PIC
port d'informació
científica

Astronomy in a Big Data platform: CosmoHub & SciPIC

Jorge Carretero, Carles Acosta, Alex Alarcón, Linda Blot, Jordi Casals, Ricard Cruz, Francisco Castander, Marc Caubet, Pablo Fosalba, Santiago Serrano and Pau Tallada

PIC
port d'informació
científica

ICE CSIC-IEEC

IFAE
25 years 1991-2016
Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

PIC
port d'informació
científica

Motivation: Galaxy catalogs

Project	Date	volume / night	Total volume	Number of objects (catalog)
SDSS	**2000 - now**	variable	116 TiB	2×10^6
MICE GC	**2013**	NA	42 TiB	5×10^8
DES	**2013 - 2018**	2.5 TiB	2 PiB	4×10^8
GAIA*	**2014 - 2019**	40 GiB	1 PiB	1.1×10^9
Euclid	**2020 - 2025**	100 GiB	580 TiB	1.5×10^9
LSST	**2022 - 2032**	15 TiB	50 PiB	1×10^{10}

* DR1 full sky star catalog

PIC
port d'informació
científica

Port d'Informació Científica (PIC)

- Founded in 2003
  - Collaboration between IFAE and CIEMAT
  - Spanish Tier-1 WLCG, Euclid SDC-ES
- Our supported projects:
  - Particle physics: LHC (Atlas, CMS, LHCb), neutrinos (T2K Japan)
  - Astrophysics: MAGIC, Cherenkov Telescope Array
  - Cosmology:

Euclid Consortium PAU MICE

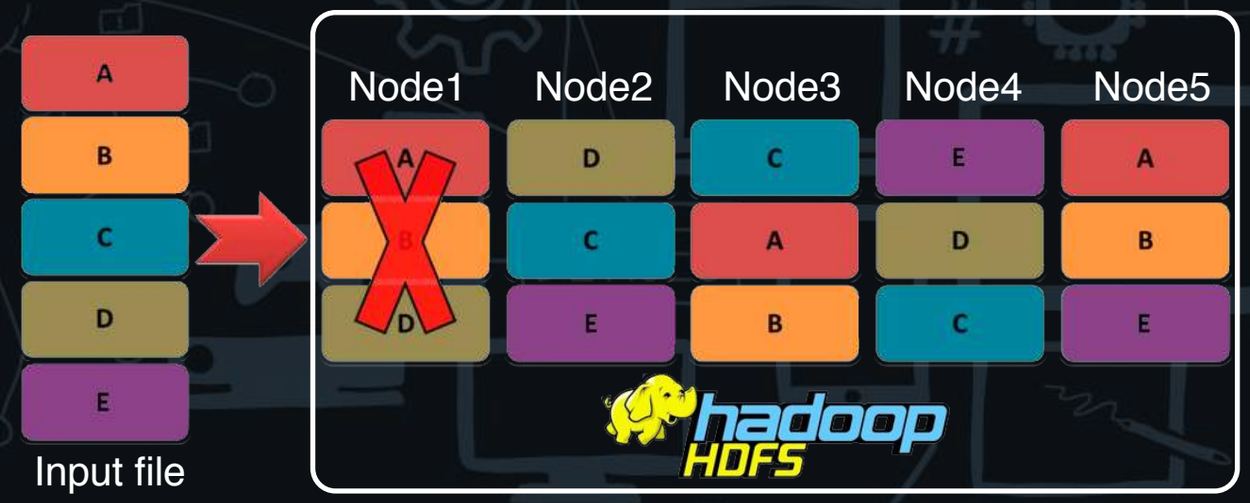
- Resources:
  - 7500 cores, 8 PiB disk, 21 PiB tape, 10Gb LAN
  - 2 x 10 Gbps WAN, optical paths to CERN and Observatorio del Roque de los Muchachos

PIC Big Data platform

- Based on Hadoop (Hortonworks HDP 2.5)
  - Open source Big Data Platform
  - <u>Distributed storage and processing</u>
    - Runs on commodity computer clusters
    - Scalable from dozens up to thousands of nodes
      - Performance scales with HW
    - Fault tolerant
      - Simple machines working together - no single point of failure

