

Evaluación de las tecnologías object storage para almacenamiento y análisis de datos climáticos

Autor: Ezequiel Cimadevilla Álvarez

Director: Antonio S. Cofiño González
Codirector: Aida Palacio Hoz

Máster en Ciencia de Datos
Universidad de Cantabria
10 Julio 2019

Grupo de Meteorología de Santander
Grupo de Computación Avanzada



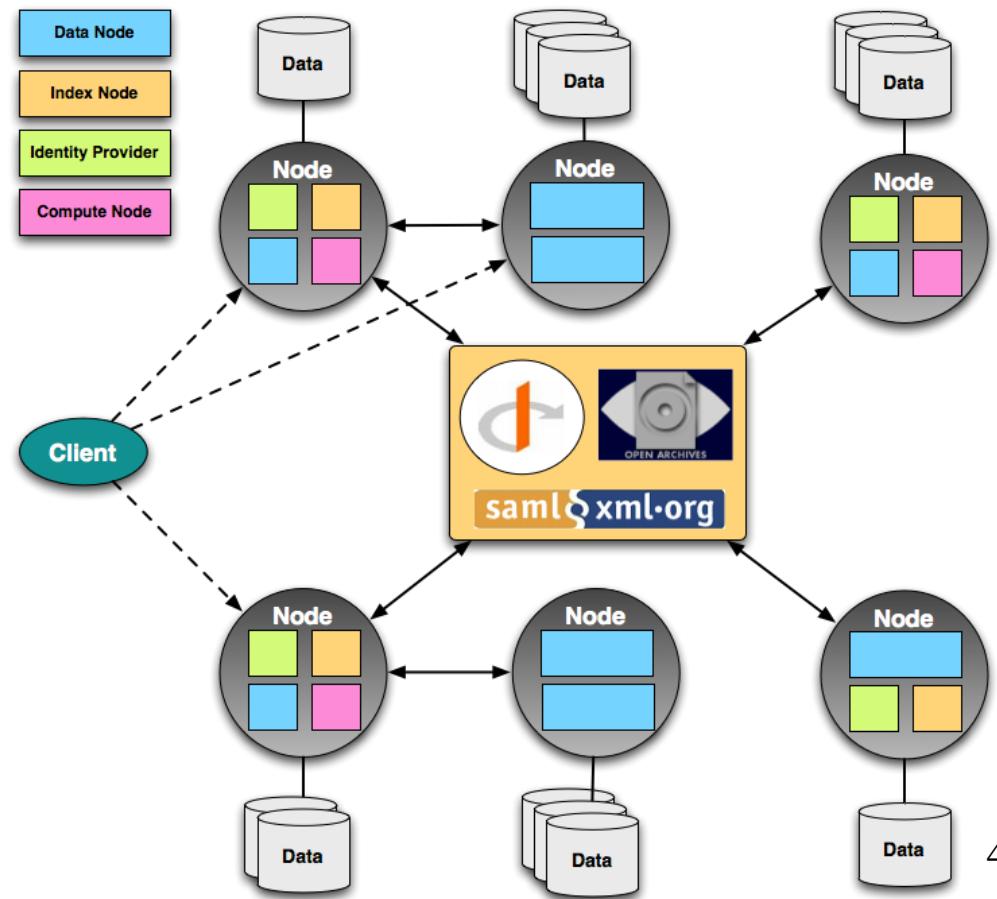
Índice

- Motivación y objetivo
- Datos climáticos
 - Introducción
 - NetCDF
 - Chunking
- Metodologías de análisis de datos
 - Descarga local
 - Servicios de análisis de datos
- Sistemas de almacenamiento
 - Sistemas de ficheros POSIX
 - Object storage
- HDF5 cloud y Zarr
- Evaluaciones HPC y cloud
- Conclusiones y trabajo futuro

Motivación y objetivo

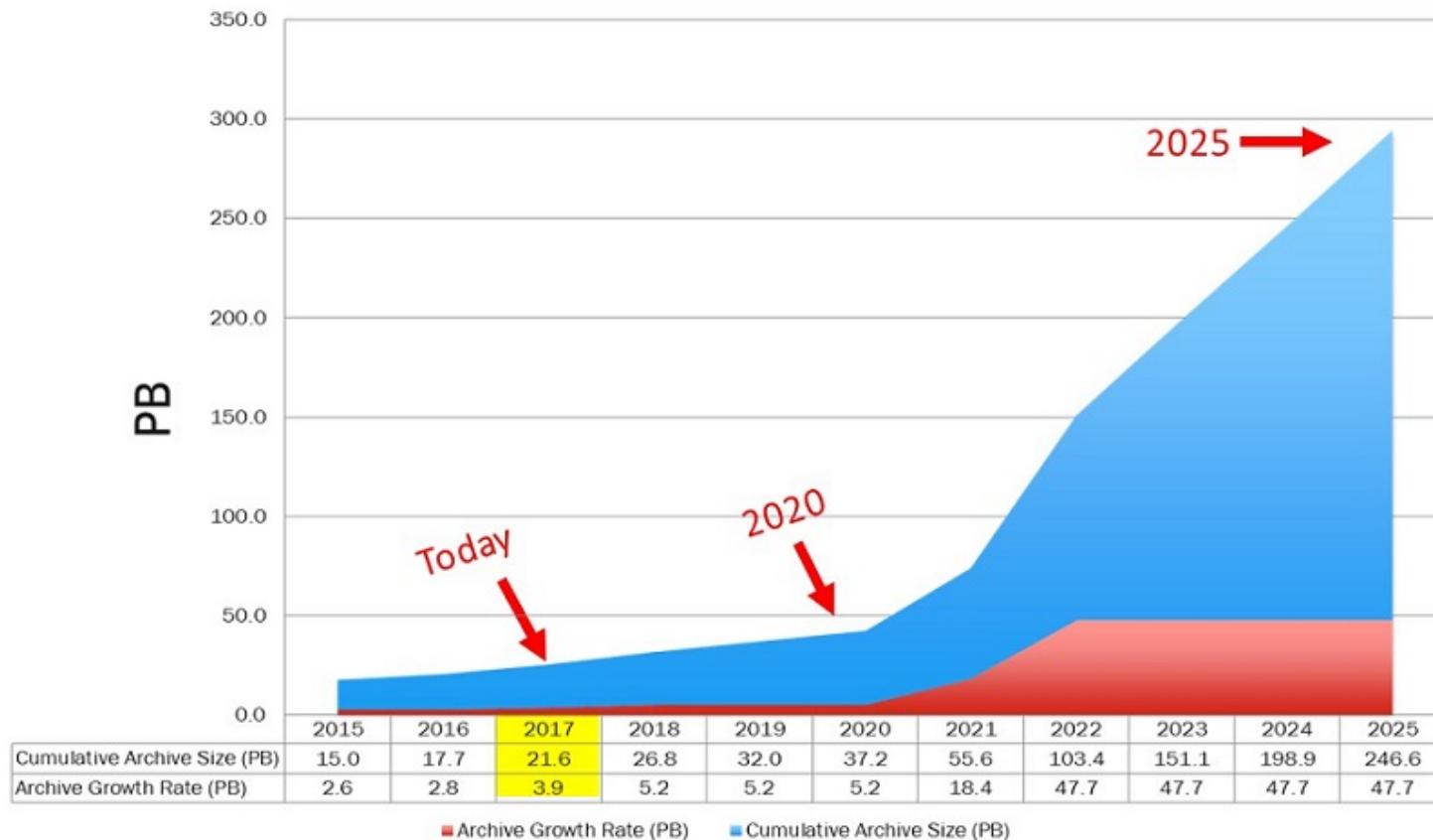
Motivación

- CMIP6 – Sexta fase del marco de trabajo para la mejora del conocimiento sobre cambio climático
 - CMIP3 – 36 TB
 - CMIP5 – 3,3 PB
 - CMIP6 – ¿100 PB?



