Analytics workflows with Ophidia/ECAS

Sandro Fiore, Ph.D.
CMCC Advanced Scientific Computing Division

3rd ENES Workshop on Workflows funded by ESiWACE
Brussels, September 13, 2018

Outline

- ECAS, EOSC-hub and Ophidia
- Ophidia
  - Architecture 1.0
    - Storage model
    - Primitives,
    - Operators
  - Architecture 2.0
    - Workflow support
      - Design language
      - Conditional and loop statements
      - Parallelism support
      - Interactive & batch
      - Inter-wf dependencies
      - Some use cases
- Future work and conclusions
    - Looking forward
    - Website and resources

ECAS, EOSC-hub and Ophidia

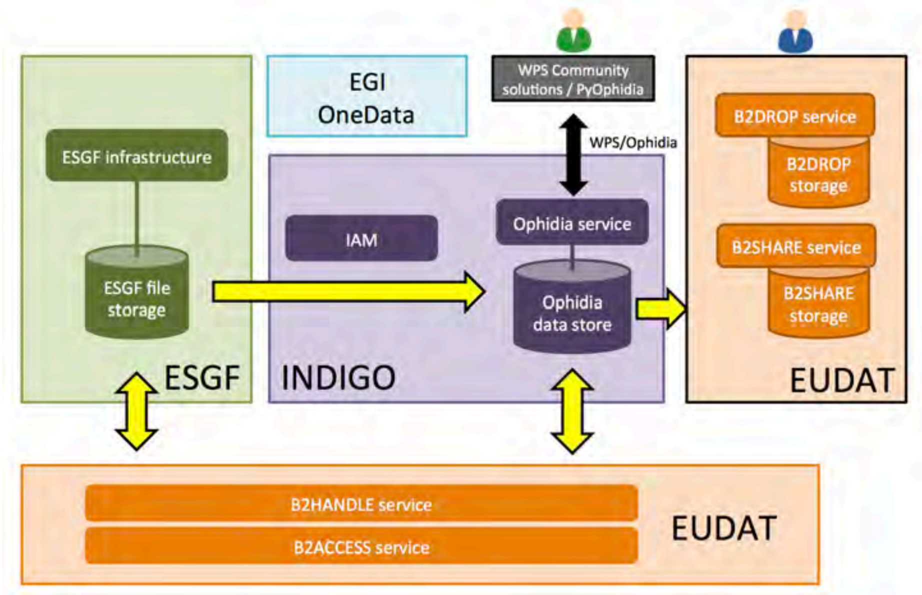
ENES Climate Analytics Service (ECAS)

Ophidia

enes
EUROPEAN NETWORK
FOR EARTH SYSTEM MODELLING

- ✓ The **ENES Climate Analytics Service (ECAS)**, is a **Thematic Services** in EOSC-hub to supports climate data analysis
- ✓ Enable **server-side** analytics workflows for Earth system researchers and beyond
- ✓ Induce cultural change: No more "**download and process at home**"
- ✓ Involved institutions: CMCC and DKRZ

ECAS builds on top of the **Ophidia big data analytics framework** integrating components from INDIGO-DataCloud, EUDAT and EGI



EOSC-hub

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.

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

Ophidia

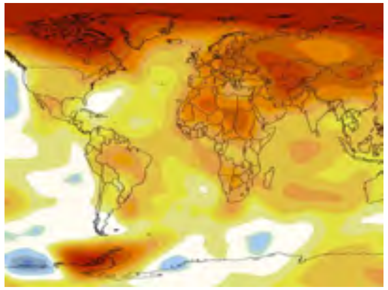
cmcc
Centro Euro-Mediterraneo sui Cambiamenti Climatici

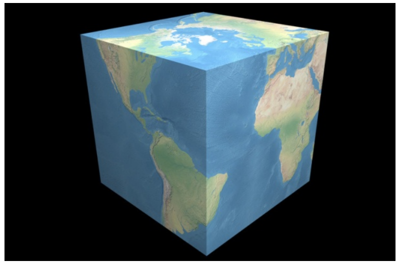
Ophidia in a nutshell

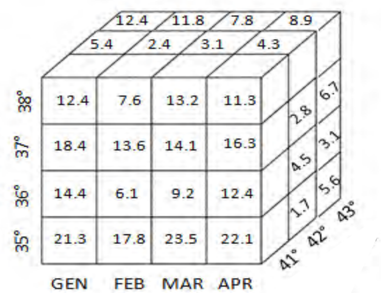
- ***Big data software stack for scientific data analysis***
- ***Features:*** *time series analysis (array-based analysis), data subsetting (by value/index), data aggregation, data intercomparison, OLAP support, etc.*
- *Use of parallel operators and parallel I/O*
- ***Support for complex workflows / 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*

