EUDAT Collaborative Data Infrastructure

PRACE

Real Use Case Session (Part I)

The ENES Climate Analytics Service

***Sandro Fiore**, Ph.D. – CMCC Foundation*

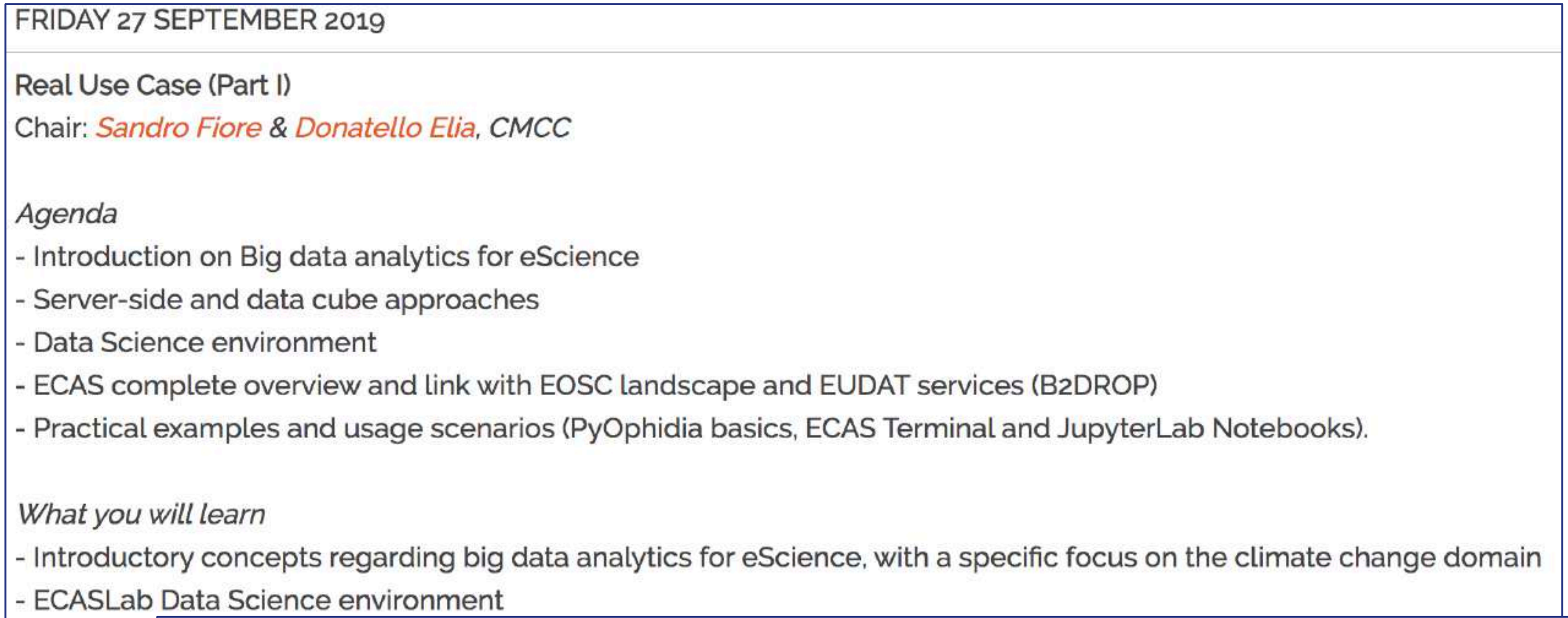
***Donatello Elia** – CMCC Foundation and University of Salento*

On behalf of the ECAS Team

cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici

DKRZ DEUTSCHES KLIMARECHENZENTRUM

Agenda



FRIDAY 27 SEPTEMBER 2019

Real Use Case (Part I)

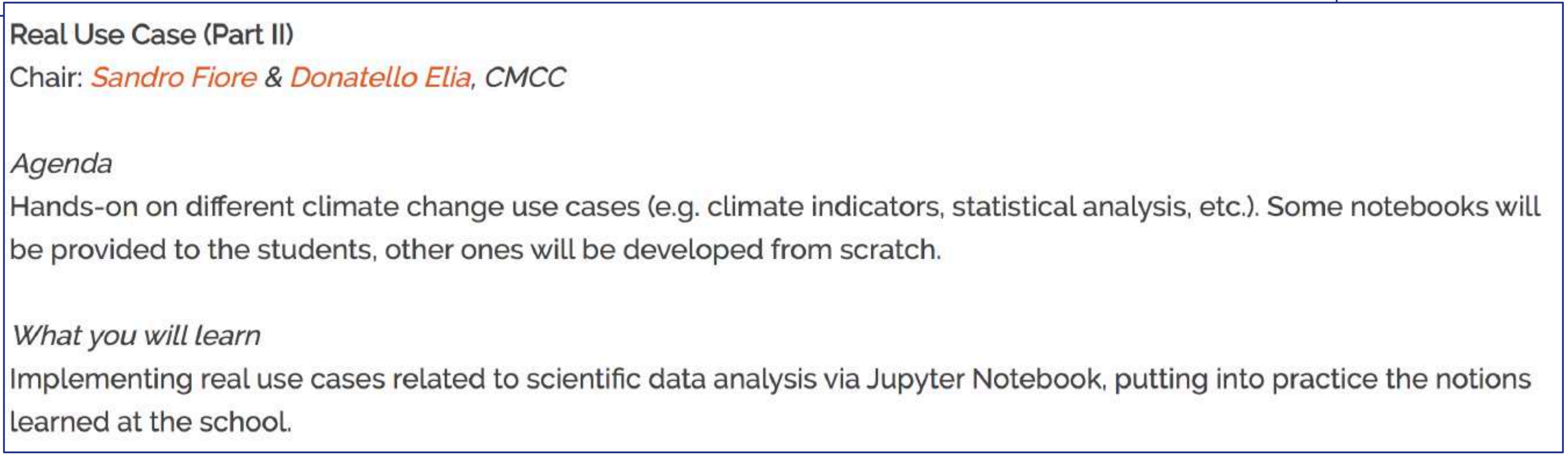
Chair: *Sandro Fiore* & *Donatello Elia*, *CMCC*

Agenda

- Introduction on Big data analytics for eScience
- Server-side and data cube approaches
- Data Science environment
- ECAS complete overview and link with EOSC landscape and EUDAT services (B2DROP)
- Practical examples and usage scenarios (PyOphidia basics, ECAS Terminal and JupyterLab Notebooks).

What you will learn

- Introductory concepts regarding big data analytics for eScience, with a specific focus on the climate change domain
- ECASLab Data Science environment



Real Use Case (Part II)

Chair: *Sandro Fiore* & *Donatello Elia*, *CMCC*

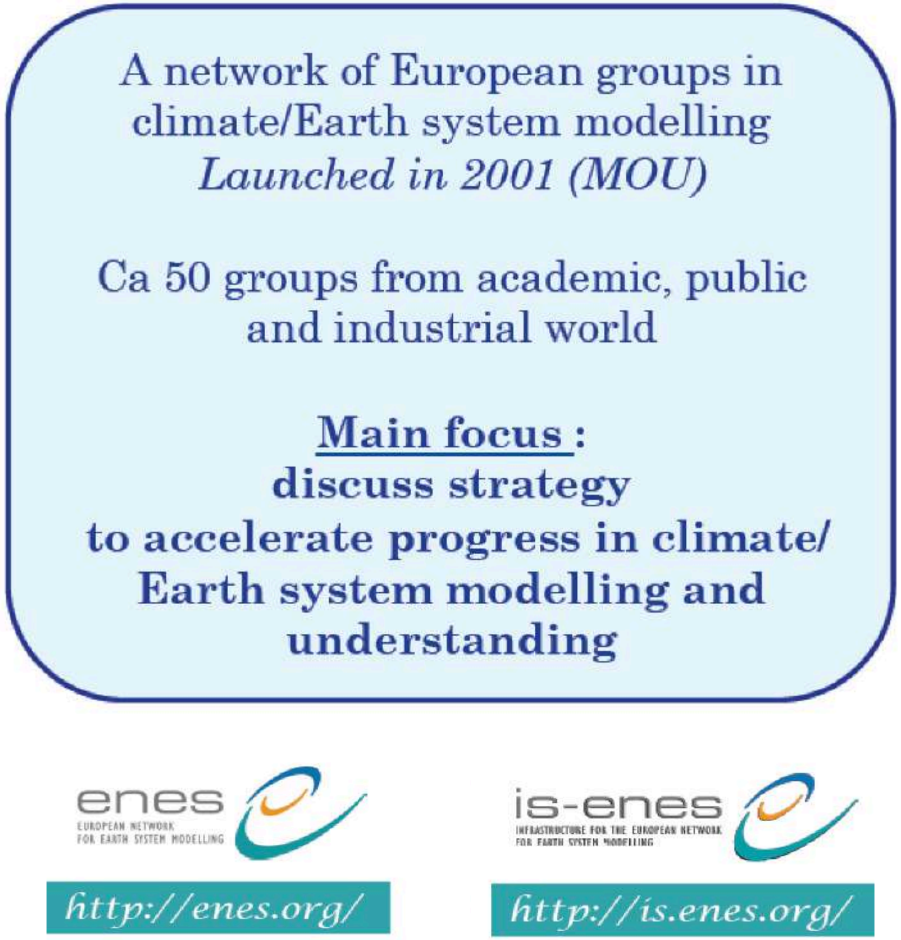
Agenda

Hands-on on different climate change use cases (e.g. climate indicators, statistical analysis, etc.). Some notebooks will be provided to the students, other ones will be developed from scratch.

What you will learn

Implementing real use cases related to scientific data analysis via Jupyter Notebook, putting into practice the notions learned at the school.

ENES
European Network for Earth System Modelling



A network of European groups in climate/Earth system modelling
Launched in 2001 (MOU)

Ca 50 groups from academic, public and industrial world

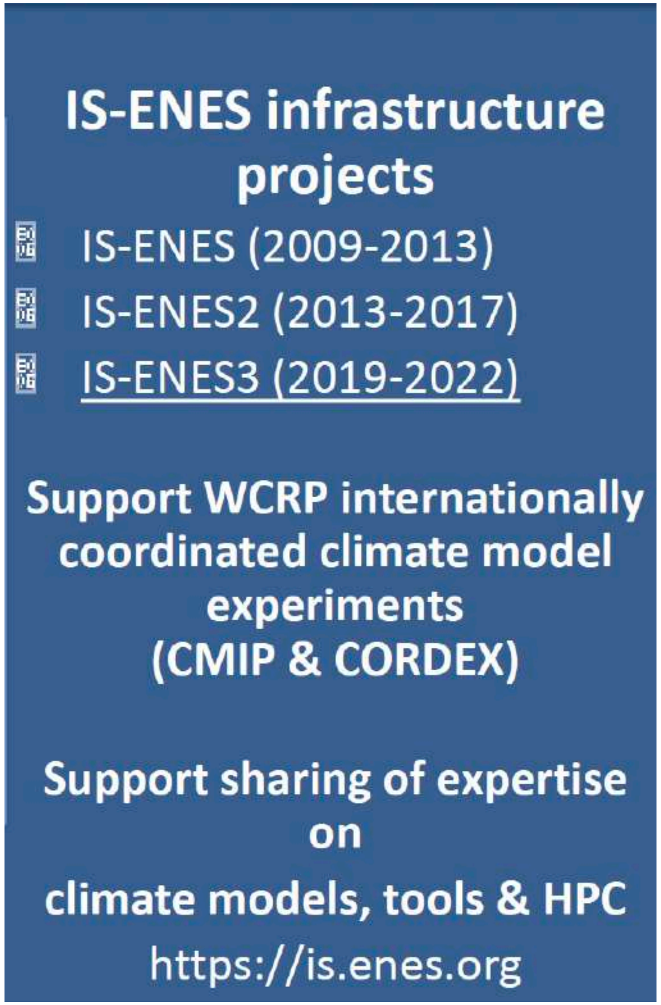
Main focus :
discuss strategy
to accelerate progress in climate/ Earth system modelling and understanding

enes
EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

http://enes.org/

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

http://is.enes.org/



IS-ENES infrastructure projects

- IS-ENES (2009-2013)
- IS-ENES2 (2013-2017)
- IS-ENES3 (2019-2022)

Support WCRP internationally coordinated climate model experiments (CMIP & CORDEX)

Support sharing of expertise on climate models, tools & HPC

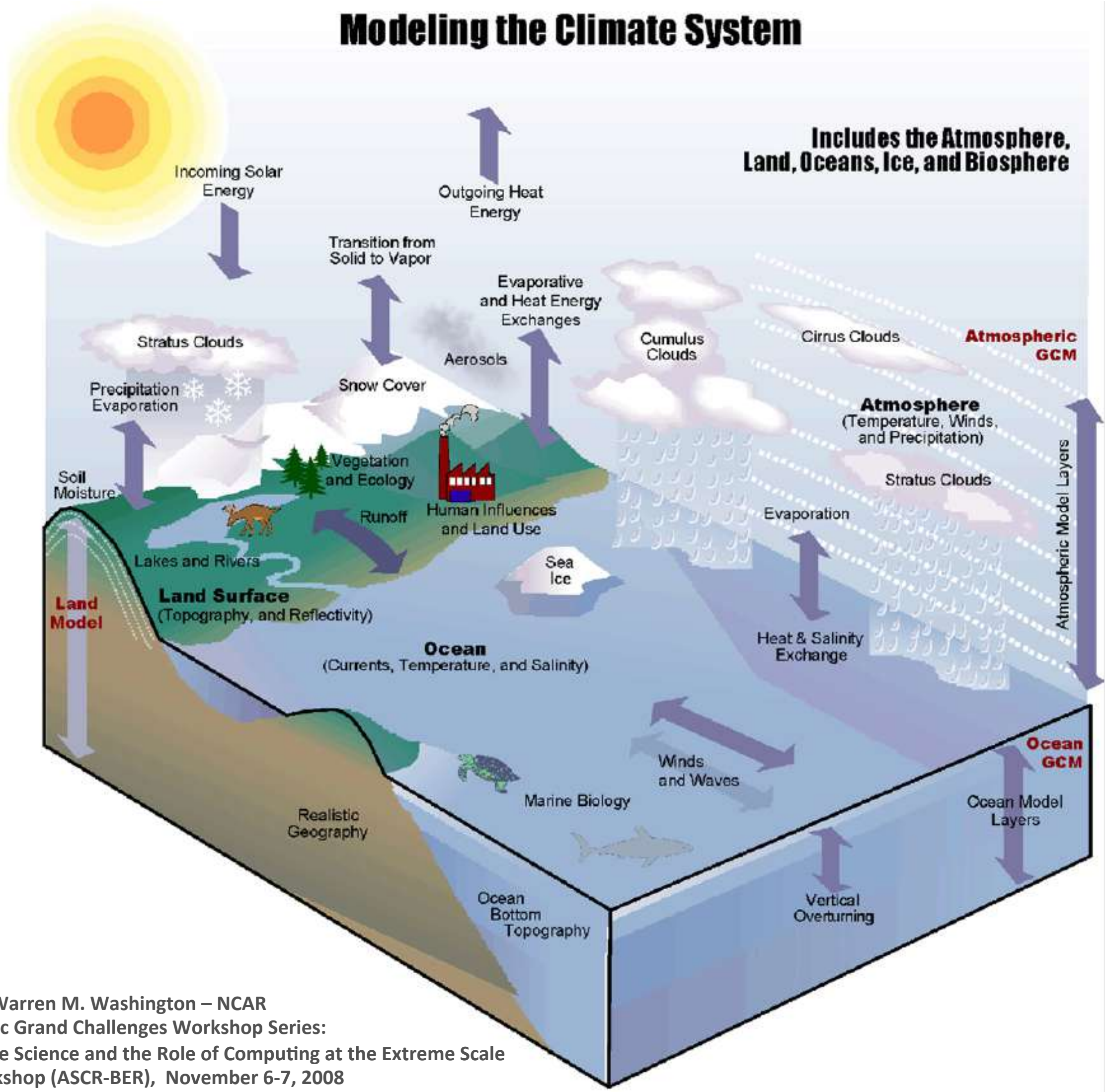
https://is.enes.org

Slide courtesy: Sylvie Joussaume

Slide courtesy: Sylvie Joussaume (IPSL)

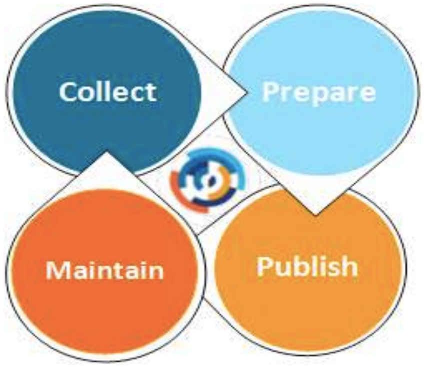
Modelling the Climate System

- Several **complex processes** to be simulated
- Several **interacting processes**
- Great range of **time scales** to be analyzed
- Great range of **spatial scales** to be considered
- Need **interdisciplinar Science** (physics, chemistry, biology, geology,…)
- Inherently **non-linear governing equations**
- Need **sophisticated numerics**
- Need a lot of **computational resources**
- **….and huge volumes of data can be produced and large datasets need to be analyzed, published, distributed, curated, ...**

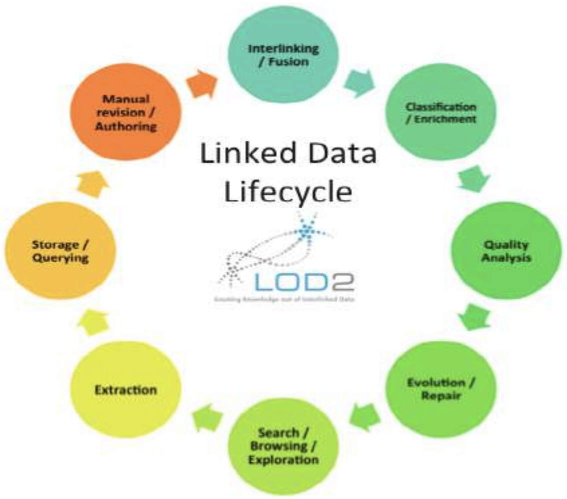


Warren M. Washington – NCAR
Scientific Grand Challenges Workshop Series:
Challenges in Climate Change Science and the Role of Computing at the Extreme Scale
DOE Workshop (ASCR-BER), November 6-7, 2008

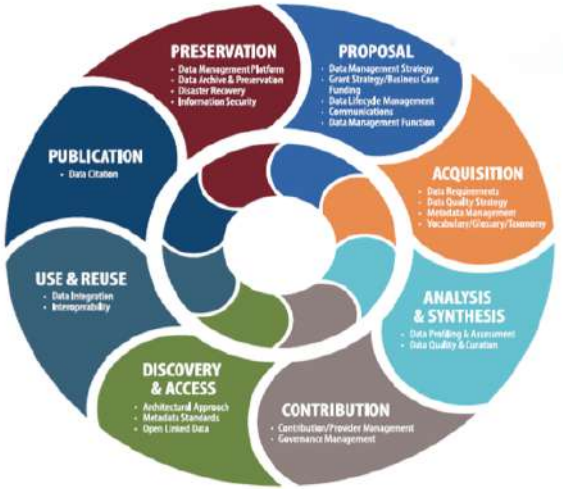
Data Lifecycles… Simple to Complex



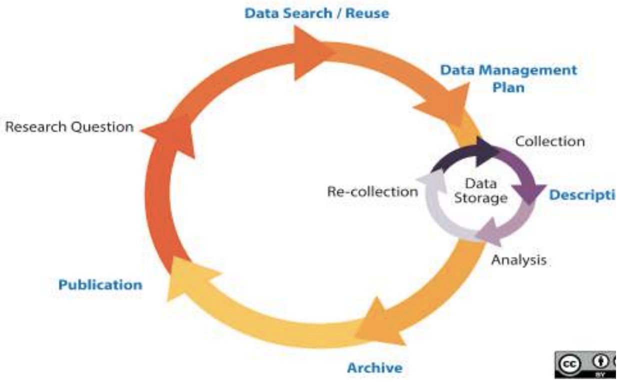
European Data Portal



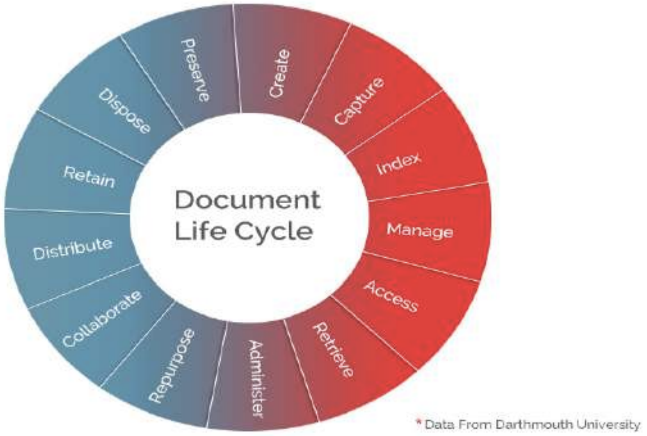
Sören Auer (2011) "The Semantic Data Web"



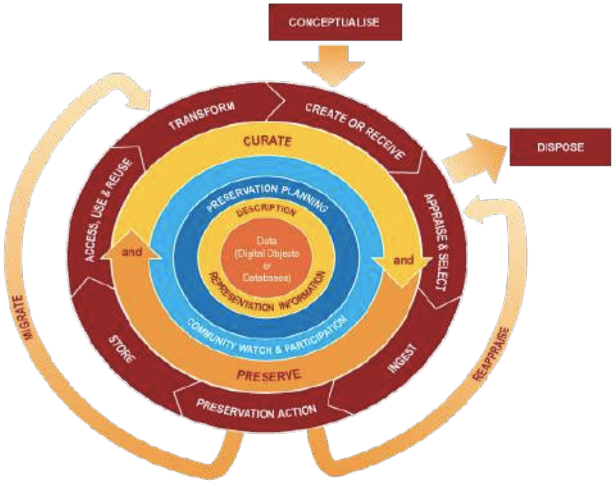
AGU Data Maturity Model



UCSC Data Lifecycle



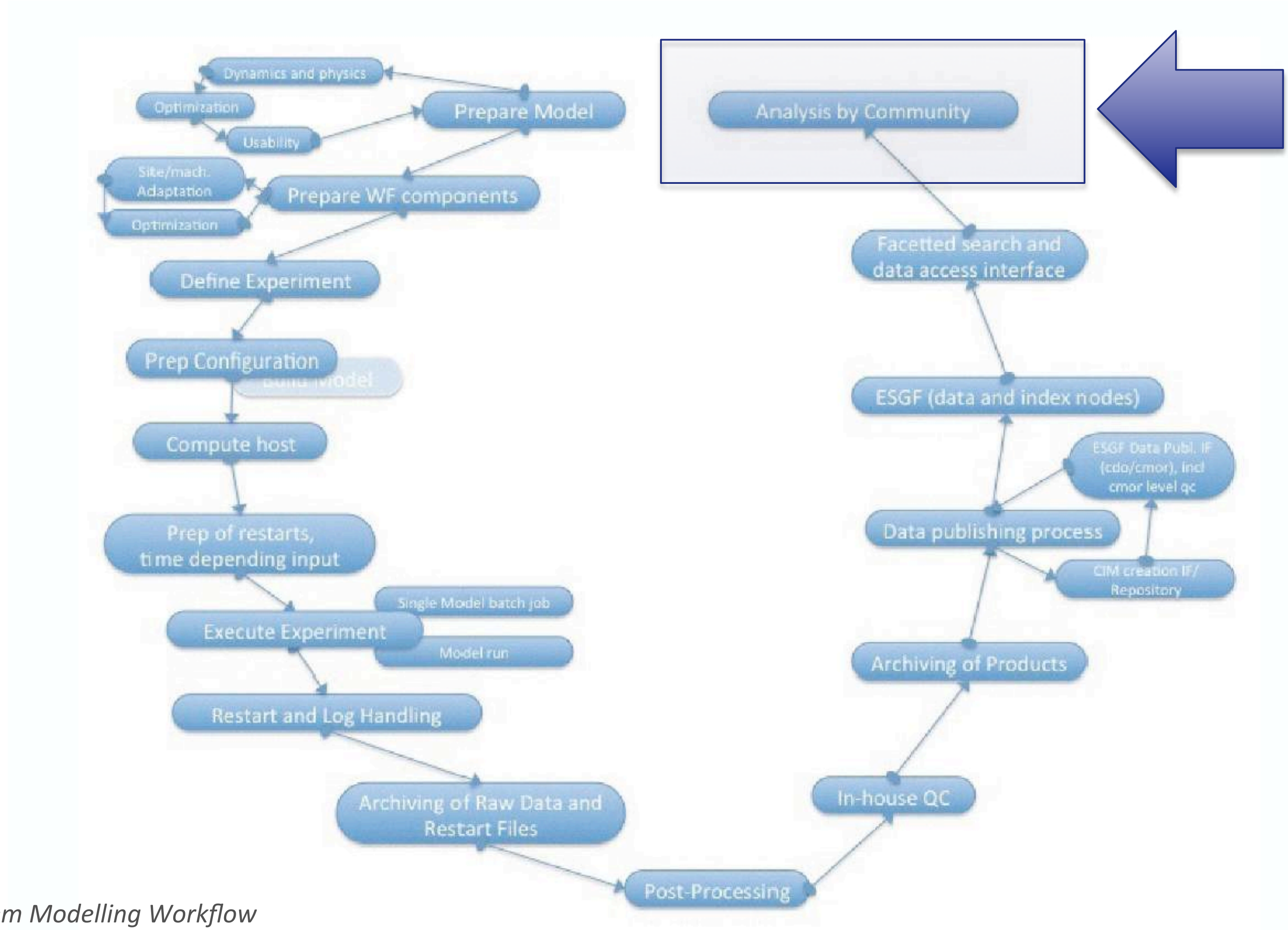
* Data From Darthmouth University



DCC Data Lifecycle

Slide courtesy: Shaun de Witt

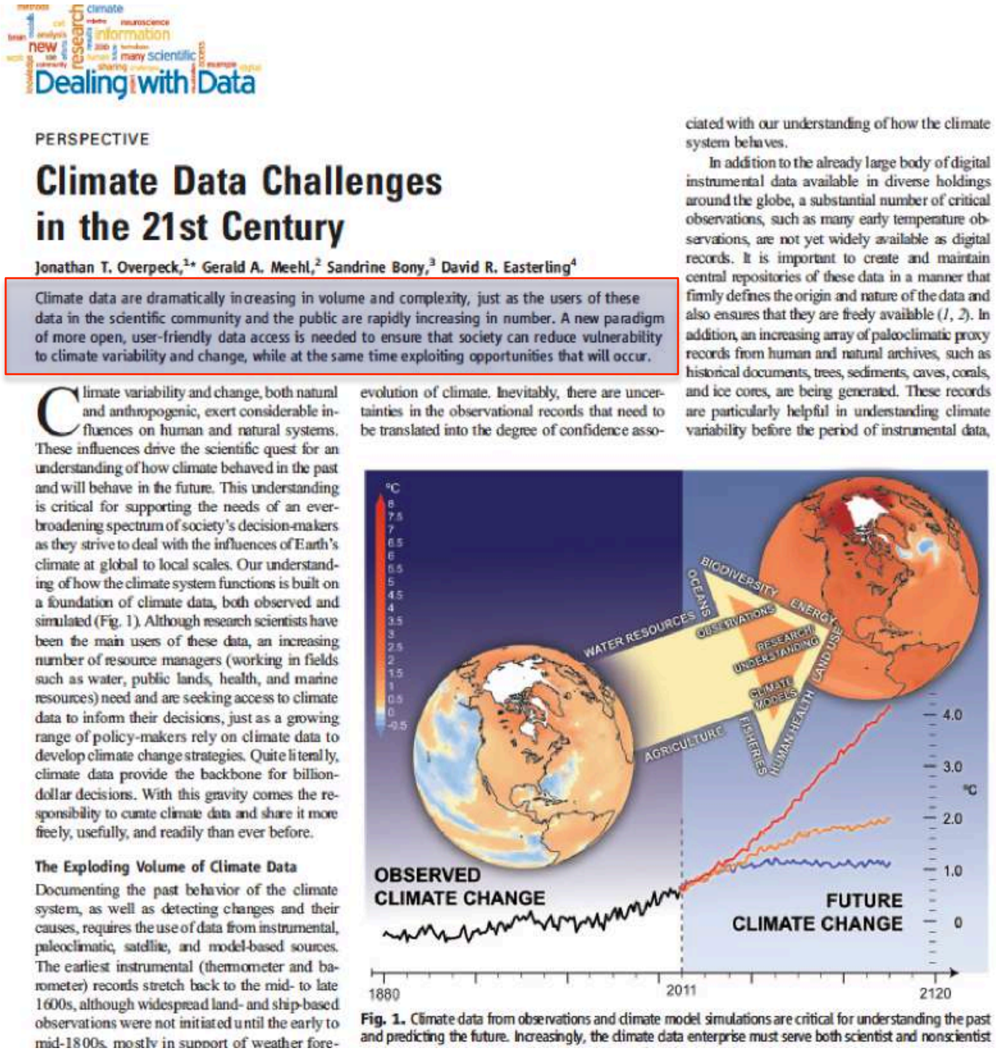
Earth System end-to-end Modelling Workflow



Today's presentation

Earth System Modelling Workflow
Source: "ISENES2 Workshop on Workflow Solutions in Earth System Modelling", by Reinhard Budich (Strategic IT Partnerships Scientific Computing Lab MPI-M) and Kerstin Fieg (Applications Deutsches Klimarechenzentrum DKRZ). June 3-5 2014, DKRZ, Hamburg.

Climate (big) Data Challenges



Dealing with Data

PERSPECTIVE

Climate Data Challenges in the 21st Century

Jonathan T. Overpeck,[1]* Gerald A. Meehl,[2] Sandrine Bony,[3] David R. Easterling[4]

Climate data are dramatically increasing in volume and complexity, just as the users of these data in the scientific community and the public are rapidly increasing in number. A new paradigm of more open, user-friendly data access is needed to ensure that society can reduce vulnerability to climate variability and change, while at the same time exploiting opportunities that will occur.

Climate variability and change, both natural and anthropogenic, exert considerable influences on human and natural systems. These influences drive the scientific quest for an understanding of how climate behaved in the past and will behave in the future. This understanding is critical for supporting the needs of an ever-broadening spectrum of society's decision-makers as they strive to deal with the influences of Earth's climate at global to local scales. Our understanding of how the climate system functions is built on a foundation of climate data, both observed and simulated (Fig. 1). Although research scientists have been the main users of these data, an increasing number of resource managers (working in fields such as water, public lands, health, and marine resources) need and are seeking access to climate data to inform their decisions, just as a growing range of policy-makers rely on climate data to develop climate change strategies. Quite literally, climate data provide the backbone for billion-dollar decisions. With this gravity comes the responsibility to curate climate data and share it more freely, usefully, and readily than ever before.

The Exploding Volume of Climate Data

Documenting the past behavior of the climate system, as well as detecting changes and their causes, requires the use of data from instrumental, paleoclimatic, satellite, and model-based sources. The earliest instrumental (thermometer and barometer) records stretch back to the mid- to late 1600s, although widespread land- and ship-based observations were not initiated until the early to mid-1800s, mostly in support of weather fore-

evolution of climate. Inevitably, there are uncertainties in the observational records that need to be translated into the degree of confidence associated with our understanding of how the climate system behaves.

In addition to the already large body of digital instrumental data available in diverse holdings around the globe, a substantial number of critical observations, such as many early temperature observations, are not yet widely available as digital records. It is important to create and maintain central repositories of these data in a manner that firmly defines the origin and nature of the data and also ensures that they are freely available (*1*, *2*). In addition, an increasing array of paleoclimatic proxy records from human and natural archives, such as historical documents, trees, sediments, caves, corals, and ice cores, are being generated. These records are particularly helpful in understanding climate variability before the period of instrumental data,

Fig. 1. Climate data from observations and climate model simulations are critical for understanding the past and predicting the future. Increasingly, the climate data enterprise must serve both scientist and nonscientist

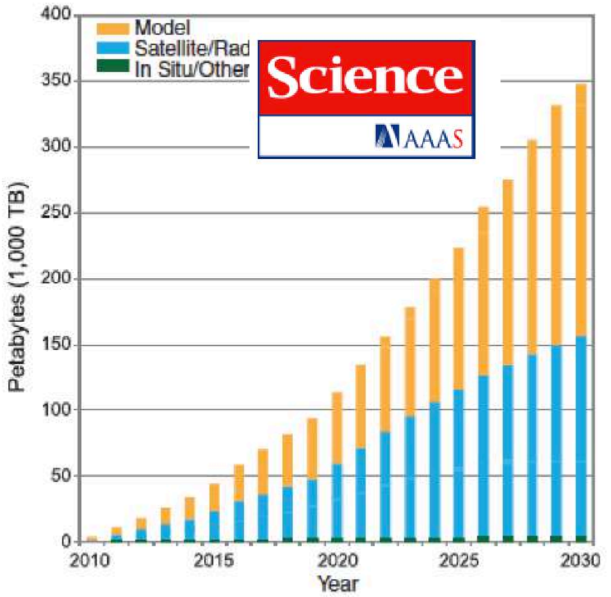
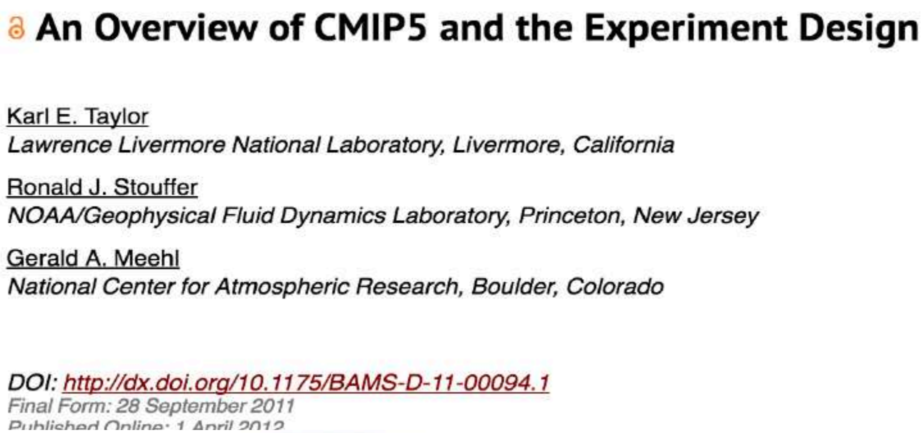


Fig. 2. The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you are not a climate scientist. The figure shows the projected increase in global climate data holdings for climate models, remotely sensed data, and in situ instrumental/proxy data.

