de mapa WcsNDMap[15], el qual es base en el sistema de coordenades WCS (World Coordinate System).

El dataflow implementat realitza un SkyMap senzill, requereix que s'especifiqui la posició i el radi a visualitzar. Un exemple de sortida d'aquest workflow és el que s'il·lustra a la figura 3, on s'agafen les dades específiques de la font astronòmica Crab Nebula (lon 83.633, lat 22.014 i radi 1.5).

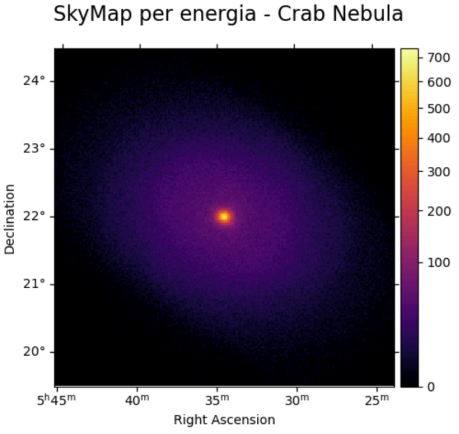


Figura 3: SkyMap del Crab Nebula

4.2 SED

La Distribució d'Energia Espectral (Spectral Energy Distribution, SED) és una representació gràfica que descriu com es distribueix el flux d'energia emès per una font astronòmica en funció de l'energia dels fotons. Aquesta eina és essencial per a l'anàlisi de les propietats físiques d'objectes d'alta energia, com ara púlsars, blàzars o restes de supernova.

2 Il·lustració dels diferents formats i transformacions d'observacions.

Font: https://docs.gammapy.org/1.3/user-guide/package.html

La implementació del SED del codi d'aquest projecte és mitjançant la magnitud E²·dN/dE en funció de l'energia. La representació, una gràfica com la de la figura 4, destaca la contribució de cada interval d'energia al flux total d'energia emesa per la font.

D'aquesta manera, el pic de la corba obtinguda permet identificar el rang energètic en què la font emet amb màxima potència, afavorint així la comparació entre diferents fonts o la validació de models teòrics.

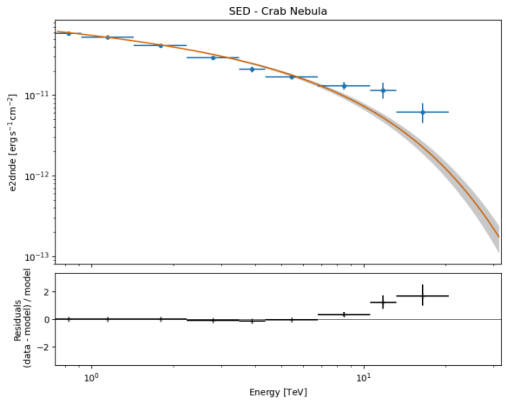


Figura 4: SED del Crab Nebula

4.3 Light Curve

La corba de llum (light curve), il·lustrada a la figura 5, és una representació gràfica que mostra la variació temporal del flux d'una font astronòmica, i permet analitzar com evoluciona l'emissió d'energia d'un objecte en funció del temps.

Aquestes corbes es construeixen a partir del nombre de fotons detectats en diferents intervals temporals, habitualment expressats com a flux, per exemple en unitats de fotons·cm⁻²·s⁻¹. La seva anàlisi permet identificar patrons de variabilitat, episodis d'augment sobtat d'emissió, així com períodes d'activitat o inactivitat, entre altres característiques rellevants per a la comprensió del comportament de la font.