Motivación

- EOSDIS - NASA's Earth Observing System Data and Information System
- 2020 - 37 PB, 2025 - 246 PB



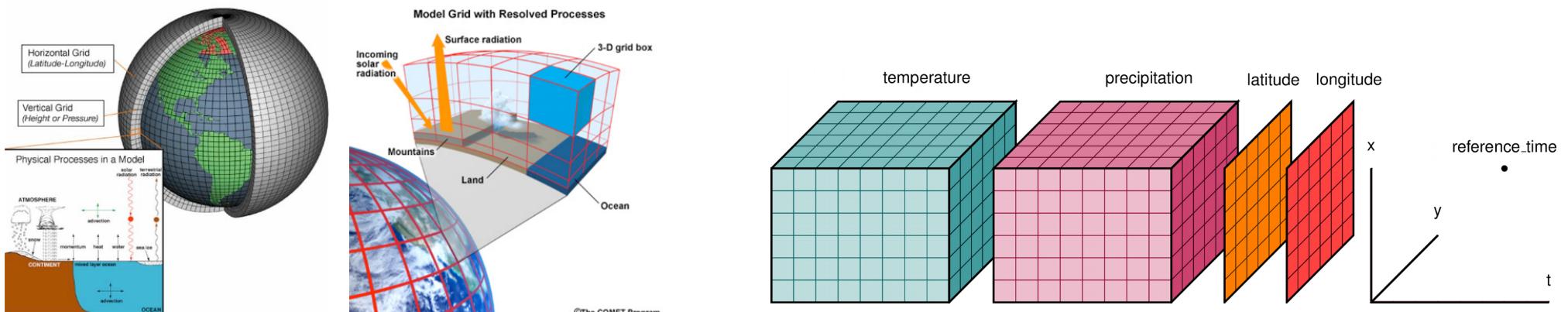
Objetivo

- Análisis del estado del arte sobre almacenamiento en object storage de datos climáticos
- Comparación entre los flujos de trabajo al realizar análisis de datos climáticos
- Despliegue de infraestructuras cloud y HPC

Datos climáticos

Datos climáticos - Introducción

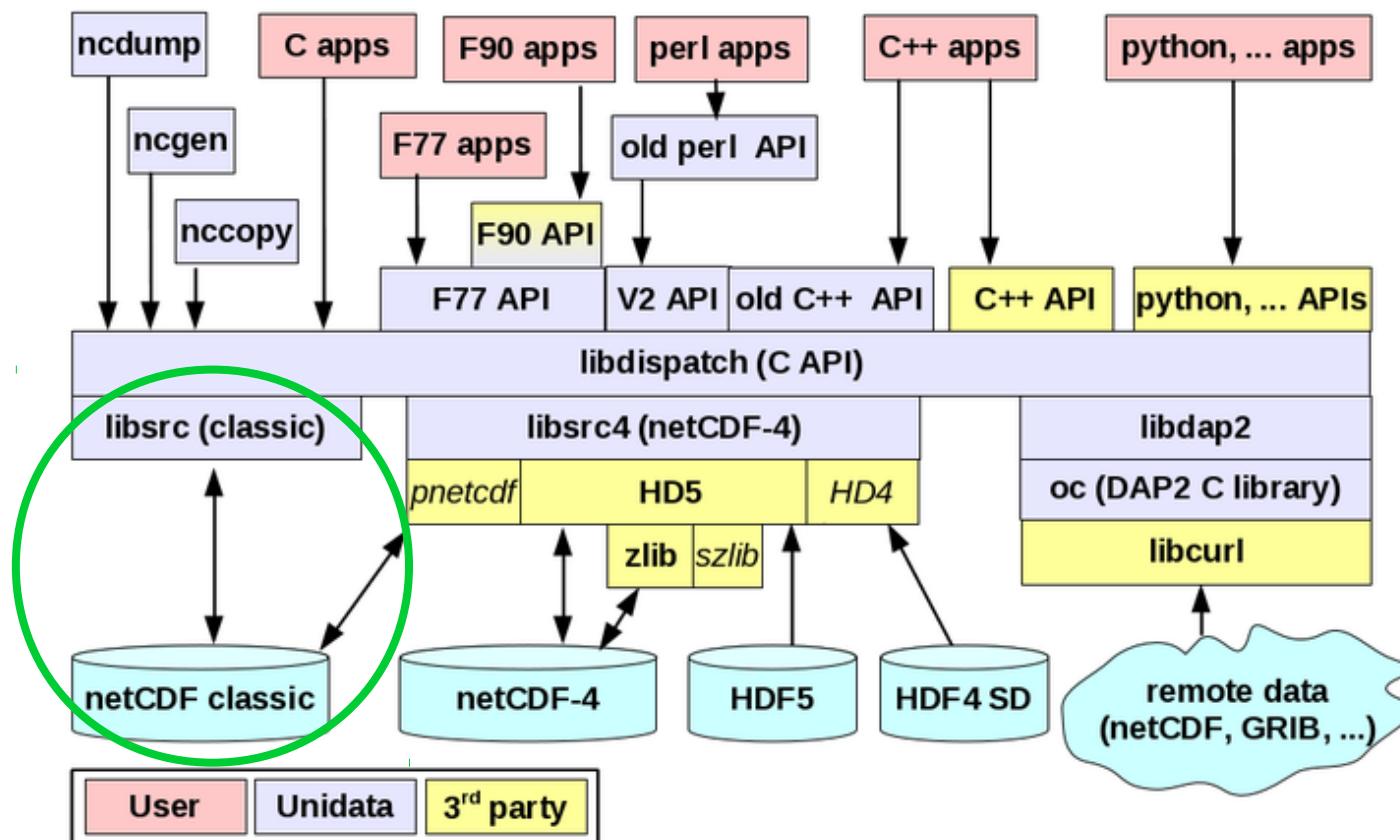
- Provienen de observaciones o son producidos por ESMs, modelos del sistema terrestre
- Son datos multidimensionales



(Image: Maslin and Austin, Nature, 2012, 486, 183)

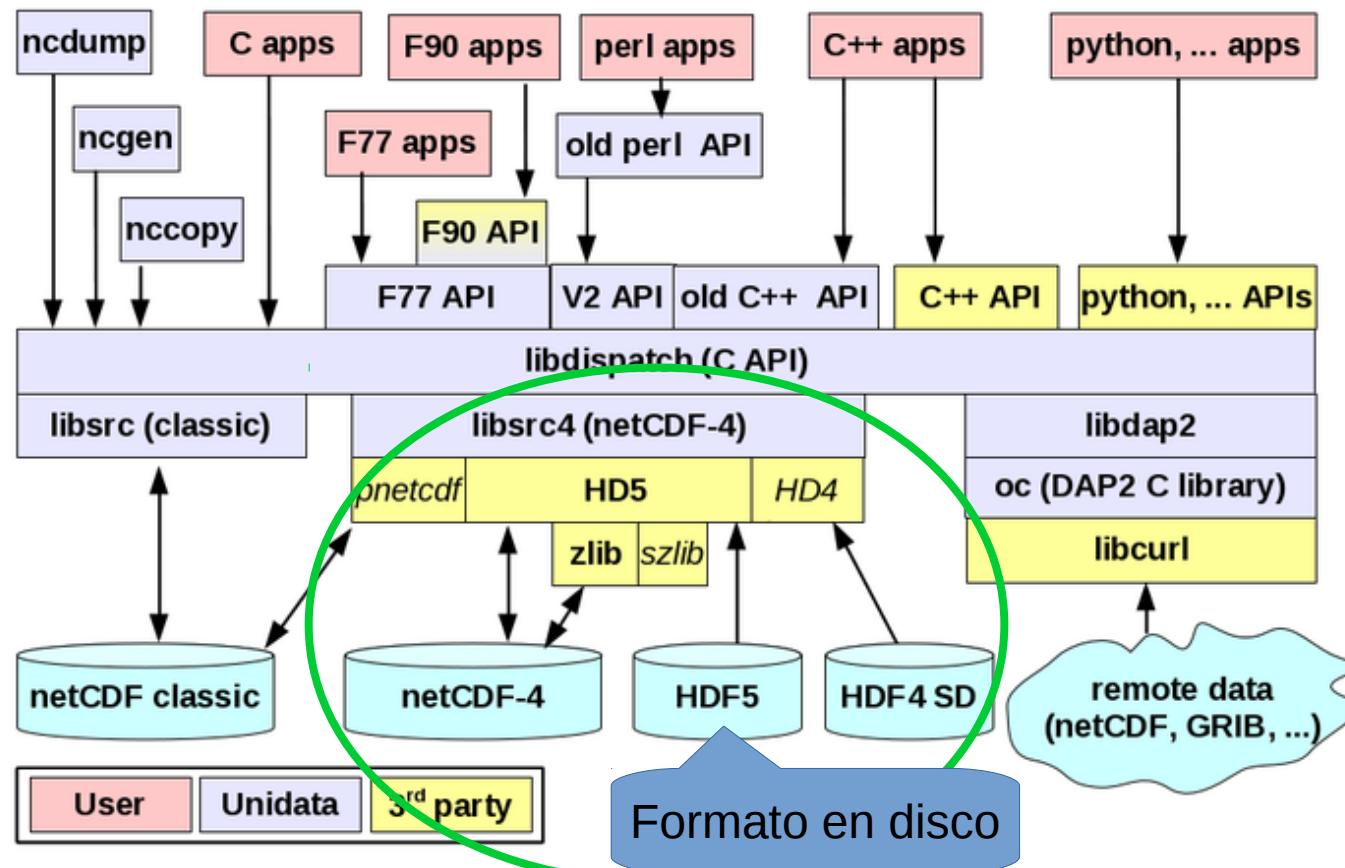
Datos climáticos - NetCDF

- Librería de referencia para leer y escribir datos climáticos, escrita en lenguaje C
- Múltiples formatos del almacenamiento y APIs