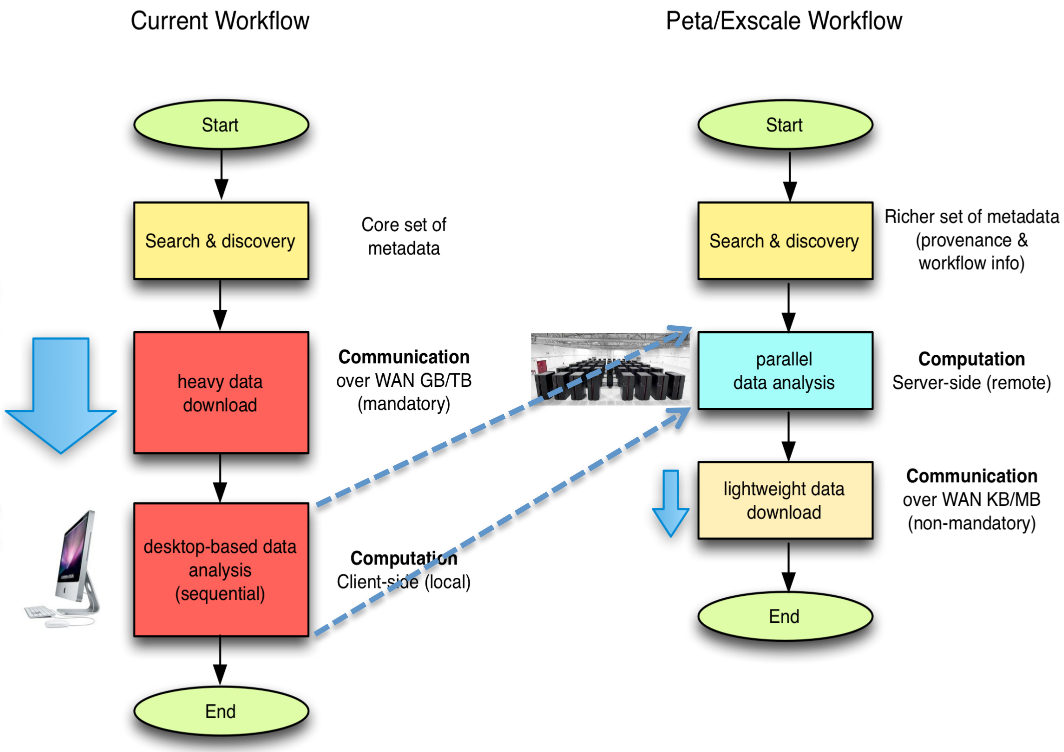
Big data and HPC convergence as a paradigm shift to large-scale data analysis experiments

- *Volume, variety, velocity are key challenges for big data in general and eScience contexts too*
- *High Performance Data Analytics solutions are becoming key to manage data analysis at scale*
- *In-memory analytics can help reducing time to solution*
- *Workflow management has to orchestrate millions of analytics jobs*



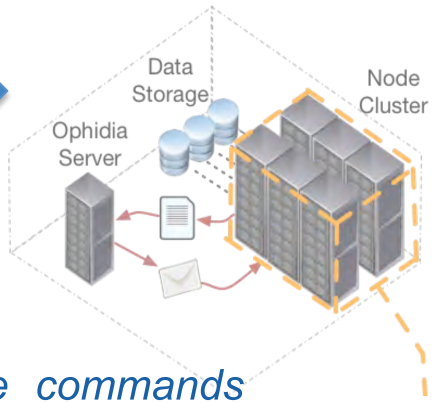


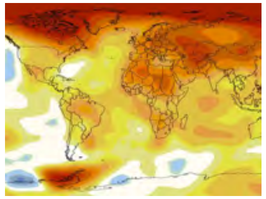


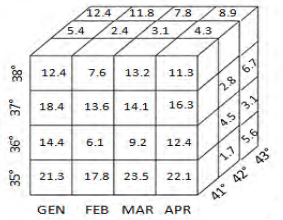


*S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D. N. Williams, G. Aloisio, "**Ophidia: toward bigdata analytics for eScience**", ICCS2013 Conference, Procedia Elsevier, Barcelona, June 5-7, 2013*