Climate Data Challenges in the 21st Century
Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony and David R. Easterling (February 10, 2011)
Science 331 (6018), 700-702. [doi: 10.1126/science.1197869]

Large scale experiments CMIP



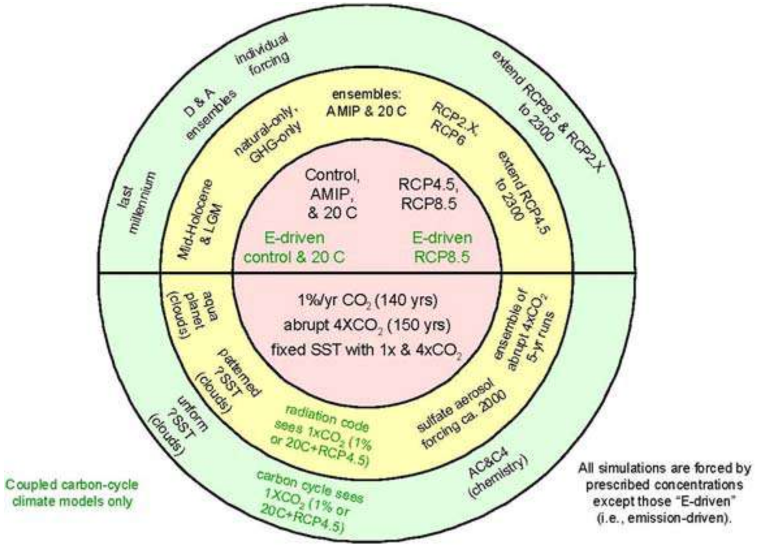
An Overview of CMIP5 and the Experiment Design

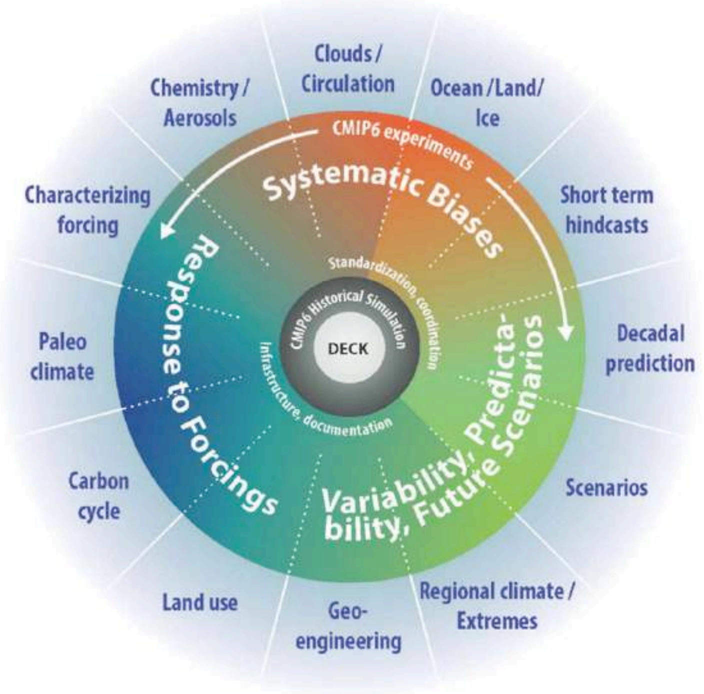
Karl E. Taylor
Lawrence Livermore National Laboratory, Livermore, California

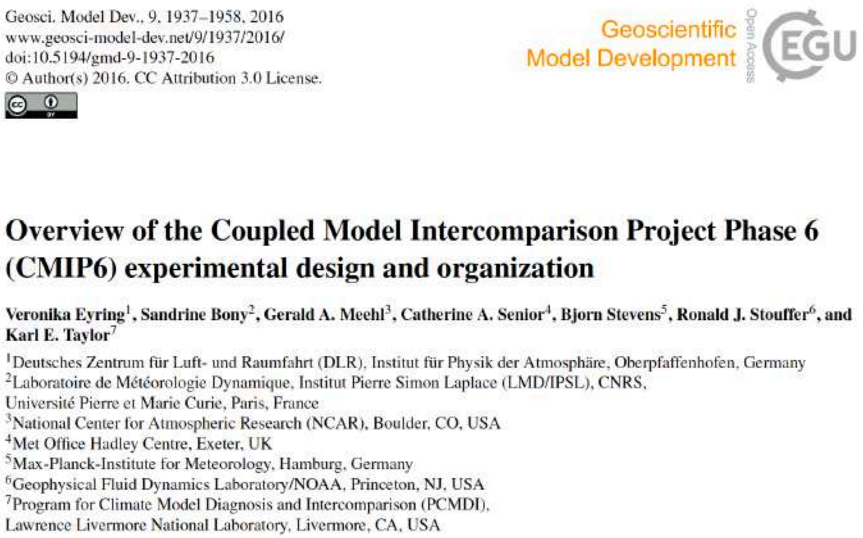
Ronald J. Stouffer
NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

Gerald A. Meehl
National Center for Atmospheric Research, Boulder, Colorado

DOI: http://dx.doi.org/10.1175/BAMS-D-11-00094.1
Final Form: 28 September 2011
Published Online: 1 April 2012


individual forcing
D & A ensembles
last millennium
ensembles: AMIP & 20 C
natural-only, GHG-only
RCP2.X, RCP6
extend RCP8.5 & RCP2.X to 2300
extend RCP4.5 to 2300
Mid-Holocene & LGM
Control, AMIP, & 20 C
RCP4.5, RCP8.5
E-driven control & 20 C
E-driven RCP8.5
1%/yr CO_2 (140 yrs)
abrupt 4XCO_2 (150 yrs)
fixed SST with 1x & 4xCO_2
aqua planet (clouds)
patterned ?SST (clouds)
unform ?SST (clouds)
ensemble of abrupt 4xCO_2 5-yr runs
radiation code sees 1xCO_2 (1% or 20C+RCP4.5)
sulfate aerosol forcing ca. 2000
carbon cycle sees 1XCO_2 (1% or 20C+RCP4.5)
AC&C4 (chemistry)
Coupled carbon-cycle climate models only
All simulations are forced by prescribed concentrations except those "E-driven" (i.e., emission-driven).





Geosci. Model Dev., 9, 1937–1958, 2016
www.geoscientific-model-dev.net/9/1937/2016/
doi:10.5194/gmd-9-1937-2016
© Author(s) 2016. CC Attribution 3.0 License.

Geoscientific Model Development
Open Access
EGU

Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization

Veronika Eyring[1], Sandrine Bony[2], Gerald A. Meehl[3], Catherine A. Senior[4], Bjorn Stevens[5], Ronald J. Stouffer[6], and Karl E. Taylor[7]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2]Laboratoire de Météorologie Dynamique, Institut Pierre Simon Laplace (LMD/IPSL), CNRS, Université Pierre et Marie Curie, Paris, France
[3]National Center for Atmospheric Research (NCAR), Boulder, CO, USA
[4]Met Office Hadley Centre, Exeter, UK
[5]Max-Planck-Institute for Meteorology, Hamburg, Germany
[6]Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ, USA
[7]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA, USA

CMIP data history and its big data evolution

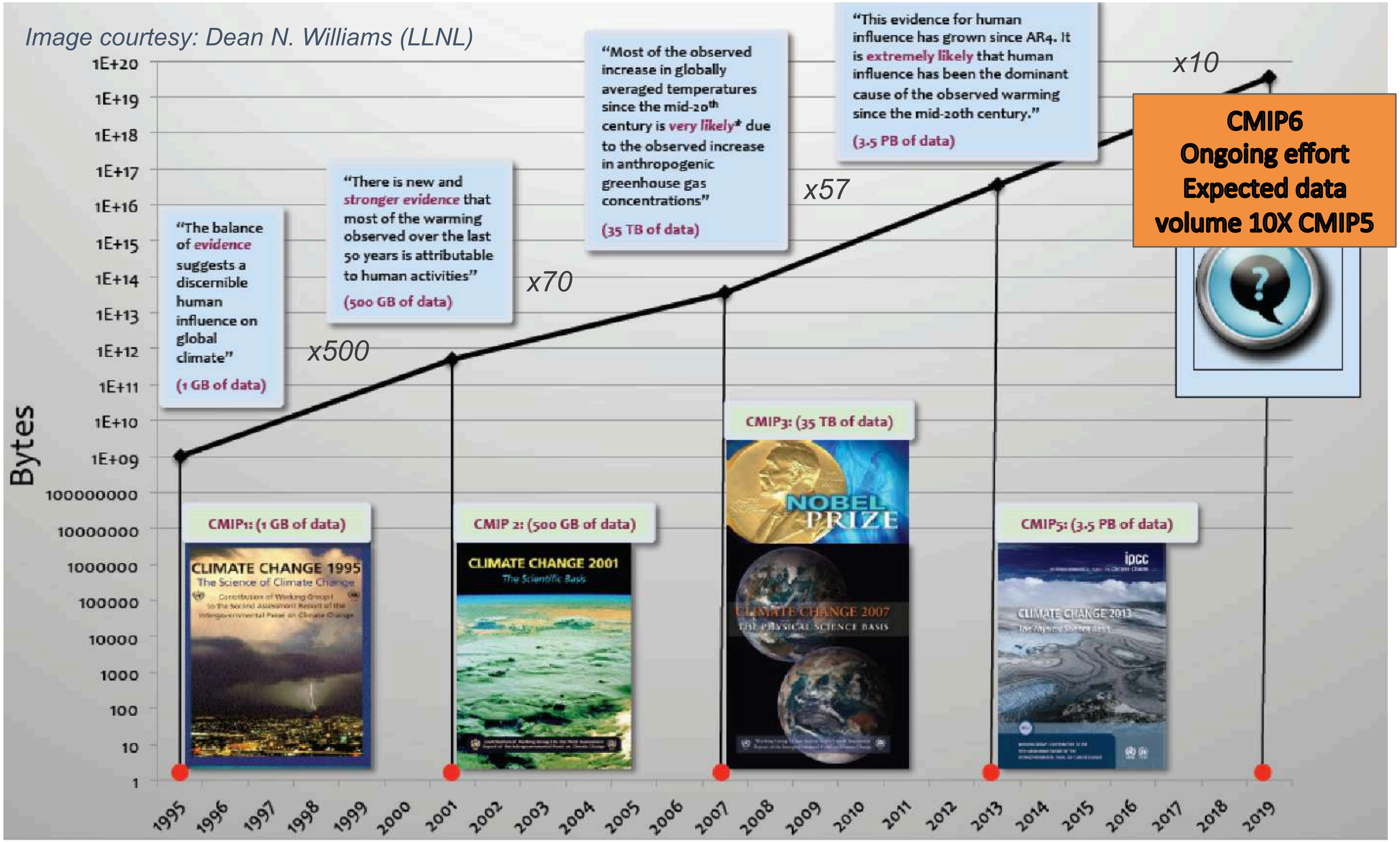


Image courtesy: Dean N. Williams (LLNL)

"The balance of evidence suggests a discernible human influence on global climate" (1 GB of data)

"There is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities" (500 GB of data)

"Most of the observed increase in globally averaged temperatures since the mid-20th century is very likely* due to the observed increase in anthropogenic greenhouse gas concentrations" (35 TB of data)

"This evidence for human influence has grown since AR4. It is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century." (3.5 PB of data)

CMIP6
Ongoing effort
Expected data volume 10X CMIP5

CMIP1: (1 GB of data)

CMIP 2: (500 GB of data)

CMIP3: (35 TB of data)

CMIP5: (3.5 PB of data)

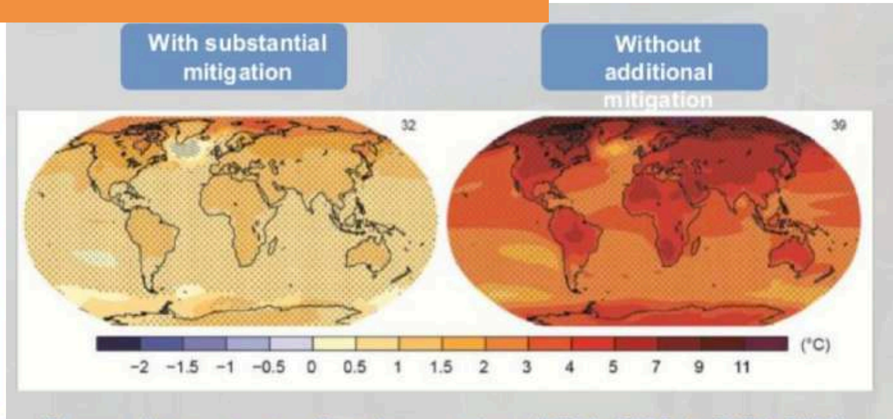
IPCC Assessment Report

Fifth Assessment Report (AR5)

The decision to prepare a Fifth Assessment Report (AR5) was taken by the members of the IPCC at its 28th Session (09-10 April 2008,

IPCC reports cover "the ***scientific***, ***technical*** and ***socio-economic*** information relevant to understanding the scientific basis of risk of **human-induced climate change**, its **potential impacts** and options for **adaptation** and **mitigation**".

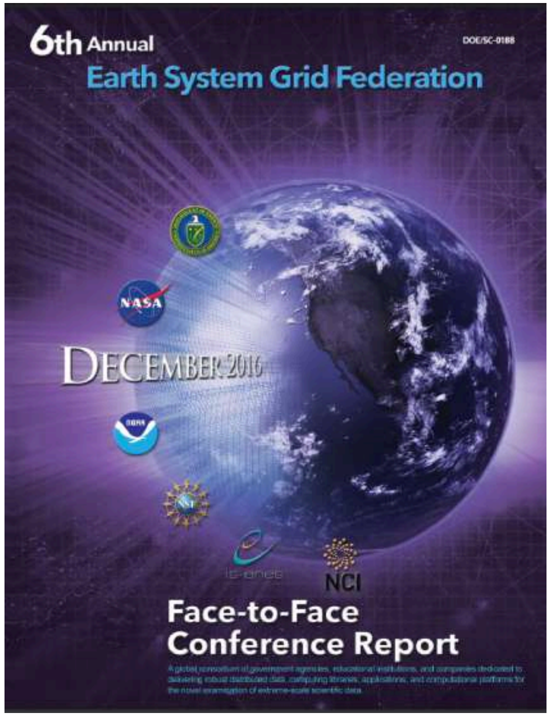
CLIMATE CHANGE 2014
Impacts, Adaptation, and Vulnerability
Volume I: Global and Sectoral Aspects
WORKING GROUP II CONTRIBUTION TO THE FIFTH ASSESSMENT REPORT OF THE INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE
WORKING GROUP I CONTRIBUTION TO THE FIFTH ASSESSMENT REPORT OF THE INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE
CLIMATE CHANGE 2014
Mitigation of Climate Change



Change in average surface temperature (1986–2005 to 2081–2100)

Getting access to CMIP data: ESGF

ESGF¹ is a coordinated multiagency, international collaboration of institutions that continually develop, deploy, and maintain software needed to facilitate and empower the study of climate.



6th Annual

DOE/SC-0188

Earth System Grid Federation

DECEMBER 2016

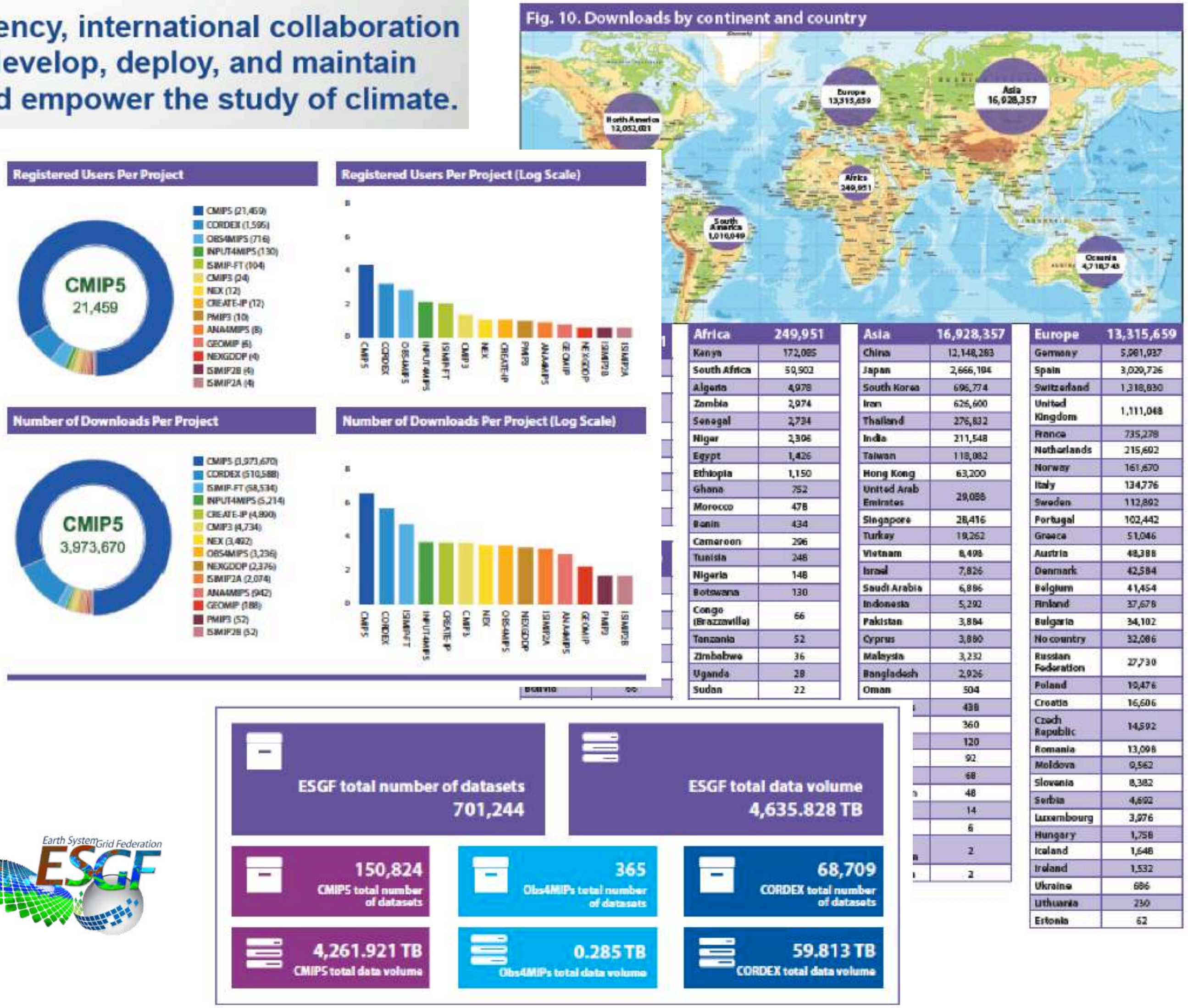
Face-to-Face Conference Report

6th Annual ESGF F2F Conference

December 5–9, 2016, Washington, D.C.

Convened by DOE, NASA, NOAA, NSF

IS-ENES, NCI



Registered Users Per Project

Registered Users Per Project (Log Scale)

Number of Downloads Per Project

Number of Downloads Per Project (Log Scale)

Fig. 10. Downloads by continent and country

Africa	249,951
Kenya	172,085
South Africa	50,902
Algeria	4,978
Zambia	2,974
Senegal	2,734
Niger	2,306
Egypt	1,426
Ethiopia	1,150
Ghana	752
Morocco	478
Benin	434
Cameroon	296
Tunisia	248
Nigeria	148
Botswana	130
Congo (Brazzaville)	66
Tanzania	52
Zimbabwe	36
Uganda	28
Sudan	22

Asia	16,928,357
China	12,148,283
Japan	2,666,184
South Korea	696,774
Iran	626,600
Thailand	276,832
India	211,548
Taiwan	118,082
Hong Kong	63,200
United Arab Emirates	29,088
Singapore	28,416
Turkey	19,262
Vietnam	8,498
Israel	7,826
Saudi Arabia	6,886
Indonesia	5,292
Pakistan	3,884
Cyprus	3,880
Malaysia	3,232
Bangladesh	2,926
Oman	504
	438
	360
	120
	92
	68
	48
	14
	6
	2
	2

Europe	13,315,659
Germany	5,981,937
Spain	3,029,726
Switzerland	1,318,830
United Kingdom	1,111,048
France	735,278
Netherlands	215,692
Norway	161,670
Italy	134,776
Sweden	112,892
Portugal	102,442
Greece	51,046
Austria	48,388
Denmark	42,584
Belgium	41,454
Finland	37,678
Bulgaria	34,102
No country	32,086
Russian Federation	27,730
Poland	19,476
Croatia	16,606
Czech Republic	14,592
Romania	13,098
Moldova	9,562
Slovenia	8,382
Serbia	4,692
Luxembourg	3,976
Hungary	1,758
Iceland	1,648
Ireland	1,532
Ukraine	686
Lithuania	230
Estonia	62

ESGF total number of datasets
701,244

ESGF total data volume
4,635.828 TB

150,824
CMIP5 total number of datasets

365
Obs4MIPs total number of datasets

68,709
CORDEX total number of datasets

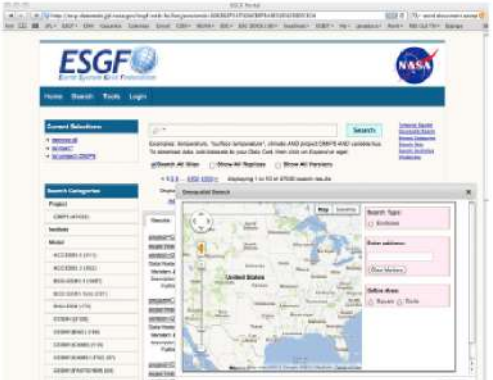
4,261.921 TB
CMIP5 total data volume

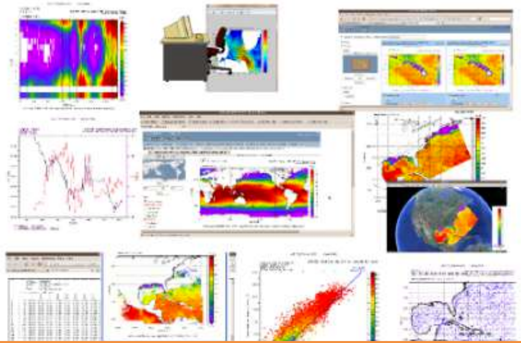
0.285 TB
Obs4MIPs total data volume

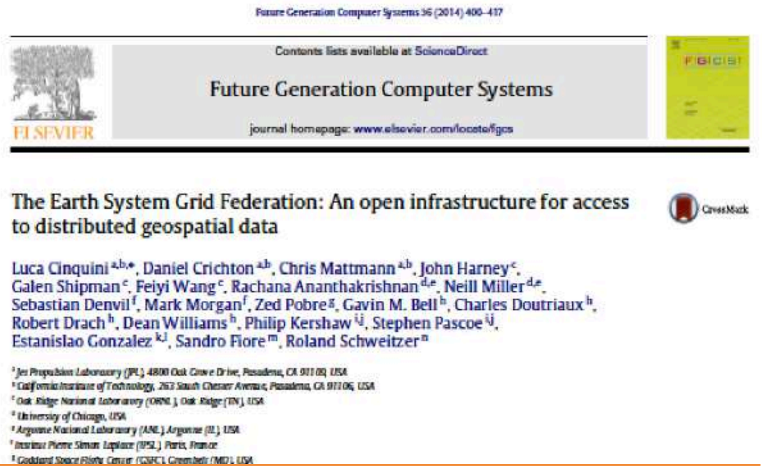
59.813 TB
CORDEX total data volume

ESGF: an open infrastructure for access to distributed geospatial data

…looking forward computing and analytics







Future Generation Computer Systems 36 (2014) 400–417

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

ELSEVIER

The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data

Luca Cinquini[a,b,*], Daniel Crichton[a,b], Chris Mattmann[a,b], John Harney[c], Galen Shipman[c], Feiyi Wang[c], Rachana Ananthakrishnan[d,e], Neill Miller[d,e], Sebastian Denvil[f], Mark Morgan[f], Zed Pobre[g], Gavin M. Bell[h], Charles Doutriaux[h], Robert Drach[h], Dean Williams[h], Philip Kershaw[i,j], Stephen Pascoe[i,j], Estanislao Gonzalez[k,l], Sandro Fiore[m], Roland Schweitzer[n]

[a] Jet Propulsion Laboratory (JPL), 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[b] California Institute of Technology, 263 South Chester Avenue, Pasadena, CA 91106, USA
[c] Oak Ridge National Laboratory (ORNL), Oak Ridge (TN), USA
[d] University of Chicago, USA
[e] Argonne National Laboratory (ANL), Argonne (IL), USA
[f] Institut Pierre Simon Laplace (IPSL), Paris, France
[g] Goddard Space Flight Center (GSFC), Greenbelt (MD), USA

The new computational platform […] will support ***parallel*** *and* ***distributed computing*** *tasks by including* ***OpenMPI, Map/Reduce*** *and streaming computing models. The new compute node will allow for* ***large-scale manipulation*** *and* ***analysis*** *of* ***data*** *[…] We intend to fully explore the possibility of providing a configurable and scalable ESGF environment that can be easily deployed on the* ***cloud*** *[…] to meet requirements such as* ***high availability*** *and* ***elastic*** *allocation of* ***computing processes****.*

Article history:
Received 30 January 2013
Received in revised form
18 June 2013
Accepted 17 July 2013
Available online 17 September 2013

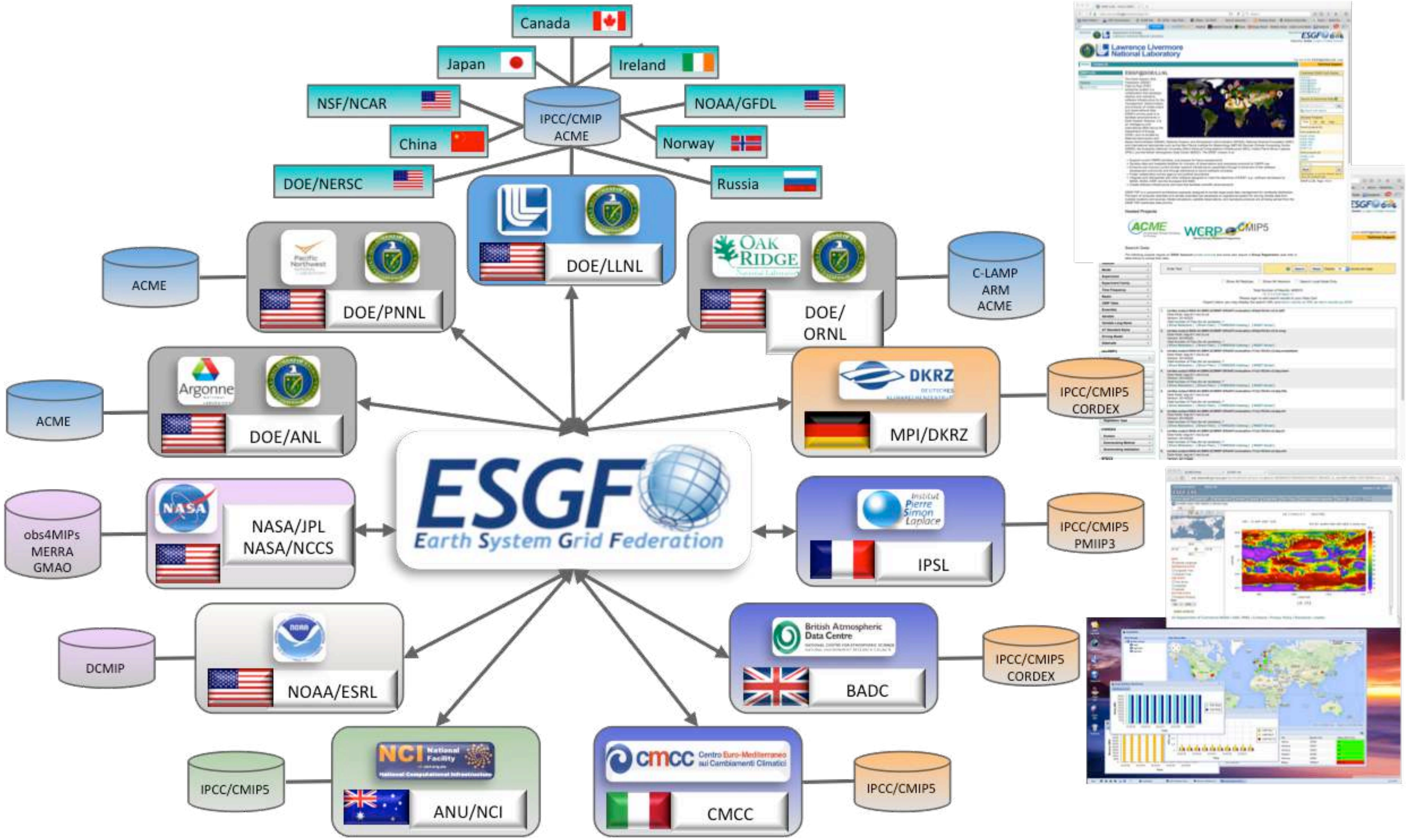
The Earth System Grid Federation (ESGF) is a multi-agency, international collaboration that aims at developing the software infrastructure needed to facilitate and empower the study of climate change on a global scale. The ESGF's architecture employs a system of geographically distributed peer nodes, which are independently administered yet united by the adoption of common federation protocols and application programming interfaces (APIs). The cornerstones of its interoperability are the peer-to-peer messaging that is continuously exchanged among all nodes in the federation; a shared architecture and API for search

[*] Corresponding author at: Jet Propulsion Laboratory (JPL), 4800 Oak Grove Drive, Pasadena, CA 91109, USA.
E-mail addresses: luca.cinquini@jpl.nasa.gov (L. Cinquini), daniel.j.crichton@jpl.nasa.gov (D. Crichton), chris.a.mattmann@jpl.nasa.gov (C. Mattmann), harneyjf@ornl.gov (J. Harney), gshipman@ornl.gov (G. Shipman), fwang2@ornl.gov (F. Wang), ranantha@mcs.anl.gov (R. Ananthakrishnan), neillm@mcs.anl.gov (N. Miller), sebastien.denvil@ipsl.jussieu.fr (S. Denvil), momipsl@ipsl.jussieu.fr (M. Morgan), zed.pobre@nasa.gov (Z. Pobre), gavin@llnl.gov (G.M. Bell), doutriaux1@llnl.gov (C. Doutriaux), drach1@llnl.gov (R. Drach), williams13@llnl.gov (D. Williams), philip.kershaw@stfc.ac.uk (P. Kershaw), stephen.pascoe@stfc.ac.uk (S. Pascoe), estanislao.gonzalez@met.fu-berlin.de (E. Gonzalez), sandro.fiore@unisalento.it (S. Fiore), Roland.Schweitzer@noaa.gov (R. Schweitzer).
[1] Present address.

