PAU Data Management at PIC

Physics of the
Accelerating
Universe

Christian Neissner & Santiago Serrano
PIC & ICE

RIA meetings - Madrid, March 23rd 2012

Outline

- Physics of the Accelerating Universe
- Port d'Informació Científica (PIC)
- Infrastructure at PIC: Computing & storage
- PAU Data Management:
  - Data Flow and Pipelines
  - Data Base
  - Pipeline Orquestration
  - Data Volume
- Summary & Outlook

Physics of the Accelerating Universe

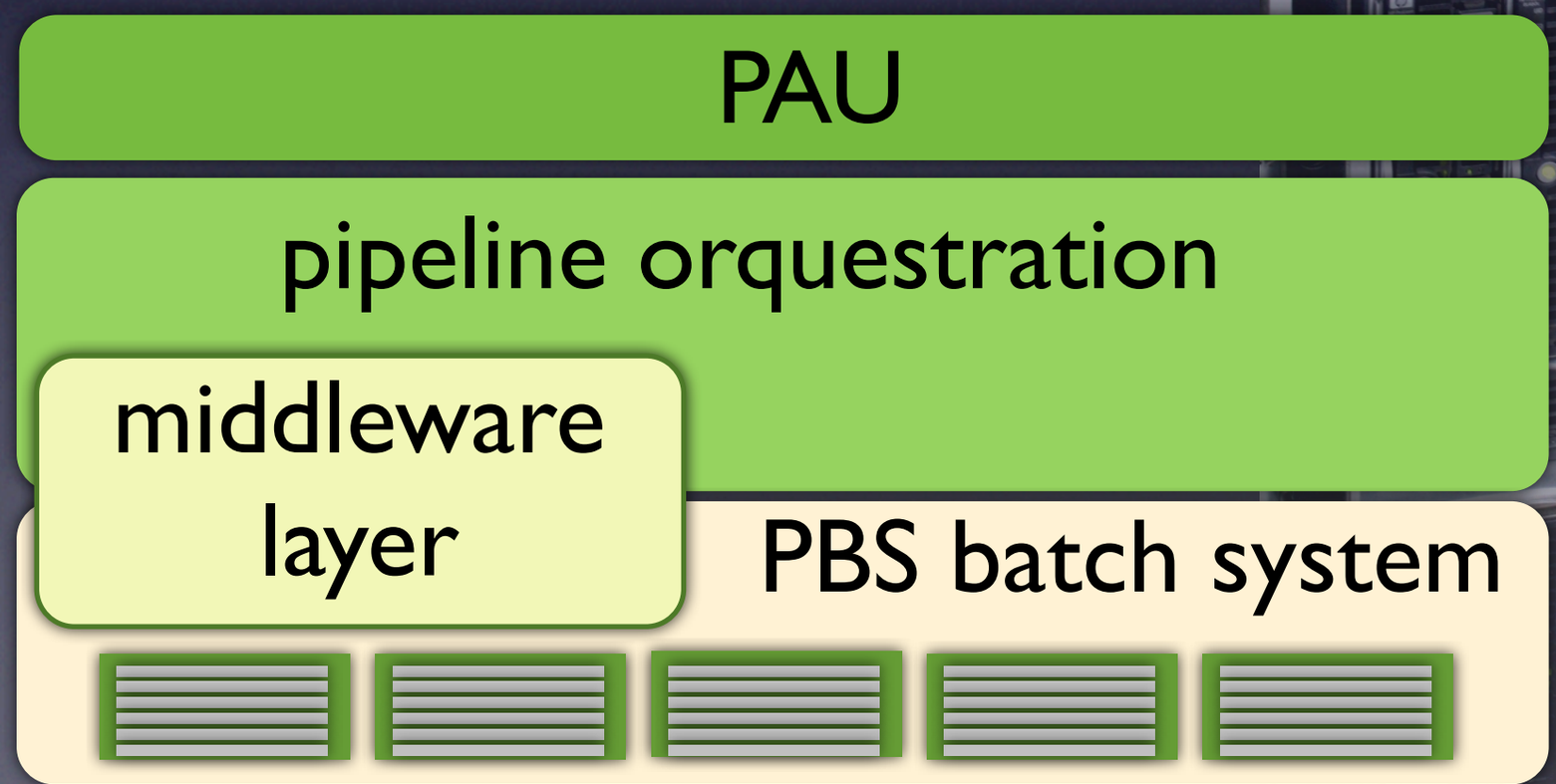
- science with data obtained by a:
  - survey of 200 deg^2 in the visible spectrum
  - camera with 8 central and 10 lateral CCD
  - 6 broad band + 42 narrow band filters
  - camera installed at the WHT at ORM

