The Ophidia big data analytics framework

Sandro Fiore, Ph.D.

Director of the Advanced Scientific Computing Division
Euro Mediterranean Center on Climate Change (CMCC)

On behalf of the Ophidia Team

2015 ESGF F2F Meeting
Monterey - December 9, 2015

Outline

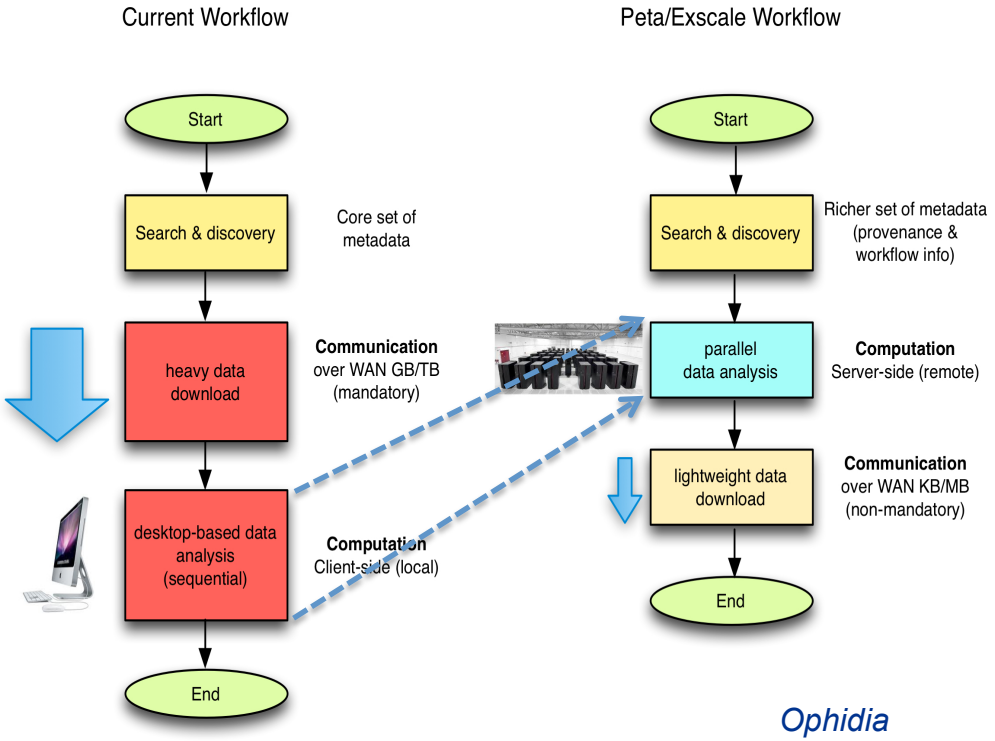
- ✔ *Ophidia introduction*
- ✔ *Workflows support*
  - ✔ *Climate indicators processing*
  - ✔ *Fire danger analysis*
  - ✔ *Cloud-based use case on climate change and biodiversity*
  - ✔ *Climate Model Intercomparison Data Analysis case study*
- ✔ *Modularity, extensibility & programmatic access*
- ✔ *Conclusions*

The Ophidia Project

Ophidia is a research effort carried out at the ***Euro Mediterranean Centre on Climate Change*** *(CMCC) to* address "big data" challenges, issues and requirements for climate change data analytics.

Requirements

- *Time series analysis*
- *Data subsetting*
- *Model intercomparison*
- *Multimodel means*
- *Massive data reduction*
- *Data transformation*
- *Climate change signal*
- *Maps generation*
- *Ensemble analysis*
- *Worflow support*
  - *Tens, hundreds of tasks*
- *Metadata management support*



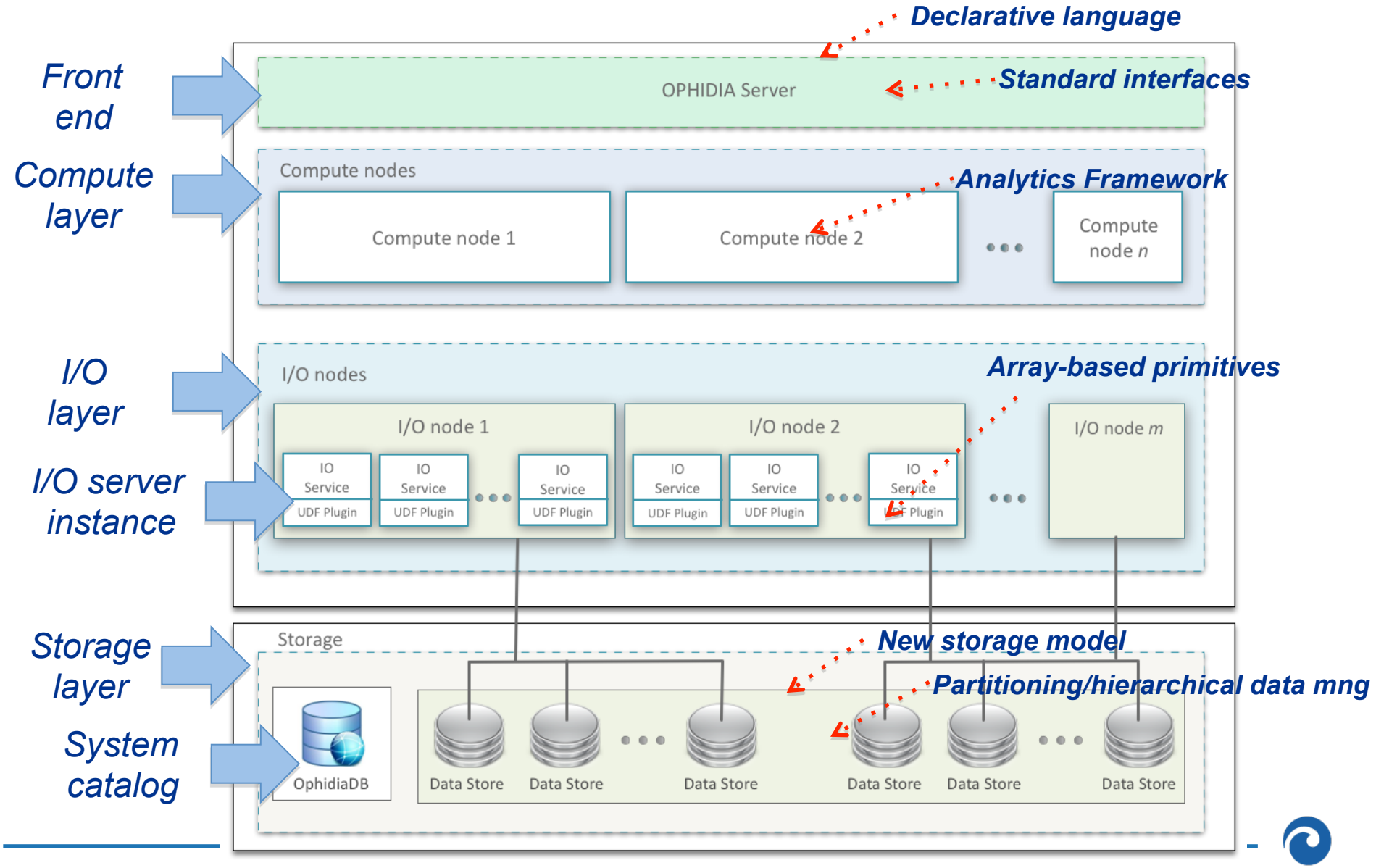
Ophidia
http://ophidia.cmcc.it/

*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*

Ophidia in a nutshell

- ✔ *Big data stack for scientific data analysis*
- ✔ *Use of parallel operators and parallel I/O*
- ✔ *Support for complex workflows / operational chains*
- ✔ *Extensible: simple API to support framework extensions like new operators and array-based primitives*
  - ✔ *currently 50+ operators and 100+ primitives provided*
- ✔ *Multiple interfaces available (WS-I, GSI/VOMS, OGC-WPS).*
- ✔ *Programmatic access via C and Python APIs*
- ✔ *Support for both batch & interactive data analysis*