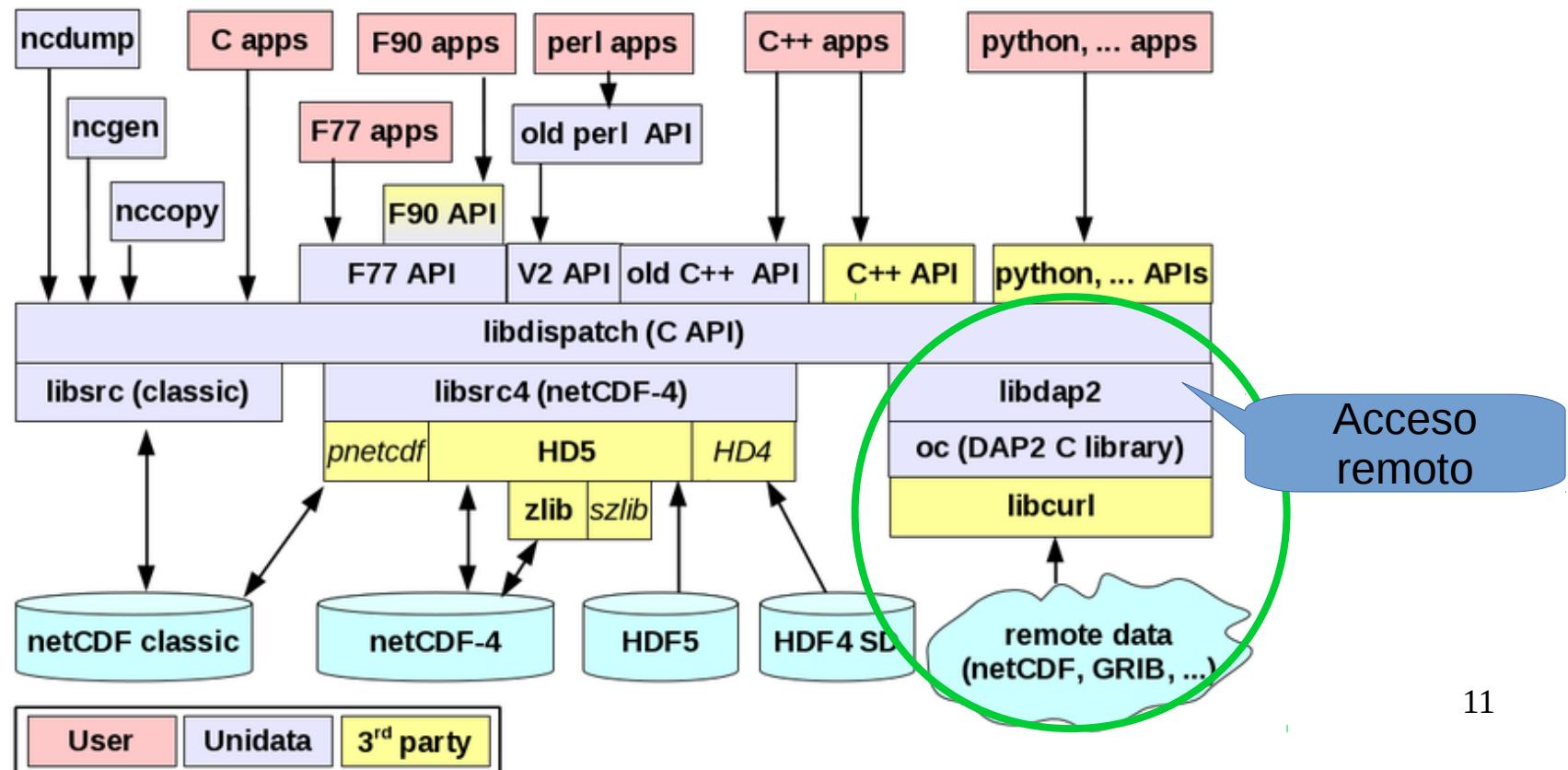
Datos climáticos - NetCDF

- Librería de referencia para leer y escribir datos climáticos, escrita en lenguaje C
- Múltiples formatos del almacenamiento y APIs



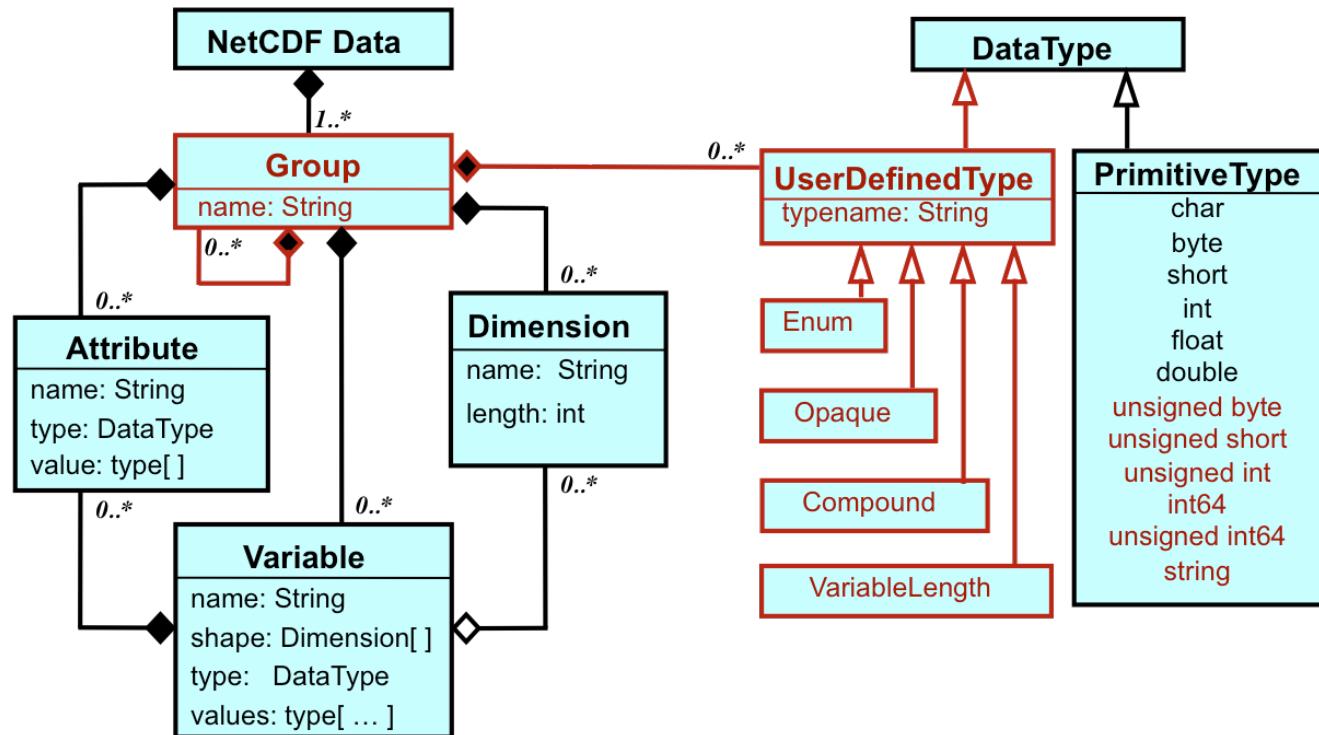
Datos climáticos - NetCDF

- Librería de referencia para leer y escribir datos climáticos, escrita en lenguaje C
- Múltiples formatos del almacenamiento y APIs



Datos climáticos - NetCDF

- Modelo de datos formado por grupos, variables multidimensionales, atributos y tipos de datos



Datos climáticos - NetCDF

```
(netcdf) [zequi@hera tap] $ ncdump -h "http://193.146.75.233:8080/thredds/dodsC/chunked/tas_AERhr_CNR
M-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330.nc"
netcdf tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330 {
dimensions:
time = UNLIMITED ; // (43824 currently)
axis_nbounds = 2 ;
lat = 128 ;
lon = 256 ;
variables:
double lat(lat) ;
lat:axis = "Y" ;
lat:standard_name = "latitude" ;
lat:long_name = "Latitude" ;
lat:units = "degrees_north" ;
double lon(lon) ;
lon:axis = "X" ;
lon:standard_name = "longitude" ;
lon:long_name = "Longitude" ;
lon:units = "degrees_east" ;
double height ;
height:name = "height" ;
height:standard_name = "height" ;
height:long_name = "height" ;
height:units = "m" ;
height:axis = "Z" ;
height:positive = "up" ;
double time(time) ;
time:axis = "T" ;
time:standard_name = "time" ;
time:long_name = "Time axis" ;
time:calendar = "gregorian" ;
time:units = "days since 1850-01-01 00:00:00" ;
time:time_origin = "1850-01-01 00:00:00" ;
time:bounds = "time_bounds" ;
time:_ChunkSizes = 2739 ;
double time_bounds(time, axis_nbounds) ;
time_bounds:_ChunkSizes = 2739, 2 ;
float tas(time, lat, lon) ;
tas:online_operation = "average" ;
tas:cell_methods = "area: time: mean" ;
tas:interval_operation = "900 s" ;
tas:interval_write = "1 h" ;
tas:_FillValue = 1.e+20f ;
tas:missing_value = 1.e+20f ;
tas:coordinates = "height" ;
tas:standard_name = "air_temperature" ;
tas:description = "Temperature at surface" ;
tas:long_name = "Surface Temperature" ;
tas:history = "none" ;
tas:units = "K" ;
tas:cell_measures = "area: areacella" ;
tas:_ChunkSizes = 2739, 8, 32 ;
// global attributes:
:Conventions = "CF-1.7 CMIP-6.2" ;
:creation_date = "2018-09-15T06:24:21Z" ;
:description = "CMIP6 historical" ;
:title = "CNRM-ESM2-1 model output prepared for CMIP6 / CMIP historical" ;
```