Port d'Informació Científica

- a scientific data center in Barcelona founded in 2003
- **HEP:** Spanish Tier-1 data center for CERNs LHC
- **Astrophysics:** Reference data center for the MAGIC collaboration
- **Cosmology:**
  - Reference data center for PAU and MICE
  - Spanish Scientific Data Center in ESAs Euclid mission

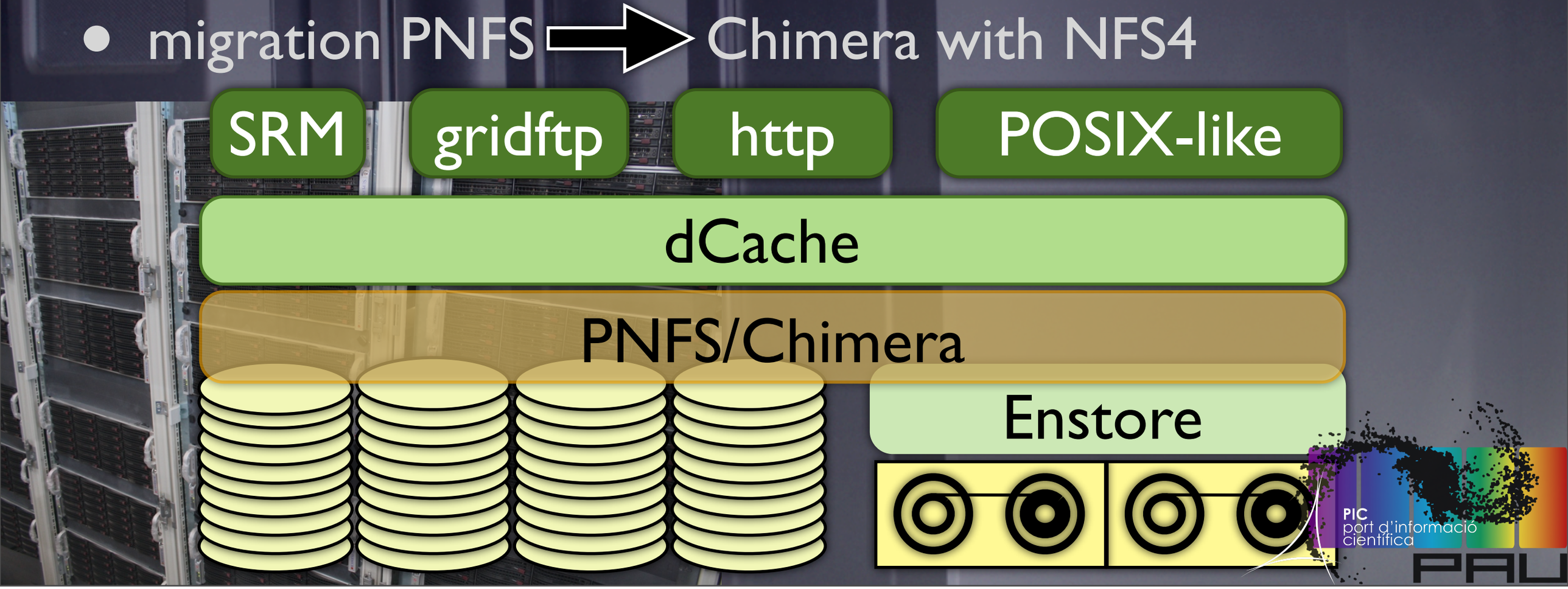
Infrastructure@PIC
Computing

- 4,000 computing cores, 150 dedicated to PAU
- PBS batch system (may change)
- gLite interfaced resources
- PYTHON services for pipeline orquestration