PIC Hadoop platform

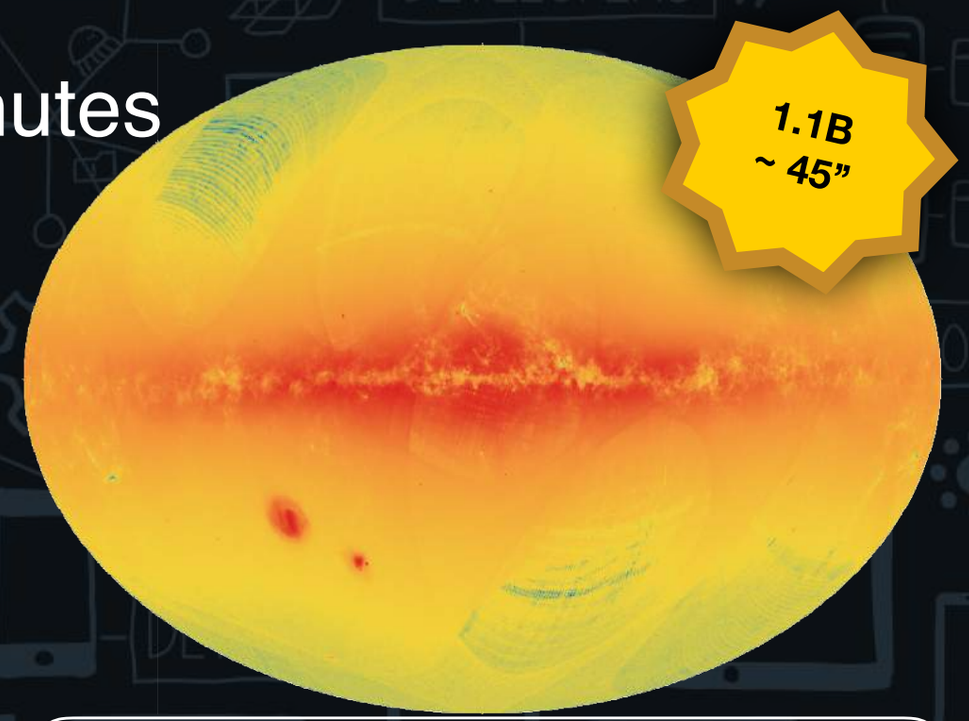
- 44 nodes + 3 head + UI
- 784 total cores
- 3 TiB total RAM
- 230 TiB HDD



PIC
port d'informació
científica

COSMO HUB

- Web portal to perform interactive exploration and distribution of massive cosmological data
- Based on Apache Hive
- Generate and download custom subsets:
  - Guided process, no SQL knowledge required
    - Expert mode available
  - Query time range: seconds to minutes
    - 85% in < 3 min
  - VAD ready to download
- Exploration (Visualization)
  - Unlimited time
  - Full dataset plots (over all rows)
    - May use sampling
    - 1D histogram & 2D heatmap



https://cosmohub.pic.es

SciPIC: Scientific pipeline at PIC

- Set of python codes/algorithms to generate synthetic galaxy catalogs using DM simulations
- Run on top of the PIC Big Data platform using Spark
- Weak Lensing (WL): through a JOIN that **involves 7.4B galaxies and 253B pixels**
- Stored in CosmoHub ready to be explored and distributed

	input		output		time	time (+WL)
	haloes	**size**	**galaxies**	**size**		
octant	5×10^9	0.7 TiB	7.4×10^9	3.2 TiB	< 2 h	< 10 h
full sky	4×10^{10}	5.5 TiB	6×10^{10}	25.5 TiB	< 16 h	< 80 h

PIC
port d'informació
científica

COSMO HUB

~ 140 active users

~ 2300 custom catalogs

~ 15 TiB hosted data

> 10^{11} objects

PIC
port d'informació
científica

Summary & Conclusions

- Study new technologies when traditional ones can't fulfill user needs.
- Hadoop has proven to be a great choice
  - Scalability
  - Fault tolerant
- CosmoHub
  - Enables powerful interactive analysis
  - Delivers custom subsets to project scientists, to external collaborators as support for Open Data Access
- SciPIC
  - In charge of generating the largest synthetic galaxy catalog
  - Very fast -> many iterations -> better results

PIC
port d'informació
científica

Team - Q&A

- CosmoHub
  Carlos Acosta, Jorge Carretero, Jordi Casals, Marc Caubet, Santiago Serrano, Pau Tallada

- SciPIC
  Alex Alarcón, Linda Blot, Jorge Carretero, Francisco Castander, Pablo Fosalba, Kai Hoffmann, Santiago Serrano, Pau Tallada

- Special thanks to
  Jordi Delgado, Martin Eriksen, Christian Neissner, Davide Piscia, Nadia Tonello, Francesc Torradeflot and the rest of PIC staff