Server-side paradigm and the datacube abstraction





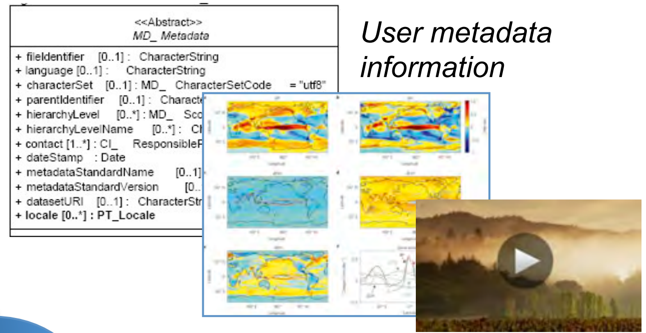


***Oph_Term**: a terlminal-like commands interpreter serving as a client for the Ophidia framework*

***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



User metadata information

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

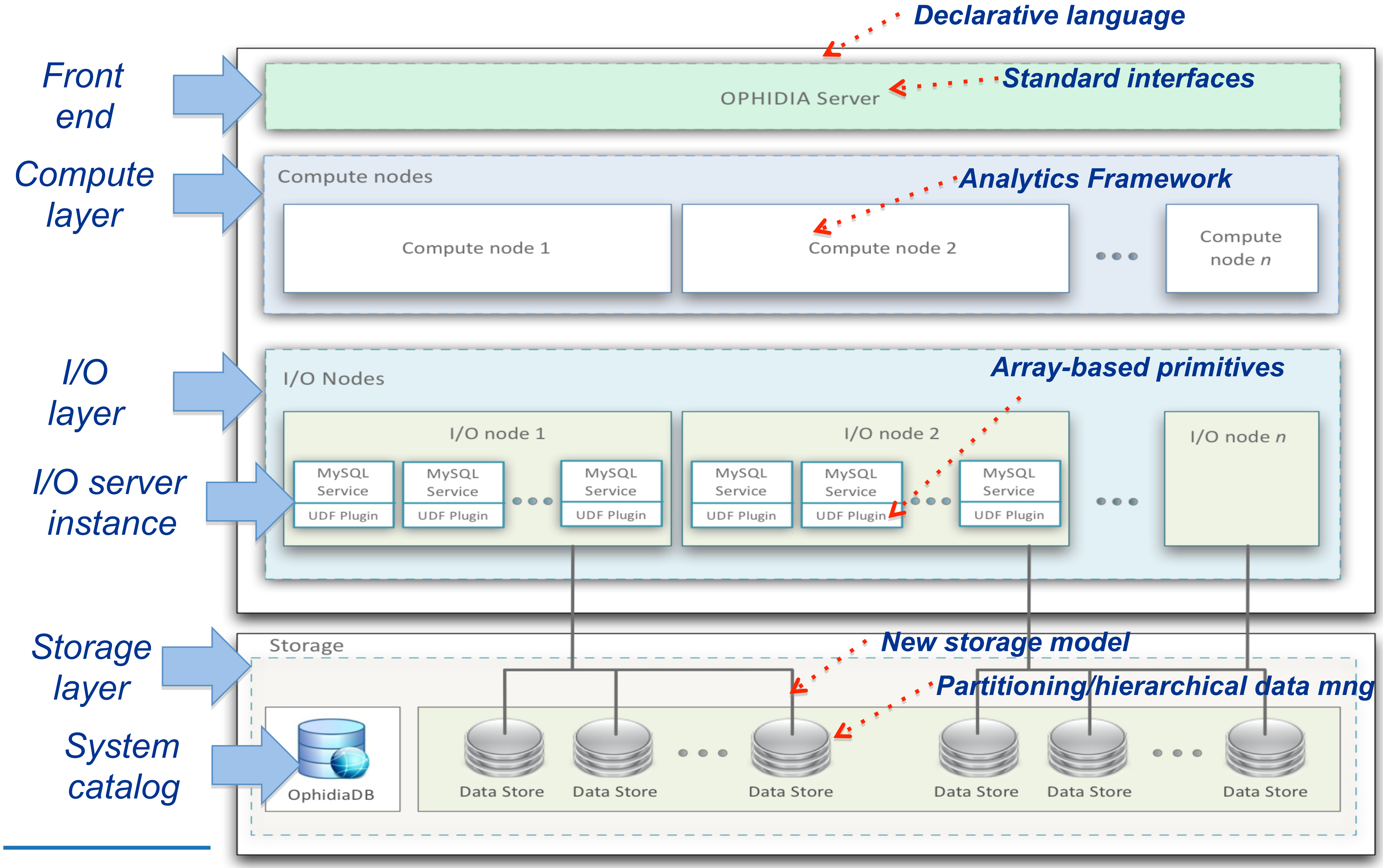
Metadata provenance

```
--> 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)
```