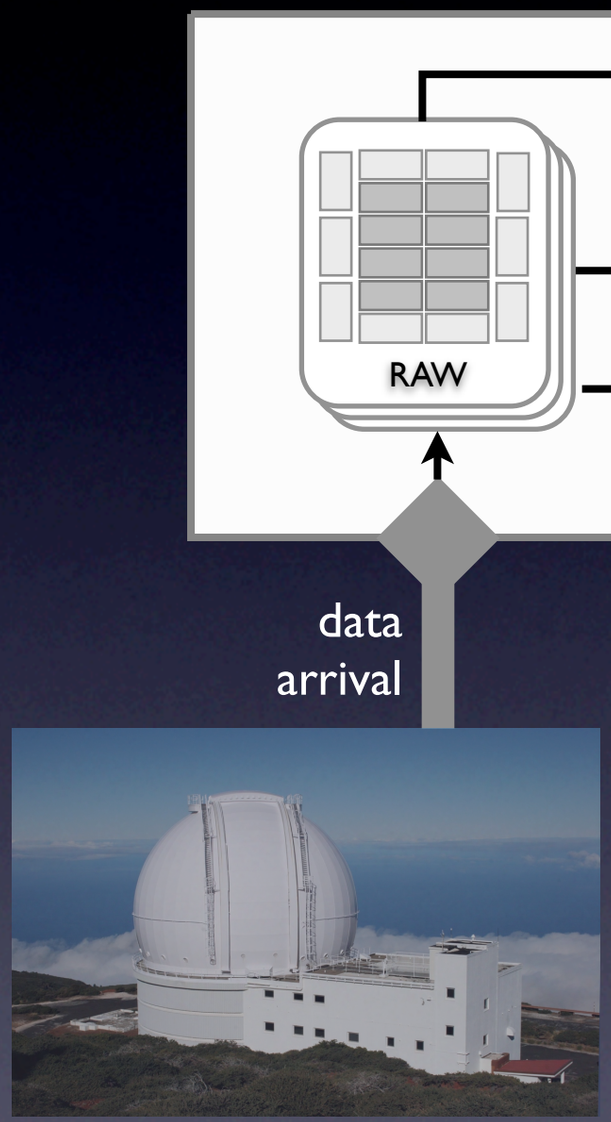


Infrastructure@PIC
Storage

- 4 PB disk + 4 PB tape
- 50 TB disk + 200 TB tape dedicated to PAU
- dCache with PNFS name space
- migration PNFS ⇨ Chimera with NFS4