Dataset: tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330.nc
Catalog: <http://193.146.75.233:8080/thredds/catalog/chunked/catalog.html>

dataSize	5745297049
id	chunked/tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330.nc

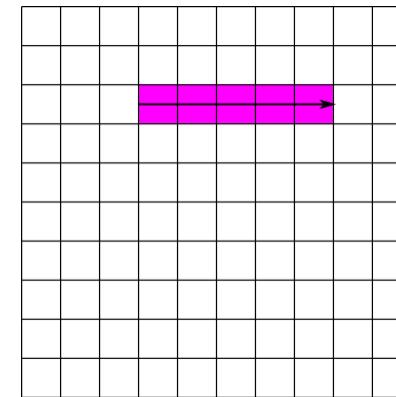
Access **Preview**

Access:

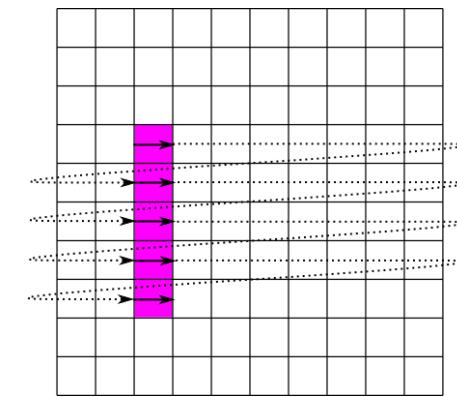
Service	Type	Description
OpenDAP	Data Access	Access dataset through OPeNDAP using the DAP2 protocol.
DAP4	Data Access	Access dataset through OPeNDAP using the DAP4 protocol.
HTTPServer	Data Access	HTTP file download.
WCS	Data Access	Supports access to geospatial data as 'coverages'.
WMS	Data Access	Supports access to georegistered map images from geoscience datasets.
NetcdfSubset	Data Access	A web service for subsetting CDM scientific datasets.
NetcdfSubset	Data Access	A web service for subsetting CDM scientific datasets.
CdmRemote	Data Access	Provides index subsetting on remote CDM datasets, using ncstream.
CdmrFeature	Data Access	Provides coordinate subsetting on remote CDM Feature Datasets, using ncstream.
ISO	Metadata	Provide ISO 19115 metadata representation of a dataset's structure and metadata.
NCML	Metadata	Provide NCML representation of a dataset.
UDDC	Metadata	An evaluation of how well the metadata contained in the dataset conforms to the NetCDF Attribute Convention for Data Discovery (NACDD)

Datos climáticos - Chunking

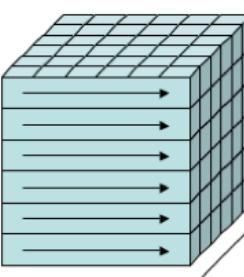
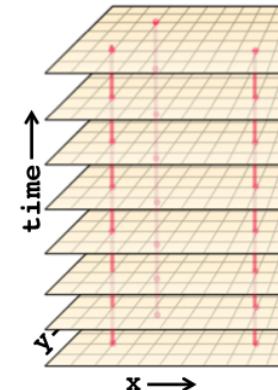
- Chunk – Unidad indivisible de acceso a disco
- Disposición de los datos multidimensionales en el almacenamiento
- Enorme variabilidad en los tiempos de acceso



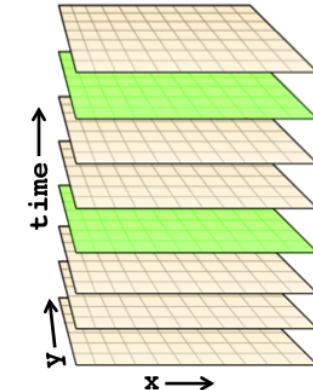
Time series access



Spatial access



index order



chunked

Metodologías de análisis de datos

Descarga local

- Metodología tradicional de análisis de datos
- Imposible de llevar a cabo cuando los datasets superan el almacenamiento local
- Requiere la instalación de software de análisis en el cliente

Dataset: tas_AERhr_CNRM-ESM2-1_hist
Catalog: <http://193.146.75.233:8080/thredds/>

dataSize	5745297049
id	chunked/tas_AERhr_CNRM-ESM2-1_hist

Access **Preview**

Access:

Service	Type
OpenDAP	Data Access
DAP4	Data Access
HTTPServer	Data Access

87% of landh.zip Completed

Retrieving the file
landh.zip from localhost

Estimated time left: 0 sec (1.66 MB of 1.91 MB copied)
Transfer rate: 690 KB/Sec

Cancel

File Edit View History Bookmarks Plot Window Help

New Ctrl+N
Open... Ctrl+O
Open Remote Dataset... Ctrl+L
Open Remote Catalog... Ctrl+Mayús+L
Close Ctrl+W
Save Image Ctrl+S
Save Image As... Ctrl+Mayús+S
Export CL Script...
Export Data Ctrl+Alt+Comilla
Export KMZ... Ctrl+Mayús+K
Export Animation... Ctrl+Mayús+A
Print... Ctrl+P
Quit Panoply Ctrl+Q

Seasonal or Monthly Climatology minus Annual Climatology (degrees Celsius)

-29.6 -23.3 -17.0 -10.7 -4.4 1.9

Data Min = -29.6, Max = 1.9, Mean = -14.1

Plot Map of Array 1 Only Interpol...
Time: 1 of 1 = 0074-12-31 -- 0075-01-01
Depth: 1 of 33 = 0.00000 m

Servicios de análisis de datos

- Acercar la computación a los datos
- Python Jupyter Notebook y JupyterHub
- Web Processing Services, OGC

The screenshot shows a Jupyter Notebook interface with two code cells and a plot.

Lorenz.ipynb:

```
File Edit View Run Kernel Tabs Settings Help
File Running Commands Cell Tools Tabs File Edit View Run Kernel Tabs Settings Help
Lorenz.ipynb x lorenz.py x
We explore the Lorenz system of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

Let's change  $(\sigma, \beta, \rho)$  with ipywidgets and examine the trajectories.
In [2]: from lorenz import solve_lorenz
w=interactive(solve_lorenz,sigma=(0.0,50.0),rho=(0.0,50.0))
w
```

Controls for sigma, beta, and rho:

sigma	10.00
beta	2.67
rho	28.00

lorenz.py:

```
def solve_lorenz(sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""
    max_time = 4.0
    N = 30

    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')

    # prepare the axes limits
    ax.set_xlim((-25, 25))
    ax.set_ylim((-35, 35))
    ax.set_zlim((5, 55))

    def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
        """Compute the time-derivative of a Lorenz system."""
        x, y, z = x_y_z
        return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]

    # Choose random starting points, uniformly distributed from -15 to 15
    np.random.seed(1)
    x0 = -15 + 30 * np.random(N, 3)

    # Solve for the trajectories
    t = np.linspace(0, max_time, int(250*max_time))
    x_t = np.asarray([integrate.odeint(lorenz_deriv, x0i, t)
                     for x0i in x0])

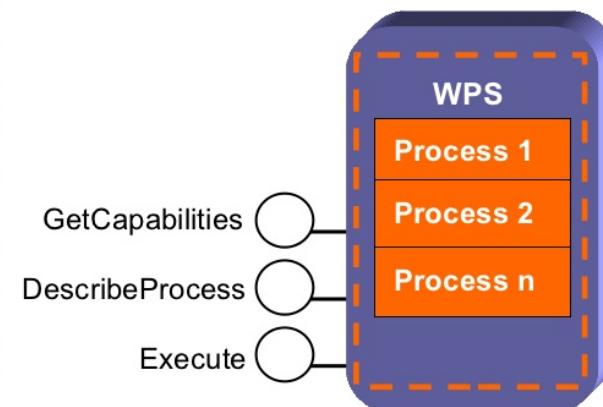
    # choose a different color for each trajectory
    colors = plt.cm.viridis(np.linspace(0, 1, N))

    for i in range(N):
        x, y, z = x_t[:, :, i].T
        lines = ax.plot(x, y, z, 'r-', c=colors[i])
        plt.setp(lines, linewidth=2)
        angle = 104
        ax.view_init(30, angle)
```

A 3D plot shows the Lorenz attractor with multiple colored trajectories.