Ophidia Architecture v0.1



Storage model (dimension-independent) & implementation
Array-based support and hierarchical storage

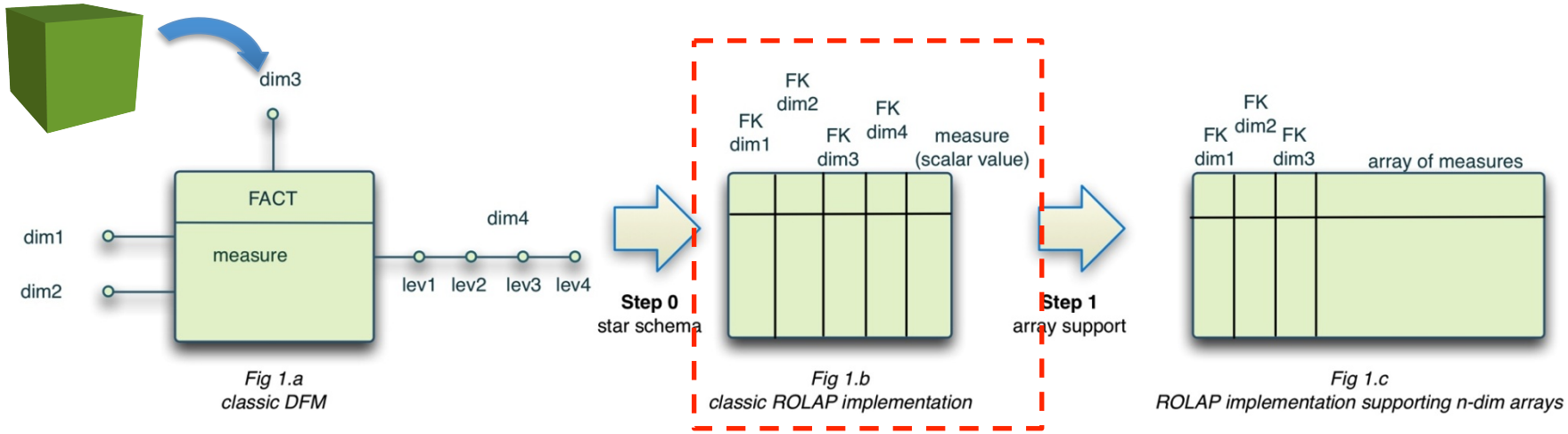


Fig 1.a
classic DFM

Fig 1.b
classic ROLAP implementation

Fig 1.c
ROLAP implementation supporting n-dim arrays

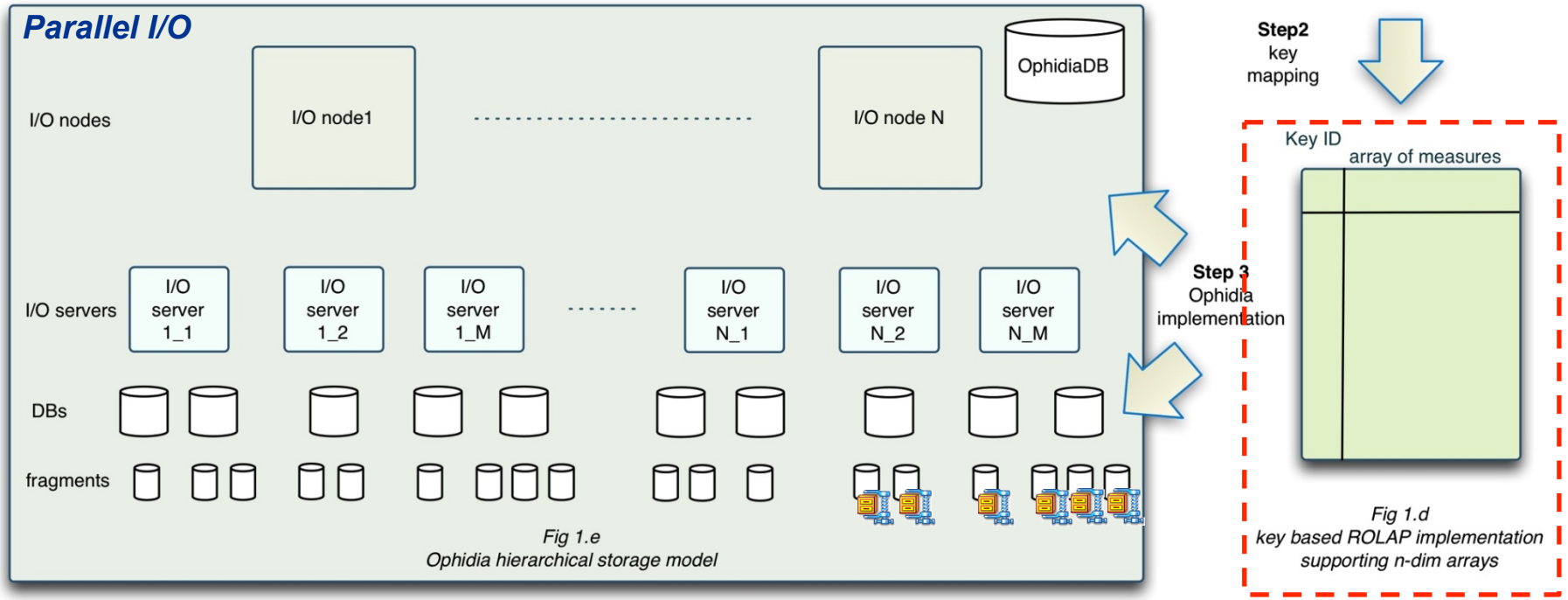
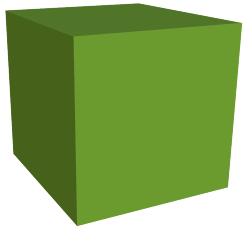
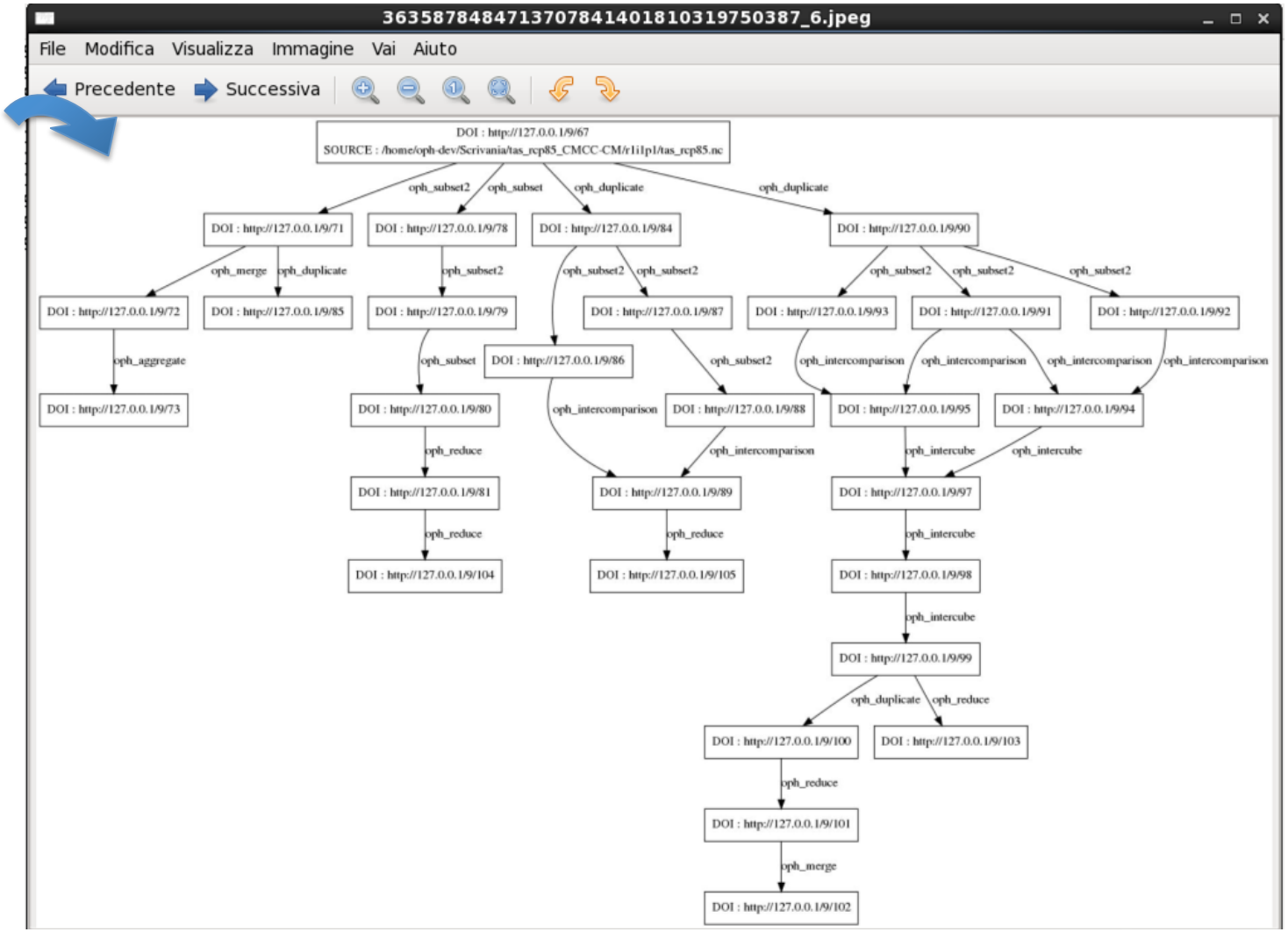