Data flow schema



data
arrival

RAW

Nightly Processing
Pipeline

Level 1
products

to
science

Storage

Data Base

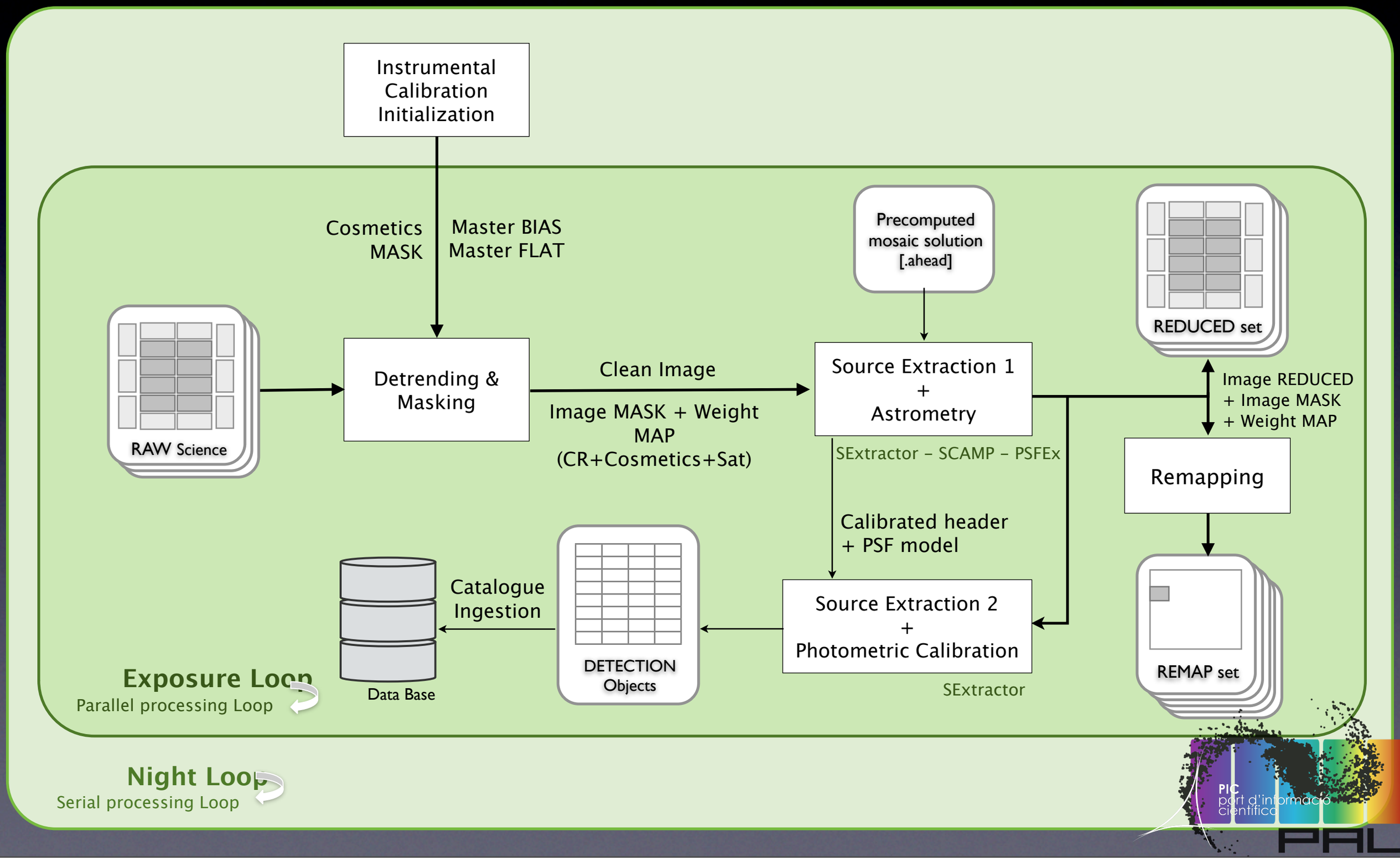
Data Monitor

Multi-band
&
Multi-epoch

Level 2
products

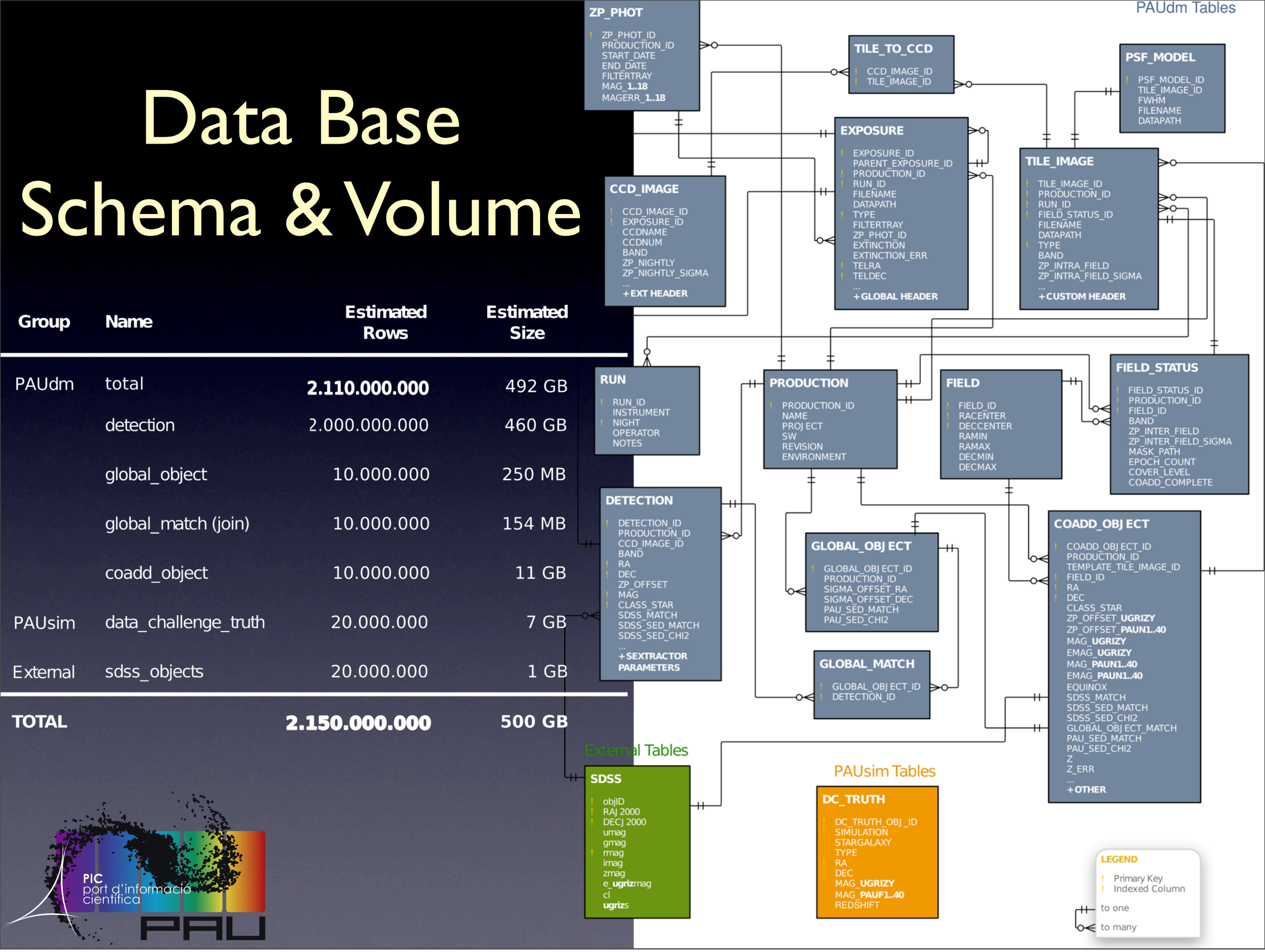
PIC data center

Nightly Single-Epoch pipeline



Data Base Schema & Volume

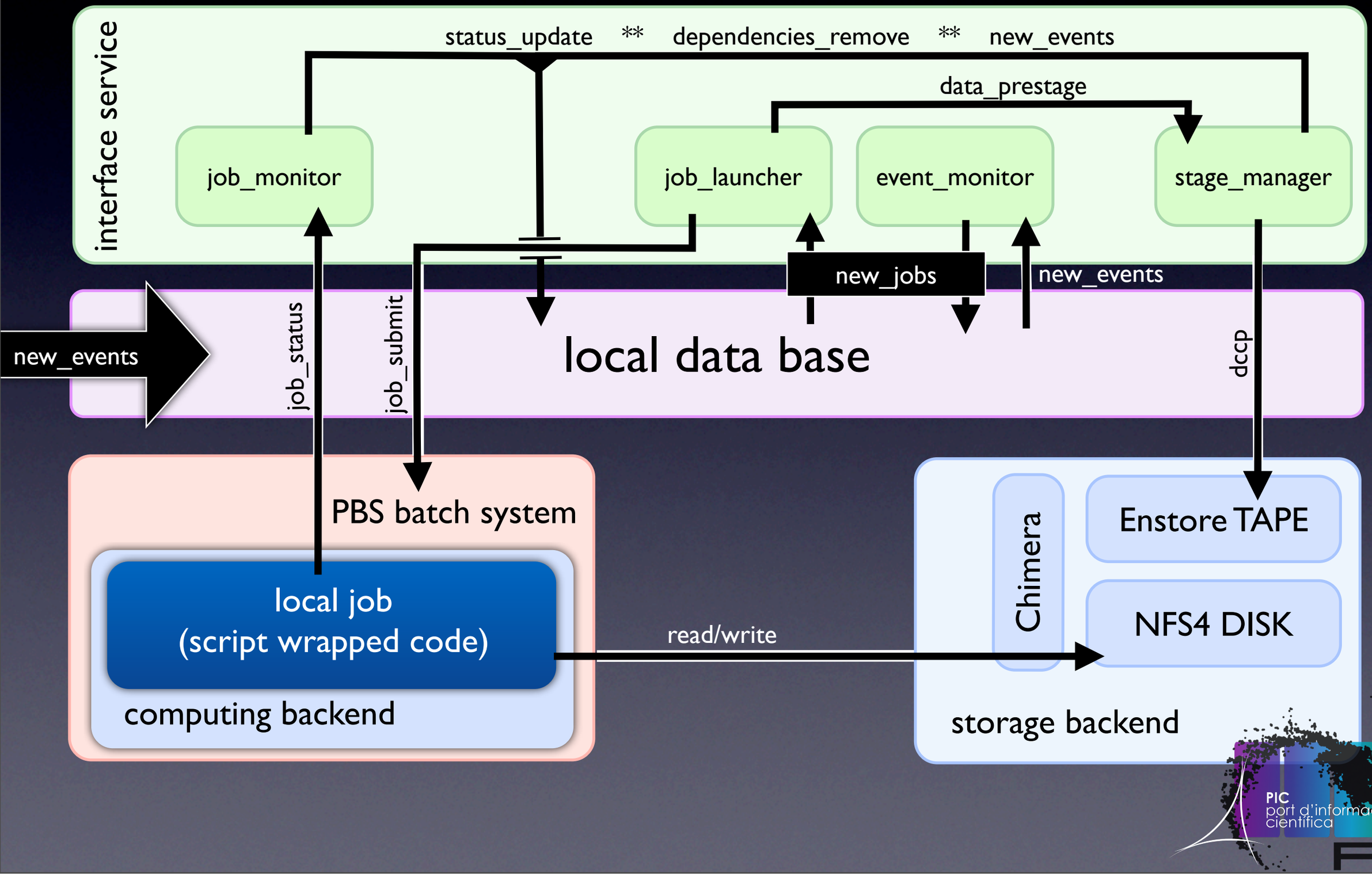
Group	Name	Estimated Rows	Estimated Size
PAUdm	total	**2.110.000.000**	492 GB
	detection	2.000.000.000	460 GB
	global_object	10.000.000	250 MB
	global_match (join)	10.000.000	154 MB
	coadd_object	10.000.000	11 GB
PAUsim	data_challenge_truth	20.000.000	7 GB
External	sdss_objects	20.000.000	1 GB
TOTAL		**2.150.000.000**	**500 GB**