Web-based Geoprocessing

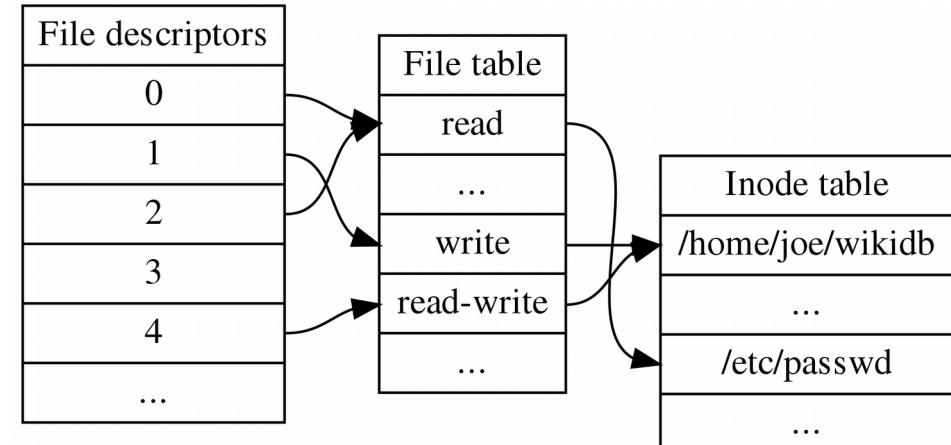
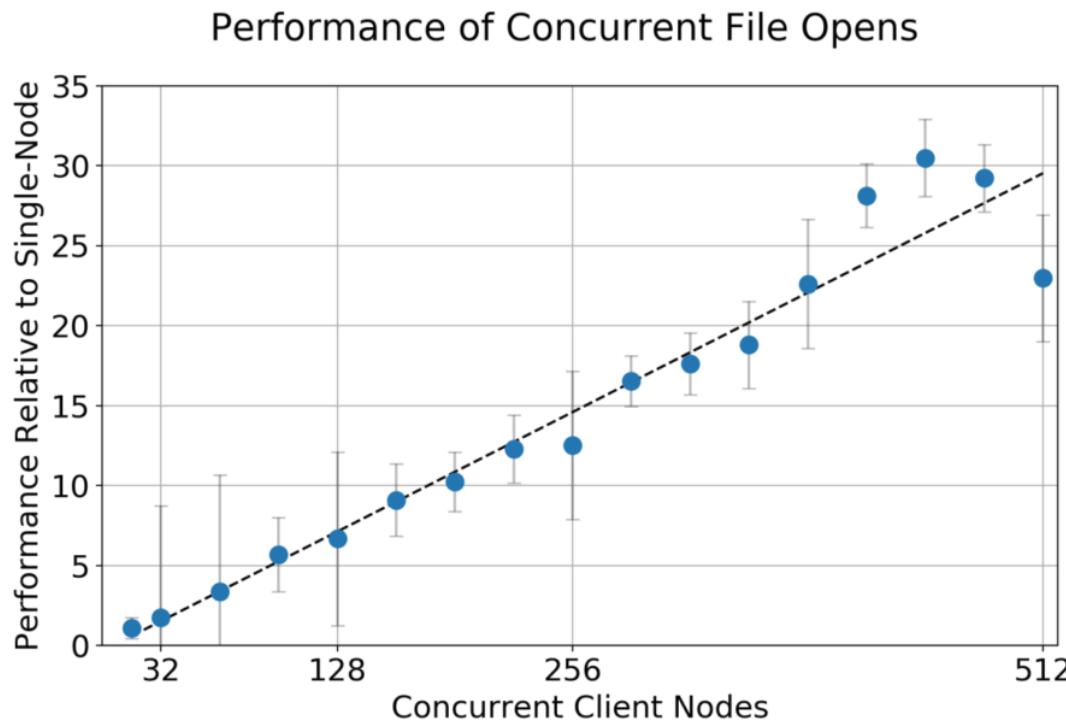
Data → Information



Sistemas de almacenamiento

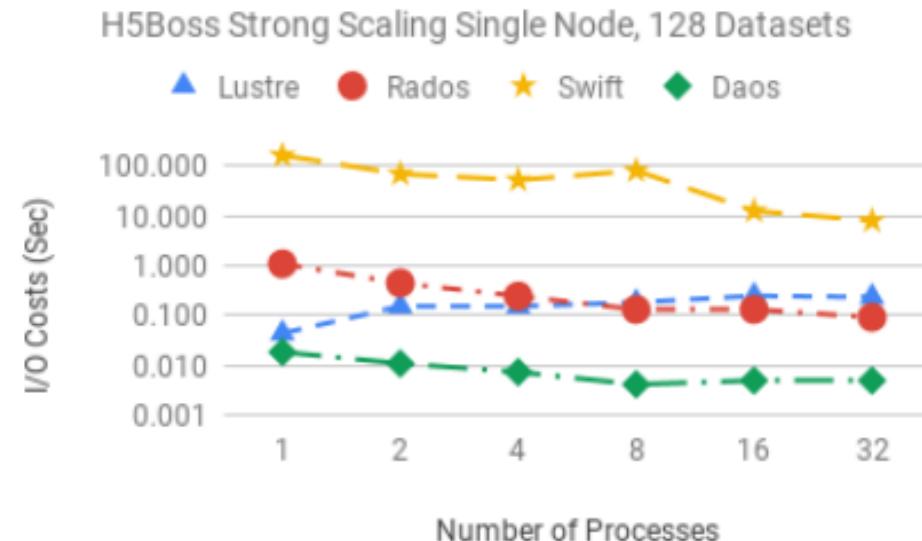
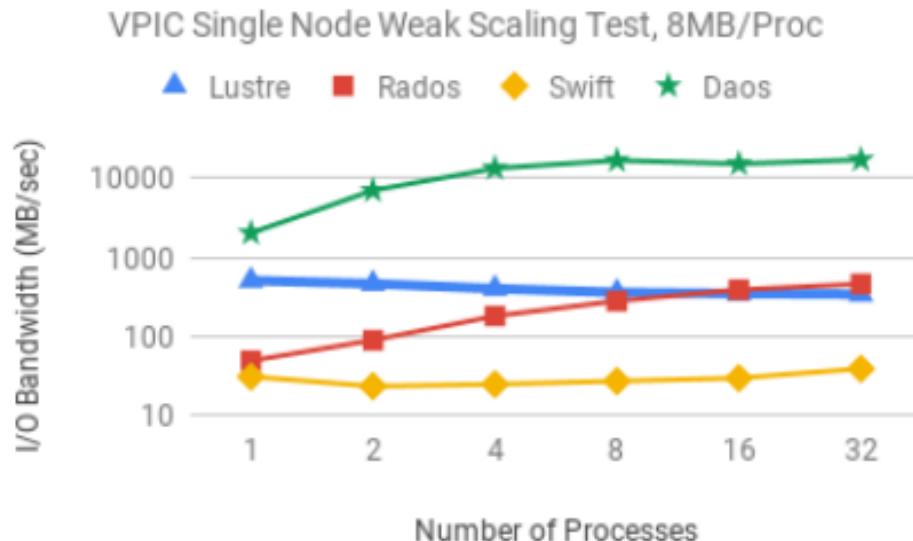
Sistemas de ficheros POSIX

- Bloqueo entre procesos paralelos debido a las semánticas de fuerte consistencia
- Limitaciones de escalabilidad en sistemas de ficheros en paralelo (Lustre, GPFS)



Object storage

- Espacio de nombres plano sin metadatos
- Acceso mediante operaciones atómicas sin estado
- Inmutabilidad de los objetos