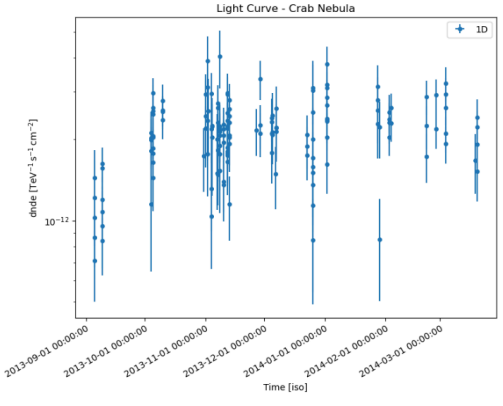


Figura 5: Light Curve del Crab Nebula

5 Resultats

En aquesta secció es presenten els resultats obtinguts de l'anàlisi de rendiment dels workflows implementats tant en versió seqüencial com en la versió paral·lelitzada integrada a CosmoHub, utilitzant Spark i el model de programació MapReduce per distribuir les tasques i optimitzar l'eficiència del processament. La base del codi seqüencial es fonamenta en el tutorial oficial de Maps de Gammapy[9] per al cas de SkyMap, mentre que per als workflows de SED i LightCurve s'han utilitzat les implementacions en sèrie desenvolupades prèviament per Yasmin en el seu treball de fi de grau, *"Paral·lelització de Workflows d'Anàlisi de Dades d'Astronomia de Raigs Gamma"*[14].

5.1 Codi Serie

S'han realitzat proves de rendiment sobre els workflows en sèrie amb l'objectiu d'avaluar el temps d'execució de cadascuna de les etapes del procés, dividit en les fases de: obtenció d'observacions, l'adquisició de les dades en format DL3; processament d'observacions, procés per transformar les dades en DL4 i posteriorment DL5; i ajust (fitting) de models, realitzat per comparar les dedes amb models astrofisics teorics.

Les proves s'han dut a terme amb el mateix conjunt de dades, les coordenades del Crab Nebula, i en el mateix entorn, utilitzant 1 CPU i 4 GB de memòria RAM. Els resultats obtinguts, expressats en segons, es presenten a la taula següent:

	Get obs	Processament	Fitting	Total
SED	0.66s	211.95s	712.22s	925.19s
Ligth Curve	0.58s	90.10s	46.47s	140.31s
SkyMap	0.27s	65.50s	-	65.78s

Taula 1: Temps d'execució del codi en sèrie

El flux de processament es pot dividir en tres etapes. Primerament, s'extrauen i filtren les dades; a continuació, es processen mitjançant els càlculs adients, i finalment, quan es vol ajustar un model teòric a les dades observades, s'executa l'etapa de fitting.

Cada representació gràfica comporta processos específics de transformació de dades i càlcul, necessaris per a la visualització adequada de les mètriques corresponents. Aquesta variabilitat en els requisits computacionals es reflecteix en diferències significatives en els temps d'execució segons el workflow analitzat. El SED es presenta com el més complex des del punt de vista computacional, amb l'etapa de fitting com a component crític, ja que representa aproximadament el 77% del temps total d'execució. En contraposició, el SkyMap és el menys

costós en termes globals; tanmateix, el processament d'observacions n'ocupa gairebé la totalitat del temps, amb un 99,6%, fet que el converteix en el principal coll d'ampolla del procés. El LightCurve es posiciona en un punt intermig, amb el processament d'observacions com a camí crític, ja que concentra el 64% del temps d'execució, davant del 33% atribuït a l'etapa de fitting.

Cal aclarir que aquestes dades s'obtenen a partir de fitxers FITS, dels quals Gammapy ja disposa d'una funció per convertir-les en un objecte manipulable, *observation*. A més, no es realitza cap filtratge de dades, ja que, sent les primeres proves, només s'ha accedit als documents provinents del Crab.

5.2 Execució paral·lela sobre CosmoHub

A partir de l'estudi de les diferents etapes del *workflow* executades en sèrie i de les possibilitats detectades per introduir paral·lelisme, els codis s'han adaptat per executar-se de manera paral·lela durant el procés de tractament de dades. En particular, s'ha optat per paral·lelitzar l'etapa de processament d'observacions, ja que es basa en un bucle en què cada observació es tracta de manera independent i, posteriorment, se'n combinen els resultats. Aquest comportament s'ajusta perfectament al patró MapReduce, proporcionat per Hadoop i Spark, i per aquest motiu s'ha escollit aquest model com a base per a la implementació distribuïda.

A més, l'accés a les dades s'ha realitzat mitjançant consultes SQL sobre el catàleg complet de CosmoHub, a través de Hive, especificant condicions per filtrar i extreure exclusivament la informació rellevant per a l'estudi. Aquestes consultes, com que són interpretades pel motor Hive, també s'executen de manera paral·lela i optimitzada, aprofitant la distribució de les dades al sistema HDFS i permetent una lectura eficient i escalable dels conjunts massius d'observacions.