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.07.002

GO-ESSP
Argonne
Sandia National Laboratories
Pacific Northwest
NCAR
Computation Institute
UC DAVIS
NOAA
janet
WDC
Institut Pierre Simon Laplace
Centro Euro-Mediterraneo per i Cambiamenti Climatici
CSIRO
NCI
FUNET
ESnet
Met Office
ESMF
British Atmospheric Data Centre
SURF NET
aarnet
• • •

ESGF and the CMIP data archive



Canada

Japan

Ireland

NSF/NCAR

NOAA/GFDL

IPCC/CMIP ACME

China

Norway

DOE/NERSC

Russia

DOE/LLNL

ACME

DOE/PNNL

DOE/ ORNL

C-LAMP ARM ACME

ACME

DOE/ANL

MPI/DKRZ

IPCC/CMIP5 CORDEX

ESGF Earth System Grid Federation

obs4MIPs MERRA GMAO

NASA/JPL NASA/NCCS

IPSL

IPCC/CMIP5 PMIIP3

DCMIP

NOAA/ESRL

BADC

IPCC/CMIP5 CORDEX

IPCC/CMIP5

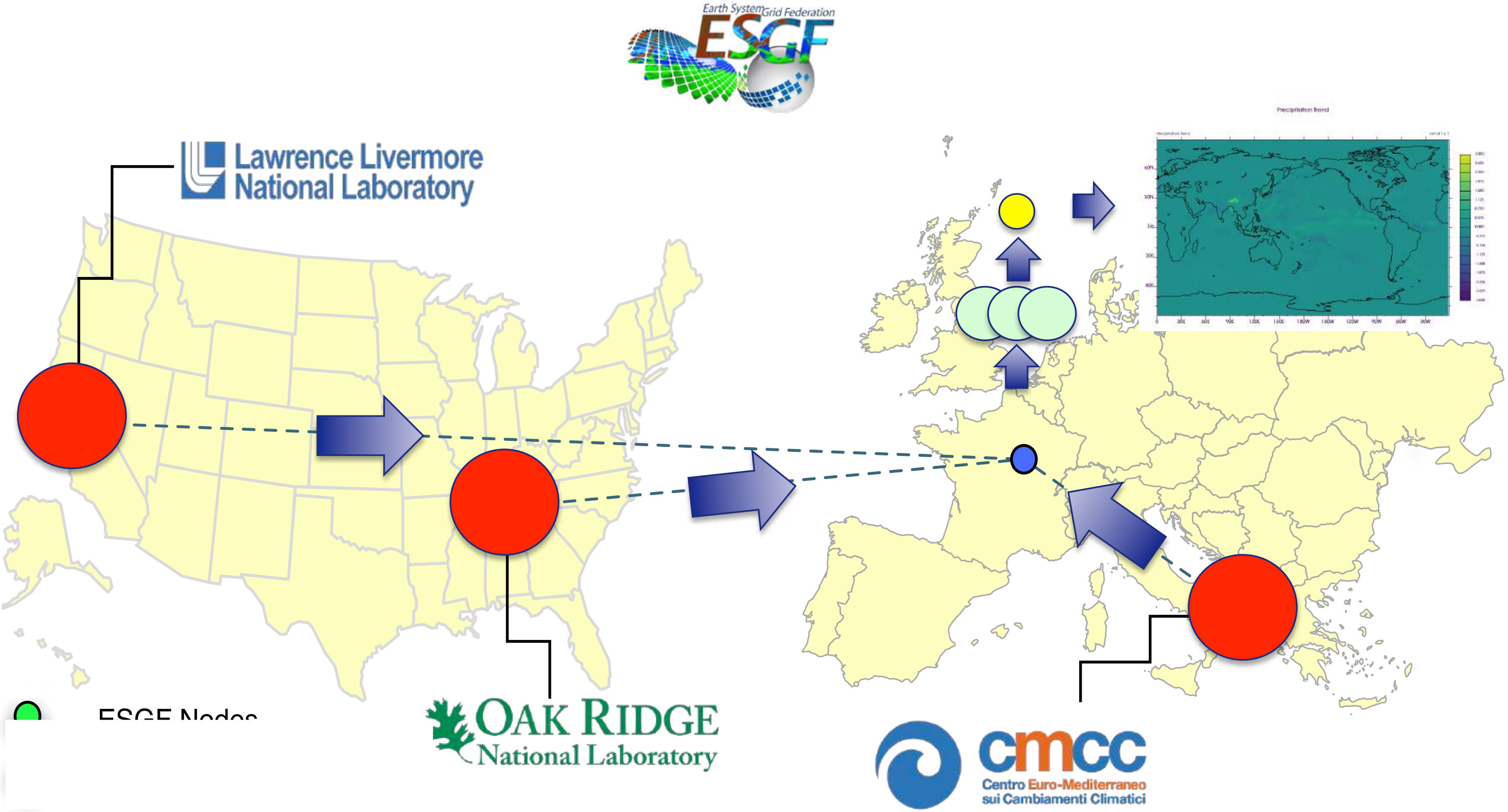
ANU/NCI

CMCC

IPCC/CMIP5

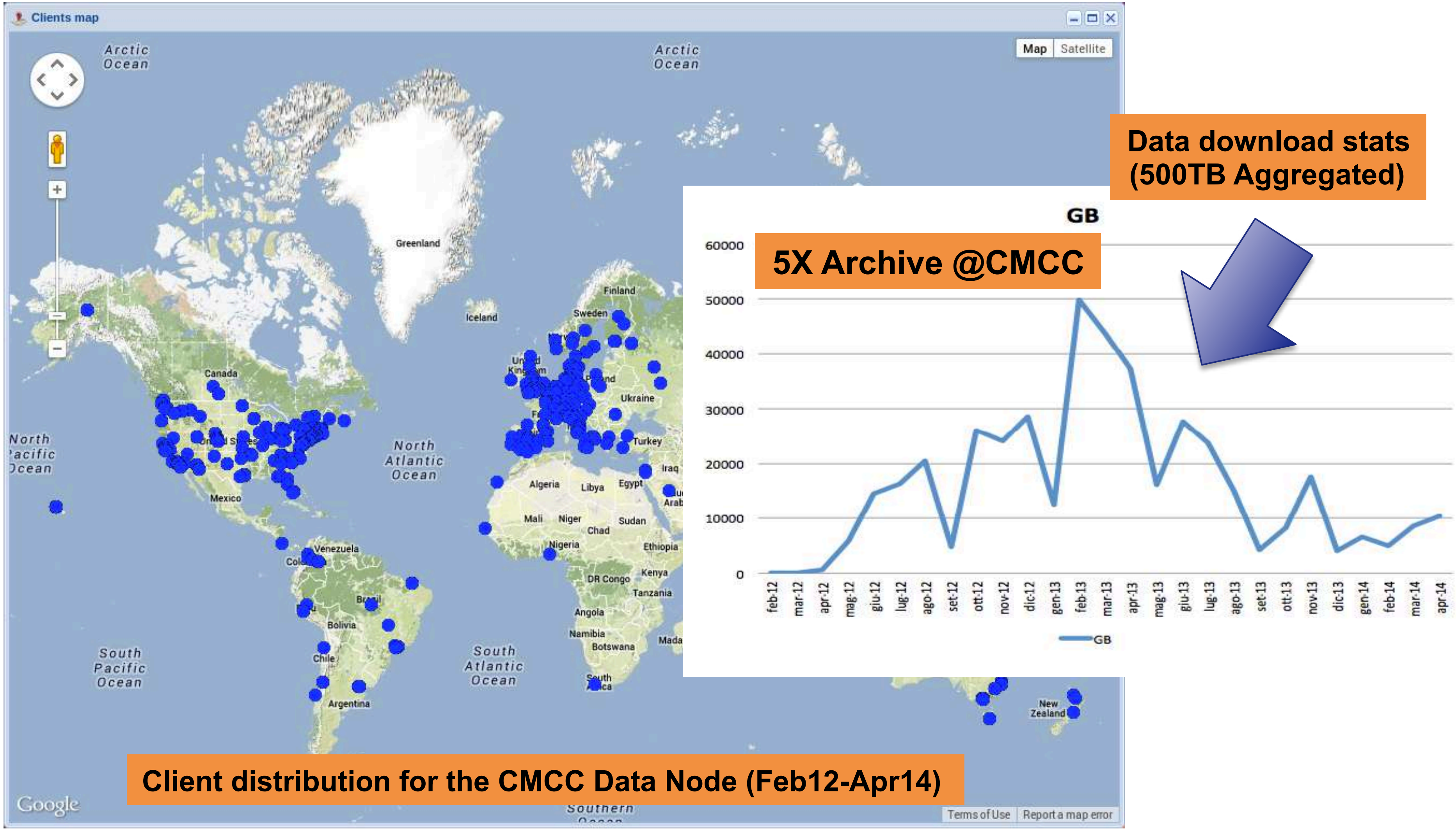
Image courtesy: Dean N. Williams (LLNL)

CMIP scientific data analysis workflow in ESGF


Earth System Grid Federation
ESGF
Lawrence Livermore National Laboratory
Precipitation trend
OAK RIDGE National Laboratory
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
ESGF Nodes

CMIP5 Data usage statistics CMCC

Data node (Feb 2012-Apr 2014)



Data download stats (500TB Aggregated)

5X Archive @CMCC

Client distribution for the CMCC Data Node (Feb12-Apr14)

Key issues and challenges regarding climate data analysis

- ESGF provides a large-scale, federated, data-sharing & access infrastructure
  - client-side and sequential nature of the current approach
  - The setup of a data analysis experiment requires that all the needed climate datasets must be downloaded from the related ESGF data nodes on the end-user’s local machine.
  - for multi-model experiments data download can take a significant amount of time (weeks!)
- The complexity of the data analysis process itself leads to the need for end-to-end workflow support solution
  - analysing large datasets involves running tens/hundreds of analytics operators in a coordinated fashion.
  - Current approaches (mostly based on bash-like scripts) requires climate scientists to take care of, implement and replicate workflow-like control logic aspects in their scripts (which are error-prone too) along with the expected application-level part.
- The large volumes of data pose additional challenges related to performance

New approaches for climate data analysis at large scale

- Dedicated data intensive facilities close to the different storage hierarchies will be needed to address high-performance scientific data management

- Server-side approaches will intrinsically and drastically reduce data movement
  - download will only relate to the final results of an analysis
  - they will foster re-usability as well as collaborative experiments
  - Need for interoperability efforts toward highly interoperable tools/envs for data analysis

- Cloud technologies will help on deploying in a flexible and dynamic manner analytics applications/tools enabling highly scalable and elastic scenarios in clouds environments

EOSC, ECAS and Ophidia

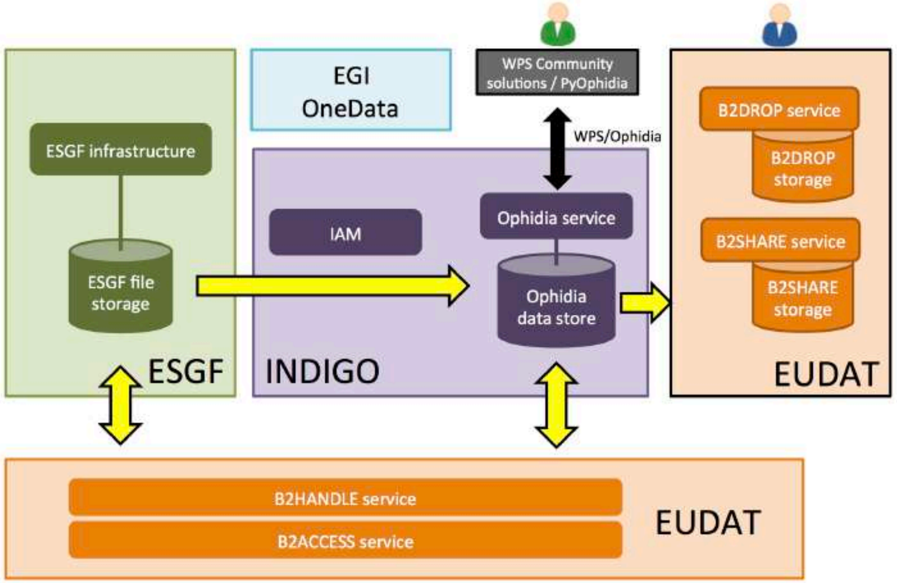
The context: European Open Science Cloud

- ✓ The **European Open Science Cloud (EOSC)** is an ambitious program will offer a **virtual environment** with **open** and **seamless services** for storage, management, **analysis** and **re-use of research data**, **across borders** and **scientifc disciplines** by federating existing scientifc data infrastructures, currently dispersed across disciplines and Member States.

- ✓ This programme will deliver an **Open Data Science Environment** that **federates existing scientific data infrastructures** to offer European science and technology researchers and practitioners seamless access to services for storage, management, analysis and re-use of research data presently restricted by geographic borders and scientific disciplines.

ENES Climate Analytics Service (ECAS)

- ✓ The **ENES Climate Analytics Service (ECAS)**, proposed by CMCC & DKRZ in EOSC-hub supports climate data analysis
- ✓ It is one of the **EOSC-Hub Thematic Services**
- ✓ ECAS builds on top of the **Ophidia big data analytics framework** with components from INDIGO-DataCloud, EUDAT and EGI
- ✓ The Analytics-Hub is a paradigm joining data and computing able to provide a **multi-model environment** for CMIP-based analytics experiments in ESGF



The European Commission launched the European Open ScienceCloud Initiative to capitalise on the data revolution. EOSC will provide European science, industry and public authorities with world-class digital infrastructure that bring state of the art computing and data storage capacity to the fingertips of any scientists and engineer in the EU.

EOSC-hub receives funding from the EU's Horizon 2020 research and innovation programme under grant agreement No. 777536.

ENES Climate Analytics Service Motivation

- ✓ Traditional approach to data analysis relies on data downloads and using local analysis tools
  - ✓ Data are now too huge to download
  - ✓ Data sharing and re-use are strongly desirable

- ✓ ECAS provides a server-side, parallel data analysis environment
  - ✓ Computationally powerful: Ophidia analytics framework
  - ✓ Easy to use: Jupyter notebooks and data sharing services

Ophidia jupyter B2SHARE B2DROP ONEDATA Infrastructure Manager

ENES Climate Analytics Service Intro

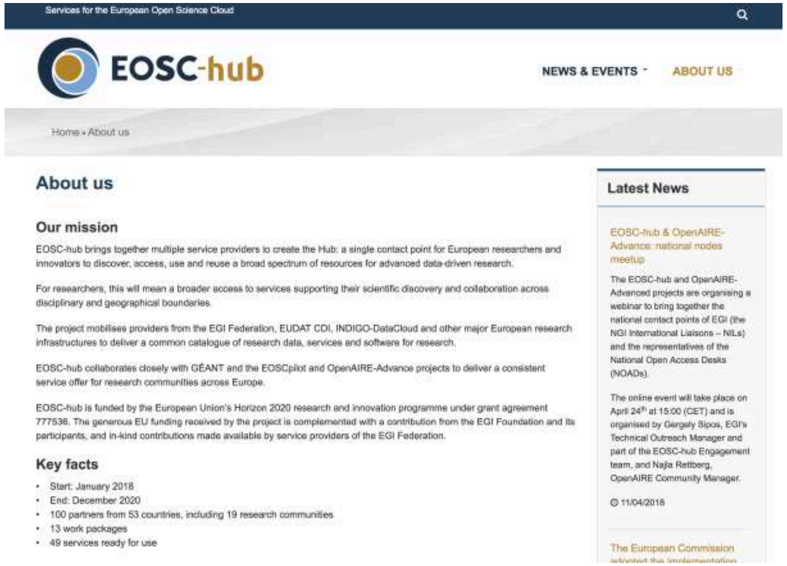
- ECAS: a **data analytics service** for EOSC
  - **ENES**: European Network for Earth System Modelling

- Involved institutions:
  - **DKRZ**: German Climate Computing Center
  - **CMCC**: Euro-Mediterranean Center on Climate Change Foundation

- Enable **server-side workflows** for Earth system researchers and beyond

- **ECASLab** is the virtual environment for ECAS
  - Integrate several **UNIDATA** software (NetCDF lib, THREDDS and IDV)

- **ECAS is based on the Ophidia big data analytics framework**



About us

Our mission

EOSC-hub brings together multiple service providers to create the Hub: a single contact point for European researchers and innovators to discover, access, use and reuse a broad spectrum of resources for advanced data-driven research.

For researchers, this will mean a broader access to services supporting their scientific discovery and collaboration across disciplinary and geographical boundaries.

The project mobilises providers from the EGI Federation, EUDAT CDI, INDIGO-DataCloud and other major European research infrastructures to deliver a common catalogue of research data, services and software for research.

EOSC-hub collaborates closely with GÉANT and the EOSCpilot and OpenAIRE-Advance projects to deliver a consistent service offer for research communities across Europe.

EOSC-hub is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement 777536. The generous EU funding received by the project is complemented with a contribution from the EGI Foundation and its participants, and in-kind contributions made available by service providers of the EGI Federation.

Key facts

- Start: January 2018
- End: December 2020
- 100 partners from 53 countries, including 19 research communities
- 13 work packages
- 49 services ready for use

Latest News

EOSC-hub & OpenAIRE-Advance: national nodes meetup

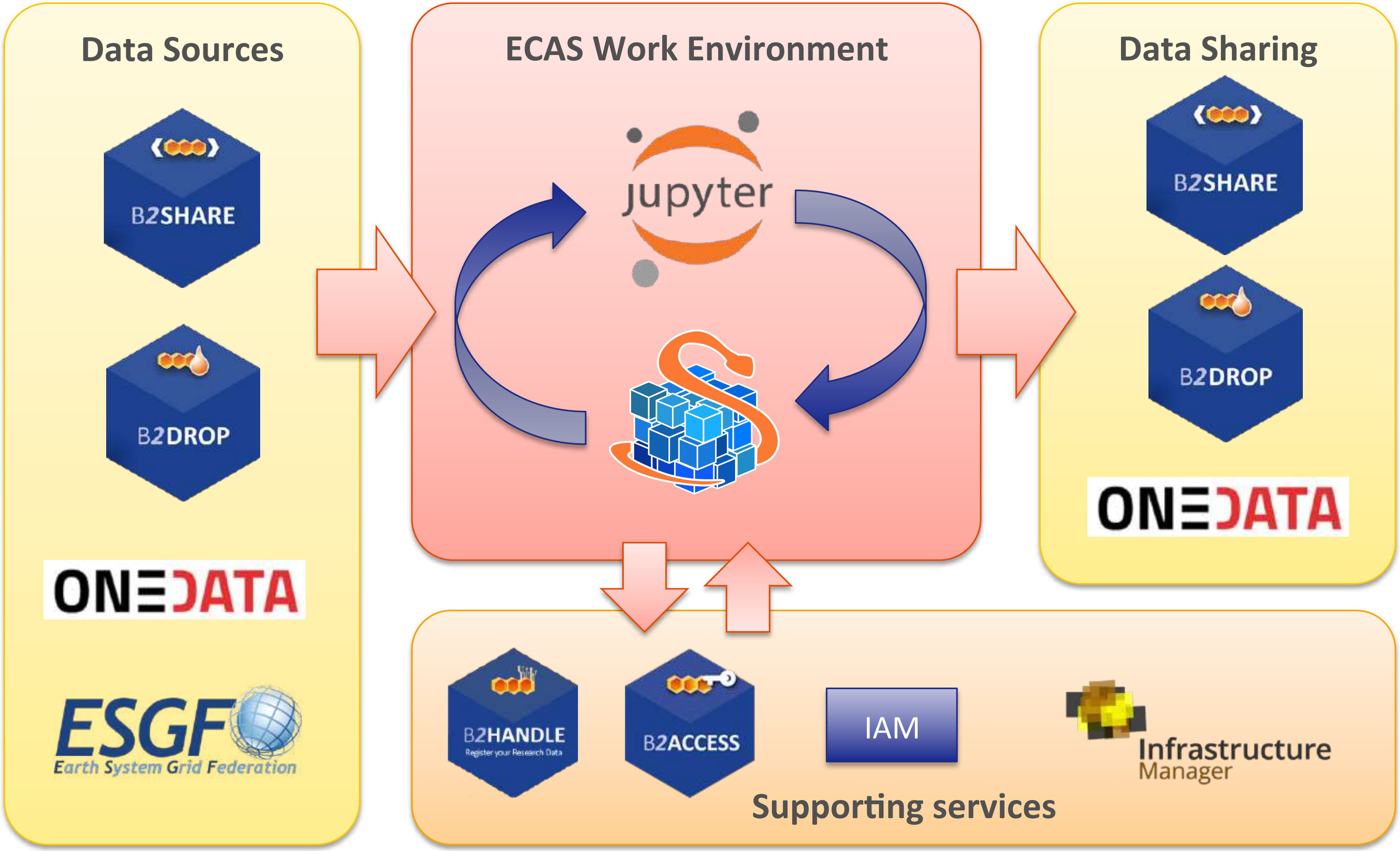
The EOSC-hub and OpenAIRE-Advanced projects are organising a webinar to bring together the national contact points of EGI (the NGI International Liaisons – NILs) and the representatives of the National Open Access Desks (NOADs).

The online event will take place on April 24th at 15:00 (CET) and is organised by Gergely Sipos, EGI's Technical Outreach Manager and part of the EOSC-hub Engagement team, and Najla Rettberg, OpenAIRE Community Manager.

11/04/2018

The European Commission adopted the implementation

ECAS Service architecture and interfaces

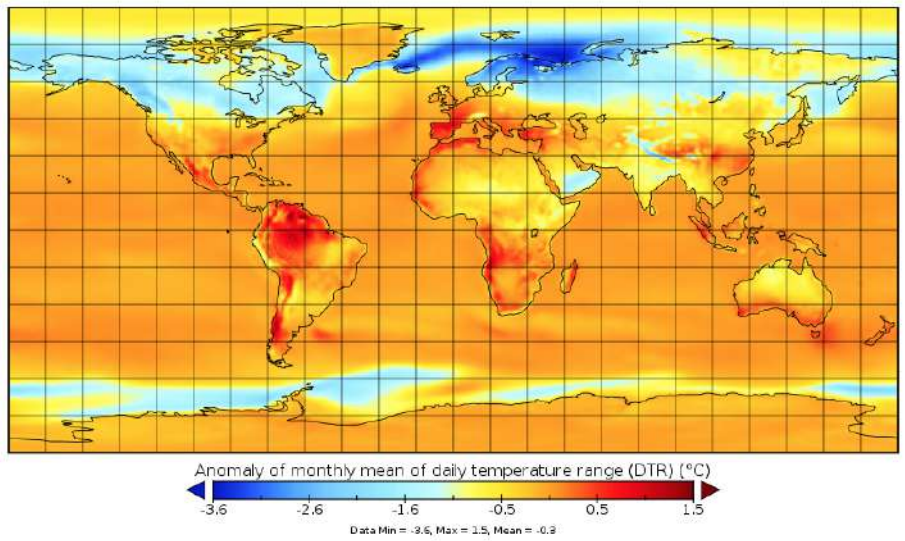


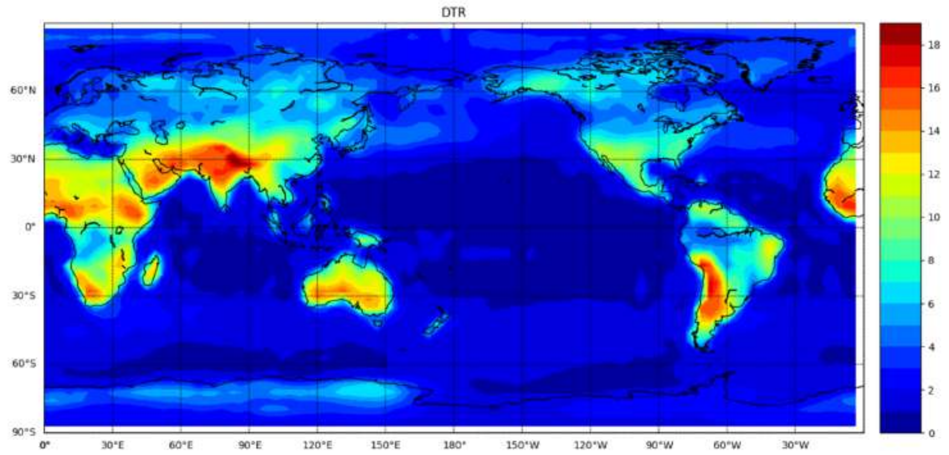
ECAS Featured use cases

Climate data sharing, access, analysis and visualization

- ✓ *Climate indicators*
  - ✓ *Integration ECAS/B2DROP*
  - ✓ *ECAS Python API extended to support pushing of results to B2DROP*
  - ✓ *Different interfaces (from file/datacube to B2DROP)*
  - ✓ *Straightforward integration of B2DROP into Notebooks*

Anomaly of monthly mean of daily temperature range (DTR)


Anomaly of monthly mean of daily temperature range (DTR) (°C)
-3.6
-2.6
-1.6
-0.5
0.5
1.5
Data Min = -3.6, Max = 1.5, Mean = -0.3



Service Access

- *DKRZ and CMCC are the current service providers for ECAS*
- *The two instances can be used by users after registration*
  - *The instances differ in the choice and amount of local data provided. Please refer to individual site documentation for details.*
- *Users can develop Jupyter notebooks with Python*
  - *Share your workflows via the ECAS workflow repository*
  - *Store results in B2DROP*
  - *Git repo at https://github.com/ECAS-Lab*

Ophidia: a scientific big data analytics framework

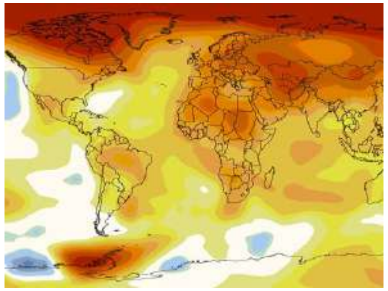
Ophidia (http://ophidia.cmcc.it) is a CMCC Foundation research project addressing fast and big data challenges for eScience

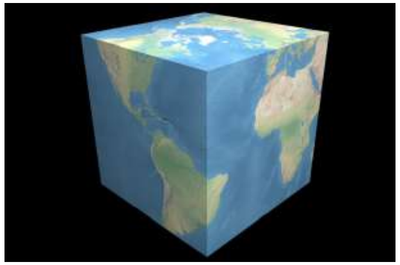
It provides support for declarative, parallel, server-side data analysis exploiting parallel computing techniques and database approaches

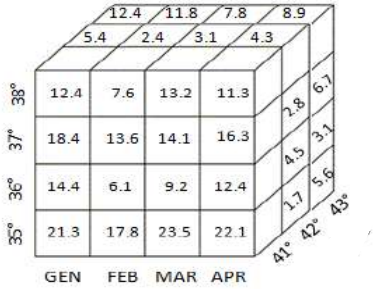
It provides end-to-end mechanisms to support complex experiments and large processing workflows on scientific datacubes

Big data challenges and the paradigm shift

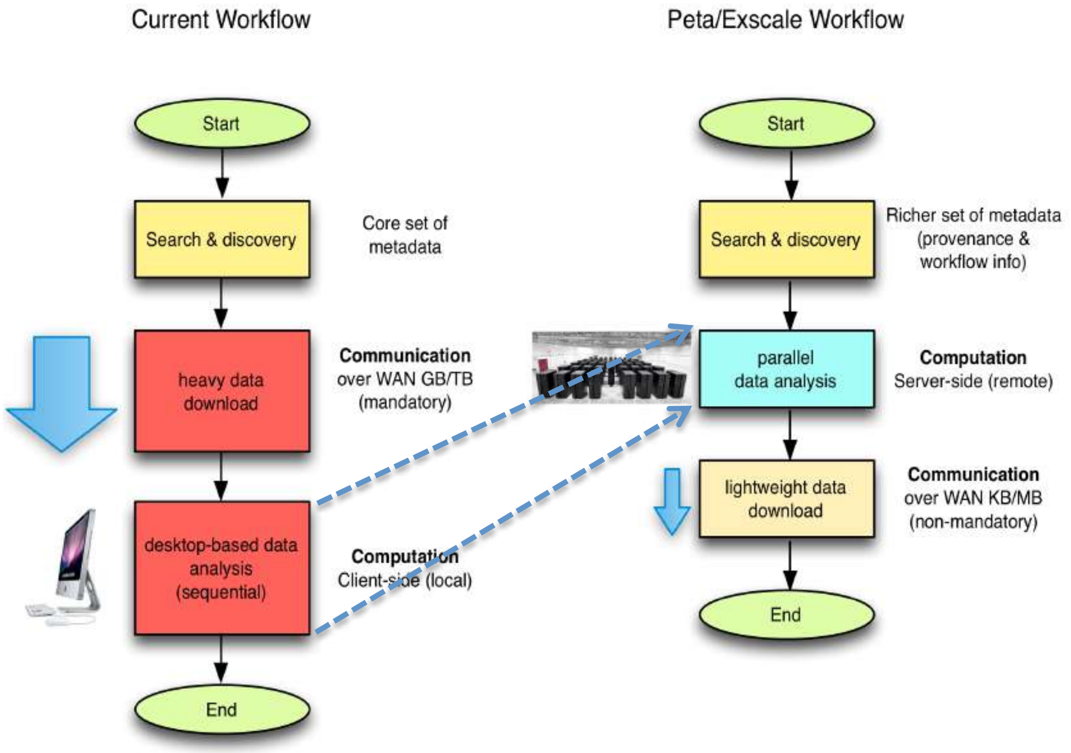
Volume, variety, velocity are key challenges for big data in general and for climate change science in particular. Client-side, sequential and disk-based workflows are three limiting factors for the current scientific data analysis tools.





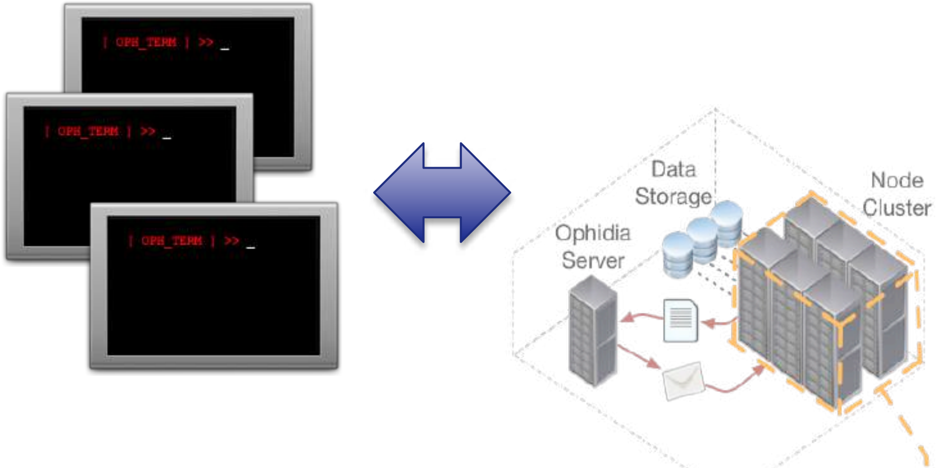


	GEN	FEB	MAR	APR
38°	12.4	7.6	13.2	11.3
37°	18.4	13.6	14.1	16.3
36°	14.4	6.1	9.2	12.4
35°	21.3	17.8	23.5	22.1


Current Workflow
Start
Search & discovery
Core set of metadata
heavy data download
Communication over WAN GB/TB (mandatory)
desktop-based data analysis (sequential)
Computation Client-side (local)
End
Peta/Exscale Workflow
Start
Search & discovery
Richer set of metadata (provenance & workflow info)
parallel data analysis
Computation Server-side (remote)
lightweight data download
Communication over WAN KB/MB (non-mandatory)
End

S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D. N. Williams, G. Aloisio, "**Ophidia: toward bigdata analytics for eScience**", ICCS2013 Conference, Procedia Elsevier, Barcelona, June 5-7, 2013

Server-side paradigm and the datacube abstraction

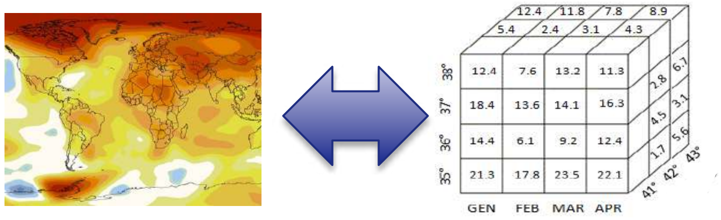


***Oph_Term**: a terlminal-like commands interpreter serving as a client for the Ophidia framework*

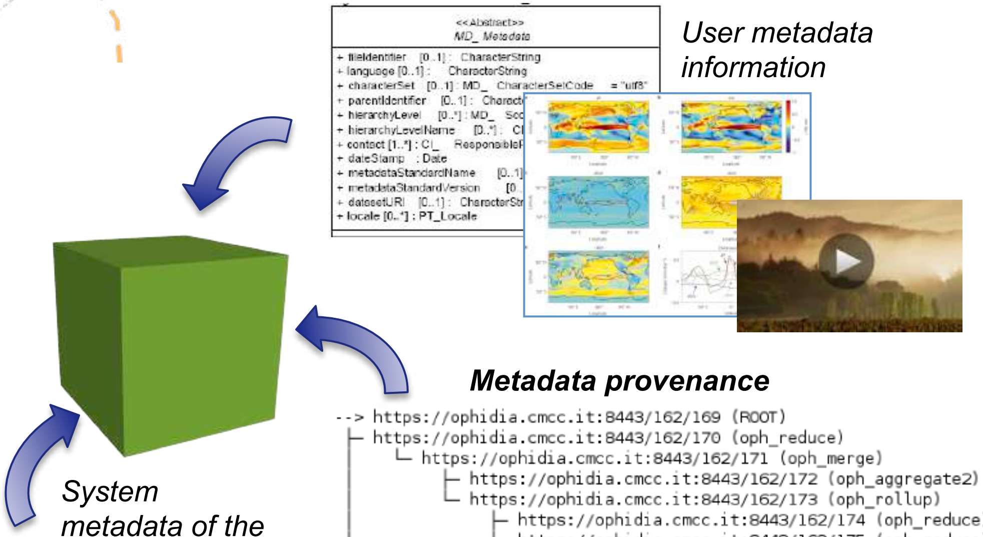
***Ophidia framework**: declarative, parallel server-side processing*

*Through the **oph_term** the user can send commands to the Ophidia framework to manipulate datasets*

Three interaction modes:
Operators, Workflows, Python Apps



	GEN	FEB	MAR	APR
38°	12.4	7.6	13.2	11.3
37°	18.4	13.6	14.1	16.3
36°	14.4	6.1	9.2	12.4
35°	21.3	17.8	23.5	22.1



<<Abstract>>
MD_ Metadata

+ fileIdentifier [0..1] : CharacterString
+ language [0..1] : CharacterString
+ characterSet [0..1] : MD_ CharacterSetCode = "utf8"
+ parentIdentifier [0..1] : Characte
+ hierarchyLevel [0..*] : MD_ Sco
+ hierarchyLevelName [0..*] : Cl
+ contact [1..*] : CI_ ResponsibleP
+ dateStamp : Date
+ metadataStandardName [0..1]
+ metadataStandardVersion [0.
+ datasetURI [0..1] : CharacterStr
+ locale [0..*] : PT_Locale

User metadata information

Metadata provenance

```
--> https://ophidia.cmcc.it:8443/162/169 (ROOT)
 ├─ https://ophidia.cmcc.it:8443/162/170 (oph_reduce)
 │     └─ https://ophidia.cmcc.it:8443/162/171 (oph_merge)
 │          ├─ https://ophidia.cmcc.it:8443/162/172 (oph_aggregate2)
 │          └─ https://ophidia.cmcc.it:8443/162/173 (oph_rollup)
 │               ├─ https://ophidia.cmcc.it:8443/162/174 (oph_reduce)
 │               └─ https://ophidia.cmcc.it:8443/162/175 (oph_reduce)
 ├─ https://ophidia.cmcc.it:8443/162/176 (oph_aggregate)
 └─ https://ophidia.cmcc.it:8443/162/177 (oph_aggregate)
```

System metadata of the datacube (size, distribution, etc.)