Fig 1.e
Ophidia hierarchical storage model

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

Provenance and PID





The analytics framework: datacube operators (about 50)

OPERATOR NAME	OPERATOR DESCRIPTION
Operators "Data processing" – Domain-agnostic	
OPH_APPLY(*datacube_in*, *datacube_out*, *array_based_primitive*)	Creates the *datacube_out* by applying the *array-based primitive* to the *datacube_in*
OPH_DUPLICATE(*datacube_in*, *datacube_out*)	Creates a copy of the *datacube_in* in the *datacube_out*
OPH_SUBSET(*datacube_in*, *subset_string*, *datacube_out*)	Creates the *datacube_out* by doing a sub-setting of the *datacube_in* by applying the *subset_string*
OPH_MERGE(*datacube_in*, *merge_param*, *datacube_out*)	Creates the *datacube_out* by merging groups of *merge_param* fragments from *datacube_in*
OPH_SPLIT(*datacube_in*, *split_param*, *datacube_out*)	Creates the *datacube_out* by splitting into groups of *split_param* fragments each fragment of the *datacube_in*
OPH_INTERCOMPARISON (*datacube_in1*, *datacube_in2*, *datacube_out*)	Creates the *datacube_out* which is the element-wise difference between *datacube_in1* and *datacube_in2*
OPH_DELETE(*datacube_in*)	Removes the *datacube_in*

Data processing (parallel operators, MPI & OpenMP based)

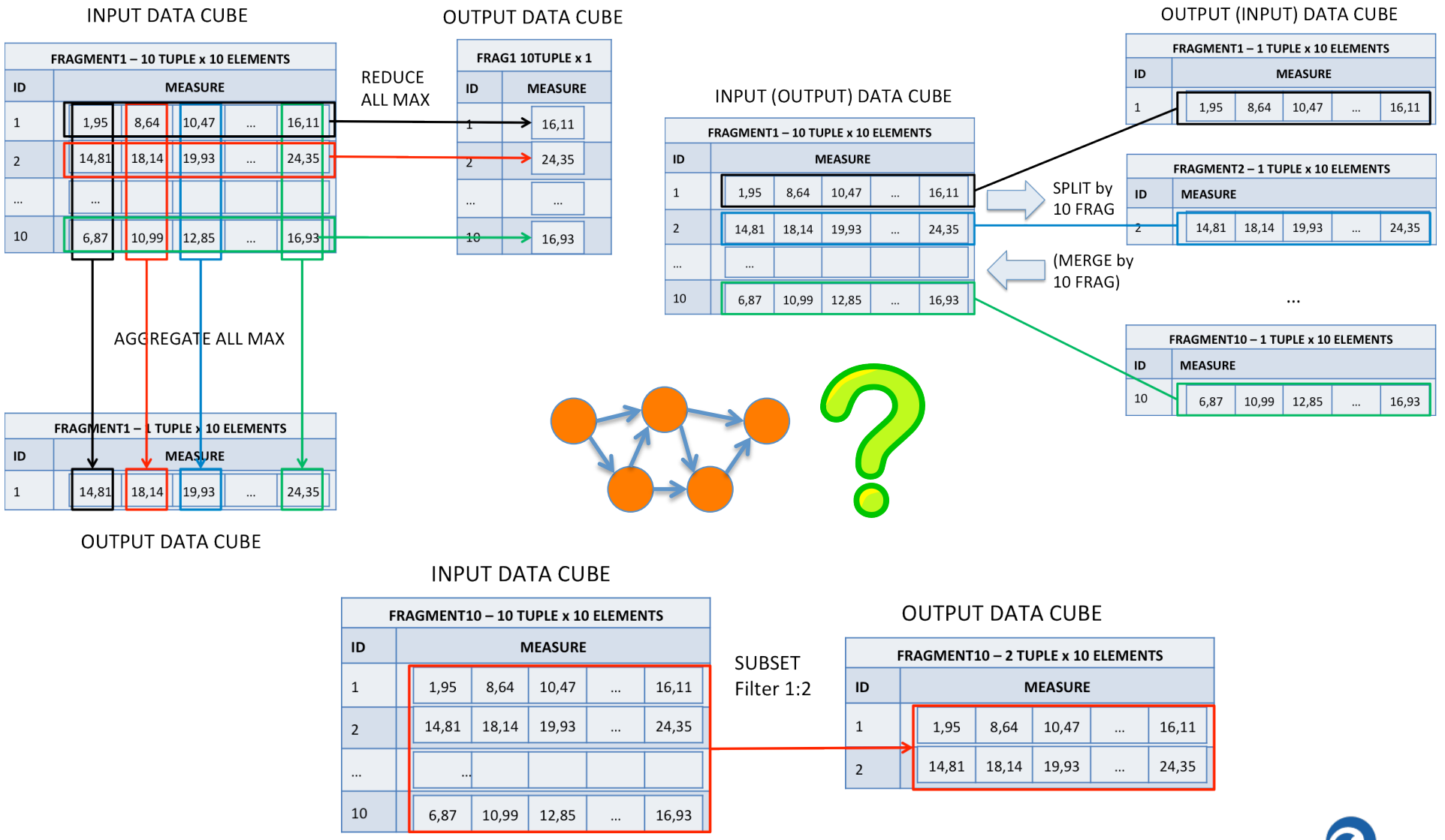
Import/Export (parallel operators)