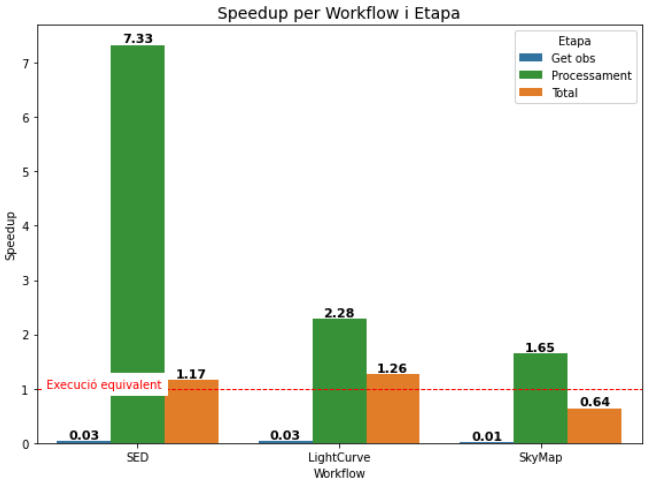
5.2.1 Comparació amb l'execució seqüencial

Així doncs, cal tenir en compte que, tot i que els codis en sèrie i paral·lel executen funcionalment el mateix procés, no són del tot comparables pel que fa al temps d'execució. La versió paral·lela introdueix diversos elements addicionals que afecten el rendiment global:

- Temps extra de configuració, derivat de la inicialització dels entorns Spark i Hive.
- Aplicació d'un filtratge global sobre tot el catàleg de CosmoHub.
- Processament més complex d'observacions, ja que la informació està estandarditzada en taules dins CosmoHub i és necessari definir des de zero la classe *observation*, indispensable per treballar amb la llibreria Gammapy.

Aquests factors introdueixen una certa sobrecàrrega que no està present en l'execució en sèrie, i per tant cal interpretar els resultats de manera contextualitzada, tenint en compte les diferències entre les dues implementacions.

Tot i això, hi ha casos en què el codi paral·lel ofereix un rendiment superior. A continuació, es presenta una comparació dels resultats obtinguts, prenent com a referència el temps més eficient assolit per a cada workflow en la seva versió paral·lela. Els resultats complets es poden consultar a l'Apèndix 1. Seguidament, es mostra la gràfica 1 amb els valors de *speedup* corresponents, que quantifiquen la millora de rendiment obtinguda en cada cas.



Gràfica 1: Speedup obtingut per workflow i etapa d'execució

Els resultats mostren que l'eficiència del paral·lelisme varia significativament en funció de l'etapa analitzada. El processament és la fase on s'obtenen els millors resultats, amb *speedups* destacats com 7,33 en el SED, 2,28 en el LightCurve i 1,65 en el SkyMap, fet que evidencia que la paral·lelització implementada és especialment eficient en tasques de càlcul intensiu. En canvi, l'etapa de Get obs presenta un rendiment molt inferior en tots els workflows, amb *speedups* per sota de 0,03, degut principalment al cost afegit de les consultes distribuïdes a Hive.

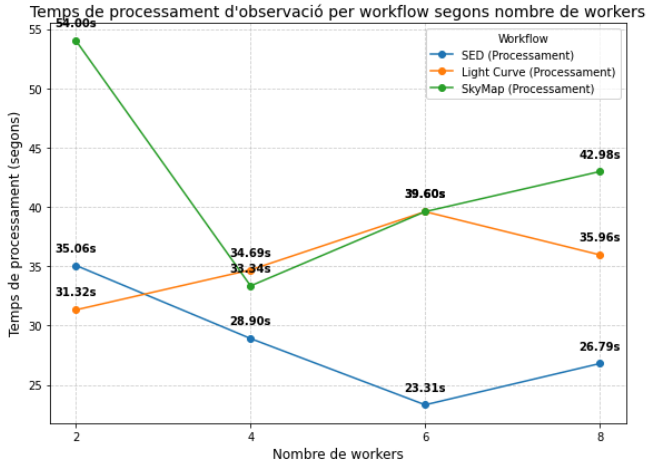
Pel que fa al rendiment global, el *speedup* total és de 1,26 per al LightCurve, 1,17 per al SED, i tan sols 0,64 per al SkyMap, sent aquest últim l'únic cas que no es beneficia del canvi respecte a la versió seqüencial. Cal tenir en compte que dins el temps total també s'inclou la fase d'inicialització de la configuració, que té un cost mitjà de 34,74 segons, i que pot afectar negativament els workflows més lleugers.