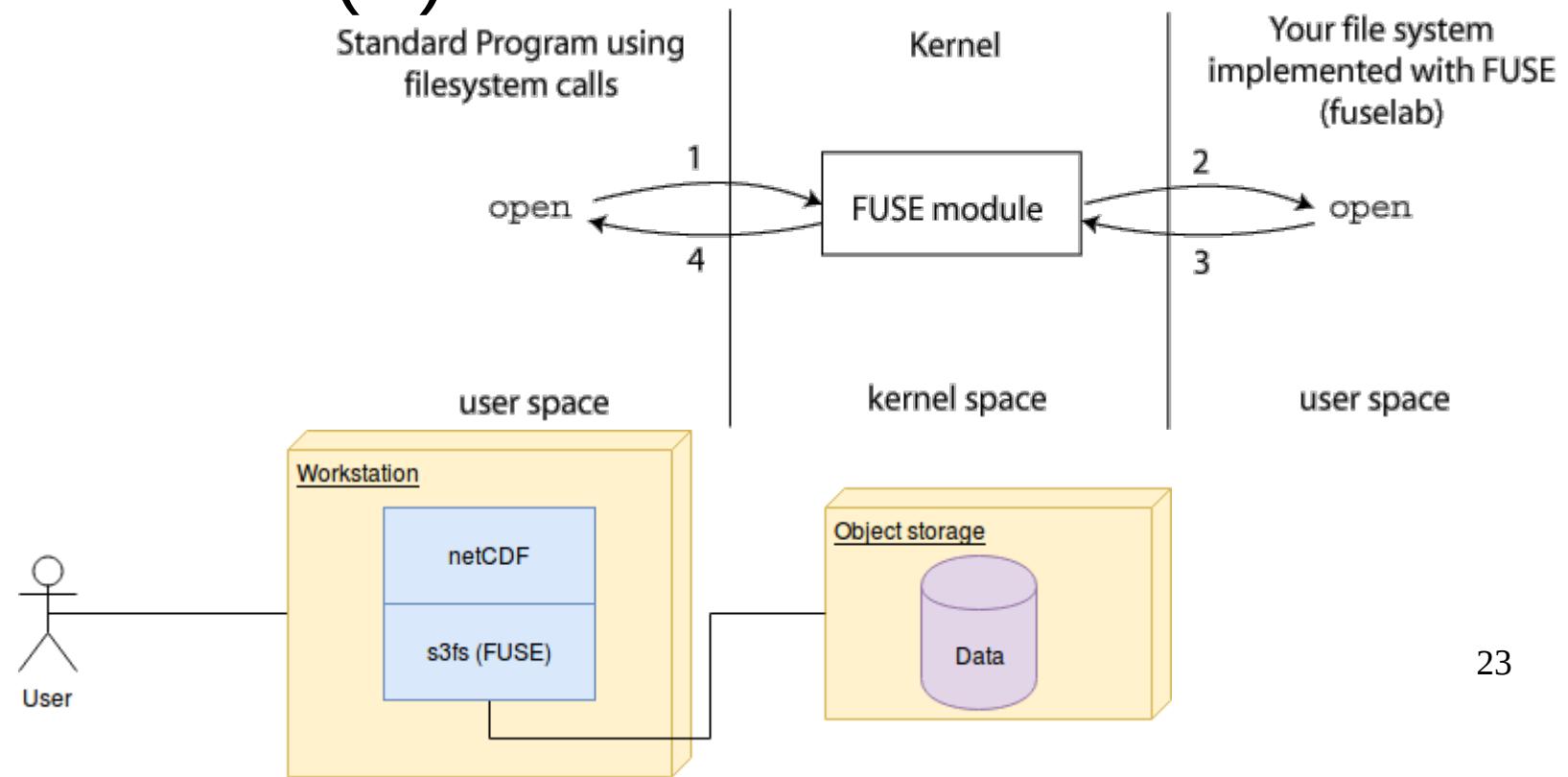
HDF5 cloud y Zarr

HDF5 cloud

- HDF5 no puede acceder a datos almacenados en object storage
- Posibles soluciones:
 - FUSE
 - Virtual Object Layer
 - HSDS

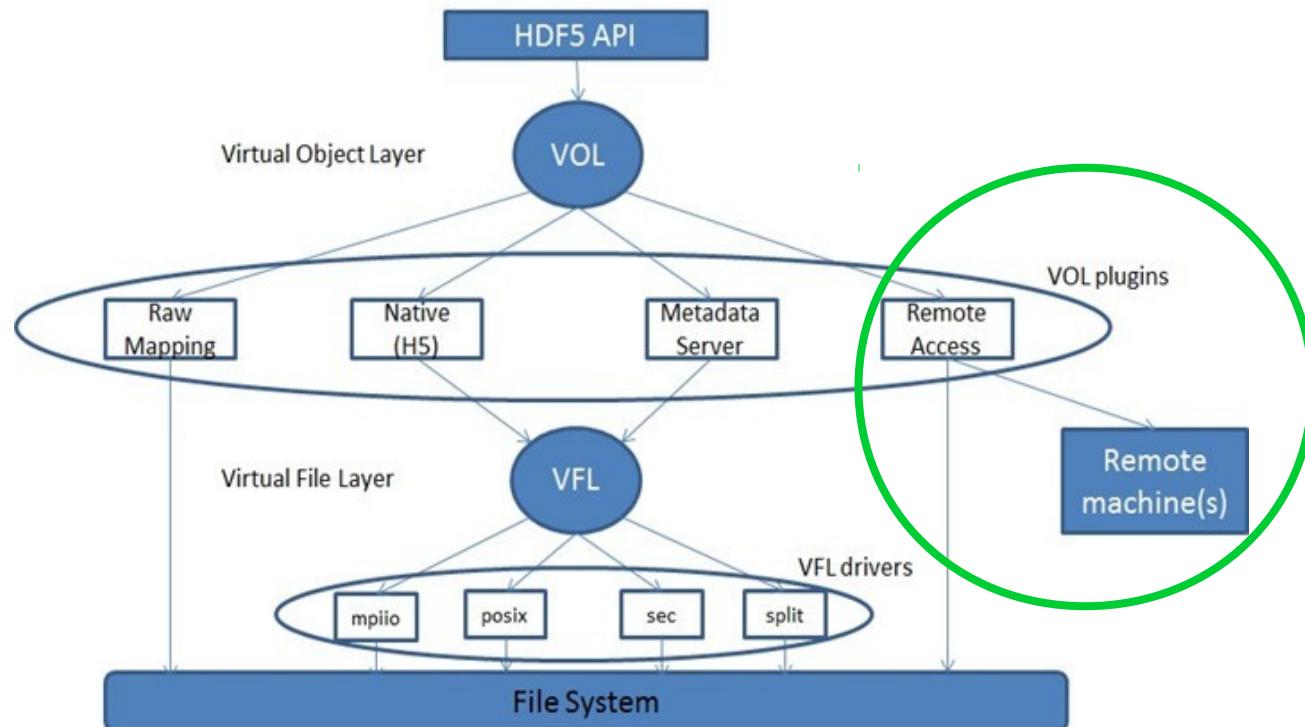
HDF5 cloud - FUSE

- Módulo que simula un sistema de ficheros sobre un object storage
- No acceso aleatorio, consistencia eventual, latencia de red (`ls`)...



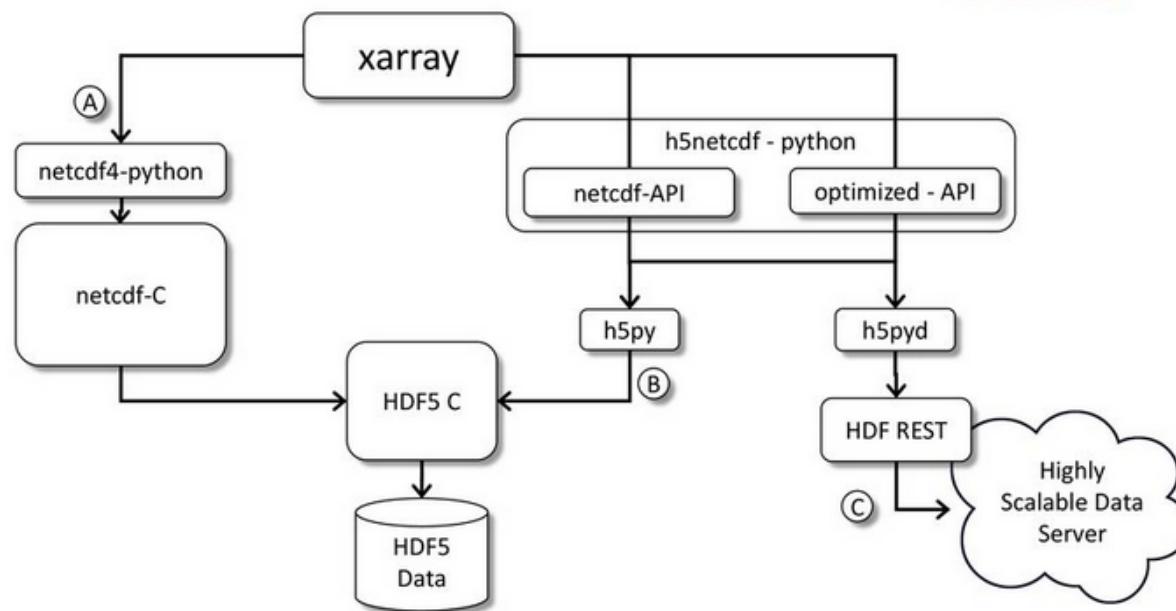
HDF5 cloud – Virtual Object Layer

- Requiere el desarrollo de un plugin que haga de interfaz con el object storage
- Solo existen pruebas de concepto, complejidad del modelo HDF5



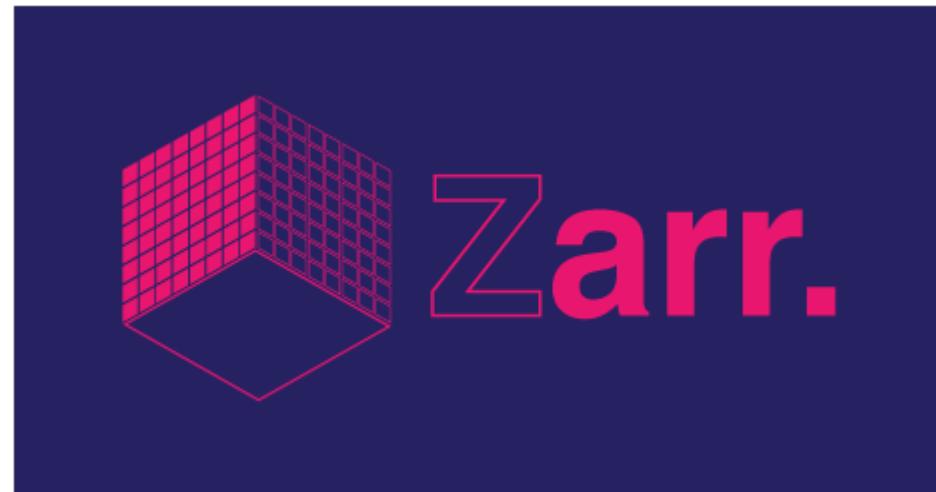
HDF5 cloud - HSDS

- API REST/HTTP que representa el modelo de datos HDF5
- Añade un intermediario entre los analistas y los datos