Data Access (sequential and parallel operators)

Metadata management (sequential and parallel operators)

OPERATOR NAME	OPERATOR DESCRIPTION
Operators "Data processing" – Domain-oriented	
OPH_EXPORT_NC (*datacube_in*, *file_out*)	Exports the *datacube_in* data into the *file_out* NetCDF file.
OPH_IMPORT_NC (*file_in*, *datacube_out*)	Imports the data stored into the *file_in* NetCDF file into the new *datacube_in* datacube
Operators "Data access"	
OPH_INSPECT_FRAG (*datacube_in*, *fragment_in*)	Inspects the data stored in the *fragment_in* from the *datacube_in*
OPH_PUBLISH(*datacube_in*)	Publishes the *datacube_in* fragments into HTML pages
Operators "Metadata"	
OPH_CUBE_ELEMENTS (*datacube_in*)	Provides the total number of the elements in the *datacube_in*
OPH_CUBE_SIZE (*datacube_in*)	Provides the disk space occupied by the *datacube_in*
OPH_LIST(void)	Provides the list of available datacubes.
OPH_CUBEIO(*datacube_in*)	Provides the provenance information related to the *datacube_in*
OPH_FIND(*search_param*)	Provides the list of datacubes matching the *search_param* criteria

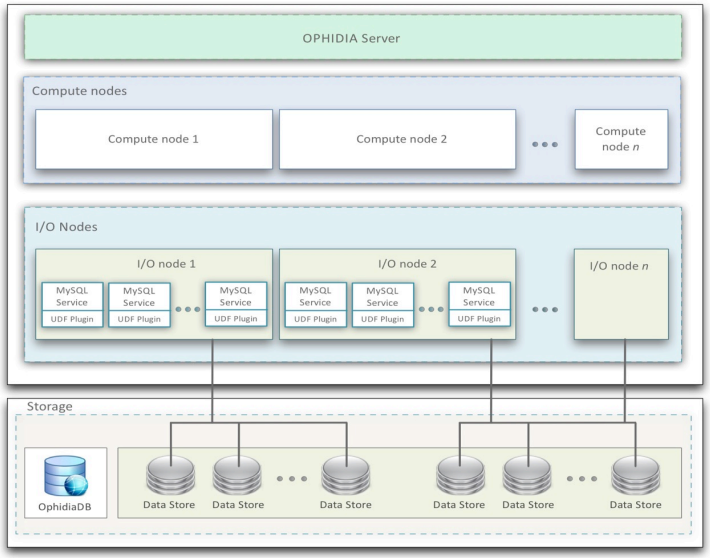
The analytics framework: datacube operators



How to submit a workflow: the Ophidia Terminal

- The Ophidia terminal provides an effective and lightweight way to interact with the Ophidia server
- Bash-like env. (commands interpreter)
- Terminal with history management, auto-completion, specific environment variables and commands with integrated help
- Easy installation as an only one executable using a small number of well-known and open-source libraries
- Simple enough for a novice and at the same time powerful enough for an expert

Let's consider a use case on climate indicators implemented in the CLIP-C project



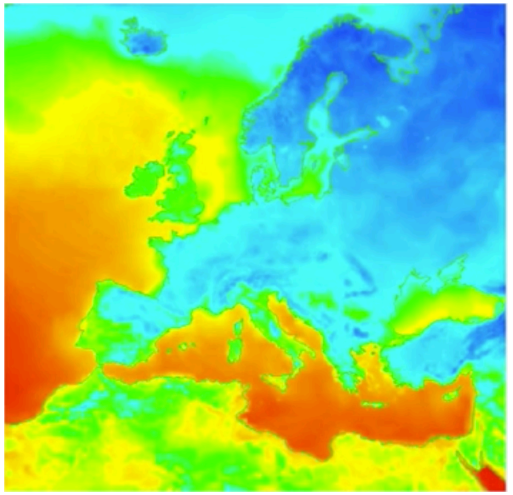
Use case on climate indicators processing

Climate Information Portal

- ✔ *In the CLIPC project, processing chains for data analysis are being implemented with Ophidia to compute **climate indicators***

- ✔ ***First set of indicators** includes: TNn, **TNx**, TXn, **TXx***
  - ✔ *Input files: 12GBs (TasMin & TasMax)*
  - ✔ *Workflows have been already implemented*
  - ✔ *Demo on Thrursday*