Ophidia in a nutshell

- eScience framework
- Server-side
- Parallel
- In-memory
- Declarative
- Datacube oriented (multi-dimensional OLAP support)

- (Shared) Sessions
- Workflows and applications
- Interactive and batch support
- HPC and HTC tasks
- Both domain-oriented (e.g. nc) and domain-agnostic support (e.g. OLAP)

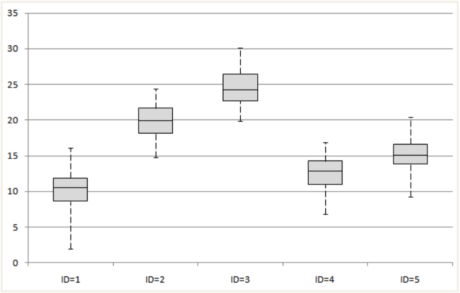
Data analytics requirements and use cases

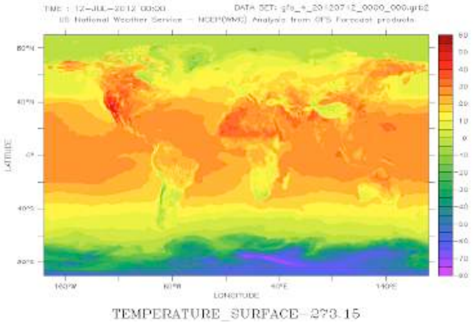
Requirements and needs focus on:

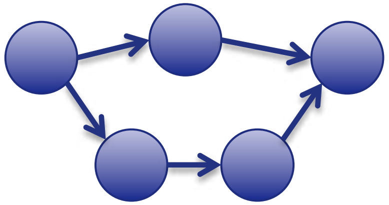
- *Data subsetting*
- *Statistical analysis*
- *Time series analysis*
- *Model intercomparison*
- *Multimodel mean*
- *Data reduction*
- *Data transformation*
- *Param. sweep experiments*
- *Maps production*
- *Workflow support*

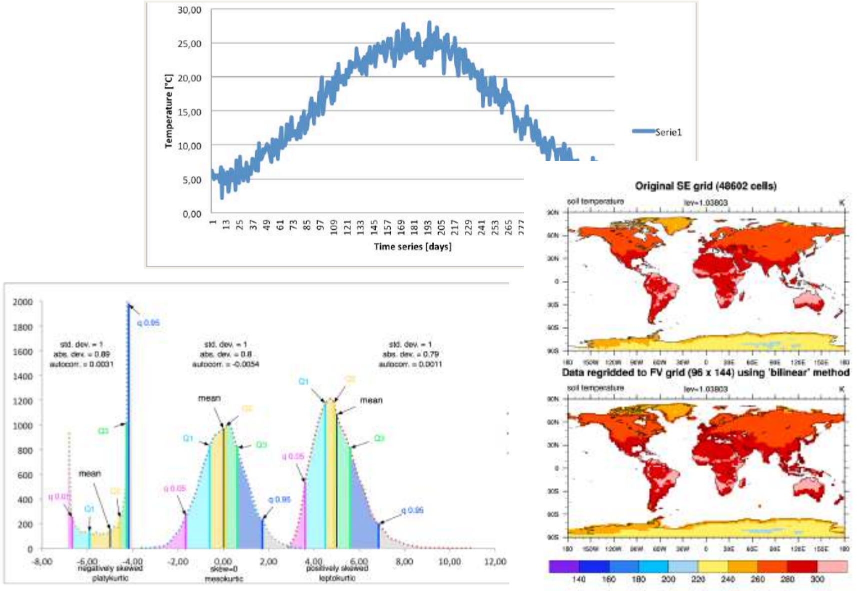
But also…

- *Performance*
- *re-usability*
- *extensibility*









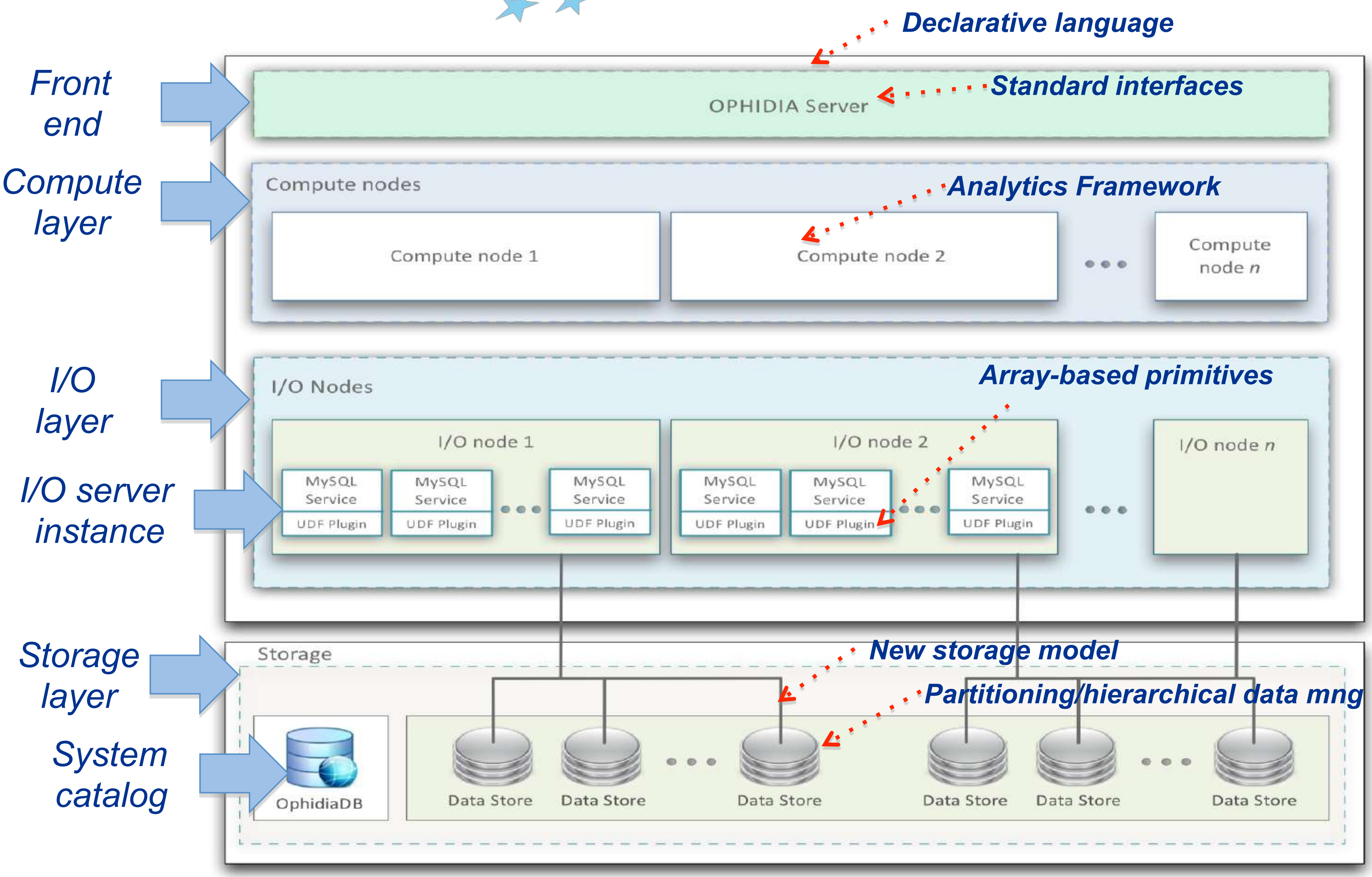
Core concepts

Storage model, primitive & operators

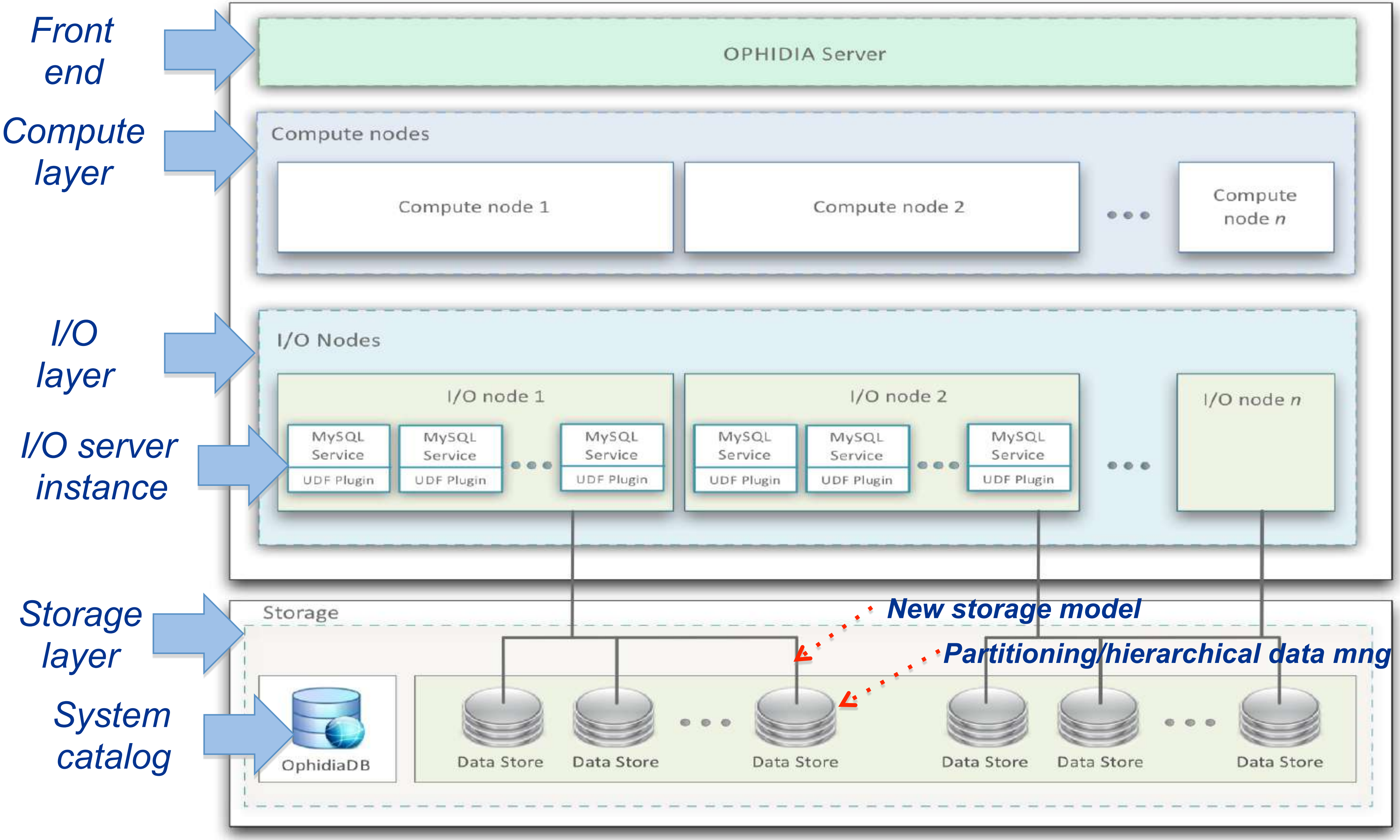
Collaborative
EUDAT Data Infrastructure

PRACE

Ophidia Architecture (sw stack view)



Storage model and chunks distribution



New storage model

Partitioning/hierarchical data mng

Ophidia storage model

- *The Ophidia storage model is a **two-step based evolution** of the **star schema** to support **scientific data management***
- *It relies on **implicit** (array-based) and **explicit** (tuple-based) **dimensions** for specific representations of data*
- *The first step includes the **support for array**-based data*
- *The second step includes a **key mapping** related to a set of foreign keys*
- *The second step makes the Ophidia storage model and implementation **independent of the number of dimensions**!*

Storage model (dimension-independent) & implementation

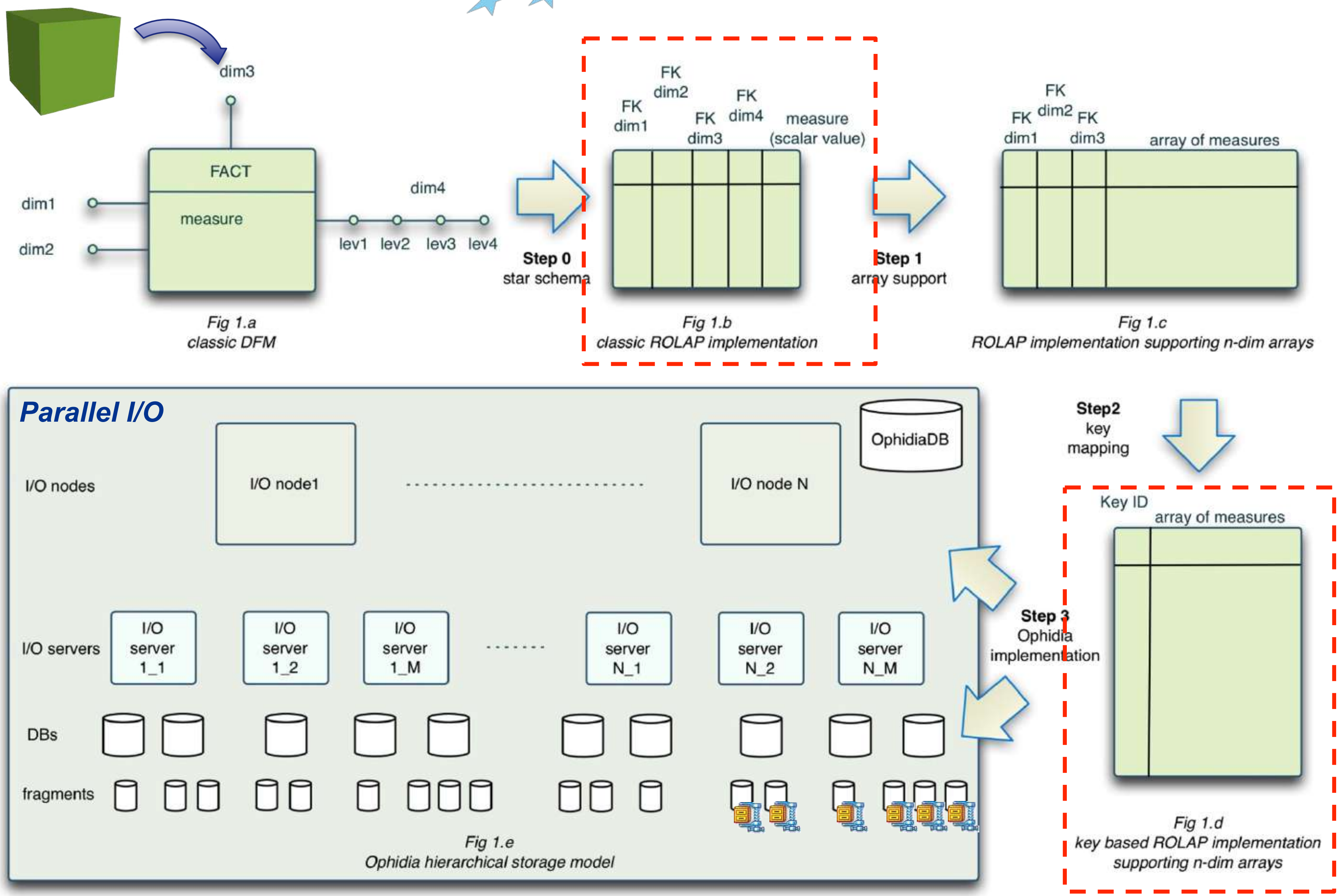


Fig 1.a
classic DFM

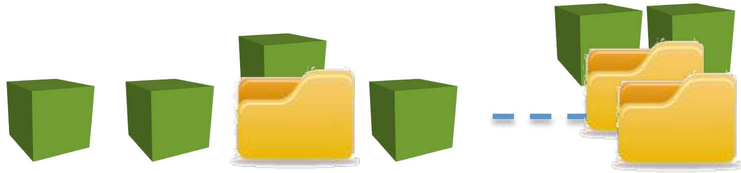
Fig 1.b
classic ROLAP implementation

Fig 1.c
ROLAP implementation supporting n-dim arrays

Fig 1.e
Ophidia hierarchical storage model

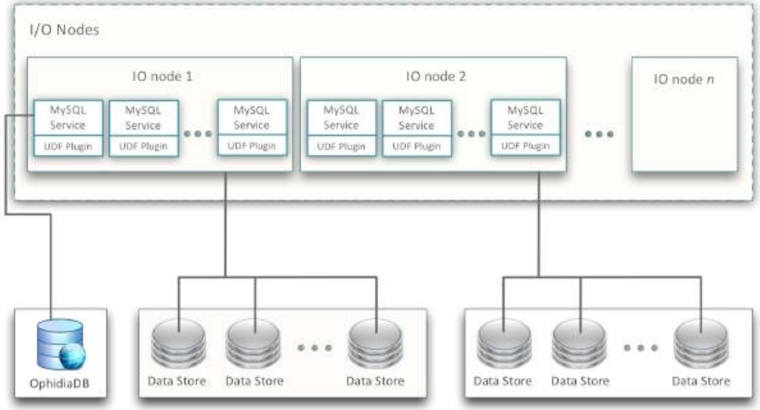
Fig 1.d
key based ROLAP implementation supporting n-dim arrays

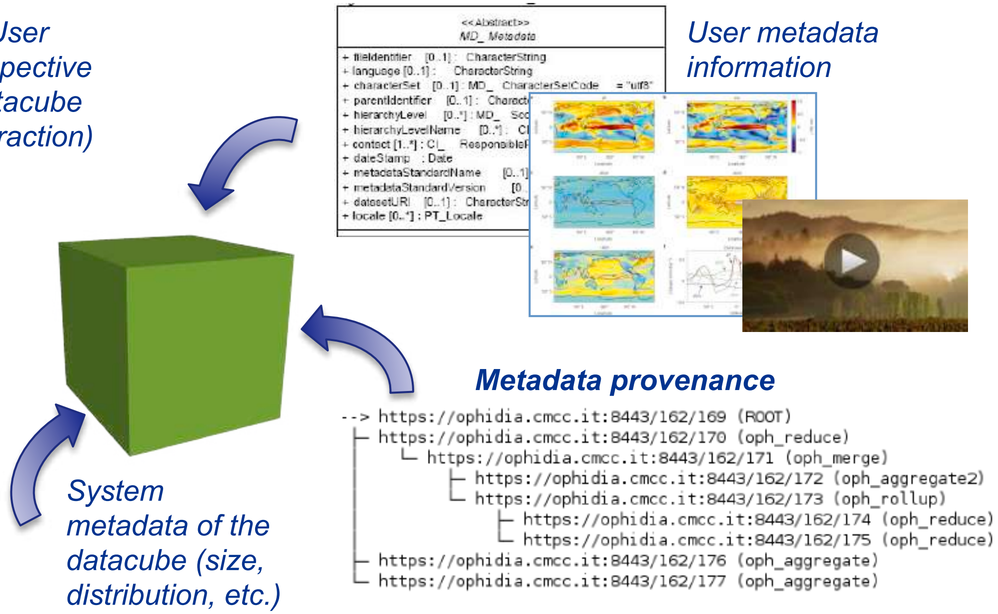
Data abstraction cube space perspective



User perspective (datacube abstraction)

System perspective (internal storage representation)





User metadata information

```
<<Abstract>>
MD_ Metadata
+ fileIdentifier   [0..1] :   CharacterString
+ language [0..1] :    CharacterString
+ characterSet   [0..1] : MD_   CharacterSetCode    = "utf8"
+ parentIdentifier   [0..1] :   Charact
+ hierarchyLevel    [0..*] : MD_   Sco
+ hierarchyLevelName    [0..*] :   Cl
+ contact [1..*] : CI_   ResponsibleP
+ dateStamp   : Date
+ metadataStandardName    [0..1]
+ metadataStandardVersion    [0
+ datasetURI   [0..1] :   CharacterStr
+ locale [0..*] : PT_Locale
```

Metadata provenance

```
--> https://ophidia.cmcc.it:8443/162/169 (ROOT)
 ├─ https://ophidia.cmcc.it:8443/162/170 (oph_reduce)
 │     └─ https://ophidia.cmcc.it:8443/162/171 (oph_merge)
 │           ├─ https://ophidia.cmcc.it:8443/162/172 (oph_aggregate2)
 │           └─ https://ophidia.cmcc.it:8443/162/173 (oph_rollup)
 │                 ├─ https://ophidia.cmcc.it:8443/162/174 (oph_reduce)
 │                 └─ https://ophidia.cmcc.it:8443/162/175 (oph_reduce)
 ├─ https://ophidia.cmcc.it:8443/162/176 (oph_aggregate)
 └─ https://ophidia.cmcc.it:8443/162/177 (oph_aggregate)
```

System metadata of the datacube (size, distribution, etc.)

Manage the Ophidia file system

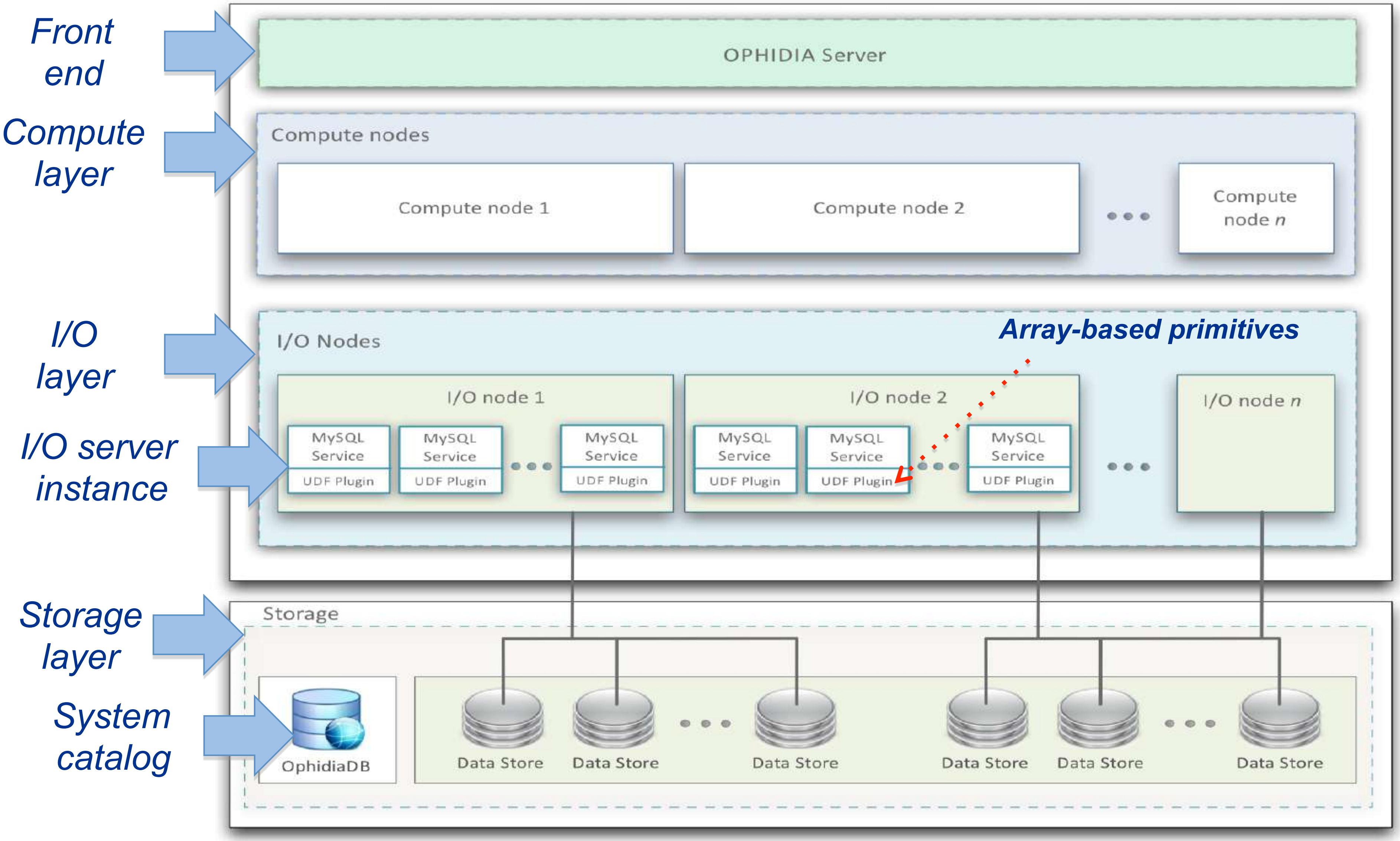
CMD	BEHAVIOR
cd	change directory
mkdir	create a new folder
rm	remove an empty folder or hide (logically delete) a container
ls	list subfolders and containers in a folder
mv	move/rename a folder or a container
…	…

Metadata associated to the datacubes

TYPE	CONTENT
Text	Plain text metadata
image	Binary string representation of an image
video	Binary string representation of a video
audio	Binary string representation of an audio stream
url	Text representing an URL

Search & Discovery

Array-based primitives

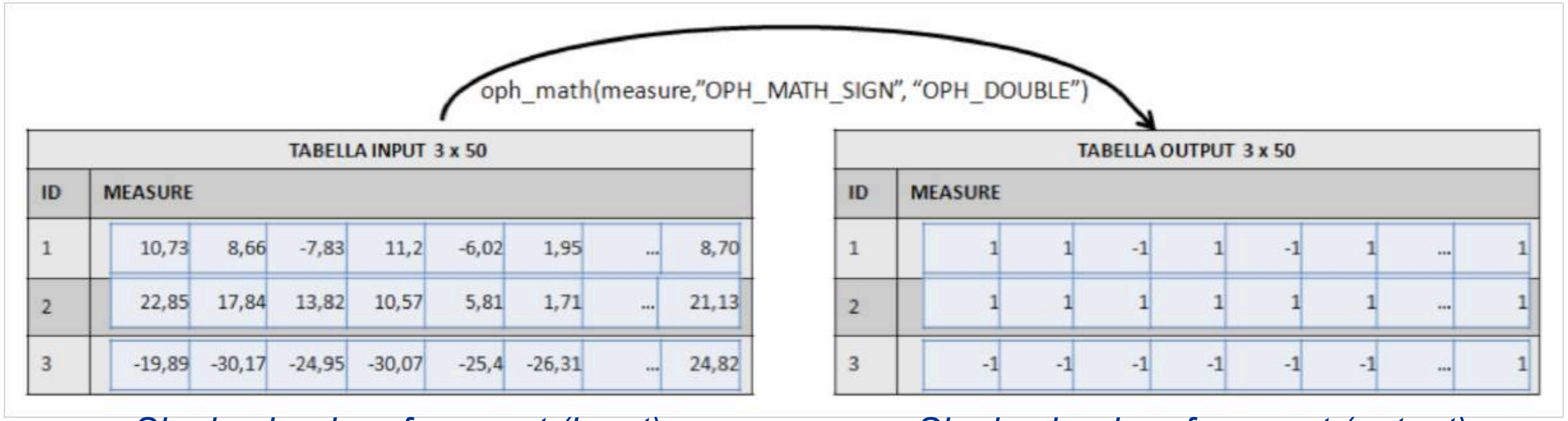


Array based primitives

- *Ophidia provides a **wide set of array-based primitives** to perform data summarization, sub-setting, predicates evaluation, statistical analysis, compression, etc.*
- *Primitives come as plugins and are applied on a single datacube chunk (fragment)*
- *They are provided both for **byte**-oriented and **bit**-oriented arrays*
- ***Primitives can be nested** to get more complex functionalities*
- ***Compression is a primitive too!***
- *New primitives can be easily integrated as additional plugins*

Array based primitives: OPH_MATH (“SIGN”)

oph_math(measure, “OPH_SIGN”, "OPH_DOUBLE”)



oph_math(measure,”OPH_MATH_SIGN”, “OPH_DOUBLE”)

TABELLA INPUT 3 x 50

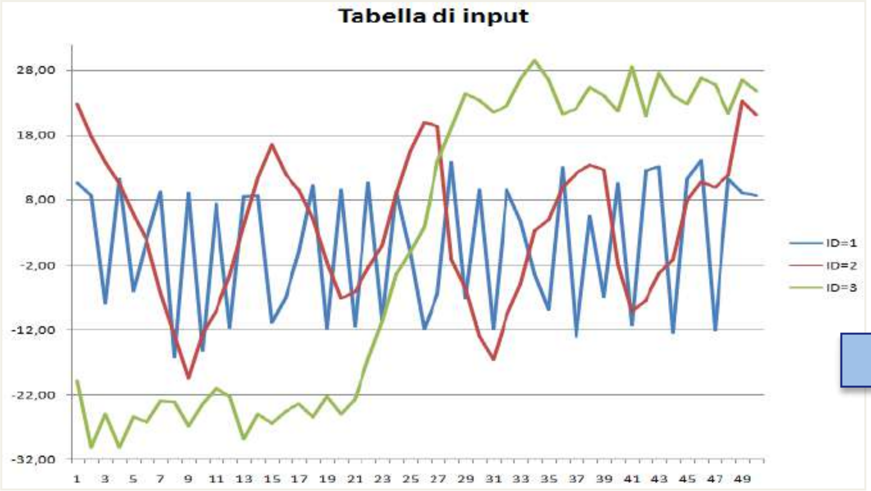
ID	MEASURE							
1	10,73	8,66	-7,83	11,2	-6,02	1,95	...	8,70
2	22,85	17,84	13,82	10,57	5,81	1,71	...	21,13
3	-19,89	-30,17	-24,95	-30,07	-25,4	-26,31	...	24,82

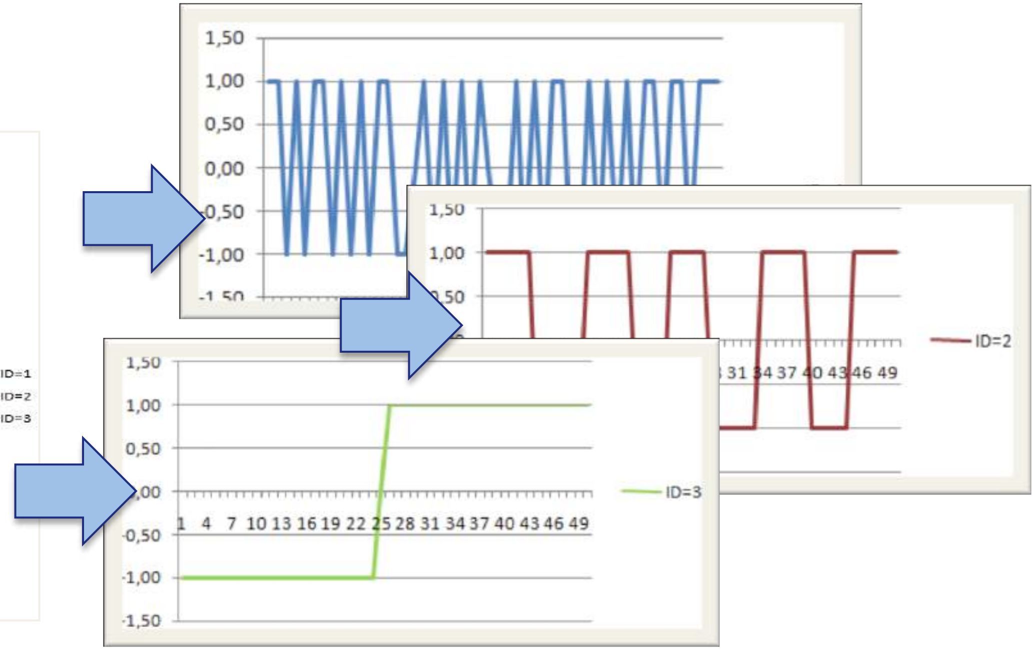
TABELLA OUTPUT 3 x 50

ID	MEASURE							
1	1	1	-1	1	-1	1	...	1
2	1	1	1	1	1	1	...	1
3	-1	-1	-1	-1	-1	-1	...	1

Single chunk or fragment (input)

Single chunk or fragment (output)





Array based primitives: OPH_BOXPLOT

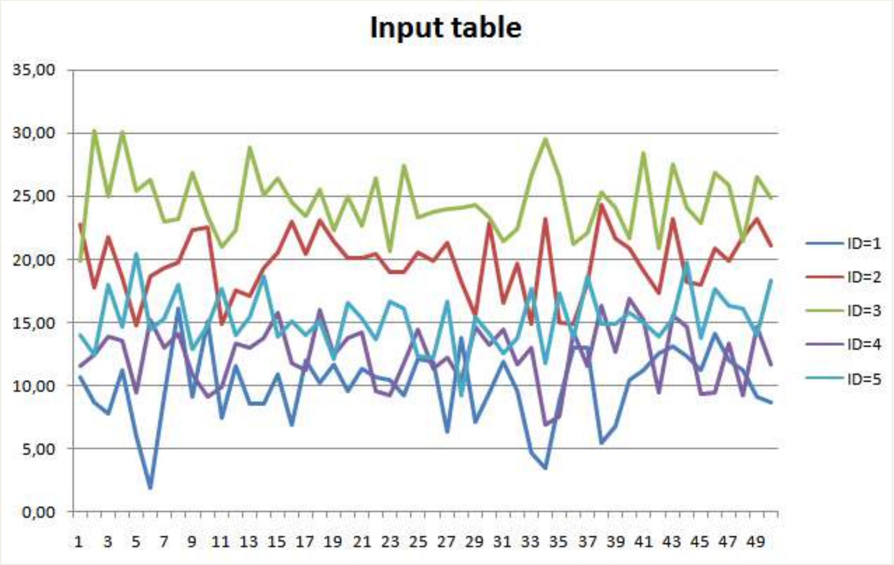
oph_boxplot*(measure, "OPH_DOUBLE”)*

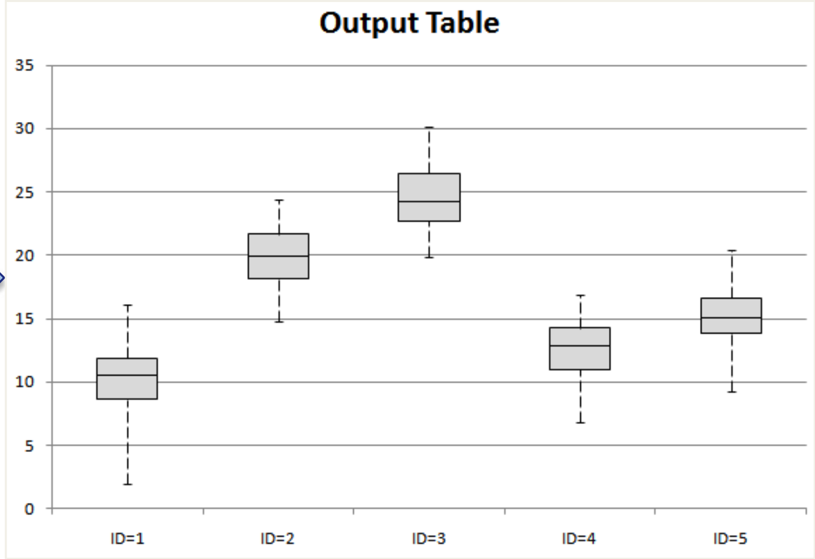
Single chunk or fragment (input)

INPUTTABLE 5 tuples x 50 elements										
ID	**MEASURE**									
1	10,73	8,66	7,83	11,20	6,02	1,95	9,25	16,11	...	8,70
2	22,85	17,84	21,82	18,57	14,81	18,71	19,31	19,83	...	21,13
3	19,89	30,17	24,95	30,07	25,40	26,31	22,95	23,18	...	24,82
4	11,60	12,49	13,91	13,53	9,48	15,27	13,05	14,17	...	11,66
5	13,94	12,43	17,95	14,70	20,41	14,46	15,37	18,00	...	18,30

Single chunk or fragment (output)

OUTPUTTABLE 5 tuples x 5 elements (summary)					
ID	**MEASURE**				
1	1,95	8,64	10,47	11,87	16,11
2	14,81	18,14	19,93	21,66	24,35
3	19,89	22,74	24,24	26,45	30,17
4	6,87	10,99	12,85	14,28	16,93
5	9,23	13,87	15,05	16,61	20,41

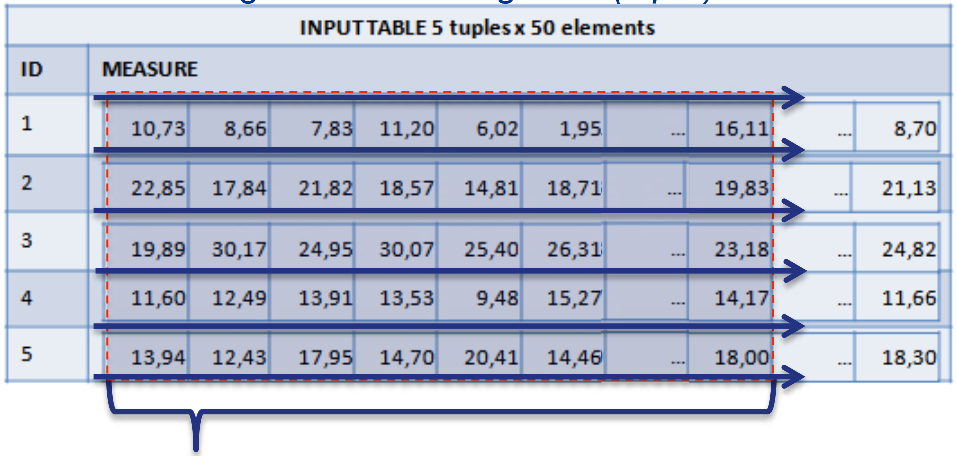




Array based primitives: nesting feature

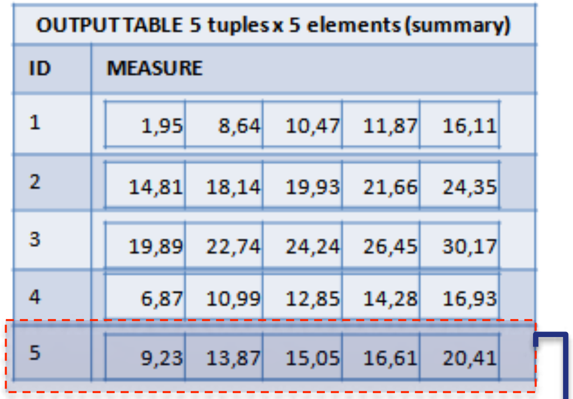
oph_boxplot(oph_subarray(oph_uncompress(measure), 1,18), "OPH_DOUBLE")

Single chunk or fragment (input)



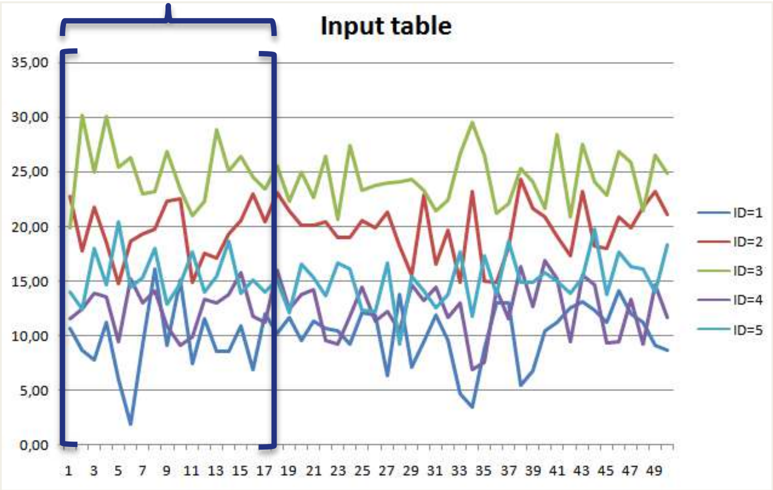
INPUTTABLE 5 tuples x 50 elements										
ID	MEASURE									
1	10,73	8,66	7,83	11,20	6,02	1,95	...	16,11	...	8,70
2	22,85	17,84	21,82	18,57	14,81	18,71	...	19,83	...	21,13
3	19,89	30,17	24,95	30,07	25,40	26,31	...	23,18	...	24,82
4	11,60	12,49	13,91	13,53	9,48	15,27	...	14,17	...	11,66
5	13,94	12,43	17,95	14,70	20,41	14,46	...	18,00	...	18,30