Zarr

- Librería escrita en Python para almacenamiento de arrays multidimensionales
- Modelo de datos muy similar a netCDF



Zarr

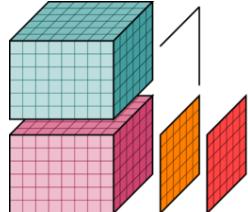
- Concepto de ‘store’ o almacén frente a fichero
- Encaja tanto en sistemas de ficheros como en object storage
- Metadatos y chunks se almacenan en objetos distintos

```
[ecimadevilla@altamira1 tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330]$ ls -a
. . . height lat lon tas time time_bounds .zattrs .zgroup
[ecimadevilla@altamira1 tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330]$ ls -a tas/
. 0.4.4 10.13.4 10.9.0 1.11.6 11.7.4 1.2.2 13.10.5 1.3.6 14.14.5 15.0.1 15.4.3 2.0.6
.. 0.4.5 10.13.5 10.9.1 11.1.6 11.7.5 12.2.0 13.10.6 13.6.0 14.14.6 15.0.2 15.4.4 2.0.7
0.0.0 0.4.6 10.13.6 10.9.2 1.11.7 11.7.6 12.2.1 13.10.7 13.6.1 14.14.7 15.0.3 15.4.5 2.1.0
0.0.1 0.4.7 10.13.7 10.9.3 11.1.7 11.7.7 12.2.2 13.1.1 13.6.2 14.1.5 15.0.4 15.4.6 2.10.0
0.0.2 0.5.0 10.1.4 10.9.4 1.1.2 11.8.0 12.2.3 13.11.0 13.6.3 14.15.0 15.0.5 15.4.7 2.10.1
```

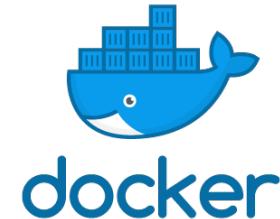
Evaluaciones HPC y cloud

Evaluaciones HPC y cloud

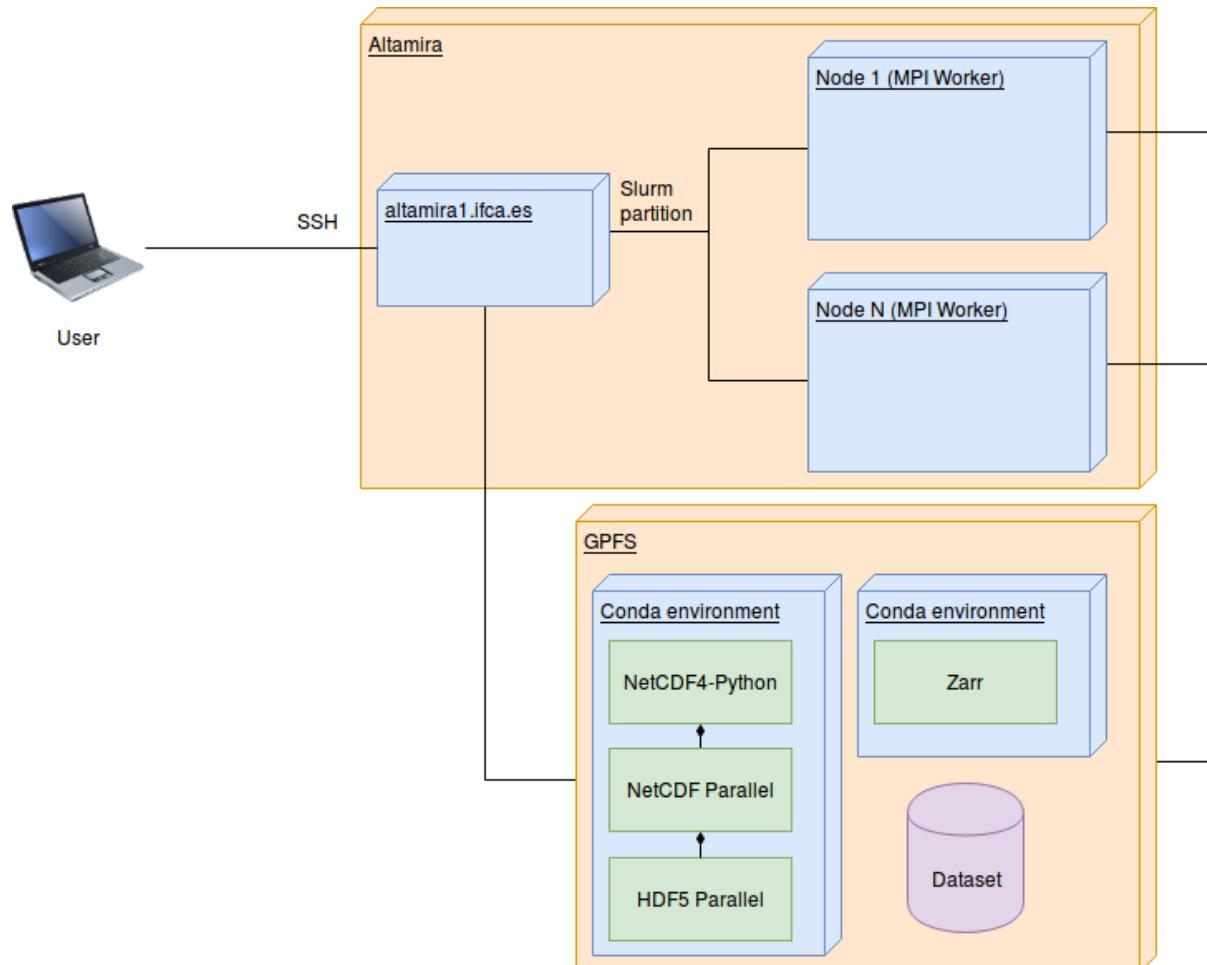
- El objetivo es desarrollar pruebas de concepto que muestren el estado del arte para cada tipo de almacenamiento
- Visualización preliminar de la eficiencia de cada tipo de acceso
- Test sobre tres infraestructuras distintas, 1 HPC y 2 cloud
- Comparación de acceso en serie y paralelo