5.2.2 Valoració de l'escalabilitat

Per analitzar l'escalabilitat del sistema, s'han dut a terme diverses proves variant el nombre de workers, tot mantenint constants les observacions utilitzades, corresponents a la Crab Nebula. L'anàlisi se centra específicament en l'etapa paral·lelitzada amb Spark, el

processament d'observacions. Les dades completes d'aquestes proves es poden consultar a l'Annex 1.

Tot i això, de manera contraintuitiva, l'increment del nombre de workers no es tradueix en una millora progressiva del rendiment. De fet, no s'observa cap tendència clara que indiqui una escalabilitat lineal o sostinguda. Aquesta manca de correlació queda reflectida a la gràfica 2.



Gràfica 2: Representació dels temps d'execució de l'etapa de processament d'observacions

Possiblement relacionada amb aquesta manca de coherència, s'ha observat també una variabilitat significativa en els temps d'execució dels mateixos blocs de codi amb les mateixes dades, amb un coeficient de variació mitjà del 14,23 %.

Els resultats obtinguts evidencien una combinació de factors que varien el rendiment. En primer lloc, el conjunt de dades emprat és relativament petit, tant en nombre d'observacions com en volum total d'informació. Aquesta limitació fa que l'ús d'un entorn distribuït com Spark no resulti eficient, ja que l'overhead associat a la seva gestió acaba superant els beneficis de la paral·lelització.

Tot i augmentar el nombre de workers, la comunicació interna del clúster i el mecanisme de divisió del treball en *splits* que utilitza Spark no optimitzen el rendiment quan es treballa amb conjunts de dades tan reduïts. Per tant, el guany potencial derivat del paral·lelisme queda anul·lat pel cost afegit de coordinació. Aquesta hipòtesi caldrà validar-la amb conjunts de dades més grans, i constitueix una línia de treball futura rellevant.

D'altra banda, la variabilitat observada en els temps d'execució, pot atribuir-se al fet que la partició de les dades no és determinista. El repartiment de la càrrega entre workers conté una certa aleatorietat, de manera que, fins i tot mantenint el mateix nombre de workers, cada execució pot distribuir les observacions de manera diferent. Aquesta variació en l'estructura interna del pla d'execució pot explicar les fluctuacions detectades en el temps de resposta.

6 Conclusió

Aquest treball ha permès avançar en la integració i optimització dels workflows d'anàlisi de raigs gamma dins la plataforma CosmoHub, aprofitant la infraestructura del PIC i el motor de processament distribuït Apache Spark. S'hi han implementat amb èxit tres workflows (SED, LightCurve i SkyMap), aconseguint una compatibilitat total entre les observacions estandarditzades en taules i les classes gestionades per la llibreria Gammapy.

Els resultats mostren que Spark pot aportar millores significatives en etapes intensives de càlcul, com el processament d'observacions. En termes globals, el speedup obtingut en comparació amb l'execució seqüencial ha estat de 1,26 per al Light Curve, 1,17 per al SED i 0,64 per al SkyMap, aquest darrer afectat per la sobrecàrrega derivada de l'adquisició de dades dins de CosmoHub.

Tanmateix, s'ha observat una inconsistència en la tendència temporal en incrementar el nombre de workers, sense una millora clara ni una escalabilitat sostinguda. Aquesta variabilitat s'atribueix al volum reduït de dades utilitzat i a la sobrecàrrega pròpia de la gestió distribuïda, que pot eclipsar els avantatges del paral·lelisme en escenaris de poca càrrega. Aquesta observació obre una línia futura de recerca, validar l'eficiència real del codi amb conjunts de dades més grans i una exigència computacional més elevada.

Cal recalcar que, durant el desenvolupament d'aquest treball, s'han presentat diversos reptes tècnics, entre els quals destaca la configuració del clúster Spark en l'entorn del PIC. Malgrat no representar una dificultat tècnica especialment complexa, aquesta tasca va endarrerir de manera notable la planificació. Per aquest motiu, la resolució aplicada s'aborda amb més detall a l'Apèndix 2.