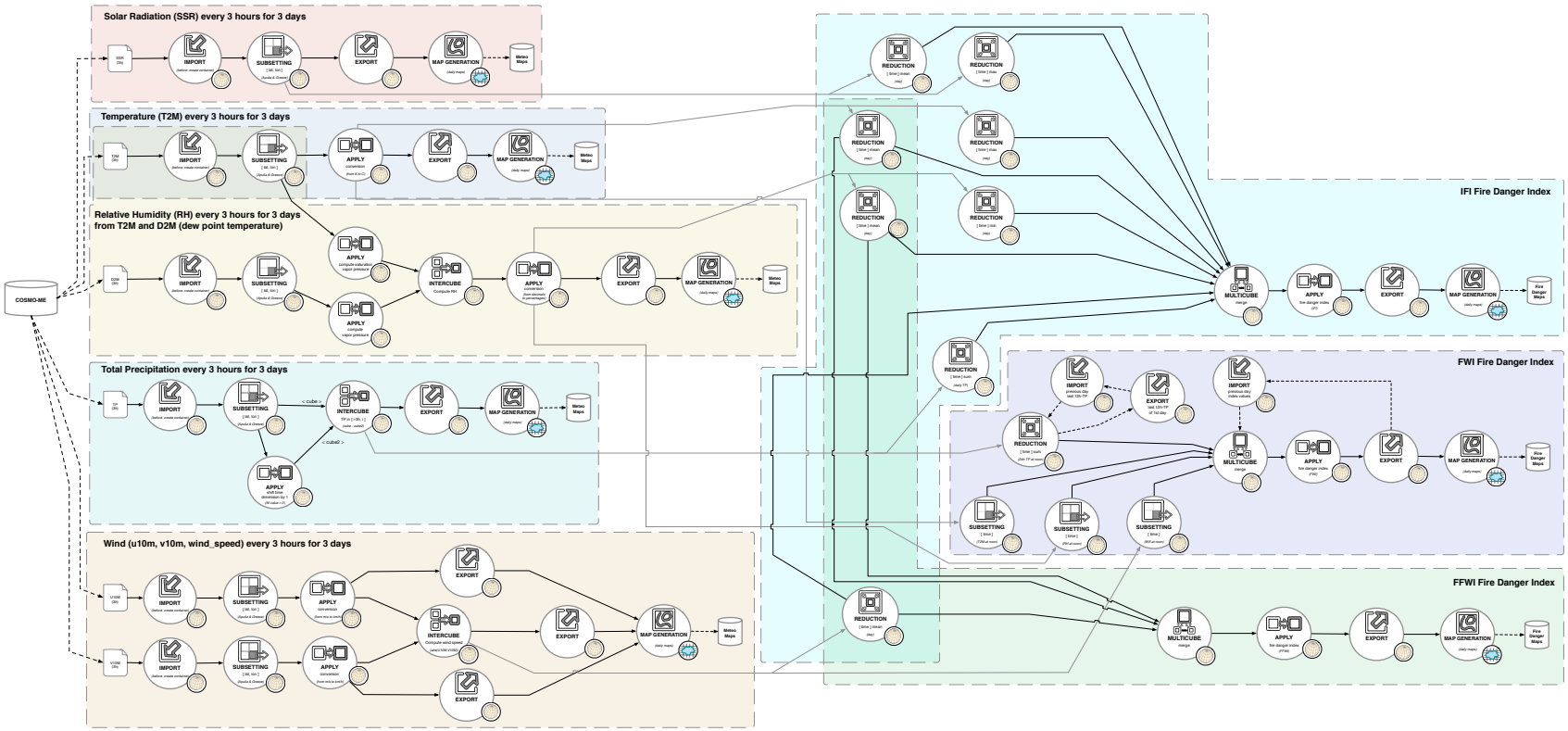
- ✔ ***Parallel approach***
  - ✔ *Inter-parallelism & Intra-parallelism*



See the demo on Thursday

Operational FIre Danger preventIon plAtform

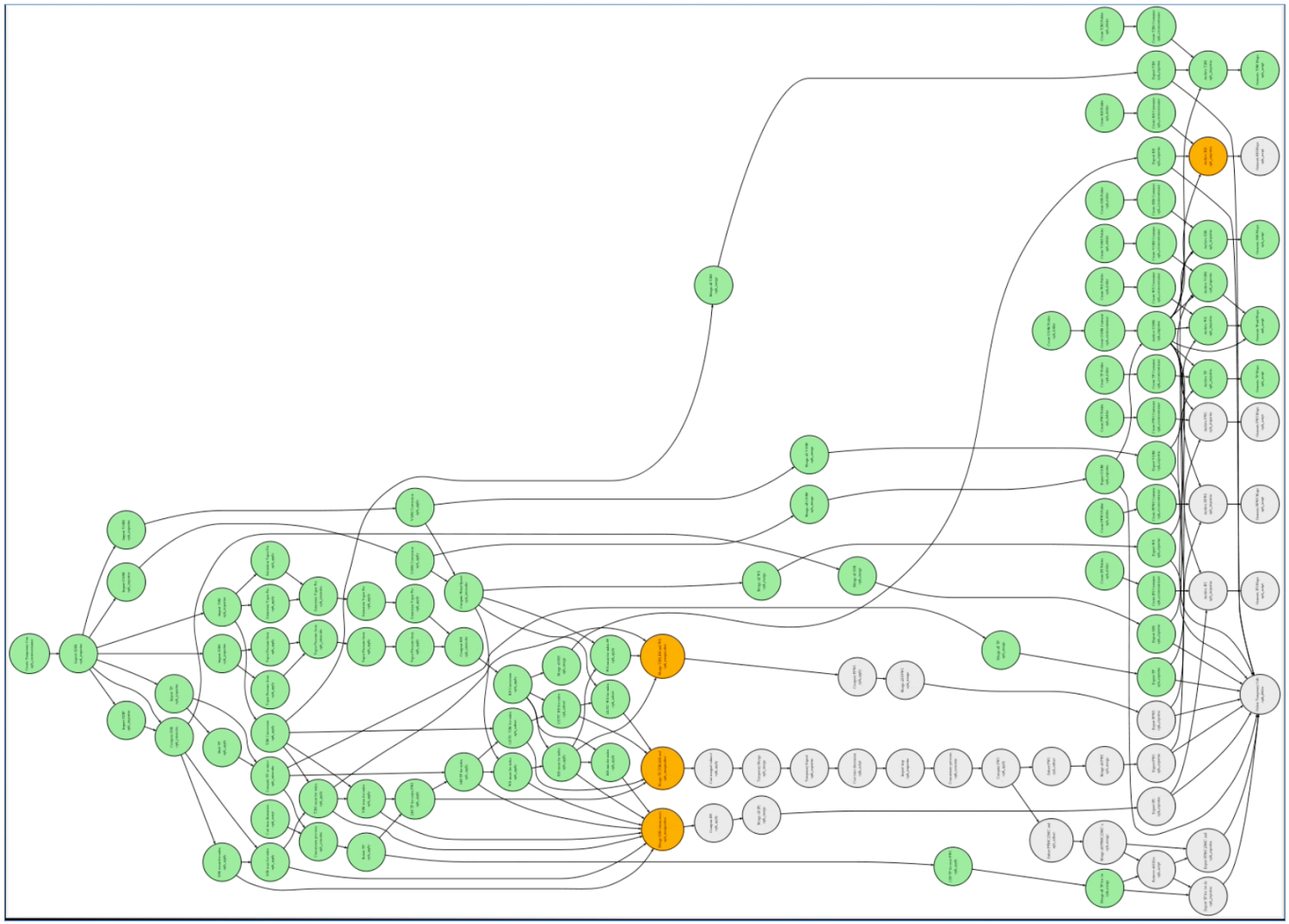
OFIDIA main objective is to build a **cross-border operational fire danger prevention infrastructure** that advances the ability of regional stakeholders across Apulia and Ioannina Regions to **detect** and **fight forest wildfires**