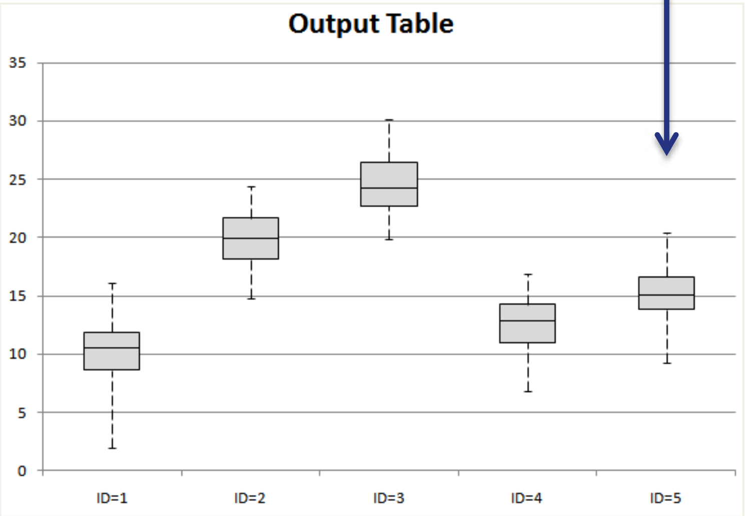
Single chunk or fragment (output)



OUTPUTTABLE 5 tuples x 5 elements (summary)					
ID	MEASURE				
1	1,95	8,64	10,47	11,87	16,11
2	14,81	18,14	19,93	21,66	24,35
3	19,89	22,74	24,24	26,45	30,17
4	6,87	10,99	12,85	14,28	16,93
5	9,23	13,87	15,05	16,61	20,41

subarray(measure, 1,18)





Array based primitives: OPH_AGGREGATE

oph_aggregate(measure,"oph_avg”)

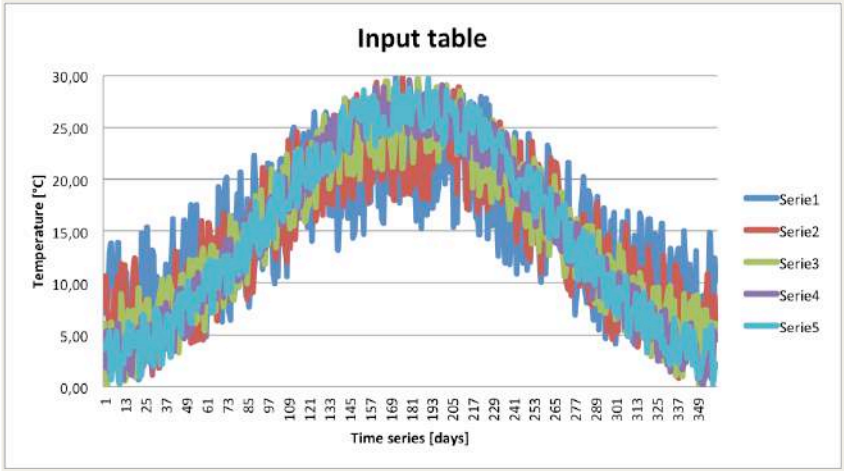
Single chunk or fragment (input)

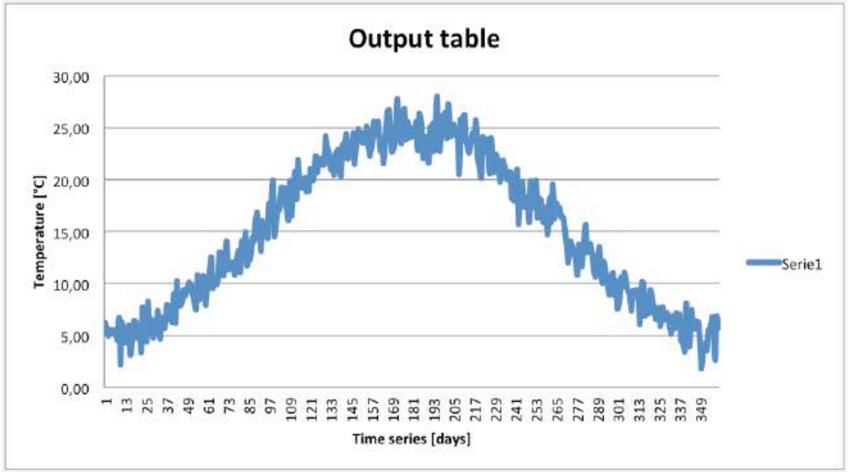
INPUT TABLE 5 tuples x 360 elements										
ID	MEASURE									
1	8,40	7,73	7,36	12,68	13,34	11,17	9,09	2,04	...	7,75
2	7,85	10,71	7,23	5,14	4,68	2,61	9,17	8,50	...	6,57
3	6,40	3,48	0,44	2,81	6,16	2,01	3,61	3,83	...	5,88
4	5,60	4,68	5,54	5,84	5,47	5,37	5,30	7,24	...	3,06
5	3,55	4,10	4,59	5,07	6,97	2,07	3,06	3,06	...	7,88

Vertical aggregation

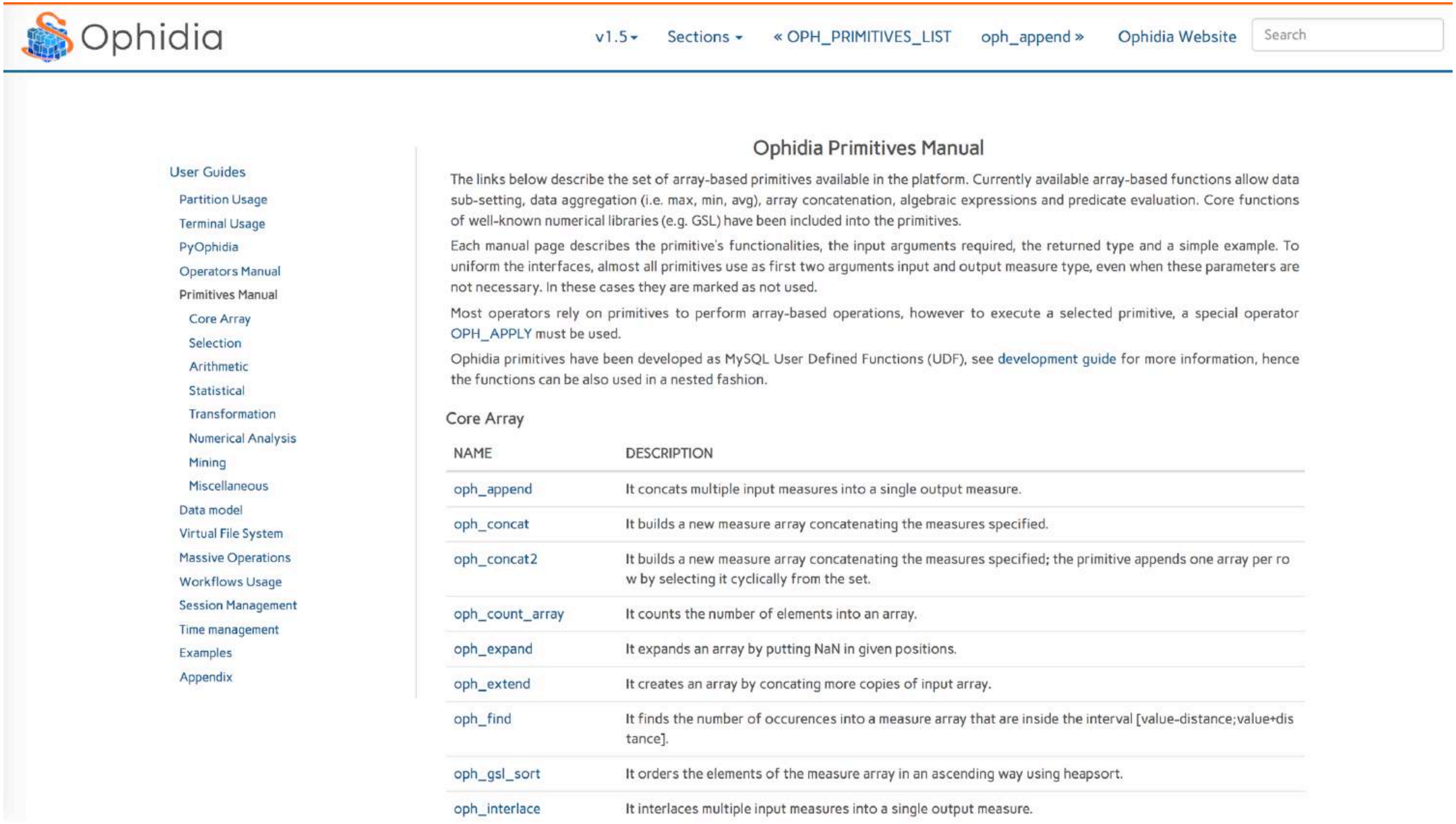
OUTPUT TABLE 1 tuple x 360 elements							
ID	MEASURE						
1	6,25	5,35	5,00	5,57	5,41	...	5,11

Single chunk or fragment (output)





Primitives documentation



Ophidia Primitives Manual

The links below describe the set of array-based primitives available in the platform. Currently available array-based functions allow data sub-setting, data aggregation (i.e. max, min, avg), array concatenation, algebraic expressions and predicate evaluation. Core functions of well-known numerical libraries (e.g. GSL) have been included into the primitives.

Each manual page describes the primitive's functionalities, the input arguments required, the returned type and a simple example. To uniform the interfaces, almost all primitives use as first two arguments input and output measure type, even when these parameters are not necessary. In these cases they are marked as not used.

Most operators rely on primitives to perform array-based operations, however to execute a selected primitive, a special operator OPH_APPLY must be used.

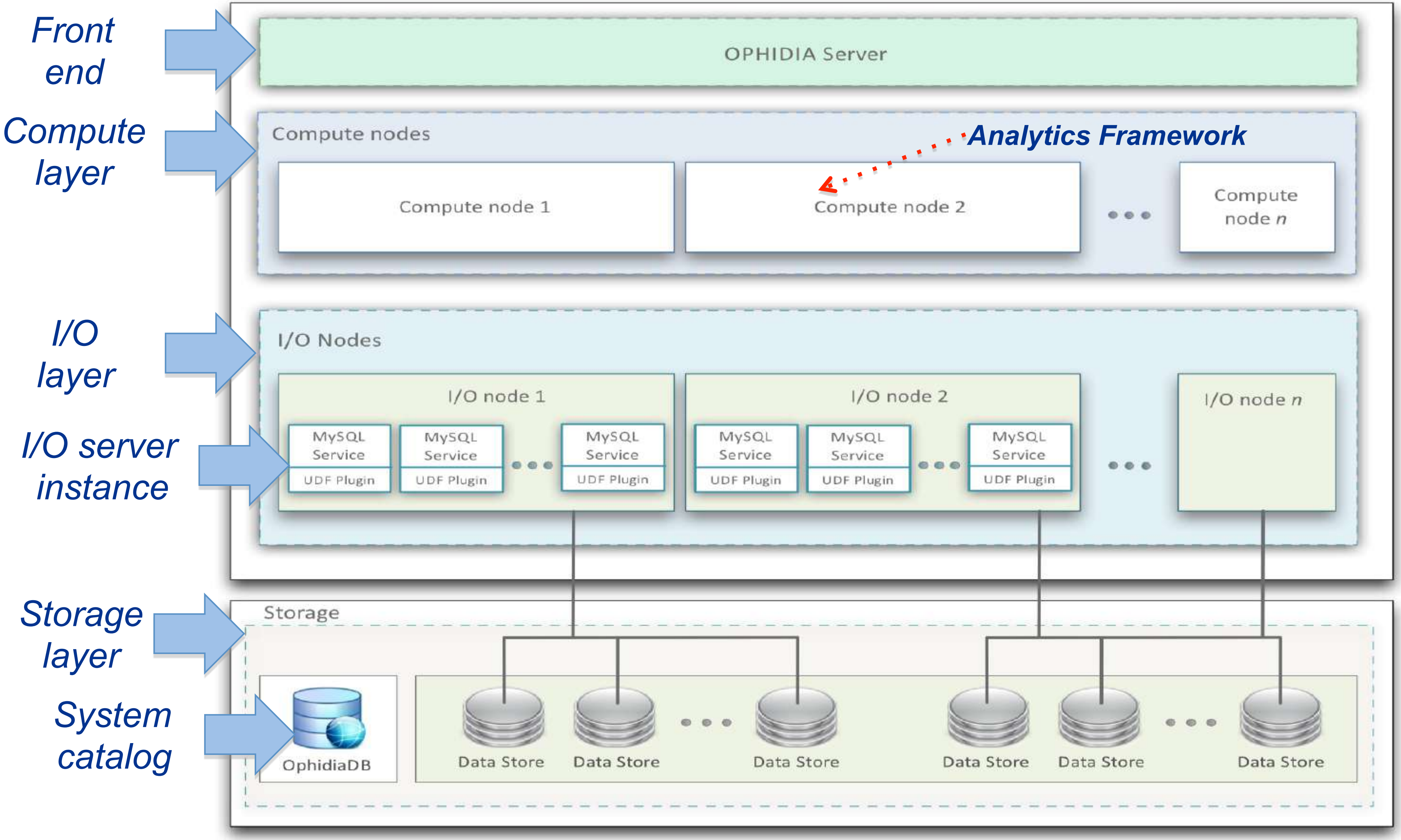
Ophidia primitives have been developed as MySQL User Defined Functions (UDF), see development guide for more information, hence the functions can be also used in a nested fashion.

Core Array

NAME	DESCRIPTION
oph_append	It concats multiple input measures into a single output measure.
oph_concat	It builds a new measure array concatenating the measures specified.
oph_concat2	It builds a new measure array concatenating the measures specified; the primitive appends one array per ro w by selecting it cyclically from the set.
oph_count_array	It counts the number of elements into an array.
oph_expand	It expands an array by putting NaN in given positions.
oph_extend	It creates an array by concating more copies of input array.
oph_find	It finds the number of occurences into a measure array that are inside the interval [value-distance;value+dis tance].
oph_gsl_sort	It orders the elements of the measure array in an ascending way using heapsort.
oph_interlace	It interlaces multiple input measures into a single output measure.

http://ophidia.cmcc.it/documentation/users/primitives/index.html

Analytics framework and datacube operators



Analytics framework and datacube operators

About 50 operators for data and metadata processing

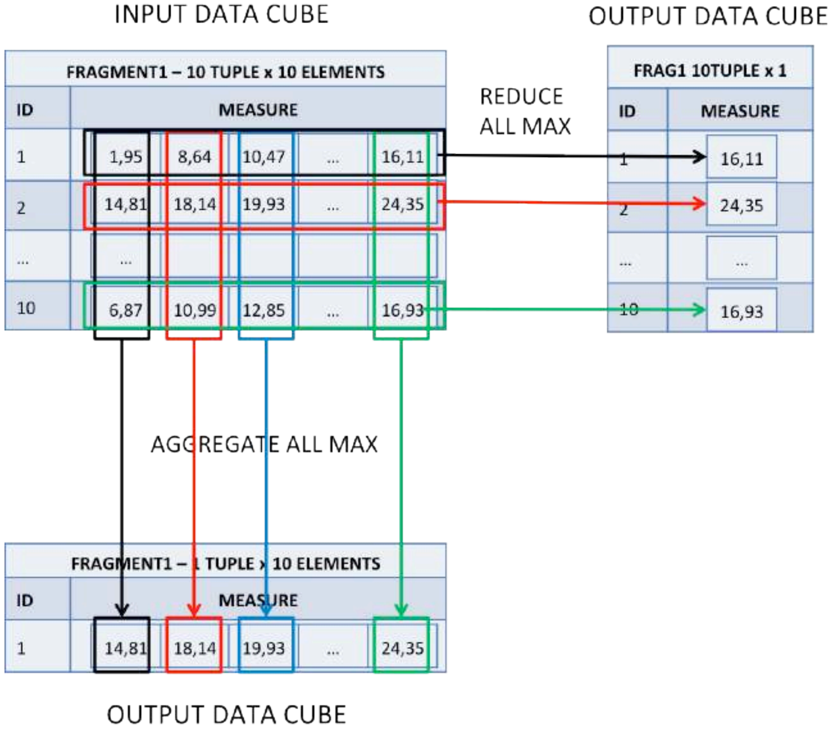
Data Operator	Description
OPH_CONCATNC	Concatenates a NetCDF file to a data cube.
OPH_DELETE	Deletes a data cube.
OPH_DUPLICATE	Duplicates a data cube.
OPH_EXPLORECUBE	Shows the content of a data cube.
OPH_EXPORTNC	Exports a whole data cube into a single NetCDF file.
OPH_IMPORTNC	Creates new a data cube importing data from a NetCDF file.
OPH_INTERCOMPARISON	Generates the difference value-by-value between two homogeneous data cubes.
OPH_INTERCUBE	It executes an operation between two data cubes and returns a new data cube as result of the specified operation applied element by element.
OPH_MERGECUBES	Merges the measures of n input data cubes creating a new data cube with the union of the n measures.
OPH_PUBLISH	Generates web pages representing the data stored in the fragments.
OPH_RANDCUBE	Creates a new data cube with random data.
OPH_REDUCE	Applies a data reduction operation along one or more implicit dimensions.
OPH_SCRIPT	Executes a bash script.
OPH_SUBSET	Extracts a subset from a data cube using the values of the dimensions.

Metadata Operator	Description
OPH_CUBEELEMENTS	Computes and displays the total number of elements contained in a data cube.
OPH_CUBEIO	Shows the provenance of a data cube.
OPH_CUBESCHEMA	Displays the metadata and dimension information associated to a data cube.
OPH_CUBESIZE	Computes and displays the total size (on disk) of a data cube.
OPH_FIND	Finds a data cube.
OPH_LIST	Displays the list of data cubes and containers available.
OPH_LOGGINGBK	Shows session and job information.
OPH_MAN	Shows a description about an operator or primitive.
OPH_METADATA	Manages metadata information.
OPH_OPERATORS_LIST	Displays the list of available operators.

oph_apply operator to run any primitive on a datacube
y=f(x) → oph_apply(oph_boxplot)

*oph_apply(oph_predicate('oph_float','oph_int',measure,'**x-298.15**','**>0**','**1**','**0**')")*

Analytics framework and datacube operators



INPUT DATA CUBE

FRAGMENT1 – 10 TUPLE x 10 ELEMENTS					
ID	MEASURE				
1	1,95	8,64	10,47	...	16,11
2	14,81	18,14	19,93	...	24,35
...	...				
10	6,87	10,99	12,85	...	16,93

REDUCE ALL MAX

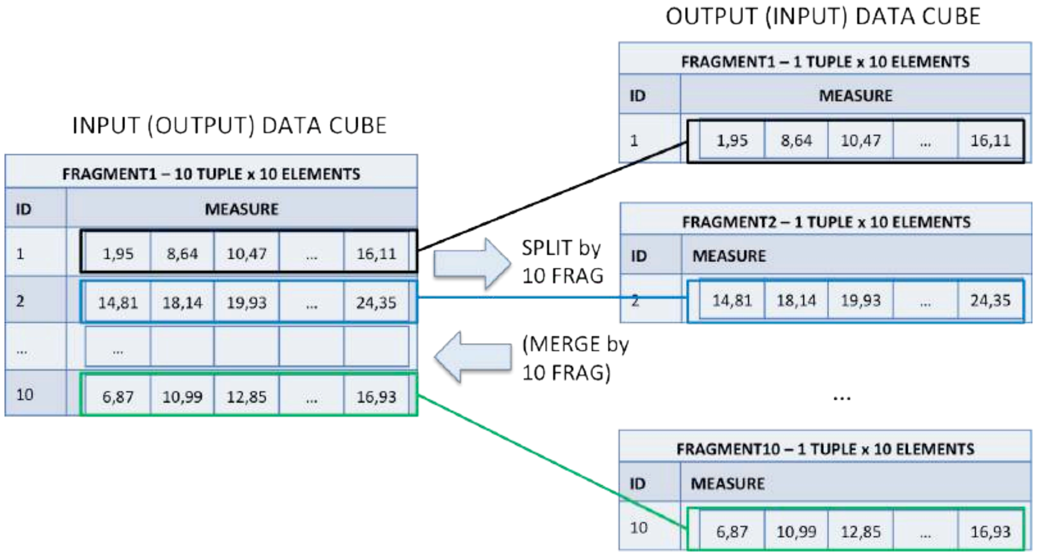
OUTPUT DATA CUBE

FRAG1 10TUPLE x 1	
ID	MEASURE
1	16,11
2	24,35
...	...
10	16,93

AGGREGATE ALL MAX

FRAGMENT1 – 1 TUPLE x 10 ELEMENTS					
ID	MEASURE				
1	14,81	18,14	19,93	...	24,35

OUTPUT DATA CUBE



INPUT (OUTPUT) DATA CUBE

FRAGMENT1 – 10 TUPLE x 10 ELEMENTS					
ID	MEASURE				
1	1,95	8,64	10,47	...	16,11
2	14,81	18,14	19,93	...	24,35
...	...				
10	6,87	10,99	12,85	...	16,93

SPLIT by 10 FRAG

(MERGE by 10 FRAG)

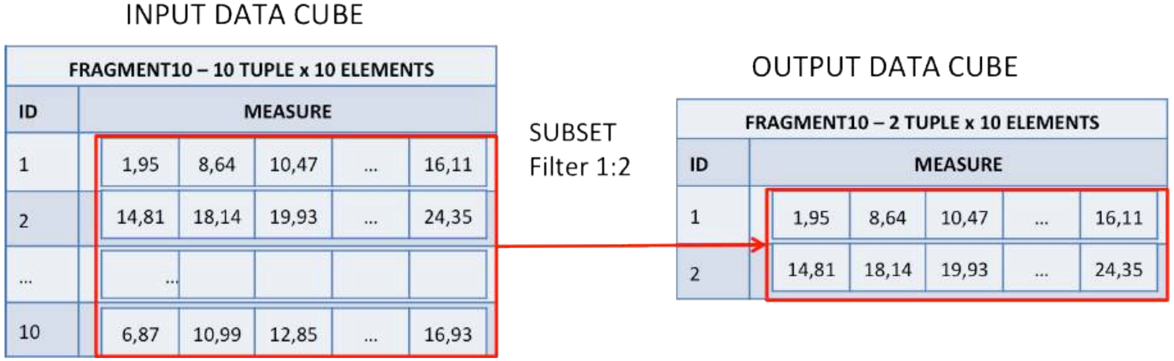
OUTPUT (INPUT) DATA CUBE

FRAGMENT1 – 1 TUPLE x 10 ELEMENTS					
ID	MEASURE				
1	1,95	8,64	10,47	...	16,11

FRAGMENT2 – 1 TUPLE x 10 ELEMENTS					
ID	MEASURE				
2	14,81	18,14	19,93	...	24,35

...

FRAGMENT10 – 1 TUPLE x 10 ELEMENTS					
ID	MEASURE				
10	6,87	10,99	12,85	...	16,93



INPUT DATA CUBE

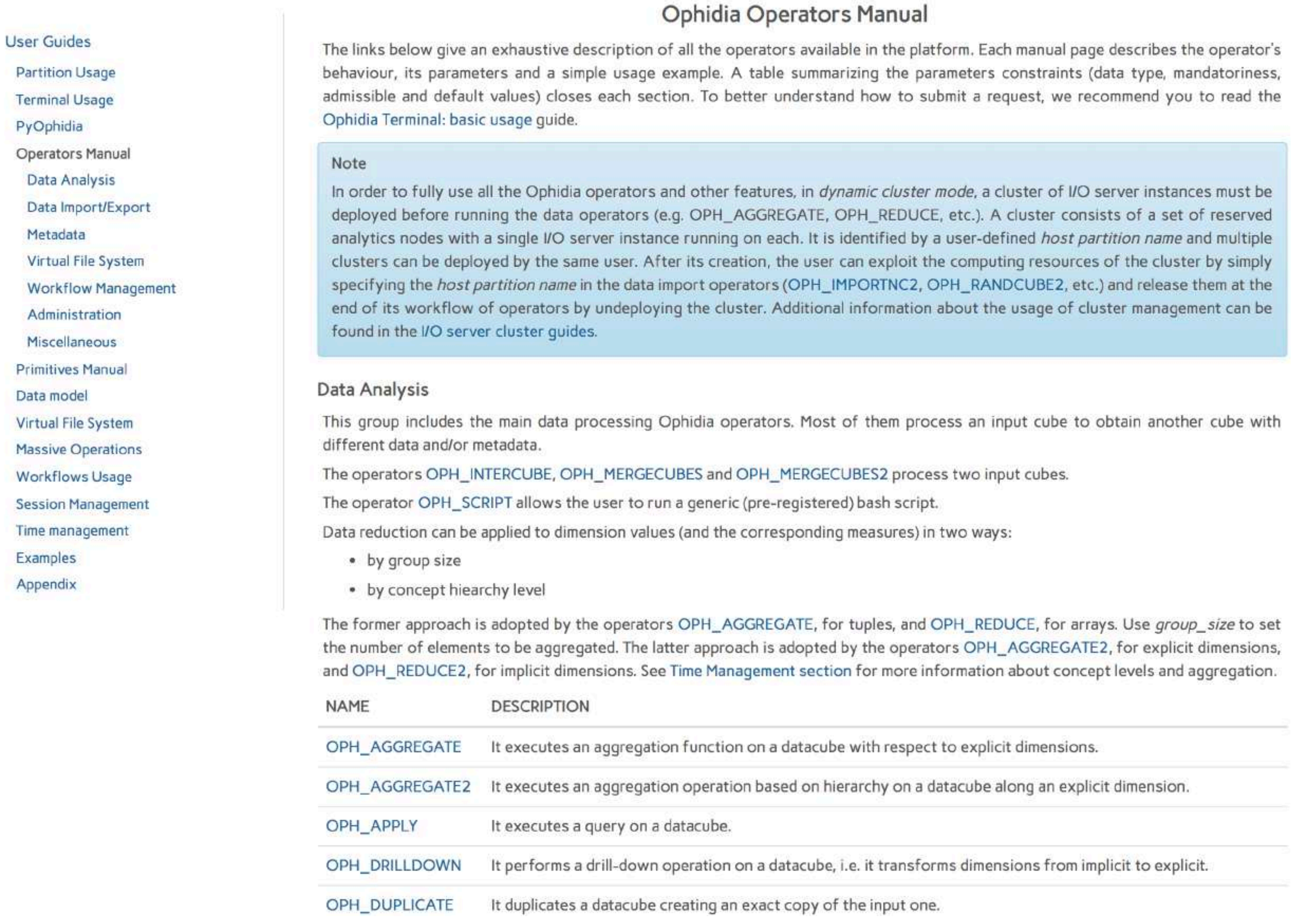
FRAGMENT10 – 10 TUPLE x 10 ELEMENTS					
ID	MEASURE				
1	1,95	8,64	10,47	...	16,11
2	14,81	18,14	19,93	...	24,35
...	...				
10	6,87	10,99	12,85	...	16,93

SUBSET Filter 1:2

OUTPUT DATA CUBE

FRAGMENT10 – 2 TUPLE x 10 ELEMENTS					
ID	MEASURE				
1	1,95	8,64	10,47	...	16,11
2	14,81	18,14	19,93	...	24,35

The analytics framework: "data" operators



- User Guides
  - Partition Usage
  - Terminal Usage
  - PyOphidia
  - Operators Manual
    - Data Analysis
    - Data Import/Export
    - Metadata
    - Virtual File System
    - Workflow Management
    - Administration
    - Miscellaneous
  - Primitives Manual
  - Data model
  - Virtual File System
  - Massive Operations
  - Workflows Usage
  - Session Management
  - Time management
  - Examples
  - Appendix

Ophidia Operators Manual

The links below give an exhaustive description of all the operators available in the platform. Each manual page describes the operator's behaviour, its parameters and a simple usage example. A table summarizing the parameters constraints (data type, mandatoriness, admissible and default values) closes each section. To better understand how to submit a request, we recommend you to read the Ophidia Terminal: basic usage guide.

> Note
>
> In order to fully use all the Ophidia operators and other features, in *dynamic cluster mode*, a cluster of I/O server instances must be deployed before running the data operators (e.g. OPH_AGGREGATE, OPH_REDUCE, etc.). A cluster consists of a set of reserved analytics nodes with a single I/O server instance running on each. It is identified by a user-defined *host partition name* and multiple clusters can be deployed by the same user. After its creation, the user can exploit the computing resources of the cluster by simply specifying the *host partition name* in the data import operators (OPH_IMPORTNC2, OPH_RANDCUBE2, etc.) and release them at the end of its workflow of operators by undeploying the cluster. Additional information about the usage of cluster management can be found in the I/O server cluster guides.

Data Analysis

This group includes the main data processing Ophidia operators. Most of them process an input cube to obtain another cube with different data and/or metadata.

The operators OPH_INTERCUBE, OPH_MERGECUBES and OPH_MERGECUBES2 process two input cubes.

The operator OPH_SCRIPT allows the user to run a generic (pre-registered) bash script.

Data reduction can be applied to dimension values (and the corresponding measures) in two ways:

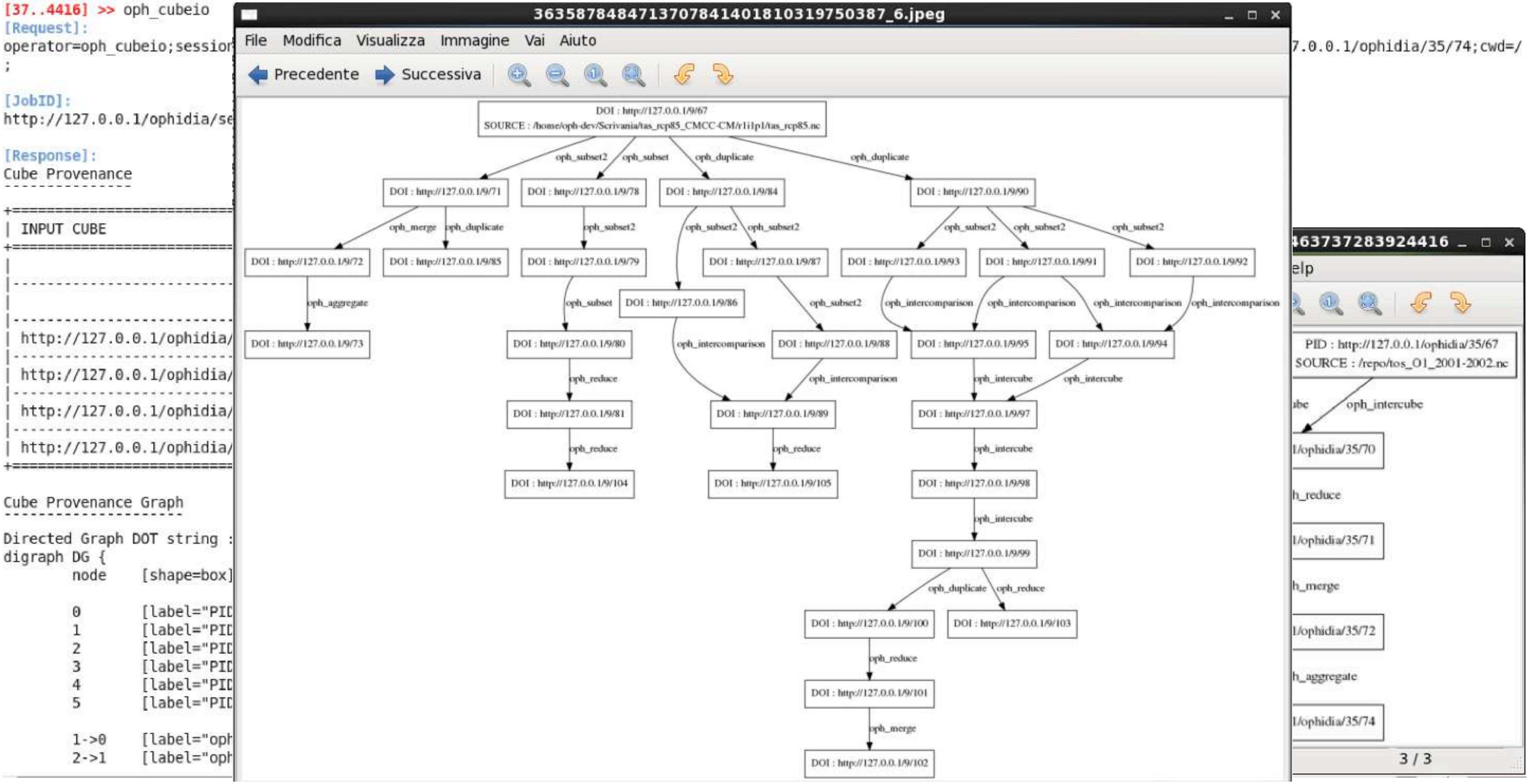
- by group size
- by concept hiearchy level

The former approach is adopted by the operators OPH_AGGREGATE, for tuples, and OPH_REDUCE, for arrays. Use *group_size* to set the number of elements to be aggregated. The latter approach is adopted by the operators OPH_AGGREGATE2, for explicit dimensions, and OPH_REDUCE2, for implicit dimensions. See Time Management section for more information about concept levels and aggregation.