En definitiva, aquest projecte posa de manifest la viabilitat d'integrar CosmoHub, Spark i GammaPy en l'anàlisi científica de dades astrofísiques. Al mateix temps, obre noves perspectives de recerca centrades en la combinació d'observacions procedents de diverses característiques, com ara raigs gamma, neutrins i ones gravitatòries, amb l'objectiu de millorar la interoperabilitat, optimitzar el rendiment i ampliar l'ecosistema d'eines disponibles per a la comunitat investigadora.

Agraïments

Aquest treball de fi de grau no hauria estat possible sense el suport i l'acompanyament dels tutors implicats, Jordi Delgado i Eduardo César Galobardes, als quals vull expressar el meu més sincer agraïment. Els agraeixo el seu suport constant i la seva orientació, que han estat clau per afrontar els reptes que han anat sorgint al llarg del projecte.

Bibliografia

[1] YouTube. (2021). UZ Estallic | Video. https://www.youtube.com/watch?v=715Njn0y0qw&t=2607s&pp=ygUZRXN0YWxsa

[2] Química.es. (s.d.). Rayos gamma. https://www.quimica.es/enciclopedia/Rayos_gamma.html

[3] Institut d'Estudis Espacials de Catalunya (IEEC). (s.d.). MAGIC. https://www.ieec.cat/es/proyecto/29/magic/

[4] Astromia.com. (s.d.). HESS – High Energy Stereoscopic System. https://www.astromia.com/fotohistoria/hess.htm

[5] VERITAS. (s.d.). VERITAS Observatory. https://veritas.sao.arizona.edu/

[6] CTA Observatory. (s.d.). Emission to discovery: Science. https://www.ctao.org/emission-to-discovery/science/

[7] Gammapy. (2021). DL3 Data Format Overview. https://docs.gammapy.org/0.18/overview/DL3.html

[8] Arxiv. (2023). Very-High-Energy Open Data Format. https://arxiv.org/abs/2308.13385

[9] Gammapy. (2021). https://docs.gammapy.org

[10] Port d'Informació Científica (PIC). (s.d.). Àrees d'activitat - Astrofísica. https://www.pic.es/areas/#astro

[11] Tallada P., Carretero J., Casals J., Acosta-Silva C., Serrano S., Caubet M., et al. (2020). CosmoHub: Interactive exploration and distribution of astronomical data on Hadoop. https://doi.org/10.1016/j.ascom.2020.100391

[12] CosmoHub. (s.d.). Home. https://cosmohub.pic.es/home

[13] Google Cloud. (s.d.). What is Apache Spark?. https://cloud.google.com/learn/what-is-apache-spark?hl=es

[14] L'Harrak El Awami, Y. (2024). Paral·lelització de workflows d'anàlisi de dades d'astronomia de raigs gamma (Treball de Fi de Grau, Grau en Enginyeria de Dades, Escola d'Enginyeria, Universitat Autònoma de Barcelona)

[15] Gammapy Project. (n.d.). Gammapy tutorials (version dev). Retrieved April 10, 2025, from https://docs.gammapy.org/dev/tutorials/

[16] Gammapy Collaboration. (s.d.). *Glossary and references* (DL3, DL4, DL5). https://docs.gammapy.org/1.0.2/user-guide/references.html

[17] Gammapy. (s.d.). Maps – API tutorial. https://docs.gammapy.org/dev/tutorials/api/maps.html

[18] International Virtual Observatory Alliance (IVOA). (s.d.). Simple Cone Search (SCS) - IVOA Recommendation. https://www.ivoa.net/documents/REC/DAL/ConeSearch-20080222.html

[19] Tallada, P., Carretero, J., Casals, J., Acosta-Silva, C., Serrano, S., Caubet, M., Castander, F. J., César, E., Crocce, M., Delfino, M., Eriksen, M., Fosalba, P., Gaztañaga, E., Merino, G., Neissner, C., & Tonello, N. (2020). CosmoHub: Interactive exploration and distribution of astronomical data on Hadoop. Astronomy and Computing, 32, 100391. https://doi.org/10.1016/j.ascom.2020.100391

[20] AWS. (n.d.). ¿Qué es Hive? - Explicación de Apache Hive. Retrieved April 19, 2025, from https://aws.amazon.com/es/what-is/apache-hive/

[21] Apache Software Foundation. (n.d.). Apache Hadoop YARN. Retrieved April 19, 2025, from https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

[22] IBM. (n.d.). ¿Qué es el sistema de archivos distribuido de Hadoop (HDFS)?. Retrieved April 19, 2025, from https://www.ibm.com/es-es/topics/hdfs

APÈNDIX

A1. Temps d'execució

A continuació es presenten els resultats de rendiment obtinguts de les execucions implementades amb Spark i integrades a CosmoHub, desglossat per etapes i nombre de workers utilitzats. Les taules 2, 3 i 4 mostren els temps d'execució, en segons, corresponents als workflows SED, Light Curve i SkyMap, respectivament.