OFIDIA: Operational FIre Danger preventIon plAtform

Co-ordinator: Prof. G. Aloisio (CMCC)
Website: http://www.ofidia.eu/

Workflow runtime execution (fire danger analysis)



https://www.youtube.com/watch?v=vxbYF1Zhpuc&feature=youtu.be

EUBrazilCC project

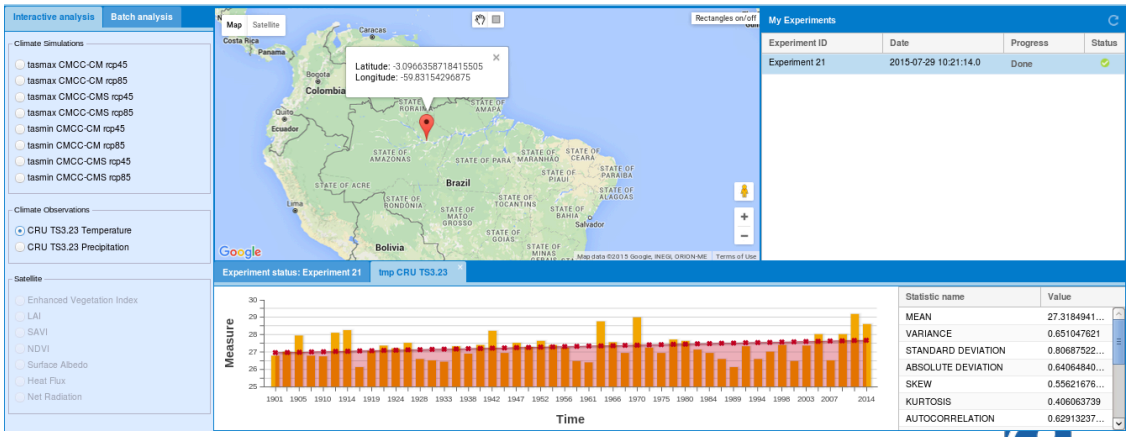
EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

- ✓ *The main objective is the creation of a **federated e-infrastructure for research using a user-centric approach**.*
- ✓ *To achieve this, we need to pursue three objectives:*
  - ✓ *Adaptation of existing applications to tackle new scenarios emerging from cooperation between **Europe** and **Brazil** relevant to both regions.*
  - ✓ *Integration of frameworks and programming models for **scientific gateways** and **complex workflows**.*
  - ✓ ***Federation of resources**, to build up a **general-purpose infrastructure** comprising existing and **heterogeneous resources***
  - ✓ ***Data analytics workflows** on heterogeneous datasets including **climate**, **remote sensing** data and **observations** (e.g. NetCDF, LANDSAT, LiDAR)*

EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

EU Coordinator
Ignacio Blanquer-Espert, iblanque@dsic.upv.es
Universitat Politècnica de València, Spain

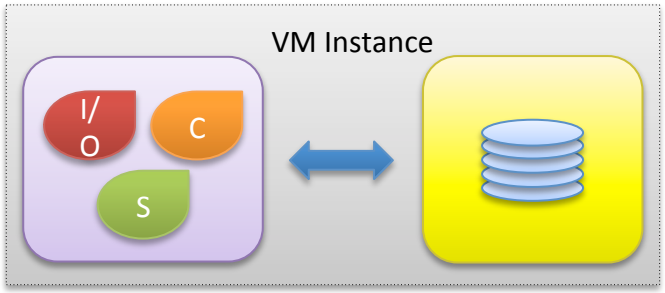
BR Coordinator
Francisco Vilar Brasileiro, fubica@dsc.ufcg.edu.br
Universidade Federal de Campina Grande, Brazil