NAME	DESCRIPTION
OPH_AGGREGATE	It executes an aggregation function on a datacube with respect to explicit dimensions.
OPH_AGGREGATE2	It executes an aggregation operation based on hierarchy on a datacube along an explicit dimension.
OPH_APPLY	It executes a query on a datacube.
OPH_DRILLDOWN	It performs a drill-down operation on a datacube, i.e. it transforms dimensions from implicit to explicit.
OPH_DUPLICATE	It duplicates a datacube creating an exact copy of the input one.

http://ophidia.cmcc.it/documentation/users/operators/index.html

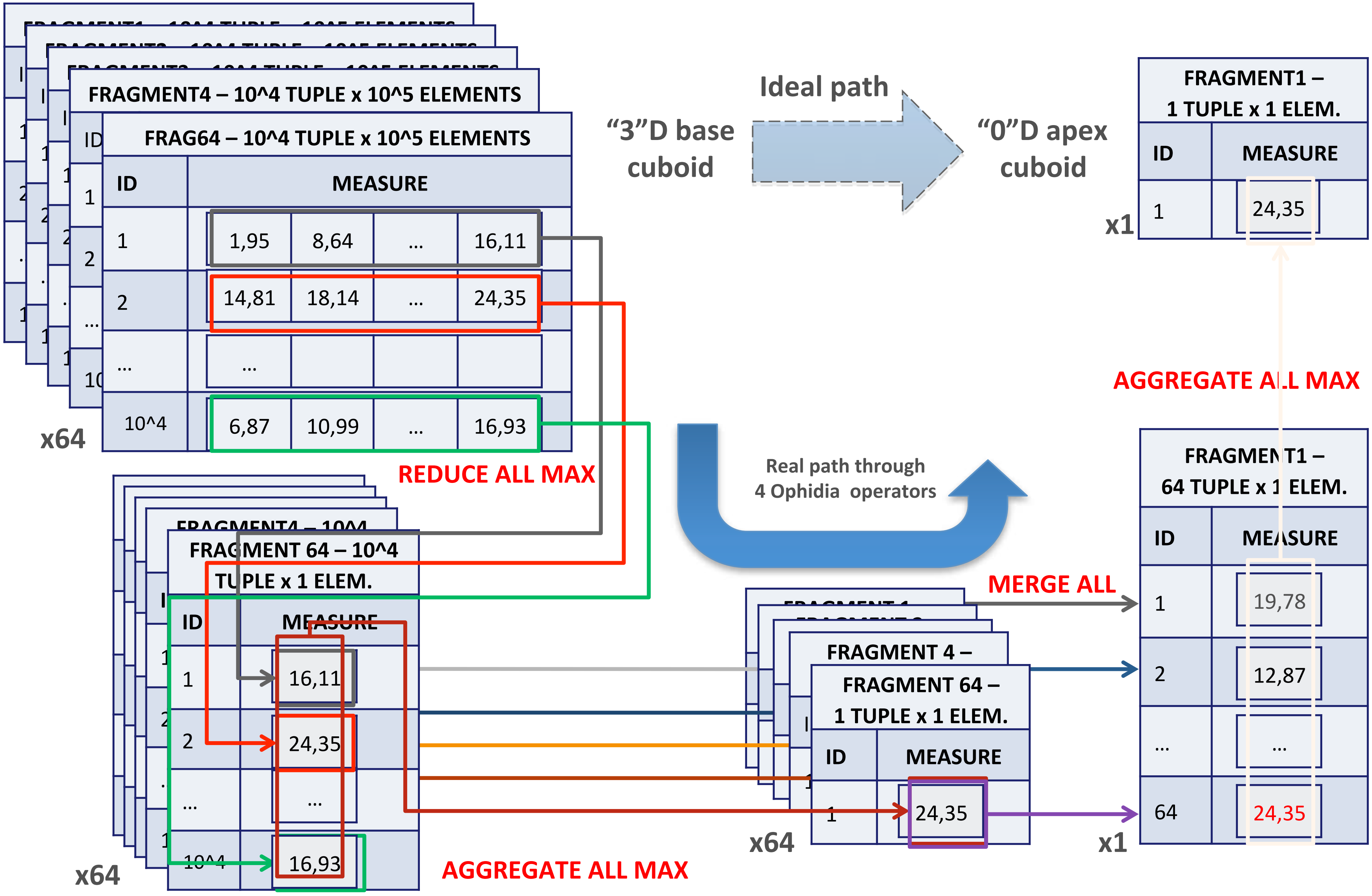
The analytics framework: metadata operators



```
[37..4416] >> oph_cubeio
[Request]:
operator=oph_cubeio;sessio                                                    7.0.0.1/ophidia/35/74;cwd=/
;

[JobID]:
http://127.0.0.1/ophidia/se

[Response]:
Cube Provenance
---------------

+===========================
| INPUT CUBE
+===========================
|
|---------------------------
|
|---------------------------
| http://127.0.0.1/ophidia/
|---------------------------
| http://127.0.0.1/ophidia/
|---------------------------
| http://127.0.0.1/ophidia/
|---------------------------
| http://127.0.0.1/ophidia/
+===========================

Cube Provenance Graph
---------------------

Directed Graph DOT string :
digraph DG {
        node    [shape=box]

        0       [label="PID
        1       [label="PID
        2       [label="PID
        3       [label="PID
        4       [label="PID
        5       [label="PID

        1->0    [label="oph
        2->1    [label="oph
```

3635878484713707841401810319750387_6.jpeg
File
Modifica
Visualizza
Immagine
Vai
Aiuto
Precedente
Successiva
DOI : http://127.0.0.1/9/67
SOURCE : /home/oph-dev/Scrivania/tas_rcp85_CMCC-CM/r1i1p1/tas_rcp85.nc
oph_subset2
oph_subset
oph_duplicate
oph_duplicate
DOI : http://127.0.0.1/9/71
DOI : http://127.0.0.1/9/78
DOI : http://127.0.0.1/9/84
DOI : http://127.0.0.1/9/90
oph_merge
oph_duplicate
oph_subset2
oph_subset2
oph_subset2
oph_subset2
oph_subset2
oph_subset2
DOI : http://127.0.0.1/9/72
DOI : http://127.0.0.1/9/85
DOI : http://127.0.0.1/9/79
DOI : http://127.0.0.1/9/87
DOI : http://127.0.0.1/9/93
DOI : http://127.0.0.1/9/91
DOI : http://127.0.0.1/9/92
oph_aggregate
oph_subset
DOI : http://127.0.0.1/9/86
oph_subset2
oph_intercomparison
oph_intercomparison
oph_intercomparison
oph_intercomparison
DOI : http://127.0.0.1/9/73
DOI : http://127.0.0.1/9/80
oph_intercomparison
DOI : http://127.0.0.1/9/88
DOI : http://127.0.0.1/9/95
DOI : http://127.0.0.1/9/94
oph_reduce
oph_intercomparison
oph_intercube
oph_intercube
DOI : http://127.0.0.1/9/81
DOI : http://127.0.0.1/9/89
DOI : http://127.0.0.1/9/97
oph_reduce
oph_reduce
oph_intercube
DOI : http://127.0.0.1/9/104
DOI : http://127.0.0.1/9/105
DOI : http://127.0.0.1/9/98
oph_intercube
DOI : http://127.0.0.1/9/99
oph_duplicate
oph_reduce
DOI : http://127.0.0.1/9/100
DOI : http://127.0.0.1/9/103
oph_reduce
DOI : http://127.0.0.1/9/101
oph_merge
DOI : http://127.0.0.1/9/102
463737283924416
elp
PID : http://127.0.0.1/ophidia/35/67
SOURCE : /repo/tos_O1_2001-2002.nc
ube
oph_intercube
1/ophidia/35/70
h_reduce
1/ophidia/35/71
h_merge
1/ophidia/35/72
h_aggregate
1/ophidia/35/74
3 / 3

Pipelining analytics operators to reduce data



FRAG64 – 10^4 TUPLE x 10^5 ELEMENTS

ID	MEASURE			
1	1,95	8,64	…	16,11
2	14,81	18,14	…	24,35
…	…			
10^4	6,87	10,99	…	16,93

FRAGMENT 64 – 10^4 TUPLE x 1 ELEM.

ID	MEASURE
1	16,11
2	24,35
…	…
10^4	16,93

FRAGMENT 64 – 1 TUPLE x 1 ELEM.

ID	MEASURE
1	24,35

FRAGMENT1 – 64 TUPLE x 1 ELEM.

ID	MEASURE
1	19,78
2	12,87
…	…
64	24,35

FRAGMENT1 – 1 TUPLE x 1 ELEM.

ID	MEASURE
1	24,35

Advanced features

Workflows management,

In-memory analytics, Python binding

Architecture evolution

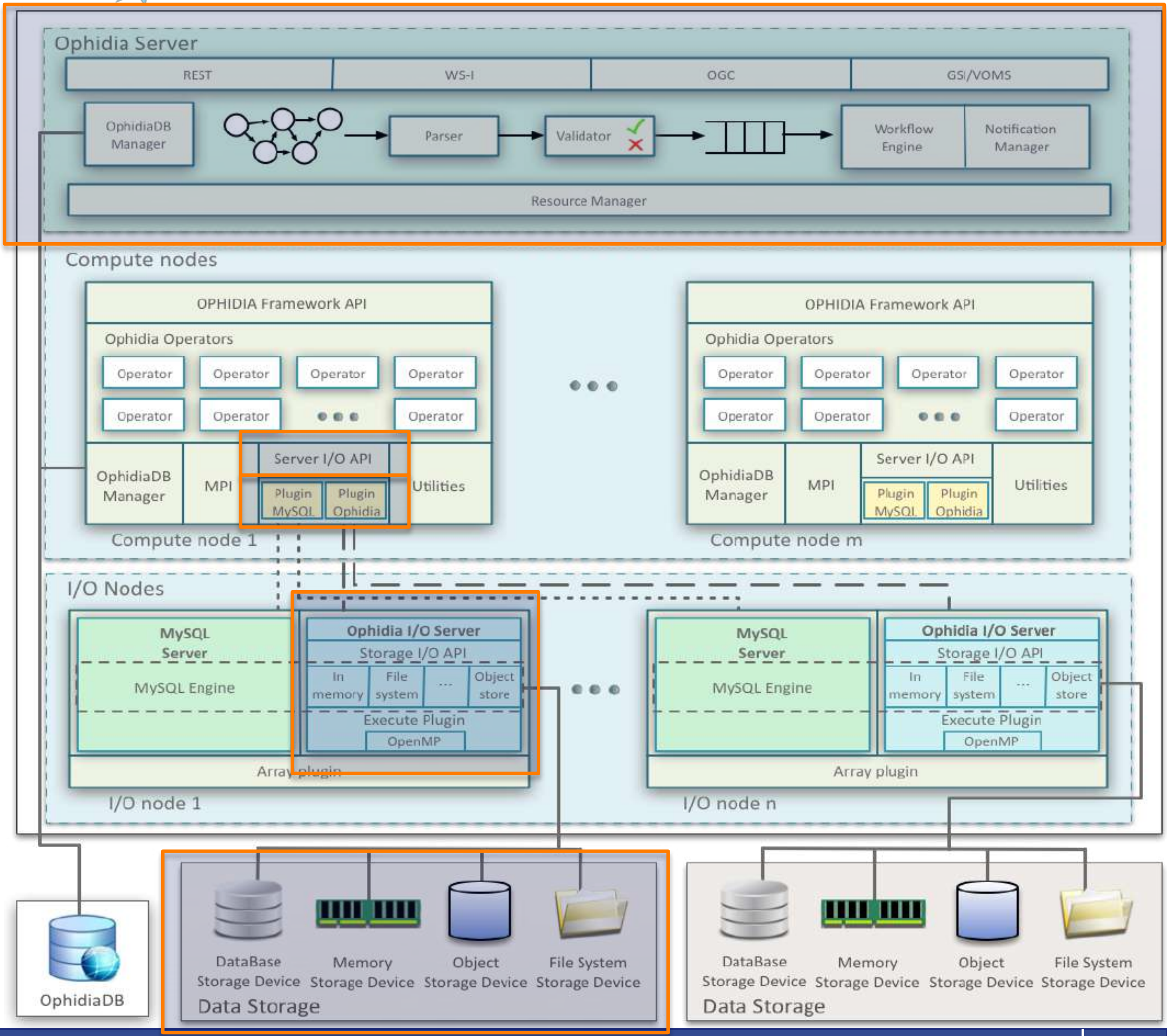
Workflow support on the server side

Separation of concerns between framework and I/O components

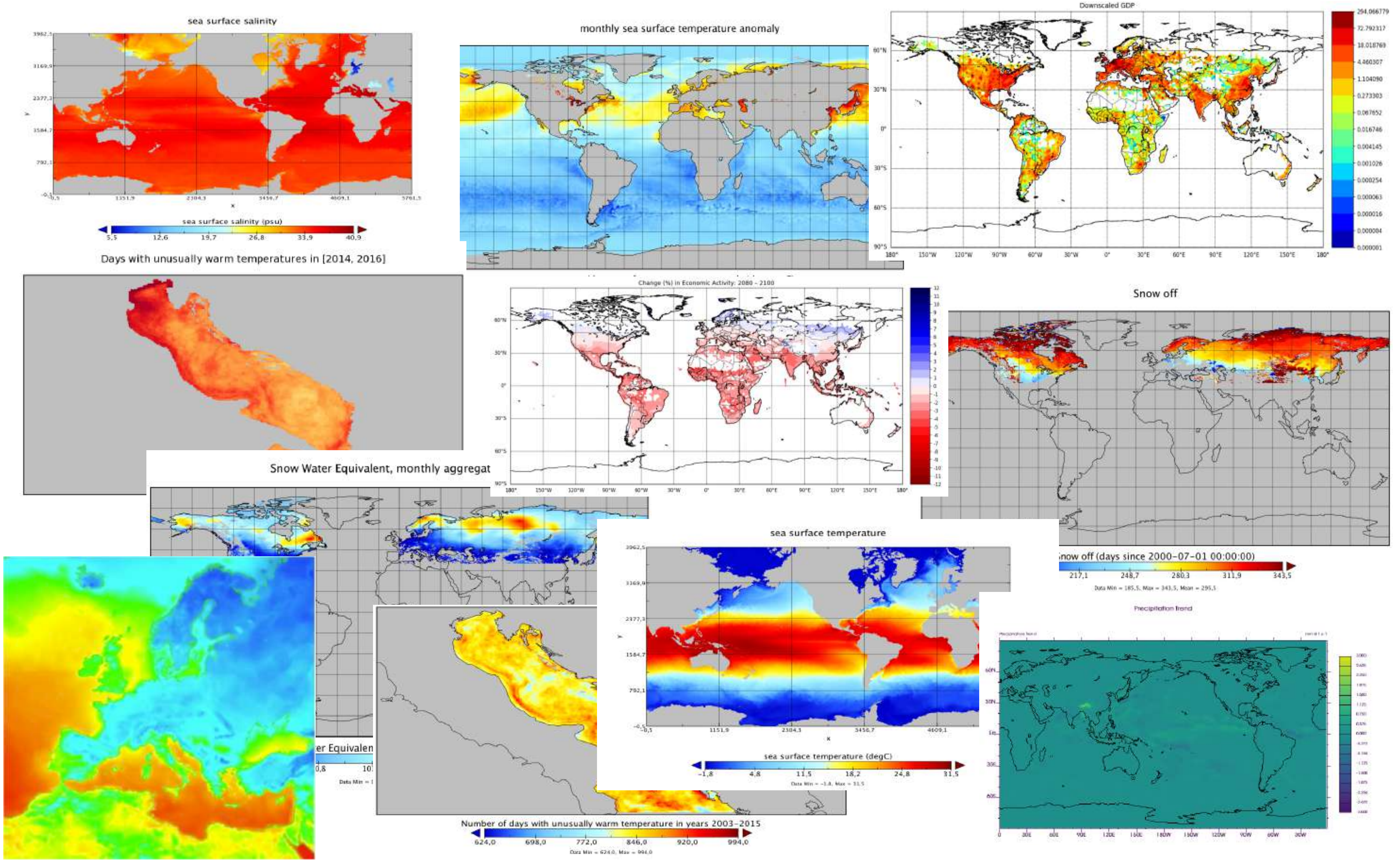
Support different **I/O servers**

Native I/O server with **parallel execution engine**

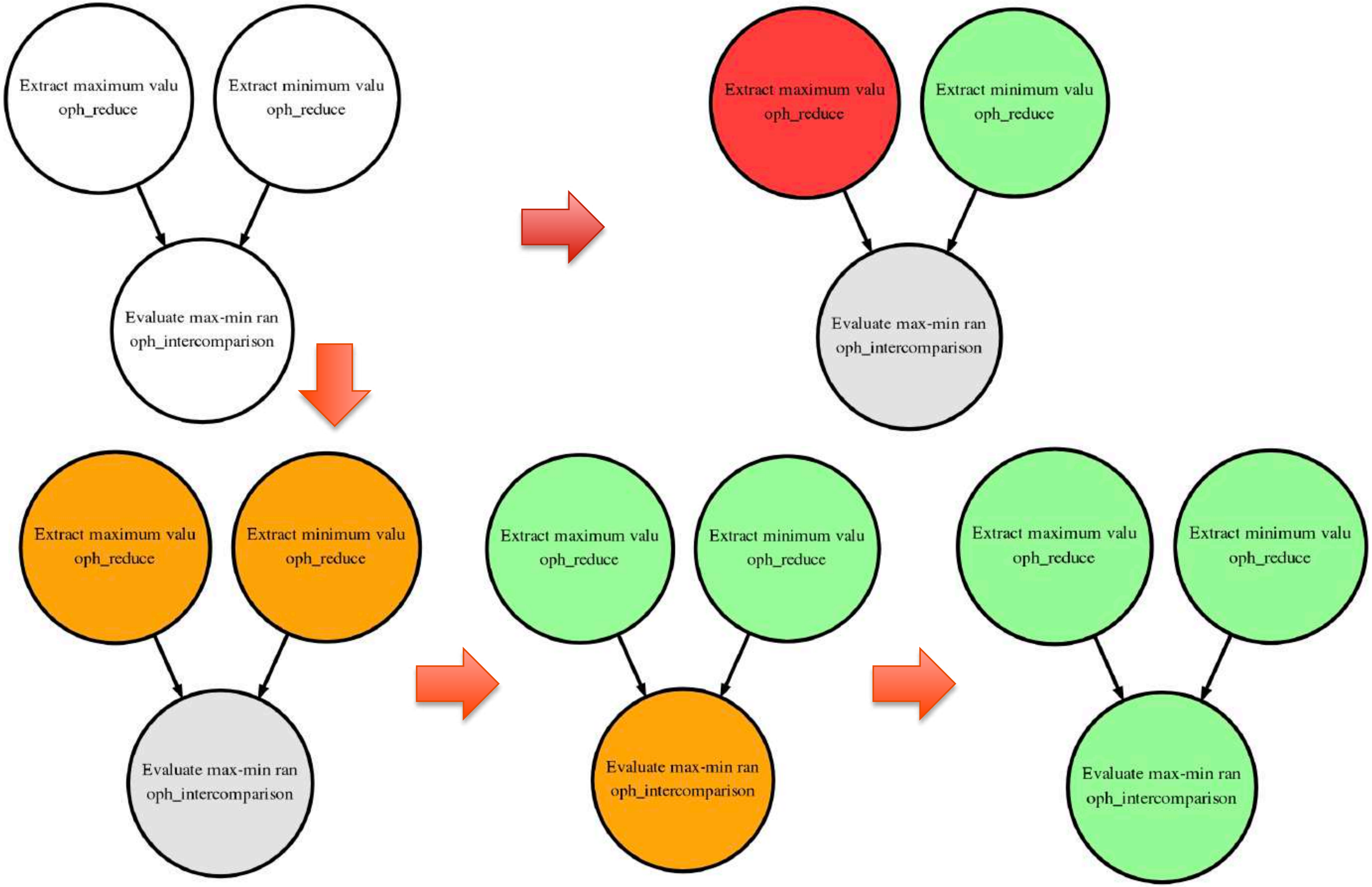
Multiple **storage systems** supported



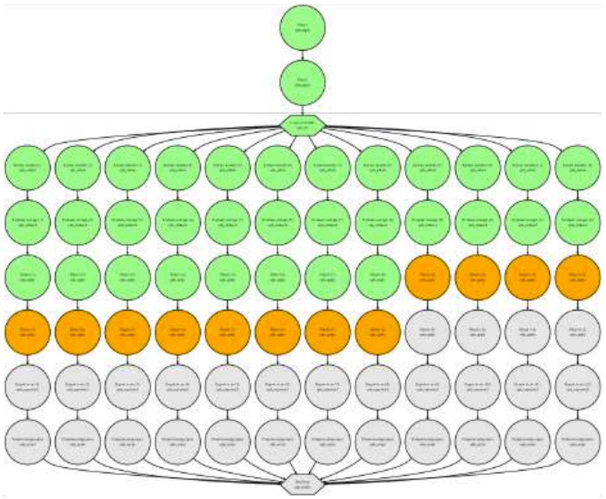
Support for scientific data analysis experiments

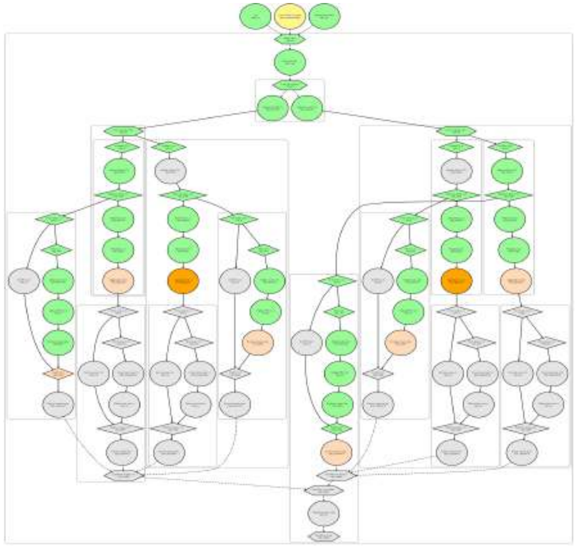

sea surface salinity
sea surface salinity (psu)
5,5
12,6
19,7
26,8
33,9
40,9
monthly sea surface temperature anomaly
Downscaled GDP
294.066779
72.792317
18.018769
4.460307
1.104090
0.273303
0.067652
0.016746
0.004145
0.001026
0.000254
0.000063
0.000016
0.000004
0.000001
Days with unusually warm temperatures in [2014, 2016]
Change (%) in Economic Activity: 2080 - 2100
Snow off
Snow Water Equivalent, monthly aggregat
sea surface temperature
sea surface temperature (degC)
-1,8
4,8
11,5
18,2
24,8
31,5
Data Min = -1,8, Max = 31,5
now off (days since 2000-07-01 00:00:00)
217,1
248,7
280,3
311,9
343,5
Data Min = 185,5, Max = 343,5, Mean = 295,5
Precipitation trend
er Equivalen
Number of days with unusually warm temperature in years 2003-2015
624,0
698,0
772,0
846,0
920,0
994,0
Data Min = 624,0, Max = 994,0

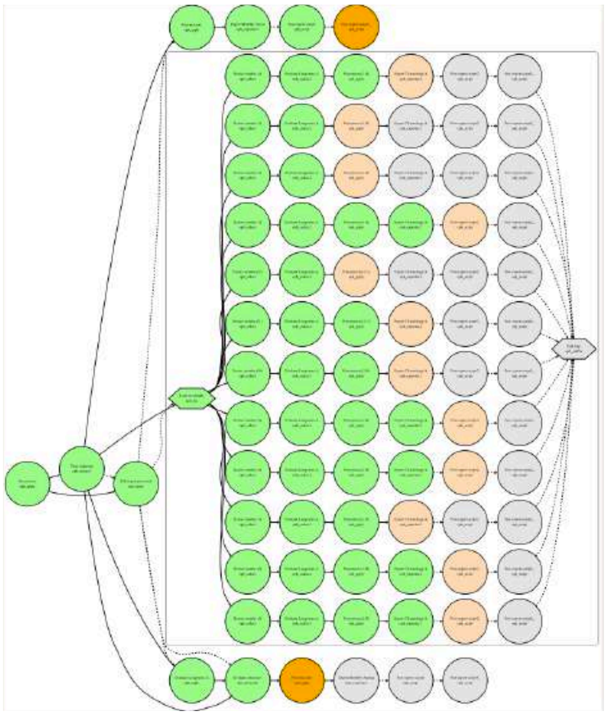
Workflow support

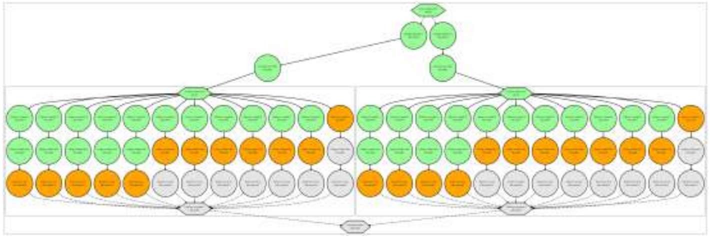


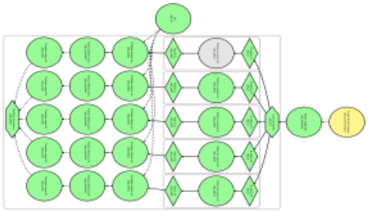
Analytics workflows support and interfaces

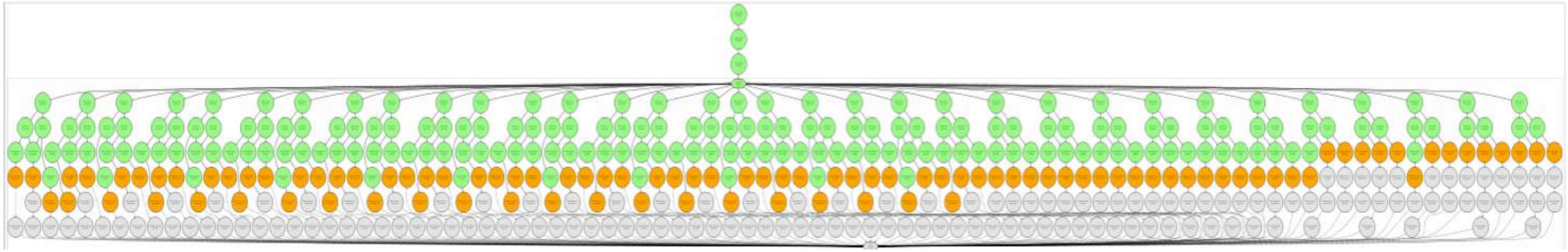




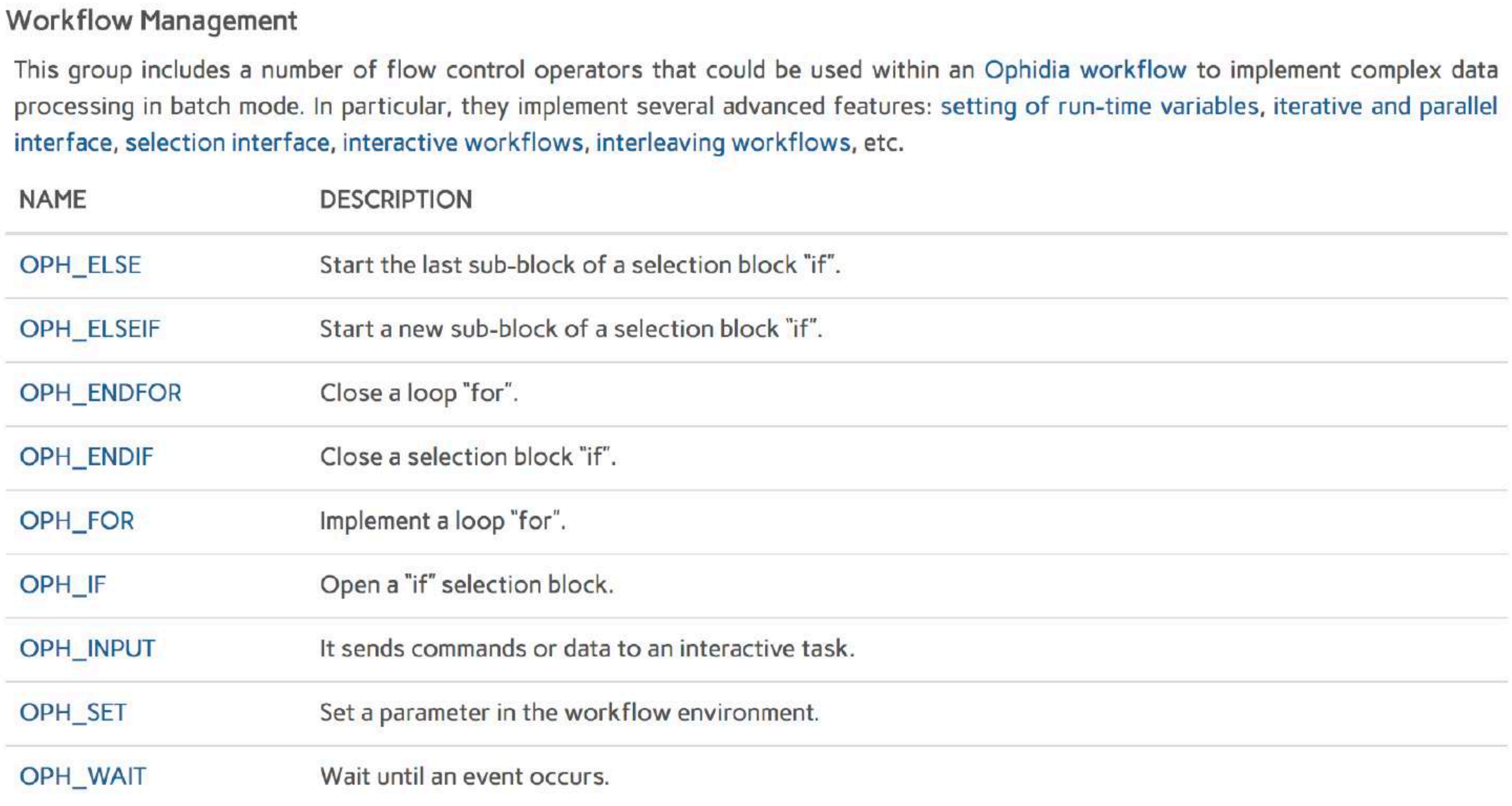








Analytics workflows support and interfaces

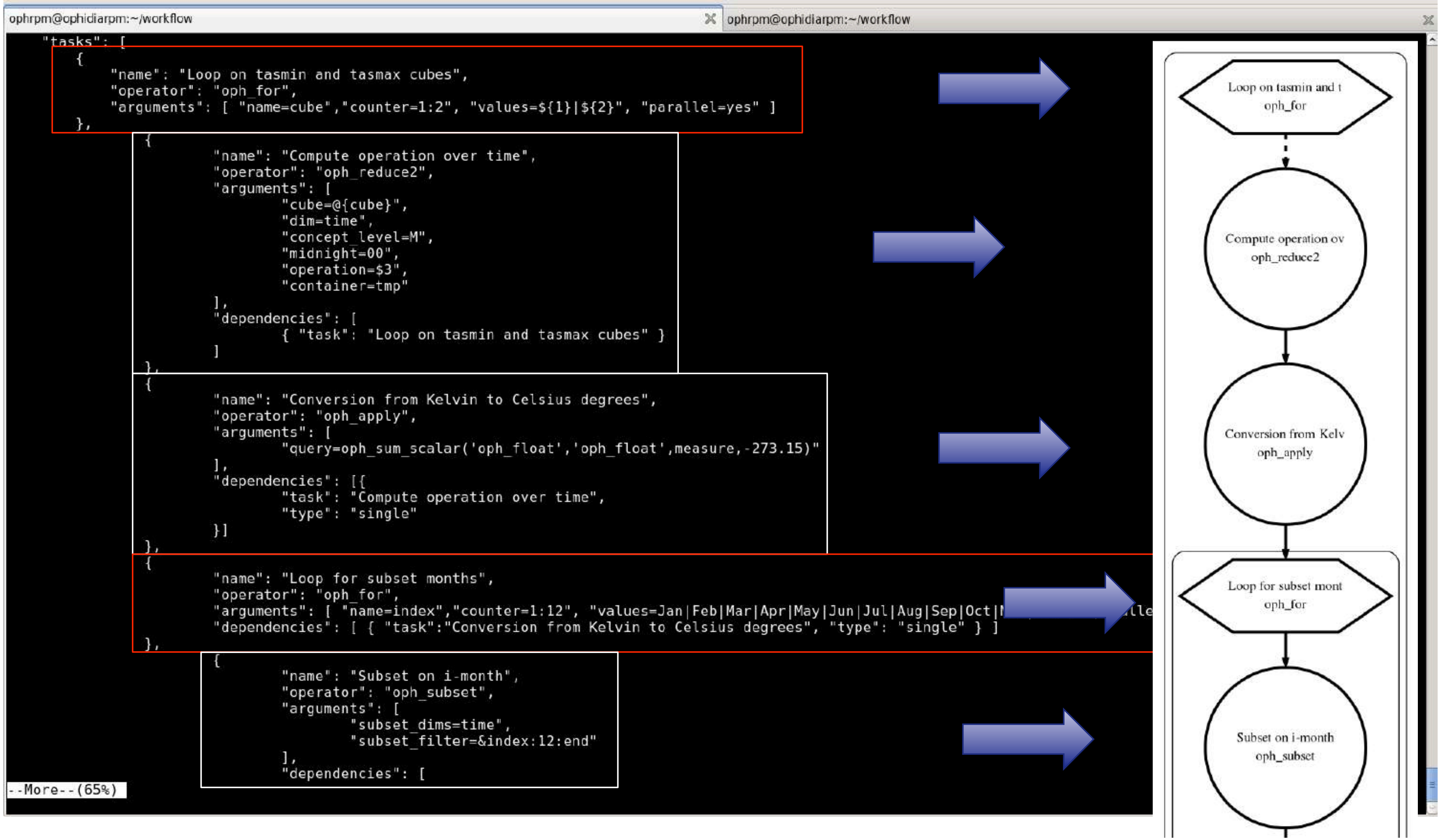


Workflow Management

This group includes a number of flow control operators that could be used within an Ophidia workflow to implement complex data processing in batch mode. In particular, they implement several advanced features: setting of run-time variables, iterative and parallel interface, selection interface, interactive workflows, interleaving workflows, etc.

NAME	DESCRIPTION
OPH_ELSE	Start the last sub-block of a selection block "if".
OPH_ELSEIF	Start a new sub-block of a selection block "if".
OPH_ENDFOR	Close a loop "for".
OPH_ENDIF	Close a selection block "if".
OPH_FOR	Implement a loop "for".
OPH_IF	Open a "if" selection block.
OPH_INPUT	It sends commands or data to an interactive task.
OPH_SET	Set a parameter in the workflow environment.
OPH_WAIT	Wait until an event occurs.