xarray

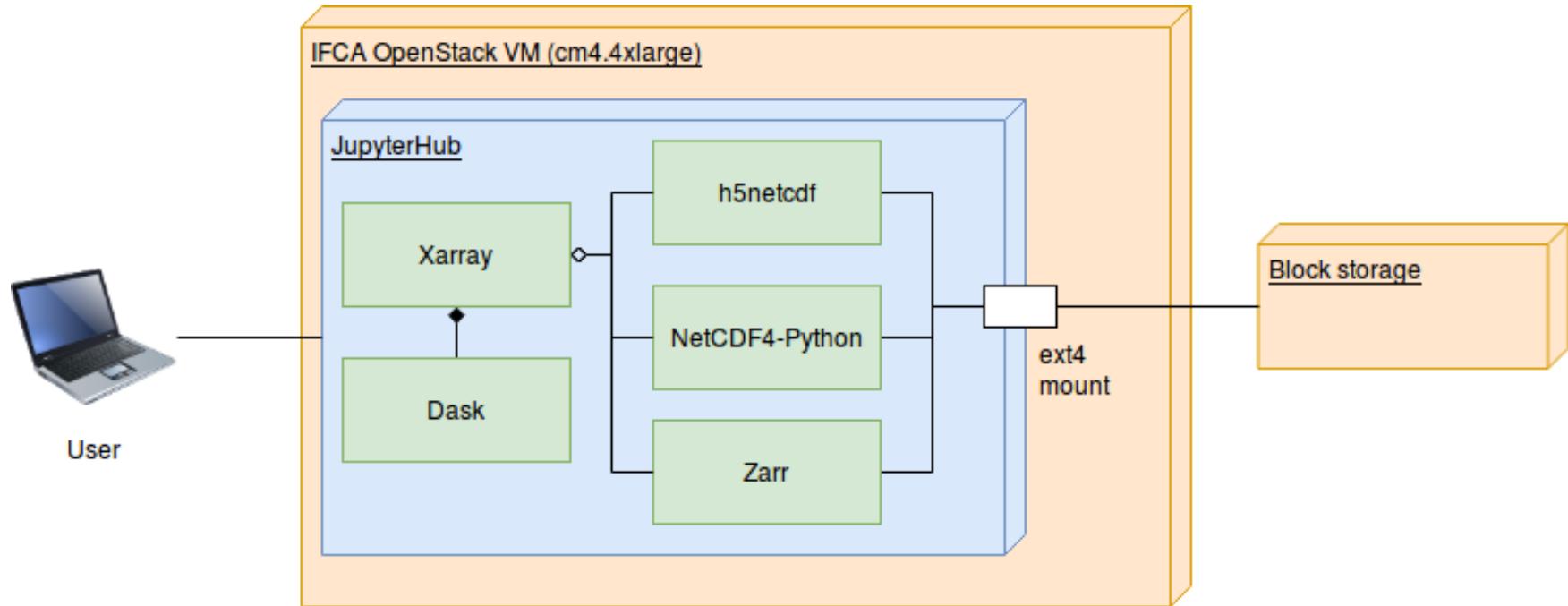


Acceso local HPC



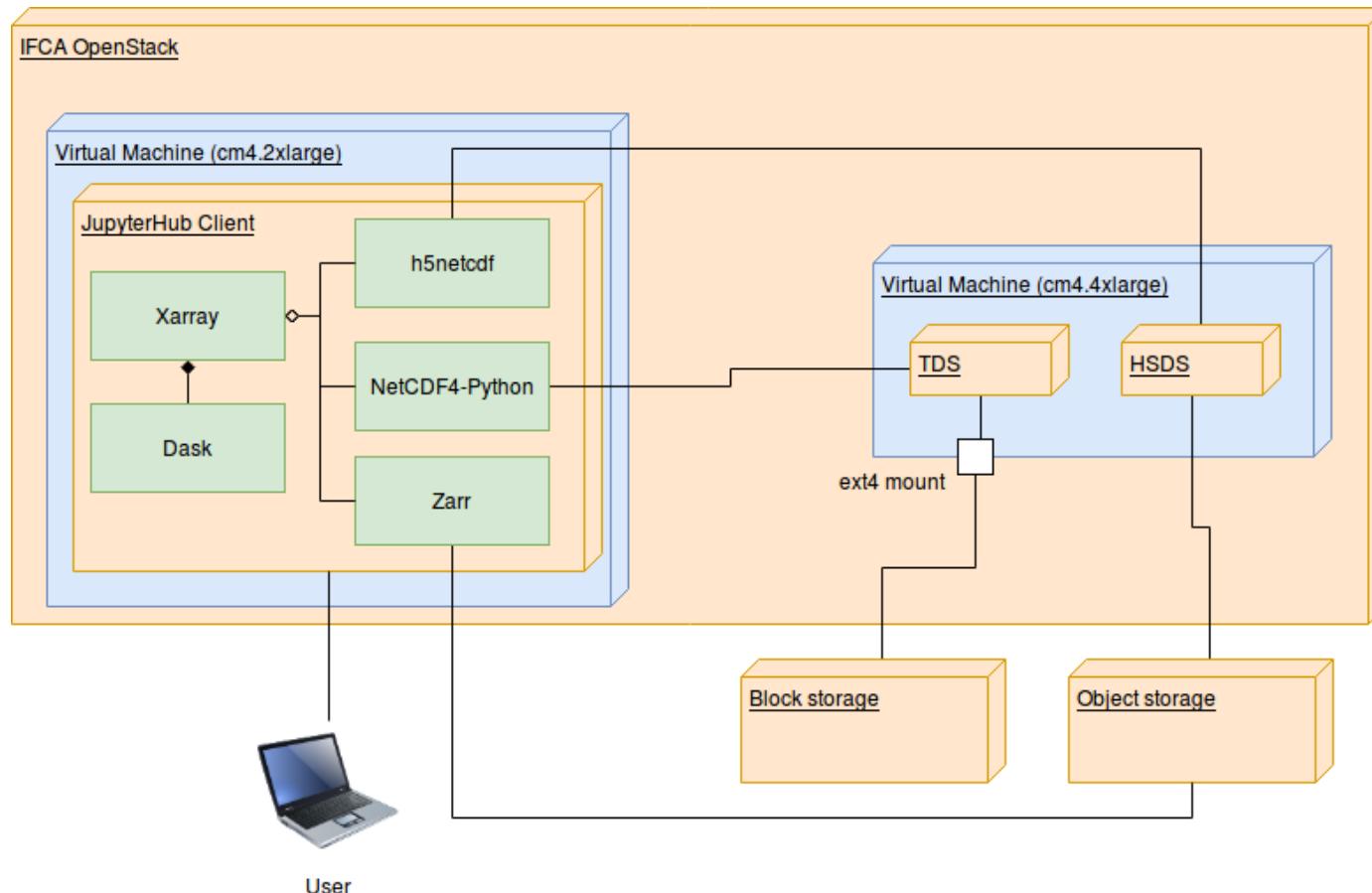
Librería / Tasks	NetCDF4 Independent	NetCDF4 Collective	Zarr
2	145,04s	-	86,82s
4	80,22s	29,52s	37,35s
8	39,73s	17,75s	14,43s

Acceso local cloud



Librería / Acceso	netCDF4	Zarr	h5netcdf
Serie	146,2s	148,8s	128,7s
Threads	90,5s	49,1s	109,1s
Speed up	1,61	3	1,17

Acceso remoto cloud



Librería / Acceso	netCDF4 - TDS	Zarr	h5netcdf - HSDS
Serie	302,9s	1273,6s	-
Threads	279,7s	97s	-
Speed up	1,08	13,12	-

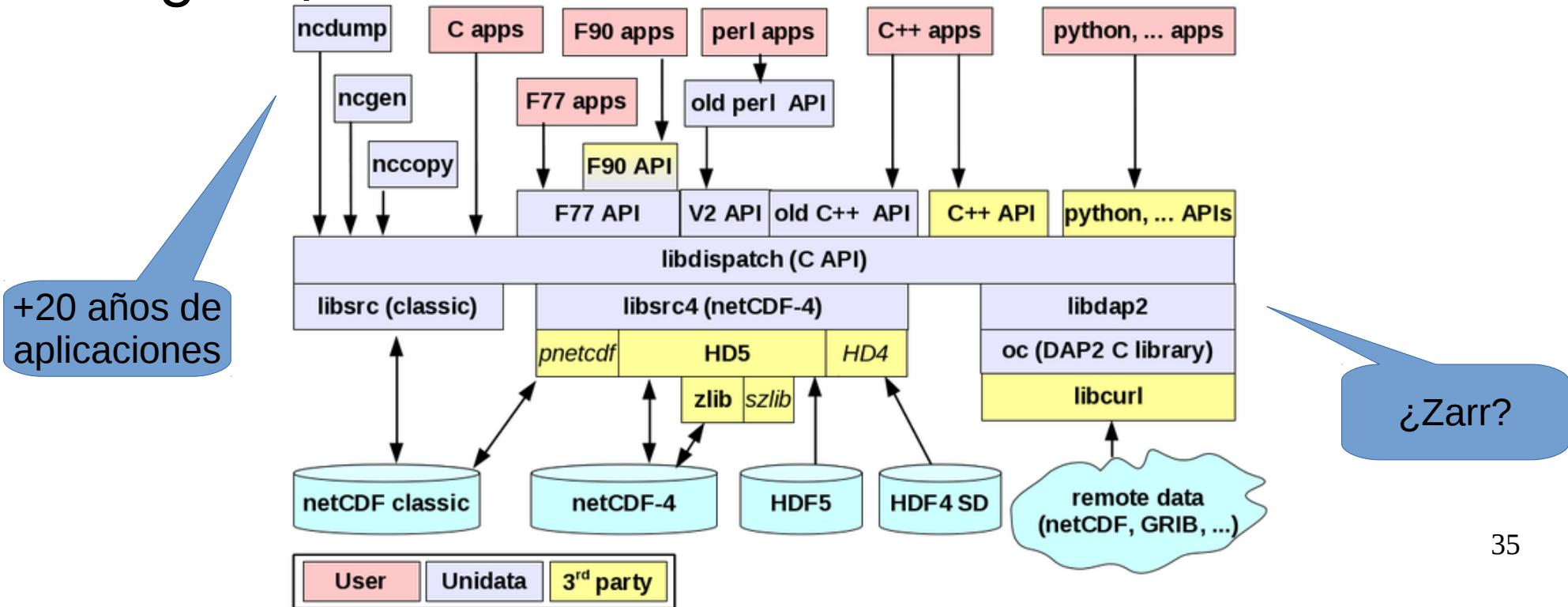
Conclusiones y trabajo futuro

Conclusiones

- Introducción a los sistemas de ficheros en paralelo y al object storage (ver [8] J. Lieu et al.)
- Mecánicas internas (partición, consistencia, POSIX)
- Compilación de librerías (HDF5, netCDF-C, netCDF4, mpi4py)
- Introducción a herramientas de análisis (xarray, Dask) (MPI)
- Agradecimiento especial a la interfaz S3 de Ceph
- Valor añadido del TFM - “Build your own Pangeo”

Conclusiones

- El movimiento de los datos a object storage (cloud) forma parte del presente
 - ¿Adaptar netCDF a object storage?
 - ¿Adoptar una nueva librería en la comunidad?



Trabajo futuro

- Explicación de las diferencias en los tiempos de acceso
- Extensión del entorno cloud a un clúster en el que realizar paralelismo distribuido
- Evaluación de casos de uso más complejos de minería de datos o machine learning
- Uso de un dataset de mayor tamaño en las pruebas

Evaluación de las tecnologías object storage para almacenamiento y análisis de datos climáticos

Autor: Ezequiel Cimadevilla Álvarez

Director: Antonio S. Cofiño González
Codirector: Aida Palacio Hoz

Máster en Ciencia de Datos
Universidad de Cantabria
10 Julio 2019

Grupo de Meteorología de Santander
Grupo de Computación Avanzada