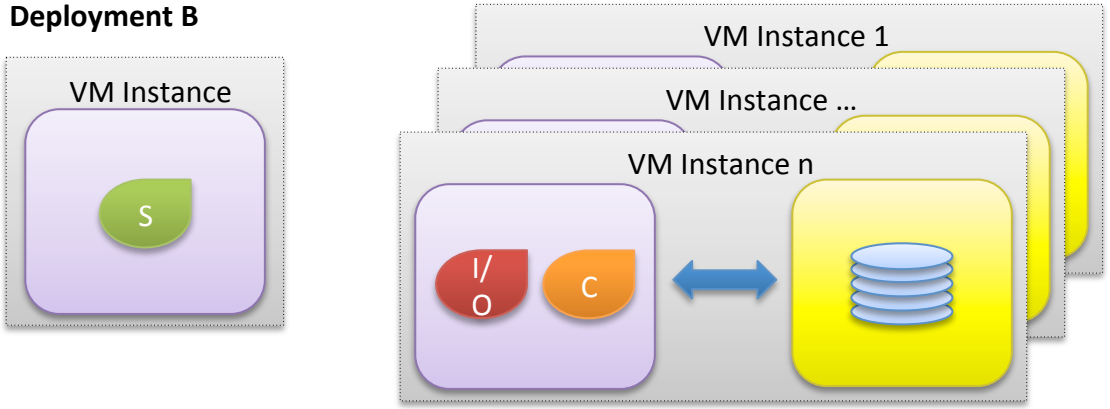
Cloud-based deployment scenarios

EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

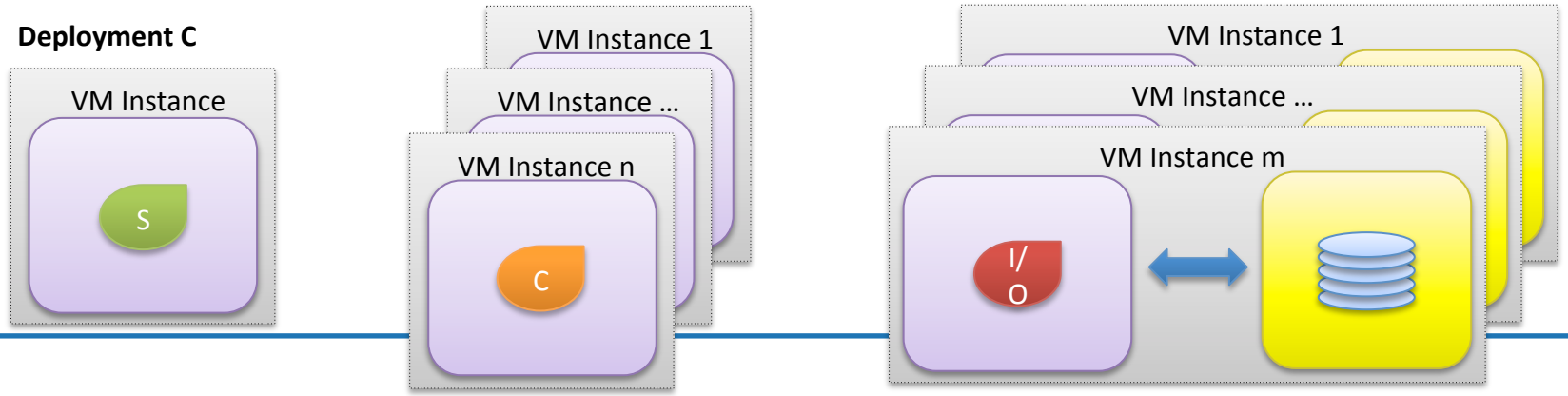
Deployment A



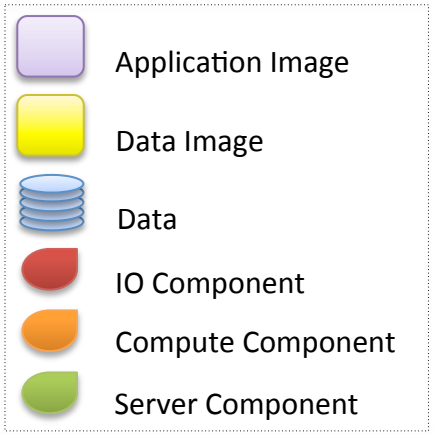
Deployment B



Deployment C



Legend



INtegrating Distributed data Infrastructures for Global ExplOitation

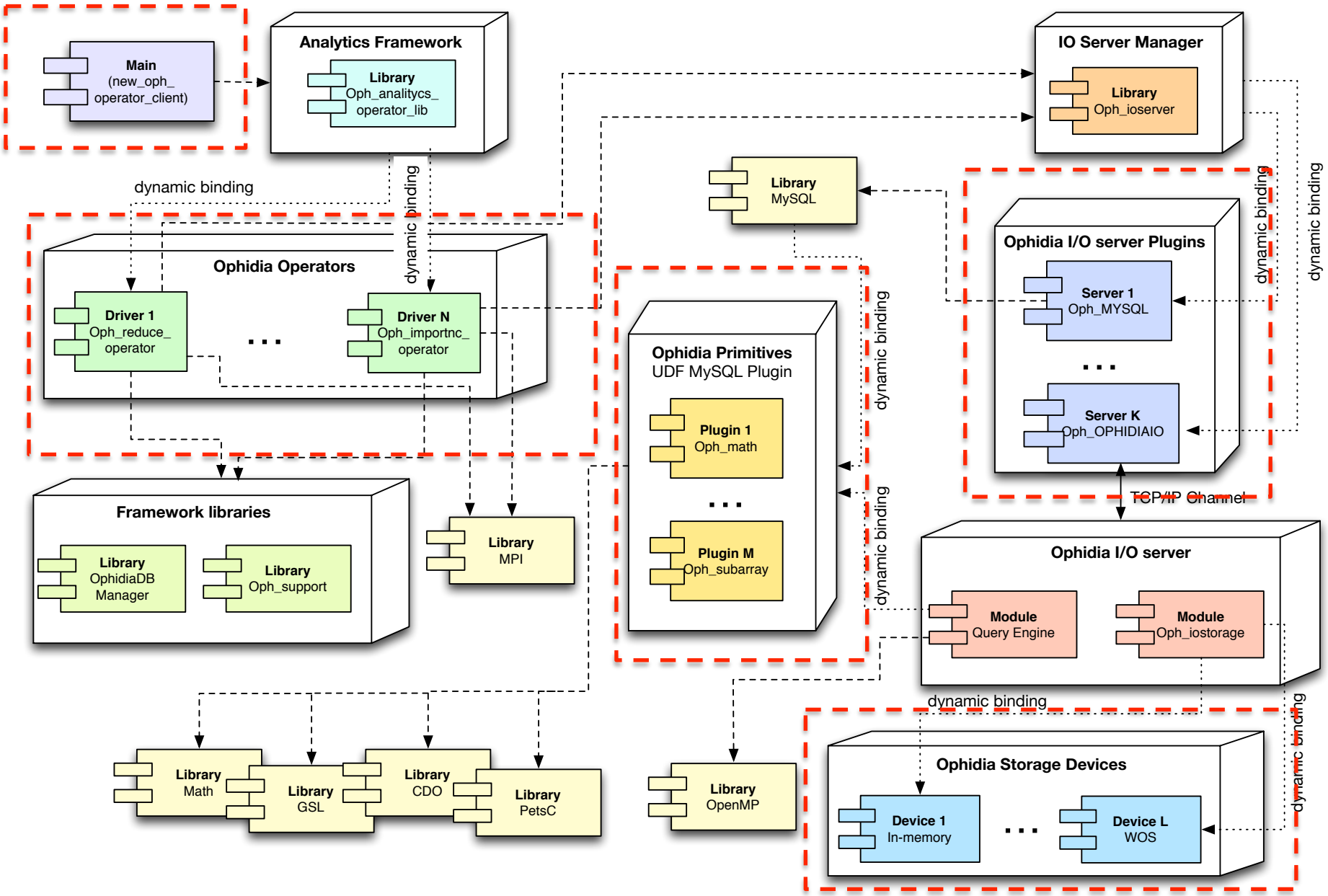
- ✔ **INDIGO-DataCloud** is a project approved within the **E-INFRA-1-2014** call of the **Horizon 2020** framework program of the European Community.
  - ✔ It aims at developing a **data/computing platform** targeting **scientific communities**, **deployable on multiple hardware** and provisioned over **hybrid** (private or public) **e-infrastructures**.

- ✔ It aims at targeting multiple case studies related to different domains

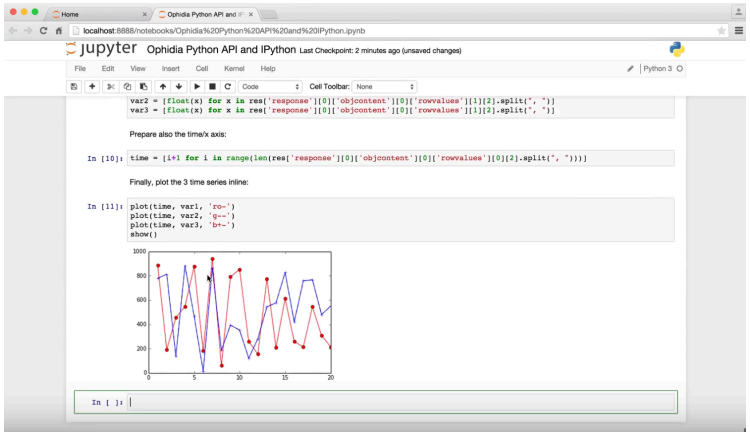
- ✔ Key points for the "*Climate Model Intercomparison Data Analysis*" case study:
  - ✔ **Interoperability** with application domain specific software and services will be addressed (e.g. IS-ENES/ESGF) – Link to the ESGF-CWT activity
  - ✔ **Server-side and parallel** approach for large scale data analysis
  - ✔ **Two-level workflows** to fully address the case study requirements
  - ✔ **Multi-site experiment** to demonstrate the feasibility of the approach at the federation level

Co-ordinator: Dr. Davide Salomoni (INFN)
Technical Director: Giacinto Donvito (INFN)
Website: https://www.indigo-datacloud.eu/