Workers	Conf	Get obs	Processament	Fitting	Total
2	35.56s	21.32s	35.06s	894.61s	986.69s
4	29.40s	23.93s	28.90s	710.38s	792.68s
6	31.18s	18.56s	23.31s	821.30s	894.42s
8	32.05s	19.86s	26.79s	848.49s	927.26s

Taula 2: Temps d'execució del SED

Workers	Conf	Get obs	Processament	Fitting	Total
2	27.97s	14.65s	31.32s	26.83s	100.85s
4	33.84s	18.48s	34.69s	29.30s	116.38s
6	25.71s	21.78s	39.45s	24.23s	111.28s
8	31.75s	18.88s	35.96s	24.87s	111.58s

Taula 3: Temps d'execució del Light Curve

Workers	Conf	Get obs	Processament	Total
2	39.75s	27.16s	54.00s	121.08s
4	47.56s	27.35s	33.34s	108.38s
6	40.07s	23.05s	39.80s	103.02s
8	42.06s	28.72s	42.98s	113.87s

Taula 4: Temps d'execució del SkyMap

Per complementar aquesta informació, la taula 5 presenta una anàlisi estadística bàsica dels temps d'execució mostrats anteriorment, que inclou la mitjana, la desviació típica i el coeficient de variació (CV) expressat en percentatge. Les etapes analitzades corresponen a l'execució d'un mateix bloc de codi, justificant que s'agrupin sense comprometre la validesa de l'anàlisi.

Etapa	Mitjana	Desviació típica	CV (%)
Configuració	34.74s	6.18s	17.79%
Get obs	21.98s	4.07s	18.52%
Processament SkyMap	42.53s	7.48s	17.59%
Processament SED i LightCurve	31.94s	5.01s	15.69%
Fitting SED	818.7s	67.8s	8.28%
Fitting LightCurve	26.31s	1.98s	7.53%

Taula 5: Mesures estadístiques

A2. Configuració d'Apache Spark

Com que el principal problema del treball va ser la inicialització del clúster Spark dins del PIC, a continuació es presenta la configuració que va funcionar. Tot i que no es descarta que hi hagi altres alternatives compatibles, però es documenta la solució aplicada en aquest projecte per facilitar la resolució de possibles incidències futures.

L'entorn on s'executa el codi ha de ser la versió Python 3.9 i no s'ha d'instal·lar el paquet *pyspark*, ja que obligaria a executar el codi en mode local. En el seu lloc, cal instal·lar únicament el paquet *findspark*, que permet inicialitzar correctament la sessió. Dins del codi, també cal especificar les rutes de les variables d'entorn corresponents a Hadoop i Hive. Els passos concrets seguits es mostren a la figura 6.

```
import findspark
findspark.init()

import os
os.environ["HADOOP_HOME"] = "/usr/local/hadoop"
os.environ["HADOOP_CONF_DIR"] = "/usr/local/hadoop/etc/hadoop"
os.environ["HIVE_HOME"] = "/usr/local/hive"
os.environ["HIVE_CONF_DIR"] = "/usr/local/hive/conf"

from pyspark.sql import SparkSession

spark = SparkSession.builder.enableHiveSupport().getOrCreate()
sc = spark.sparkContext
spark.sql('use cosmohub')
```

Figura 6: Fragment de codi per a la inicialització d'Apache Spark

La connexió amb el clúster serà correcta si la variable master adopta un valor semblant a *v3.4.5:7077*, en cap cas s'ha de fixar com *local[*]*. A més, és imprescindible disposar de permisos sobre el projecte on es treballa per extreure dades de CosmoHub i utilitzar Hive; en cas contrari, no serà possible accedir als conjunts de dades. Finalment, cal executar la comanda *kinit* <usuari> per obtenir el corresponent *token* de Kerberos.