Behind the scene: workflow JSON representation



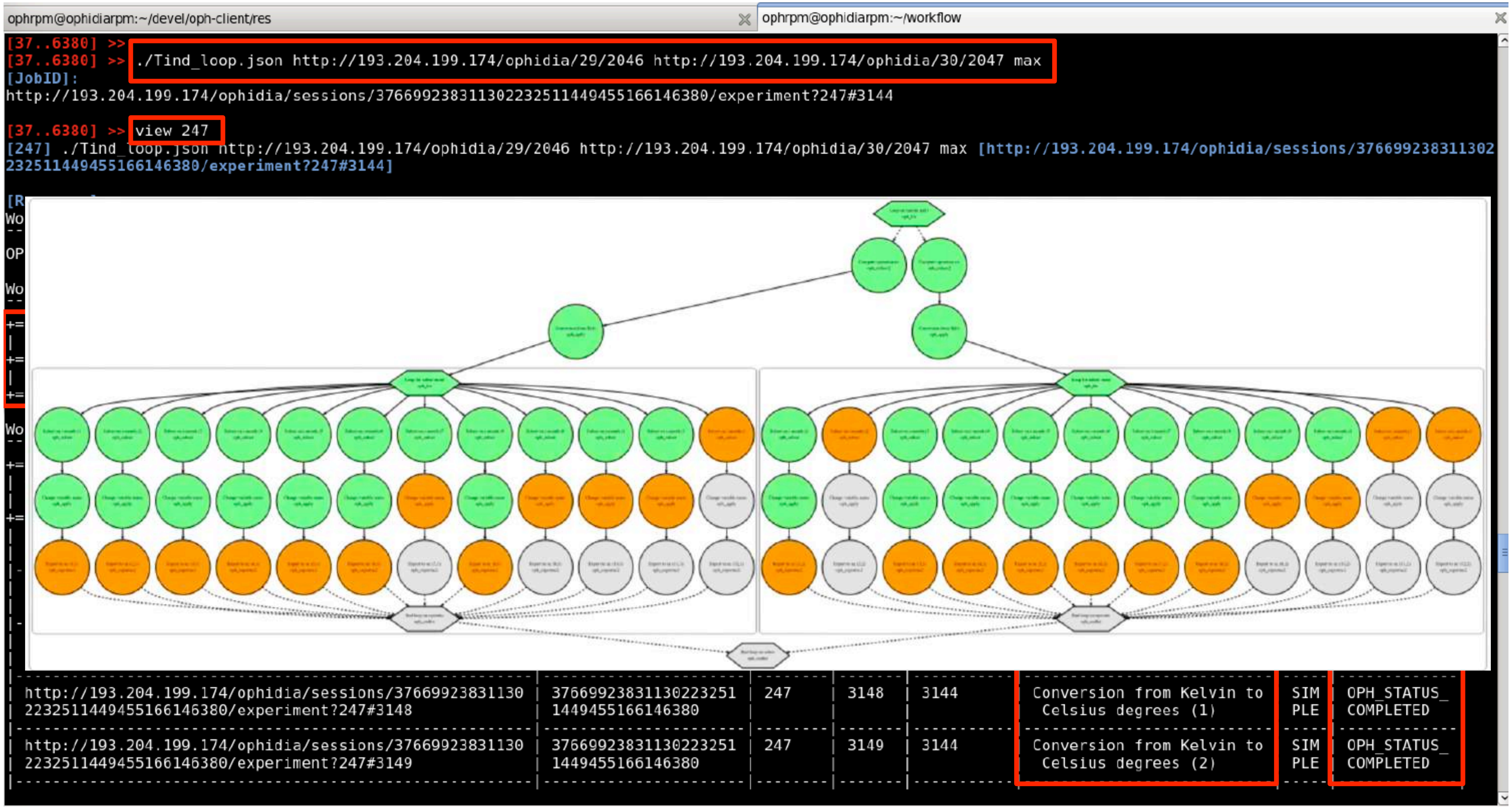
```
    "tasks": [
        {
            "name": "Loop on tasmin and tasmax cubes",
            "operator": "oph_for",
            "arguments": [ "name=cube","counter=1:2", "values=${1}|${2}", "parallel=yes" ]
        },
                {
                        "name": "Compute operation over time",
                        "operator": "oph_reduce2",
                        "arguments": [
                                "cube=@{cube}",
                                "dim=time",
                                "concept_level=M",
                                "midnight=00",
                                "operation=$3",
                                "container=tmp"
                        ],
                        "dependencies": [
                                { "task": "Loop on tasmin and tasmax cubes" }
                        ]
                },
                {
                        "name": "Conversion from Kelvin to Celsius degrees",
                        "operator": "oph_apply",
                        "arguments": [
                                "query=oph_sum_scalar('oph_float','oph_float',measure,-273.15)"
                        ],
                        "dependencies": [{
                                "task": "Compute operation over time",
                                "type": "single"
                        }]
                },
                {
                        "name": "Loop for subset months",
                        "operator": "oph_for",
                        "arguments": [ "name=index","counter=1:12", "values=Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|N          lle
                        "dependencies": [ { "task":"Conversion from Kelvin to Celsius degrees", "type": "single" } ]
                },
                        {
                                "name": "Subset on i-month",
                                "operator": "oph_subset",
                                "arguments": [
                                        "subset_dims=time",
                                        "subset_filter=&index:12:end"
                                ],
                                "dependencies": [
--More--(65%)
```

***Youtube video**: https://www.youtube.com/watch?v=PTZkw60YCNU*

Collaborative EUDAT Data Infrastructure

PRACE

Workflow submission


ophrpm@ophidiarpm:~/devel/oph-client/res
ophrpm@ophidiarpm:~/workflow

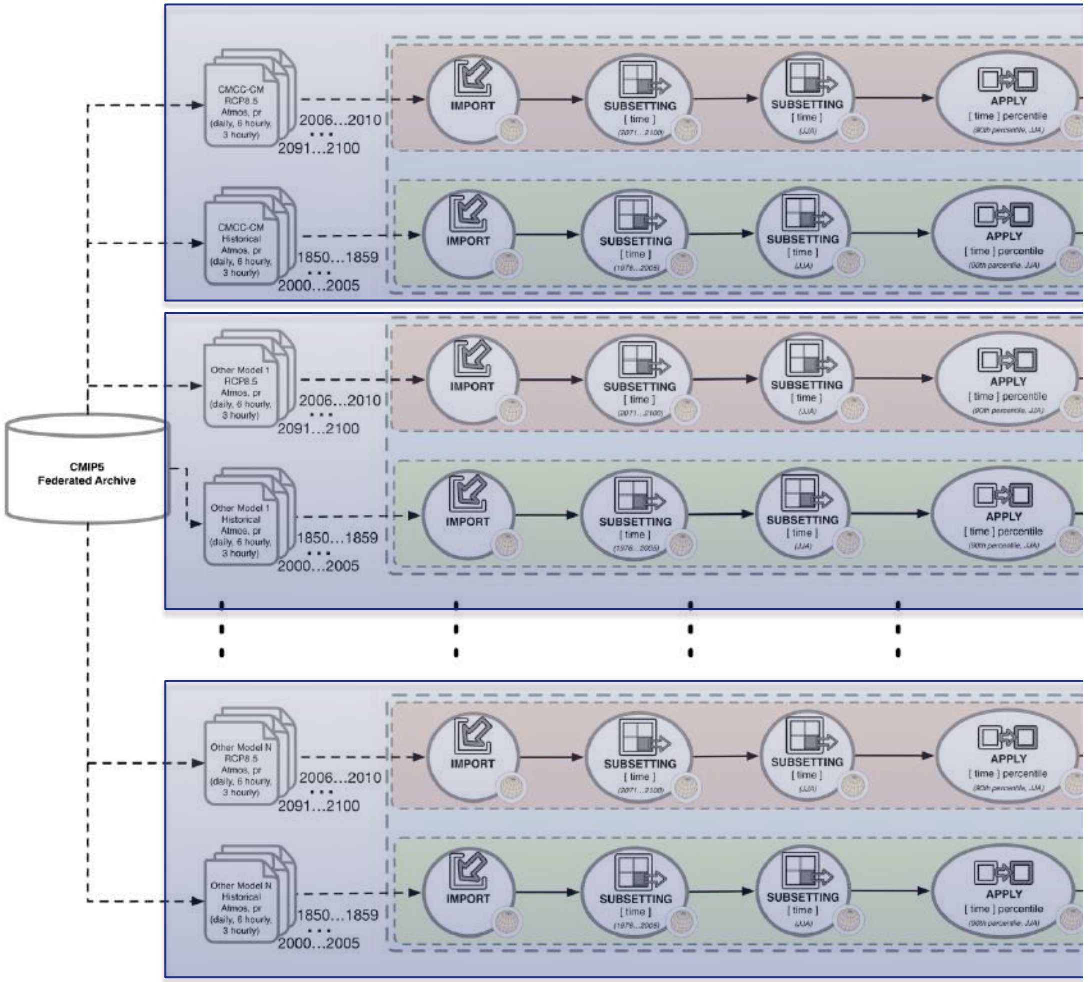
```
[37..6380] >>
[37..6380] >> ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max
[JobID]:
http://193.204.199.174/ophidia/sessions/376699238311302232511449455166146380/experiment?247#3144

[37..6380] >> view 247
[247] ./Tind_loop.json http://193.204.199.174/ophidia/29/2046 http://193.204.199.174/ophidia/30/2047 max [http://193.204.199.174/ophidia/sessions/376699238311302
232511449455166146380/experiment?247#3144]

[R
Wo
--
OP

Wo
--
+=
|
+=
|
+=

Wo
--
+=
|
|
+=
|
|
|-
|
|
|-
|
|
|----------------------------------------------------|------------------------|---------|--------|--------|------------------------|-----|-------------
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247     | 3148   | 3144   | Conversion from Kelvin to | SIM | OPH_STATUS_
| 2232511449455166146380/experiment?247#3148             | 1449455166146380     |         |        |        |  Celsius degrees (1)      | PLE | COMPLETED
|----------------------------------------------------|------------------------|---------|--------|--------|------------------------|-----|-------------
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247     | 3149   | 3144   | Conversion from Kelvin to | SIM | OPH_STATUS_
| 2232511449455166146380/experiment?247#3149             | 1449455166146380     |         |        |        |  Celsius degrees (2)      | PLE | COMPLETED
|----------------------------------------------------|------------------------|---------|--------|--------|------------------------|-----|-------------
```

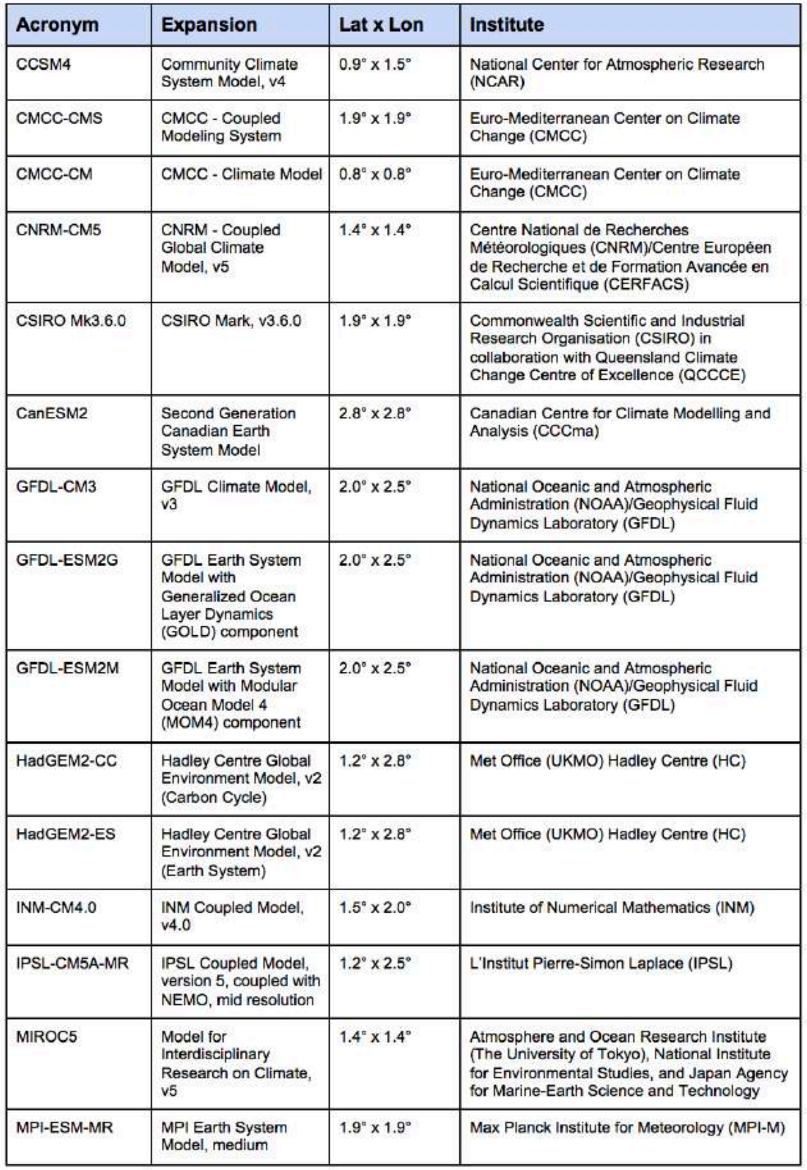
Collaborative
EUDAT Data Infrastructure

PRACE

Workflow example: Multi-model analysis (CMIP)

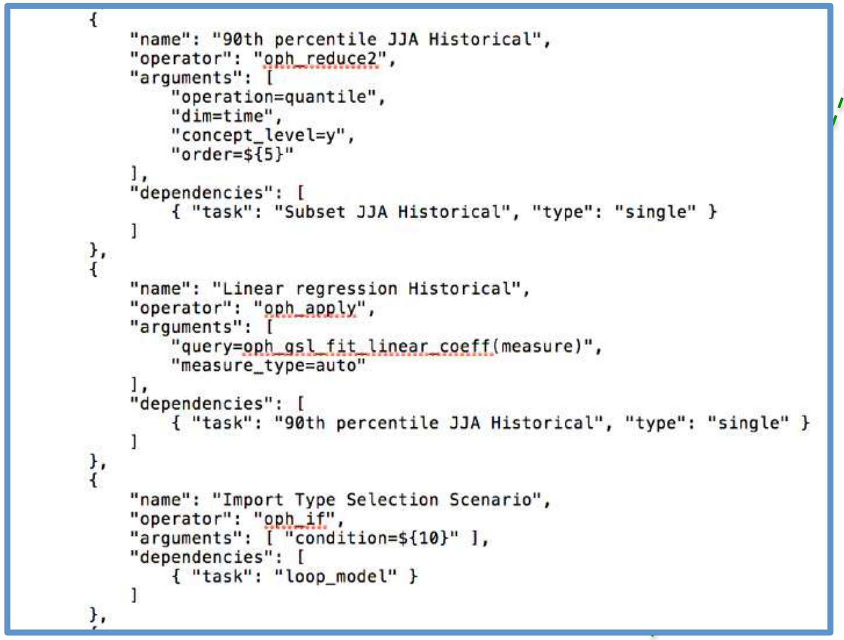
Single model precipitation trend a


CMIP5 Federated Archive
CMCC-CM RCP8.5 Atmos, pr (daily, 6 hourly, 3 hourly)
2006…2010
…
2091…2100
CMCC-CM Historical Atmos, pr (daily, 6 hourly, 3 hourly)
1850…1859
…
2000…2005
Other Model 1 RCP8.5 Atmos, pr (daily, 6 hourly, 3 hourly)
Other Model 1 Historical Atmos, pr (daily, 6 hourly, 3 hourly)
Other Model N RCP8.5 Atmos, pr (daily, 6 hourly, 3 hourly)
Other Model N Historical Atmos, pr (daily, 6 hourly, 3 hourly)
IMPORT
SUBSETTING [time] (2071…2100)
SUBSETTING [time] (1976…2005)
SUBSETTING [time] (JJA)
APPLY [time] percentile (90th percentile, JJA)

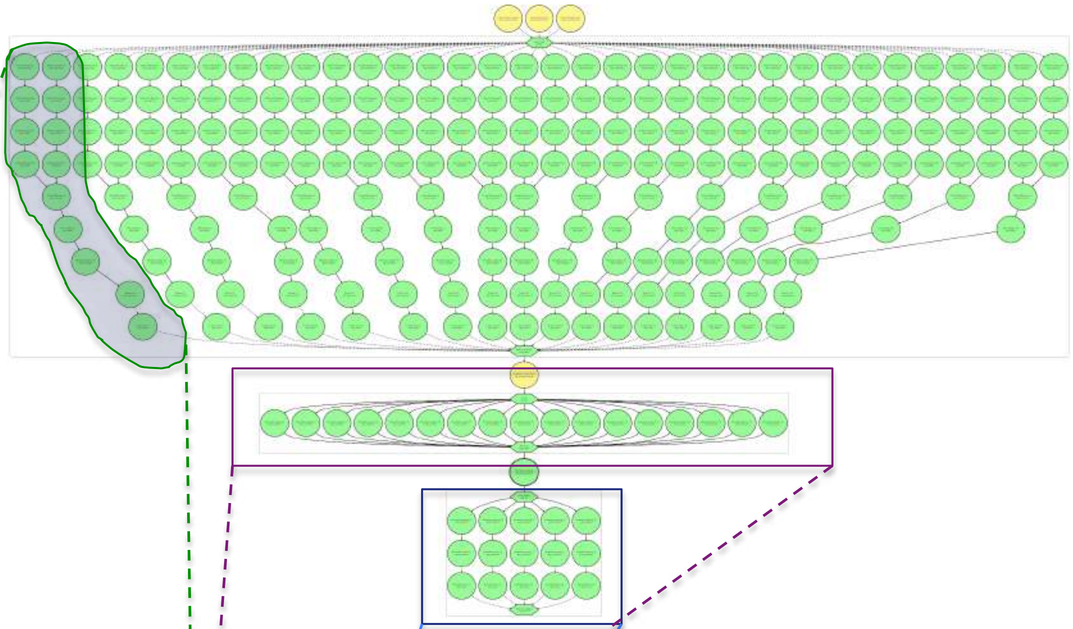


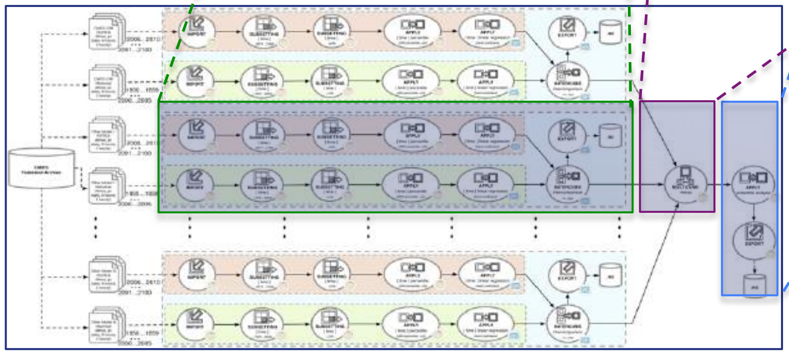
Acronym	Expansion	Lat x Lon	Institute
CCSM4	Community Climate System Model, v4	0.9° x 1.5°	National Center for Atmospheric Research (NCAR)
CMCC-CMS	CMCC - Coupled Modeling System	1.9° x 1.9°	Euro-Mediterranean Center on Climate Change (CMCC)
CMCC-CM	CMCC - Climate Model	0.8° x 0.8°	Euro-Mediterranean Center on Climate Change (CMCC)
CNRM-CM5	CNRM - Coupled Global Climate Model, v5	1.4° x 1.4°	Centre National de Recherches Météorologiques (CNRM)/Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS)
CSIRO Mk3.6.0	CSIRO Mark, v3.6.0	1.9° x 1.9°	Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with Queensland Climate Change Centre of Excellence (QCCCE)
CanESM2	Second Generation Canadian Earth System Model	2.8° x 2.8°	Canadian Centre for Climate Modelling and Analysis (CCCma)
GFDL-CM3	GFDL Climate Model, v3	2.0° x 2.5°	National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL)
GFDL-ESM2G	GFDL Earth System Model with Generalized Ocean Layer Dynamics (GOLD) component	2.0° x 2.5°	National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL)
GFDL-ESM2M	GFDL Earth System Model with Modular Ocean Model 4 (MOM4) component	2.0° x 2.5°	National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL)
HadGEM2-CC	Hadley Centre Global Environment Model, v2 (Carbon Cycle)	1.2° x 2.8°	Met Office (UKMO) Hadley Centre (HC)
HadGEM2-ES	Hadley Centre Global Environment Model, v2 (Earth System)	1.2° x 2.8°	Met Office (UKMO) Hadley Centre (HC)
INM-CM4.0	INM Coupled Model, v4.0	1.5° x 2.0°	Institute of Numerical Mathematics (INM)
IPSL-CM5A-MR	IPSL Coupled Model, version 5, coupled with NEMO, mid resolution	1.2° x 2.5°	L'Institut Pierre-Simon Laplace (IPSL)
MIROC5	Model for Interdisciplinary Research on Climate, v5	1.4° x 1.4°	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology
MPI-ESM-MR	MPI Earth System Model, medium	1.9° x 1.9°	Max Planck Institute for Meteorology (MPI-M)

Workflow example: Multi-model analysis



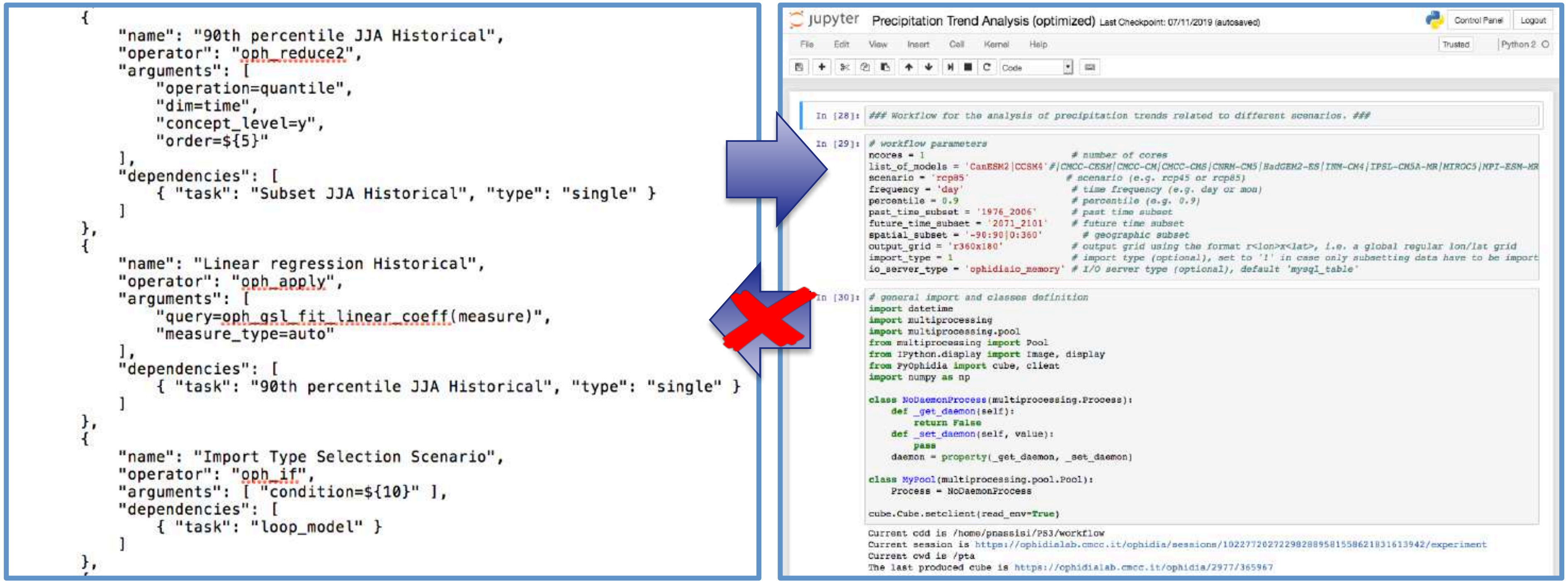
```
{
    "name": "90th percentile JJA Historical",
    "operator": "oph_reduce2",
    "arguments": [
        "operation=quantile",
        "dim=time",
        "concept_level=y",
        "order=${5}"
    ],
    "dependencies": [
        { "task": "Subset JJA Historical", "type": "single" }
    ]
},
{
    "name": "Linear regression Historical",
    "operator": "oph_apply",
    "arguments": [
        "query=oph_gsl_fit_linear_coeff(measure)",
        "measure_type=auto"
    ],
    "dependencies": [
        { "task": "90th percentile JJA Historical", "type": "single" }
    ]
},
{
    "name": "Import Type Selection Scenario",
    "operator": "oph_if",
    "arguments": [ "condition=${10}" ],
    "dependencies": [
        { "task": "loop_model" }
    ]
},
```





The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191 http://www.esiwace.eu

Multi-model analytics workflow
Two implementation Strategies



```
{
    "name": "90th percentile JJA Historical",
    "operator": "oph_reduce2",
    "arguments": [
        "operation=quantile",
        "dim=time",
        "concept_level=y",
        "order=${5}"
    ],
    "dependencies": [
        { "task": "Subset JJA Historical", "type": "single" }
    ]
},
{
    "name": "Linear regression Historical",
    "operator": "oph_apply",
    "arguments": [
        "query=oph_gsl_fit_linear_coeff(measure)",
        "measure_type=auto"
    ],
    "dependencies": [
        { "task": "90th percentile JJA Historical", "type": "single" }
    ]
},
{
    "name": "Import Type Selection Scenario",
    "operator": "oph_if",
    "arguments": [ "condition=${10}" ],
    "dependencies": [
        { "task": "loop_model" }
    ]
},
```

jupyter
Precipitation Trend Analysis (optimized) Last Checkpoint: 07/11/2019 (autosaved)
Control Panel
Logout
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 2
Code

```
In [28]: ### Workflow for the analysis of precipitation trends related to different scenarios. ###

In [29]: # workflow parameters
         ncores = 1                          # number of cores
         list_of_models = 'CanESM2|CCSM4' #|CMCC-CESM|CMCC-CM|CMCC-CMS|CNRM-CM5|HadGEM2-ES|INM-CM4|IPSL-CM5A-MR|MIROC5|MPI-ESM-MR
         scenario = 'rcp85'                 # scenario (e.g. rcp45 or rcp85)
         frequency = 'day'                   # time frequency (e.g. day or mon)
         percentile = 0.9                    # percentile (e.g. 0.9)
         past_time_subset = '1976_2006'      # past time subset
         future_time_subset = '2071_2101'    # future time subset
         spatial_subset = '-90:90|0:360'       # geographic subset
         output_grid = 'r360x180'            # output grid using the format r<lon>x<lat>, i.e. a global regular lon/lat grid
         import_type = 1                     # import type (optional), set to '1' in case only subsetting data have to be import
         io_server_type = 'ophidiaio_memory' # I/O server type (optional), default 'mysql_table'

In [30]: # general import and classes definition
         import datetime
         import multiprocessing
         import multiprocessing.pool
         from multiprocessing import Pool
         from IPython.display import Image, display
         from PyOphidia import cube, client
         import numpy as np

         class NoDaemonProcess(multiprocessing.Process):
             def _get_daemon(self):
                 return False
             def _set_daemon(self, value):
                 pass
             daemon = property(_get_daemon, _set_daemon)

         class MyPool(multiprocessing.pool.Pool):
             Process = NoDaemonProcess

         cube.Cube.setclient(read_env=True)

         Current cdd is /home/pnassisi/PS3/workflow
         Current session is https://ophidialab.cmcc.it/ophidia/sessions/102277202722982889581558621831613942/experiment
         Current cwd is /pta
         The last produced cube is https://ophidialab.cmcc.it/ophidia/2977/365967
```

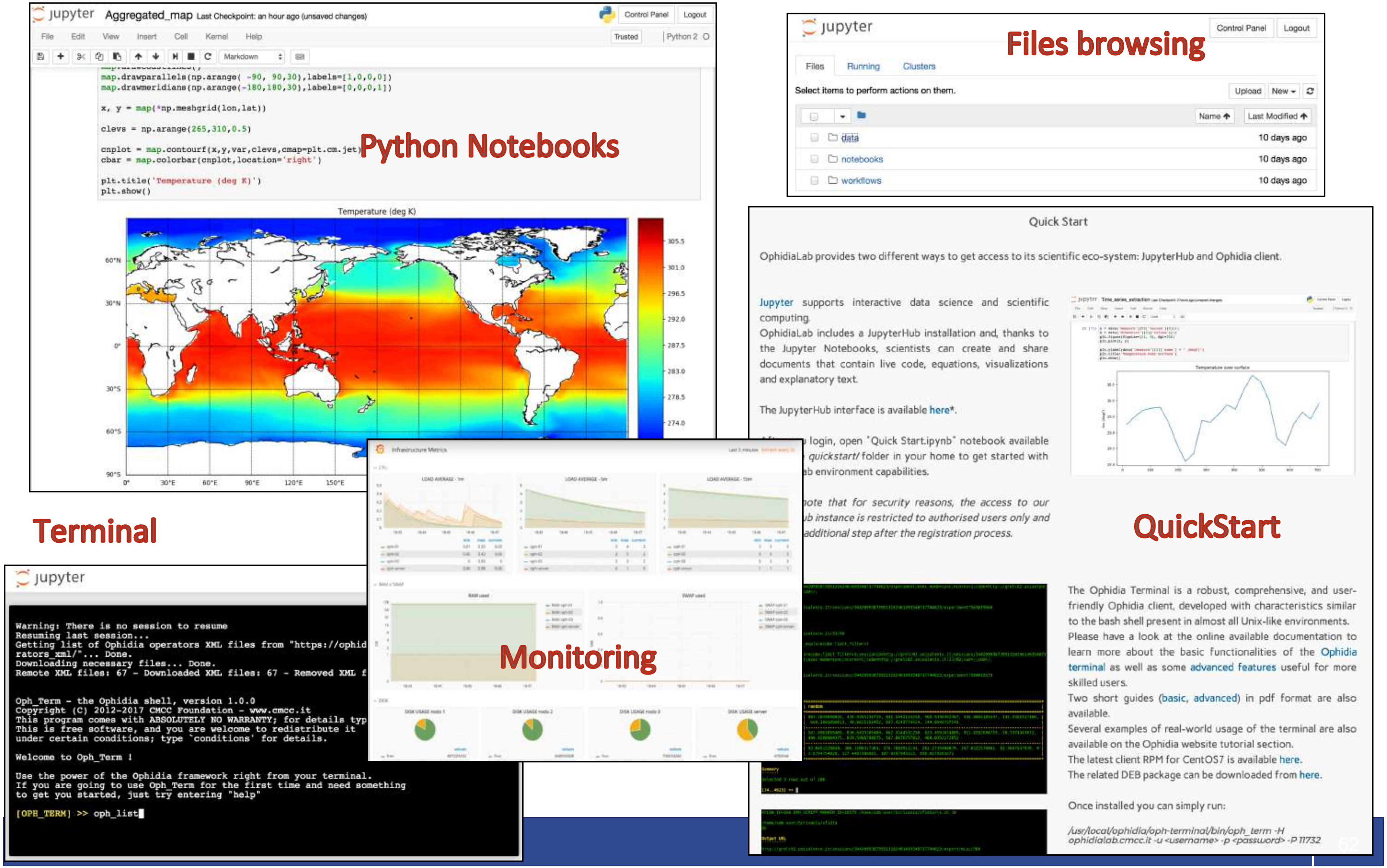
	Approach	Mode	Library	Code	LoC	ExecTime
Workflow	SS - SI*	Batch	Ophida WF	JSON	544	**199 (1.6x)**
Notebook	SS - MI*	Interactive	PyOphidia	Python	122	319

** SS: Server Side; SI: Single Interaction, MI: Multiple Interactions*

ECASLab

Data Science environment

ECASLab in a nutshell



```
map.drawparallels(np.arange( -90, 90,30),labels=[1,0,0,0])
map.drawmeridians(np.arange(-180,180,30),labels=[0,0,0,1])

x, y = map(*np.meshgrid(lon,lat))

clevs = np.arange(265,310,0.5)

cnplot = map.contourf(x,y,var,clevs,cmap=plt.cm.jet)
cbar = map.colorbar(cnplot,location='right')

plt.title('Temperature (deg K)')
plt.show()
```

```
Warning: There is no session to resume
Resuming last session...
Getting list of Ophidia operators XML files from "https://ophid
rators_xml/"... Done.
Downloading necessary files... Done.
Remote XML files: 67 - Downloaded XML files: 67 - Removed XML f


Oph_Term - the Ophidia shell, version 1.0.0
Copyright (C) 2012-2017 CMCC Foundation - www.cmcc.it
This program comes with ABSOLUTELY NO WARRANTY; for details typ
This is free software, and you are welcome to redistribute it
under certain conditions; type `conditions' for details.

Welcome to Oph_Term !

Use the power of the Ophidia framework right from your terminal.
If you are going to use Oph_Term for the first time and need something
to get you started, just try entering "help"

[OPH_TERM] >> oph_list
```

Quick Start

OphidiaLab provides two different ways to get access to its scientific eco-system: JupyterHub and Ophidia client.

Jupyter supports interactive data science and scientific computing.
OphidiaLab includes a JupyterHub installation and, thanks to the Jupyter Notebooks, scientists can create and share documents that contain live code, equations, visualizations and explanatory text.

The JupyterHub interface is available here*.

u login, open "Quick Start.ipynb" notebook available quickstart/ folder in your home to get started with ab environment capabilities.

note that for security reasons, the access to our ub instance is restricted to authorised users only and additional step after the registration process.

The Ophidia Terminal is a robust, comprehensive, and user-friendly Ophidia client, developed with characteristics similar to the bash shell present in almost all Unix-like environments.
Please have a look at the online available documentation to learn more about the basic functionalities of the Ophidia terminal as well as some advanced features useful for more skilled users.
Two short guides (basic, advanced) in pdf format are also available.
Several examples of real-world usage of the terminal are also available on the Ophidia website tutorial section.
The latest client RPM for CentOS7 is available here.
The related DEB package can be downloaded from here.