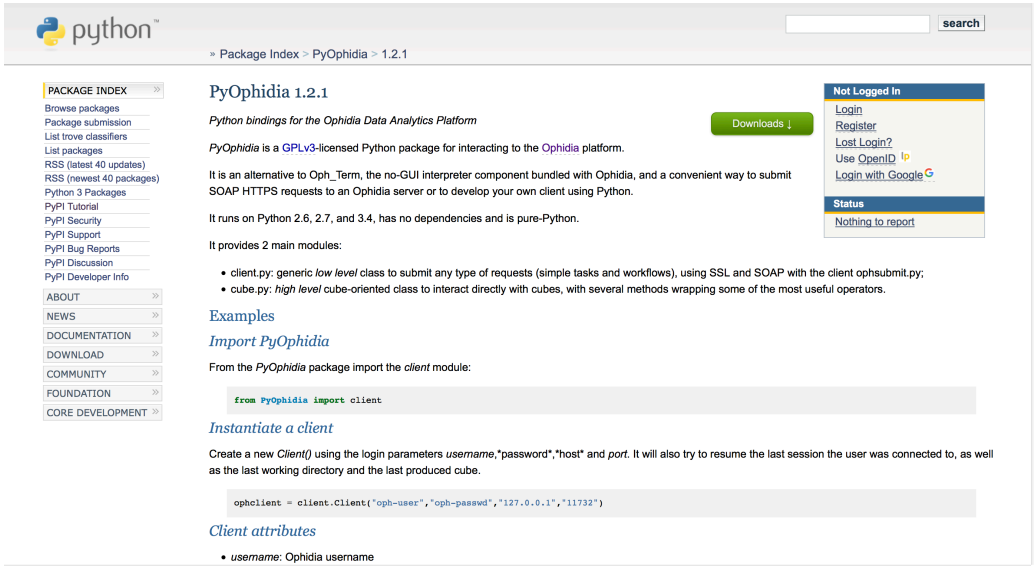
Modularity and extensibility: APIs and dynamic bindings



Programmatic access through the PyOphidia class



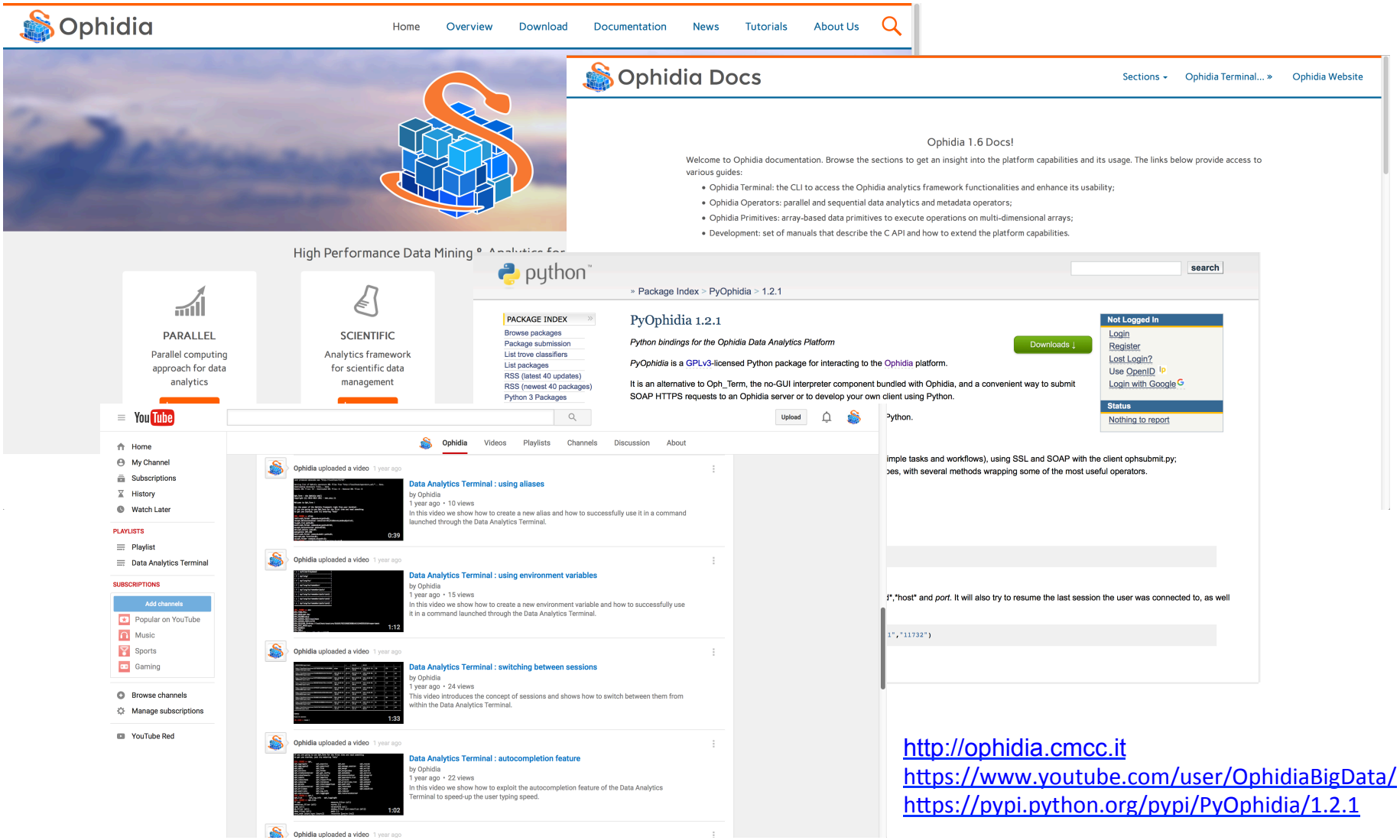
https://pypi.python.org/pypi/PyOphidia/1.2.1



https://www.youtube.com/watch?v=8pcrBXboF6U&feature=youtu.be

- ✓ **PyOphidia** provides a Python interface to submit commands to the Ophidia Server and to retrieve/deserialize the results
- ✓ Two classes implemented:
  - ✓ **Client class**: connect to the server, navigate into the ophidia file system, submit workflows, manage sessions, etc.
  - ✓ **Cube class**: manipulate cubes (reduce, subset, operations between cubes, intercomparison, etc.), get information on cubes (schema, dimensions, metadata, etc.)

Useful resources



http://ophidia.cmcc.it
https://www.youtube.com/user/OphidiaBigData/
https://pypi.python.org/pypi/PyOphidia/1.2.1

Conclusions

- ✔ *Ophidia is a big data analytics framework for eScience*
  - ✔ *OLAP approach for big data – multidimensional data model*
- ✔ *Multiple use cases for data analysis in different domains/contexts have been implemented*
  - ✔ *Sea situational awareness, fire danger prevention, climate indicators, couple model intercomparison data analysis, biodiversity and climate change*
- ✔ *It provides access via CLI (end-users) and API (devel users)*
- ✔ *Programmatic access via C and Python APIs*
- ✔ *Several deployment scenarios in the cloud have been implemented*
  - ✔ *Mainly in the EUBrazilCC project*
  - ✔ *Publication of the VMIs on the EGI AppDB expected before the end of the year (final testing are ongoing)*
- ✔ *Roadmap towards ESGF: WPS support already available*
- ✔ *Official Release v1.0 is coming: January 2016*

Thanks

http://ophidia.cmcc.it

@OphidiaBigData

www.youtube.com/user/OphidiaBigData

Do you want to join us?