PIC port d'informació científica

PAU

Pipeline orquestration



Data Volume

Level	Name	Type	Unit Size	Estimated Number	Estimated Size
0	RAW Science	mosaic fits	600 MB	53.500	32.000 GB
	RAW Bias	mosaic fits	600 MB	7.500	4.500 GB
	RAW Flat	mosaic fits	600 MB	7.500	4.500 GB
1	Reduced Science	mosaic fits	600 MB	53.500	32.000 GB
	Weight Map	mosaic fits	600 MB	53.500	32.000 GB
	Image Mask	mosaic fits	300 MB	53.500	16.000 GB
	Master Bias*	mosaic fits	600 MB	375	225 GB
	Master Flat*	mosaic fits	600 MB	375	225 GB
2	Remap Science*	remap fits	18 MB	5.670.000	102.000 GB
	Remap Weight Map*	remap fits	18 MB	5.670.000	102.000 GB
	Remap Image Mask*	remap fits	9 MB	5.670.000	51.000 GB
	Coadd Science	remap fits	18 MB	440.000	8.000 GB
	Coadd Image Mask	remap fits	9 MB	440.000	4.000 GB
	Depth Mask	remap fits	18 MB	440.000	8.000 GB
	PSF Model	ascii	1 MB	6.000.000	6.000 GB
TOTAL				**24.559.750**	**402.450 GB**

PIC port d'informació científica

PAU

Summary & Outlook

- Why using this kind of infrastructure for a 500 TB data management project ?
  - It scales, it provides intrinsic redundancy and it is highly reliable.
- finishing the level-1 to level-2 pipeline
- running data challenges: functionality, performence, ...
- prepare the input archives for Euclid:
  - PAU data, mock catalogs from MICE simulations, ...
- Euclid: supposed distributed data volume
  - 300 PB of legacy archive, hundreds of TB of DBs

Thank you!