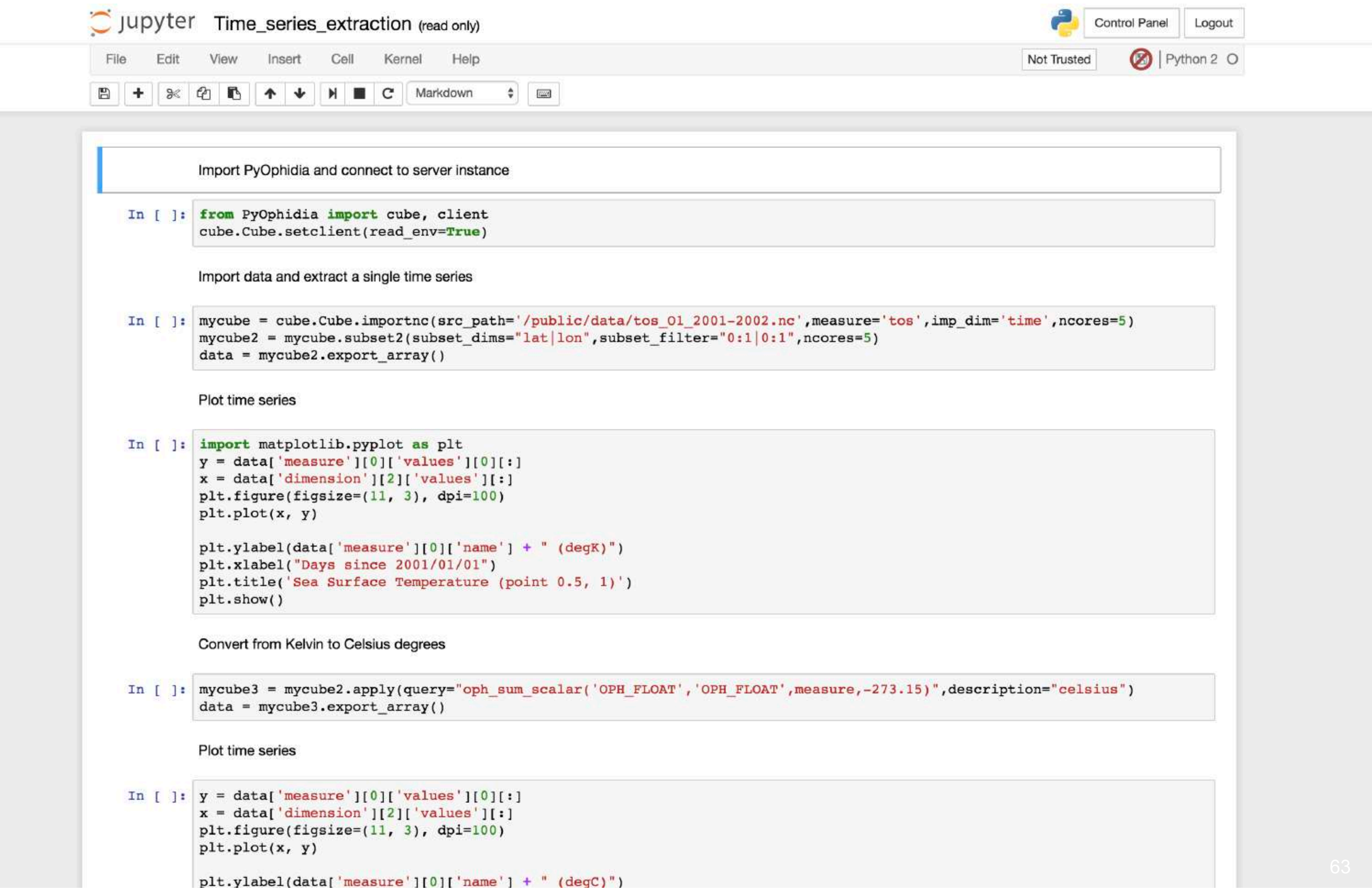
Once installed you can simply run:

/usr/local/ophidia/oph-terminal/bin/oph_term -H ophidialab.cmcc.it -u <username> -p <password> -P 11732

Collaborative Data Infrastructure
EUDAT

PRACE

Jupyter notebooks



Import PyOphidia and connect to server instance

```
In [ ]: from PyOphidia import cube, client
        cube.Cube.setclient(read_env=True)
```

Import data and extract a single time series

```
In [ ]: mycube = cube.Cube.importnc(src_path='/public/data/tos_O1_2001-2002.nc',measure='tos',imp_dim='time',ncores=5)
        mycube2 = mycube.subset2(subset_dims="lat|lon",subset_filter="0:1|0:1",ncores=5)
        data = mycube2.export_array()
```

Plot time series

```
In [ ]: import matplotlib.pyplot as plt
        y = data['measure'][0]['values'][0][:]
        x = data['dimension'][2]['values'][:]
        plt.figure(figsize=(11, 3), dpi=100)
        plt.plot(x, y)

        plt.ylabel(data['measure'][0]['name'] + " (degK)")
        plt.xlabel("Days since 2001/01/01")
        plt.title('Sea Surface Temperature (point 0.5, 1)')
        plt.show()
```

Convert from Kelvin to Celsius degrees

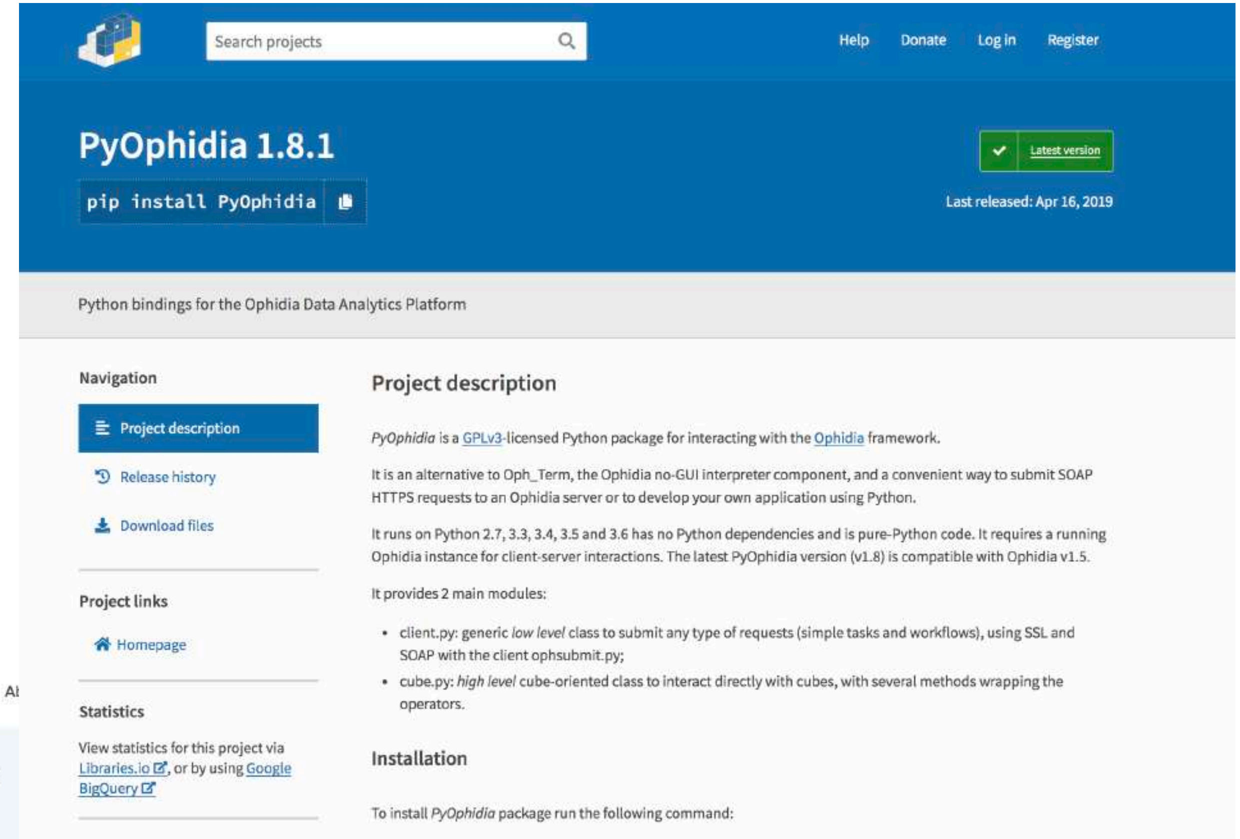
```
In [ ]: mycube3 = mycube2.apply(query="oph_sum_scalar('OPH_FLOAT','OPH_FLOAT',measure,-273.15)",description="celsius")
        data = mycube3.export_array()
```

Plot time series

```
In [ ]: y = data['measure'][0]['values'][0][:]
        x = data['dimension'][2]['values'][:]
        plt.figure(figsize=(11, 3), dpi=100)
        plt.plot(x, y)

        plt.ylabel(data['measure'][0]['name'] + " (degC)")
```

Python eco-system and the PyOphidia class

- Programmatic support for data science applications
- Python binding to Ophidia
- Based on two Python classes
- Available on conda-forge



Search projects

PyOphidia 1.8.1

```
pip install PyOphidia
```

Last released: Apr 16, 2019

Python bindings for the Ophidia Data Analytics Platform

Navigation

Project links

Statistics

View statistics for this project via Libraries.io, or by using Google BigQuery

Project description

PyOphidia is a GPLv3-licensed Python package for interacting with the Ophidia framework.

It is an alternative to Oph_Term, the Ophidia no-GUI interpreter component, and a convenient way to submit SOAP HTTPS requests to an Ophidia server or to develop your own application using Python.

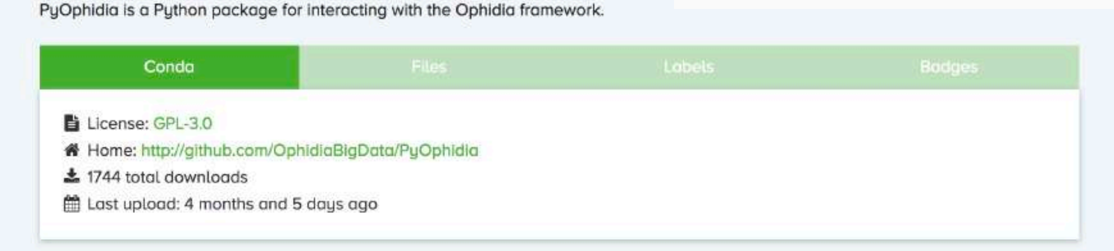
It runs on Python 2.7, 3.3, 3.4, 3.5 and 3.6 has no Python dependencies and is pure-Python code. It requires a running Ophidia instance for client-server interactions. The latest PyOphidia version (v1.8) is compatible with Ophidia v1.5.

It provides 2 main modules:

- client.py: generic *low level* class to submit any type of requests (simple tasks and workflows), using SSL and SOAP with the client ophsubmit.py;
- cube.py: *high level* cube-oriented class to interact directly with cubes, with several methods wrapping the operators.

Installation

To install *PyOphidia* package run the following command:



conda-forge / packages / pyophidia 1.7.0

PyOphidia is a Python package for interacting with the Ophidia framework.

License: GPL-3.0

Home: http://github.com/OphidiaBigData/PyOphidia

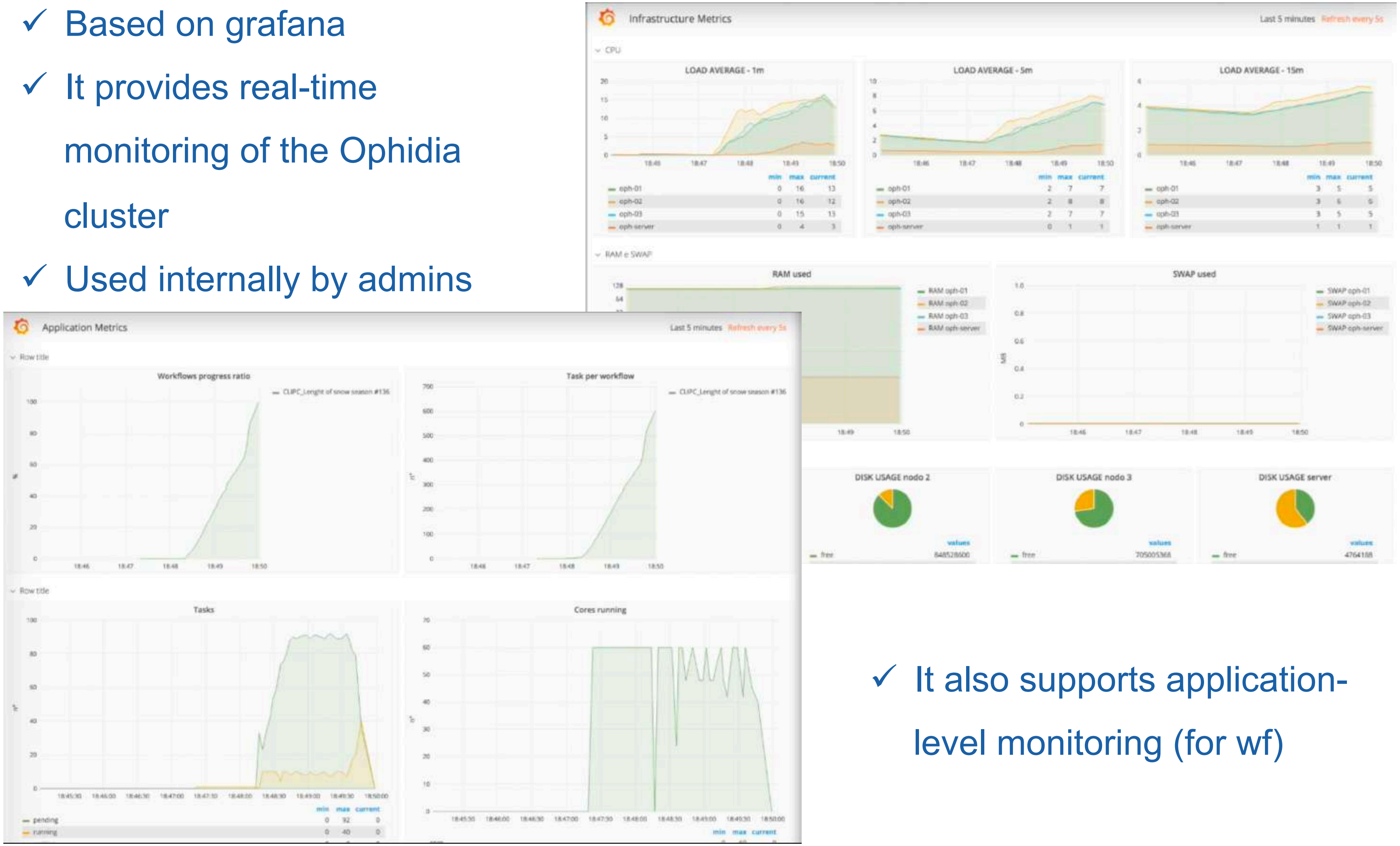
1744 total downloads

Last upload: 4 months and 5 days ago

https://pypi.org/project/PyOphidia/

Grafana monitoring interface

- ✓ Based on grafana
- ✓ It provides real-time monitoring of the Ophidia cluster
- ✓ Used internally by admins



- ✓ It also supports application-level monitoring (for wf)

Levels of parallelism and HPC deployment

Three levels of parallelism

Datacube-level parallelism

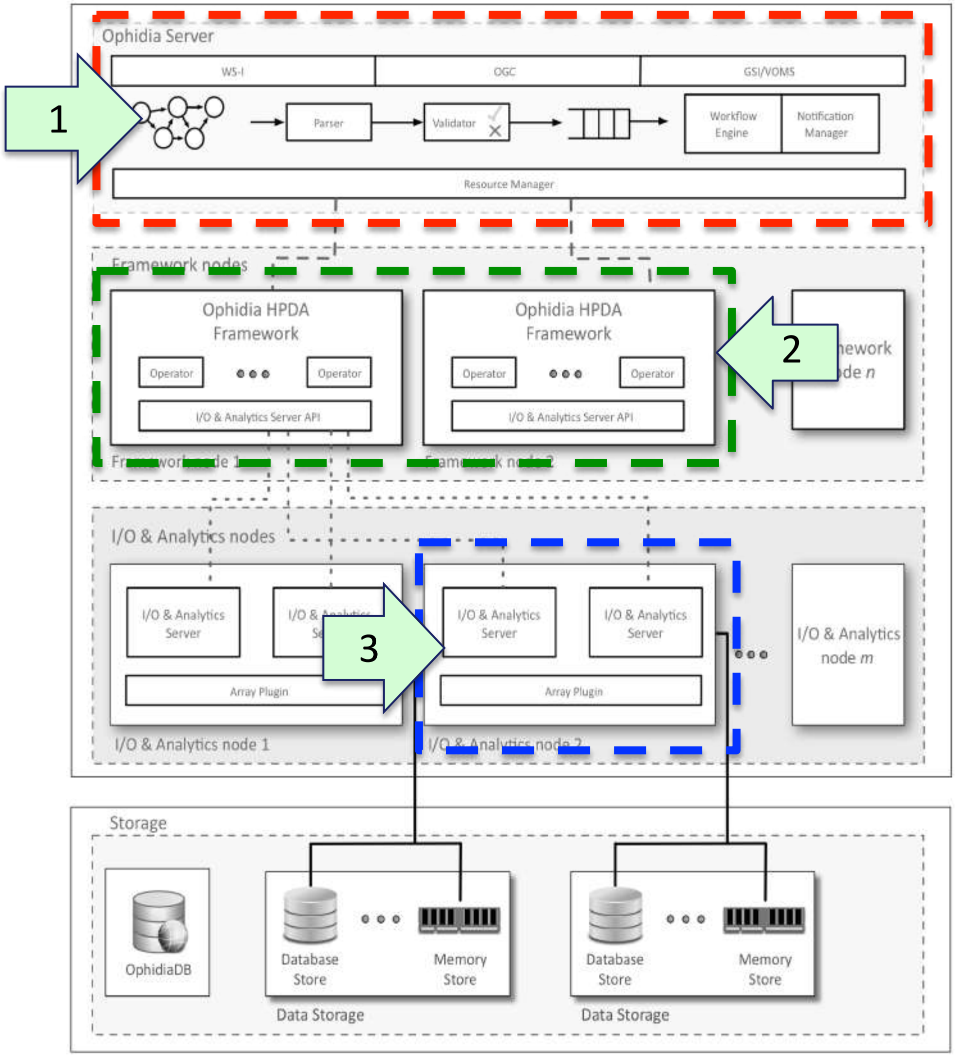
- HTC paradigm
- At the front-end level
- Based on the "massive" operator concept

Framework-level parallelism

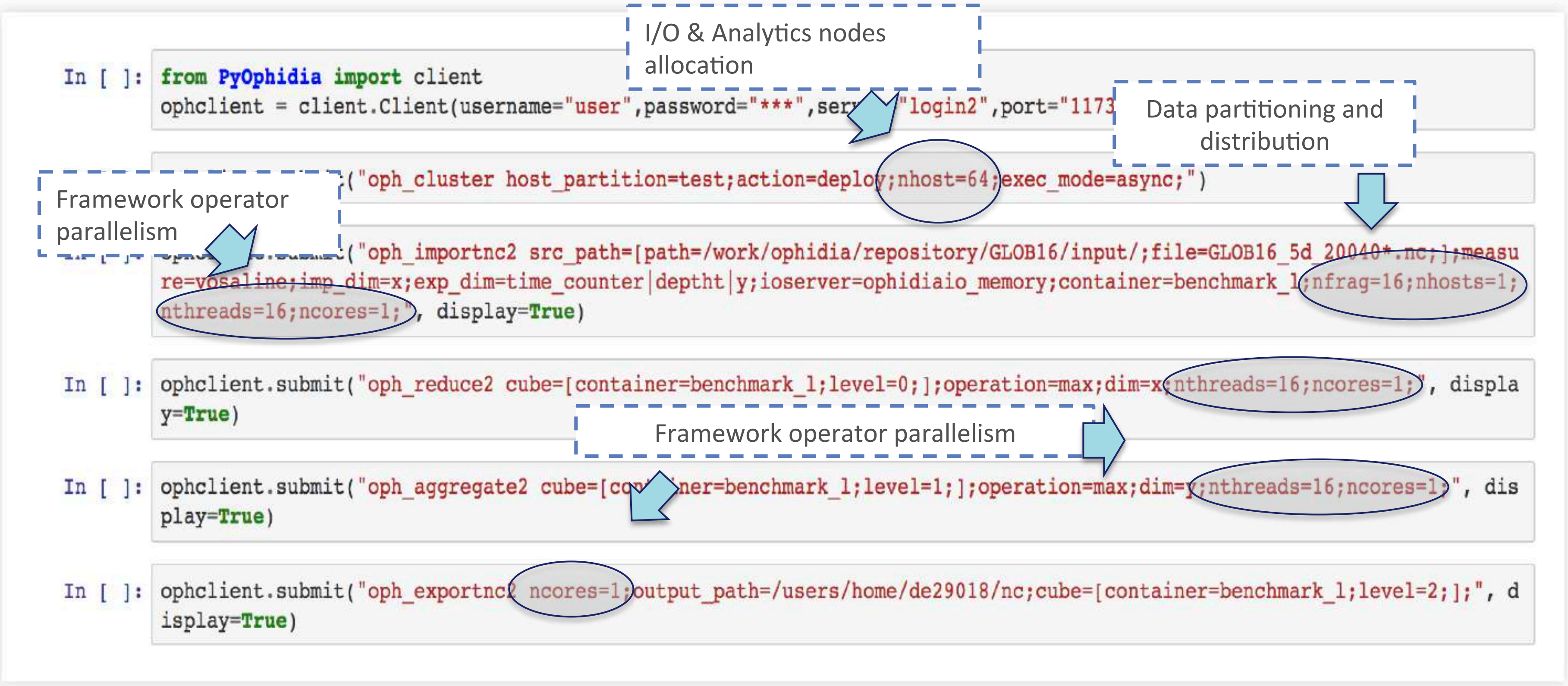
- HPC paradigm
- MPI/Pthread
- At the HPDA framework level

Fragment-level parallelism

- OpenMP based
- At the I/O & analytics server level



Python code and HPC infrastructure transparency



I/O & Analytics nodes allocation

Data partitioning and distribution

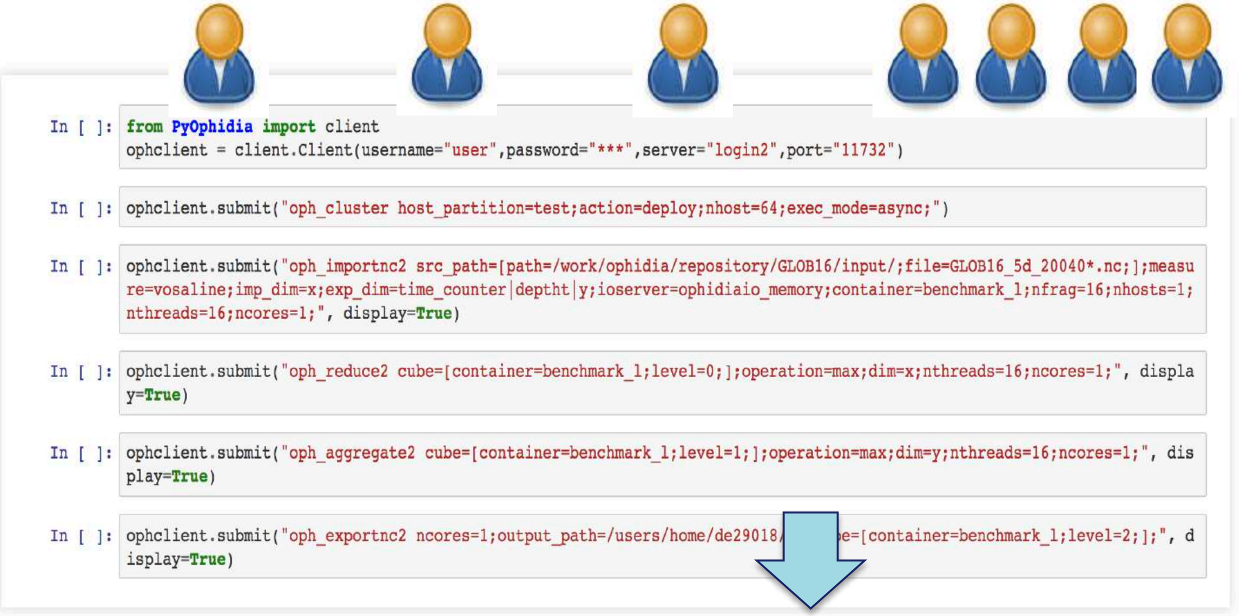
Framework operator parallelism

Framework operator parallelism

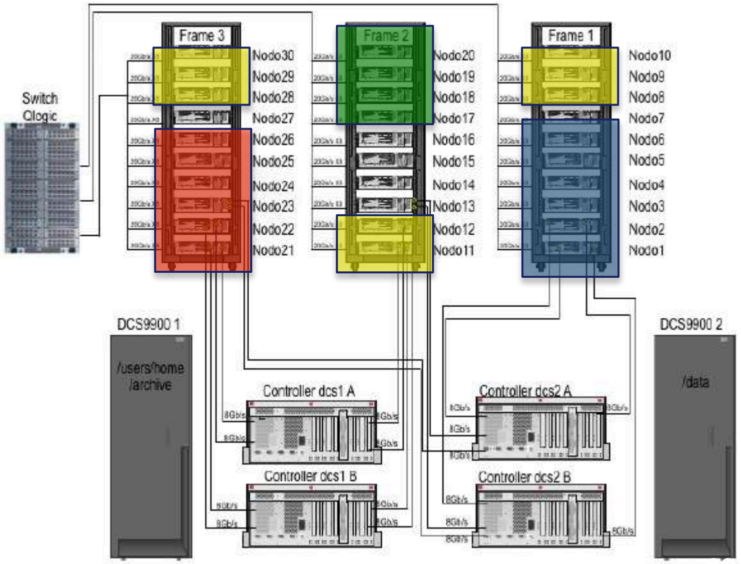
```
In [ ]: from PyOphidia import client
        ophclient = client.Client(username="user",password="***",ser[illegible]"login2",port="1173

In [ ]: [illegible]("oph_cluster host_partition=test;action=deploy;nhost=64;exec_mode=async;")

In [ ]: [illegible]("oph_importnc2 src_path=[path=/work/ophidia/repository/GLOB16/input/;file=GLOB16_5d_20040*.nc;];measu
        re=vosaline;imp_dim=x;exp_dim=time_counter|deptht|y;ioserver=ophidiaio_memory;container=benchmark_1;nfrag=16;nhosts=1;
        nthreads=16;ncores=1;", display=True)

In [ ]: ophclient.submit("oph_reduce2 cube=[container=benchmark_1;level=0;];operation=max;dim=x;nthreads=16;ncores=1;", displa
        y=True)

In [ ]: ophclient.submit("oph_aggregate2 cube=[co[illegible]ner=benchmark_1;level=1;];operation=max;dim=y;nthreads=16;ncores=1;", dis
        play=True)

In [ ]: ophclient.submit("oph_exportnc2 ncores=1;output_path=/users/home/de29018/nc;cube=[container=benchmark_1;level=2;];", d
        isplay=True)
```

On-demand instantiation of an Ophidia custer

- Target environment
  - HPC machines
- Athena Cluster
  - 482 nodes
  - Intel Xeon E5-2670 Sandybridge 2,6GHz, 16 cores/node
  - 7712 cores total, 160 TFLOPS
- Deployment statement
  - ophclient.submit("oph_cluster host_partition=test; action=**deploy**;nhost=64;")
- It allocates a set of nodes on the HPC cluster as I/O & analytics servers



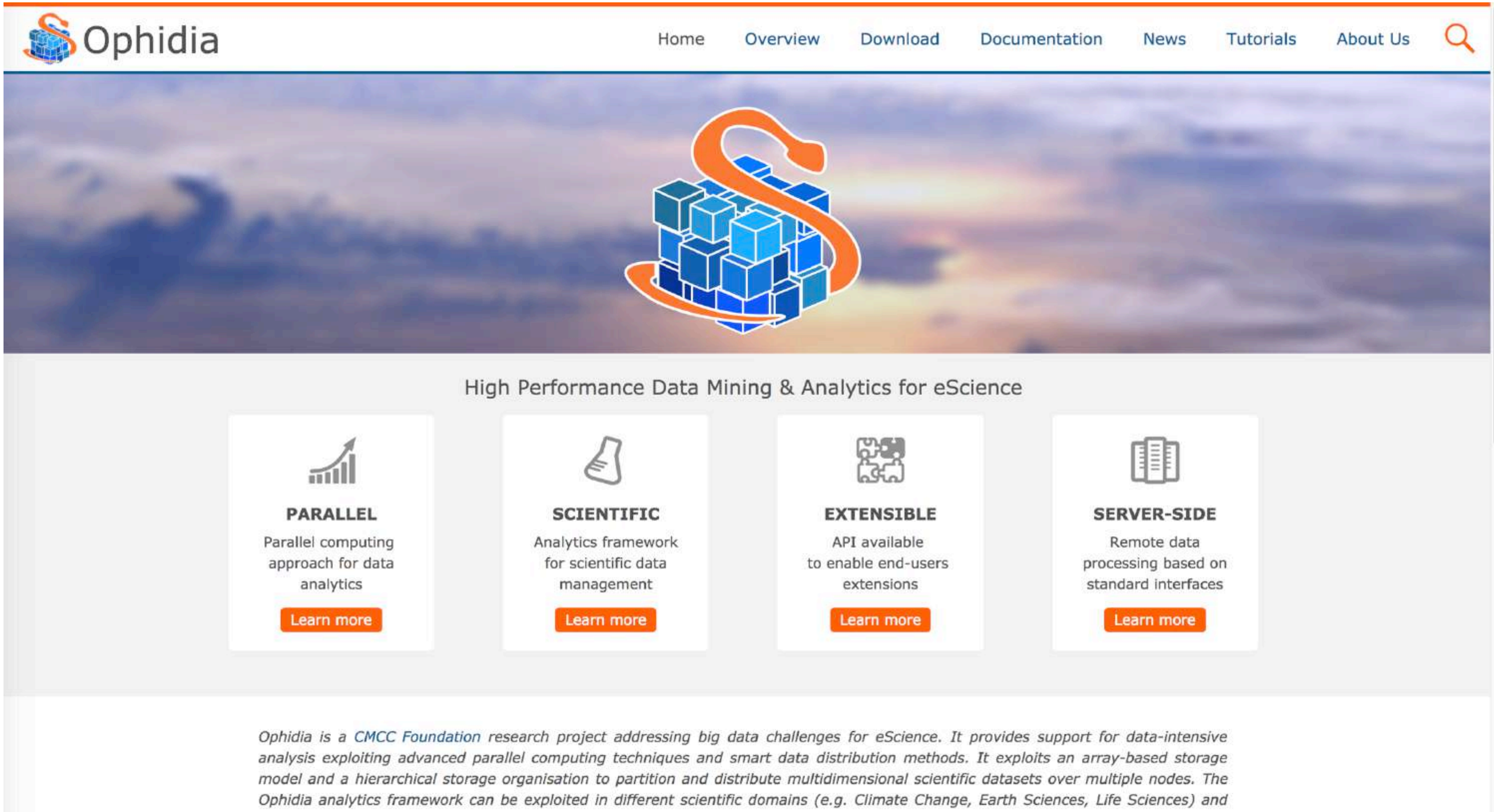
```
In [ ]: from PyOphidia import client
        ophclient = client.Client(username="user",password="***",server="login2",port="11732")

In [ ]: ophclient.submit("oph_cluster host_partition=test;action=deploy;nhost=64;exec_mode=async;")

In [ ]: ophclient.submit("oph_importnc2 src_path=[path=/work/ophidia/repository/GLOB16/input/;file=GLOB16_5d_20040*.nc;];measu
        re=vosaline;imp_dim=x;exp_dim=time_counter|deptht|y;ioserver=ophidiaio_memory;container=benchmark_1;nfrag=16;nhosts=1;
        nthreads=16;ncores=1;", display=True)

In [ ]: ophclient.submit("oph_reduce2 cube=[container=benchmark_1;level=0;];operation=max;dim=x;nthreads=16;ncores=1;", displa
        y=True)

In [ ]: ophclient.submit("oph_aggregate2 cube=[container=benchmark_1;level=1;];operation=max;dim=y;nthreads=16;ncores=1;", dis
        play=True)

In [ ]: ophclient.submit("oph_exportnc2 ncores=1;output_path=/users/home/de29018/      e=[container=benchmark_1;level=2;];", d
        isplay=True)
```



Useful resources and final remarks

Website

Website: http://ophidia.cmcc.it



High Performance Data Mining & Analytics for eScience

PARALLEL
Parallel computing approach for data analytics

SCIENTIFIC
Analytics framework for scientific data management

EXTENSIBLE
API available to enable end-users extensions

SERVER-SIDE
Remote data processing based on standard interfaces

Ophidia is a CMCC Foundation research project addressing big data challenges for eScience. It provides support for data-intensive analysis exploiting advanced parallel computing techniques and smart data distribution methods. It exploits an array-based storage model and a hierarchical storage organisation to partition and distribute multidimensional scientific datasets over multiple nodes. The Ophidia analytics framework can be exploited in different scientific domains (e.g. Climate Change, Earth Sciences, Life Sciences) and

Some nice readings

[9] S. Fiore, C. Palazzo, A. D'Anca, D. Elia, E. Londero, C. Knapic, S. Monna, N. M. Marcucci, F. Aguilar, M. Płóciennik, J. E. M. De Lucas, G. Aloisio, "Big Data Analytics on Large-Scale Scientific Datasets in the INDIGO-DataCloud Project". In Proceedings of the ACM International Conference on Computing Frontiers (CF '17), May 15-17, 2017, Siena, Italy, pp. 343-348

[8] A. D'Anca, C. Palazzo, D. Elia, S. Fiore, I. Bistinas, K. Böttcher, V. Bennett, G. Aloisio, "On the Use of In-memory Analytics Workflows to Compute eScience Indicators from Large Climate Datasets," 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, May 14-17, 2017, pp. 1035-1043.

[7] S. Fiore, M. Plóciennik, C. M. Doutriaux, C. Palazzo, J. Boutte, T. Zok, D. Elia, M. Owsiak, A. D'Anca, Z. Shaheen, R. Bruno, M. Fargetta, M. Caballer, G. Moltó, I. Blanquer, R. Barbera, M. David, G. Donvito, D. N. Williams, V. Anantharaj, D. Salomoni, G. Aloisio, "Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system". In Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016. p. 2911-2918.

[6] D. Elia, S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D. N. Williams, G. Aloisio, "An in-memory based framework for scientific data analytics". In Proceedings of the ACM International Conference on Computing Frontiers (CF '16), May 16-19, 2016, Como, Italy, pp. 424-429

[5] C. Palazzo, A. Mariello, S. Fiore, A. D'Anca, D. Elia, D. N. Williams, G. Aloisio, "A Workflow-Enabled Big Data Analytics Software Stack for eScience", The Second International Symposium on Big Data Principles, Architectures & Applications (BDAA 2015), HPCS 2015, Amsterdam, The Netherlands, July 20-24, 2015, pp. 545-552

[4] S. Fiore, A. D'Anca, D. Elia, C. Palazzo, I. Foster, D. Williams, G. Aloisio, "Ophidia: A Full Software Stack for Scientific Data Analytics", proc. of the 2014 International Conference on High Performance Computing & Simulation (HPCS 2014), July 21 – 25, 2014, Bologna, Italy, pp. 343-350, ISBN: 978-1-4799-5311-0

[3] S. Fiore, C. Palazzo, A. D'Anca, I. T. Foster, D. N. Williams, G. Aloisio, "A big data analytics framework for scientific data management", IEEE BigData Conference 2013: 1-8

[2] S. Fiore, A. D'Anca, C. Palazzo, I. T. Foster, D. N. Williams, G. Aloisio, "Ophidia: Toward Big Data Analytics for eScience", ICCS 2013, June 5-7, 2013 Barcelona, Spain, ICCS, volume 18 of Procedia Computer Science, page 2376-2385. Elsevier, 2013

[1] G. Aloisio, S. Fiore, I. Foster, D. Williams , "Scientific big data analytics challenges at large scale", Big Data and Extreme-scale Computing (BDEC), April 30 to May 01, 2013, Charleston, South Carolina, USA (position paper).

Useful Resources

- *ECASLab: https://ecaslab.cmcc.it/web/home.html*
- *JupyterHub: https://ecaslab.cmcc.it/jupyter/hub/login*
- *Website: https://ophidia.cmcc.it*
- *Documentation : http://ophidia.cmcc.it/documentation*
- *The Ophidia code is available on GitHub under GPLv3 license at https://github.com/OphidiaBigData*
- *RPMs are also available for CentOS6 at the following repo: http://download.ophidia.cmcc.it/rpm*
- *Youtube Channel https://www.youtube.com/user/OphidiaBigData/*
- *To get started in a few minutes with Ophidia, a Virtual Machine Image (OVA format) is also available at https://download.ophidia.cmcc.it/vmi_desktop/*

What have we learned today?

- *Community experiments in the climate domain: the CMIP use case*
- *Needs and challenges for analyzing climate (big) data*
- *ECAS: a solutions for server-side, parallel data analysis in the EOSC landscape*
- *In-depth view of the ECAS core framework (Ophidia)*
  - *Architecture, datacube abstraction, storage back-end, primitives and operators*
  - *Link with EUDAT B2* services*
  - *Workflows and Python apps*
- *ECASLab: a Data Science eco-system for climate data analysis*

EUDAT Collaborative Data Infrastructure

PRACE

Thanks

http://ophidia.cmcc.it

@OphidiaBigData

www.youtube.com/user/OphidiaBigData

EOSC-hub

EOSC-hub receives funding from the EU's Horizon 2020 research and innovation programme under grant agreement No. 777536.

is-enes INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

IS-ENES3 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084