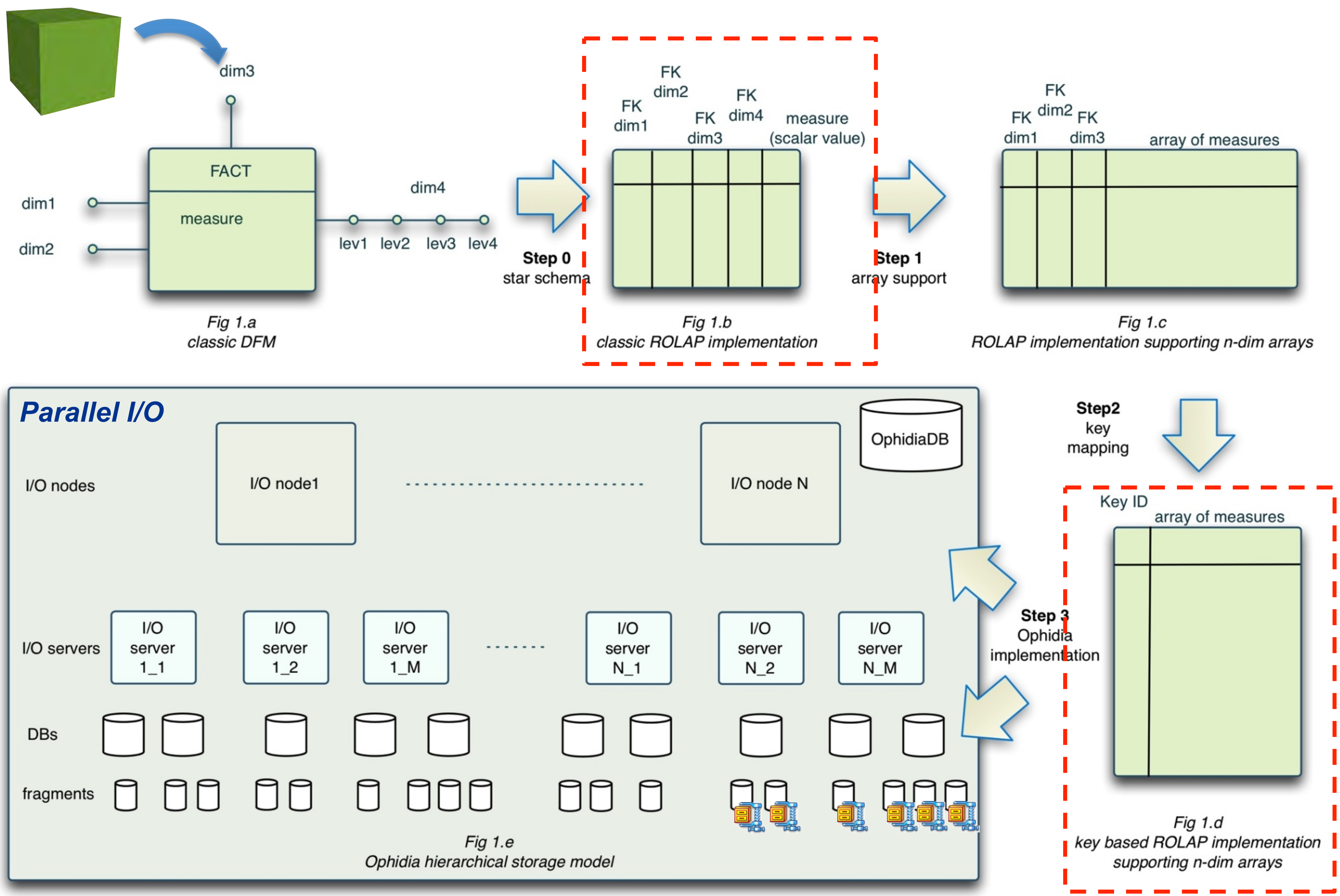
Ophidia architecture 1.0

Storage model, primitive & operators

Ophidia Architecture (sw stack view)



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



Array based primitives (single chunk level)

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

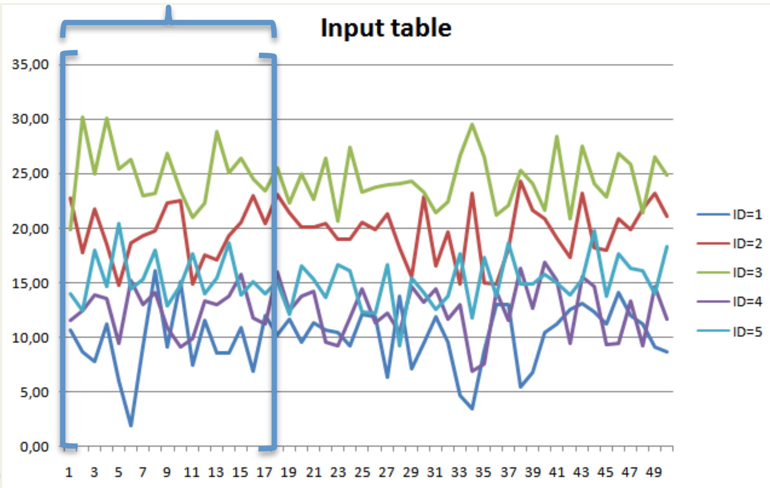
Single chunk or fragment (input)

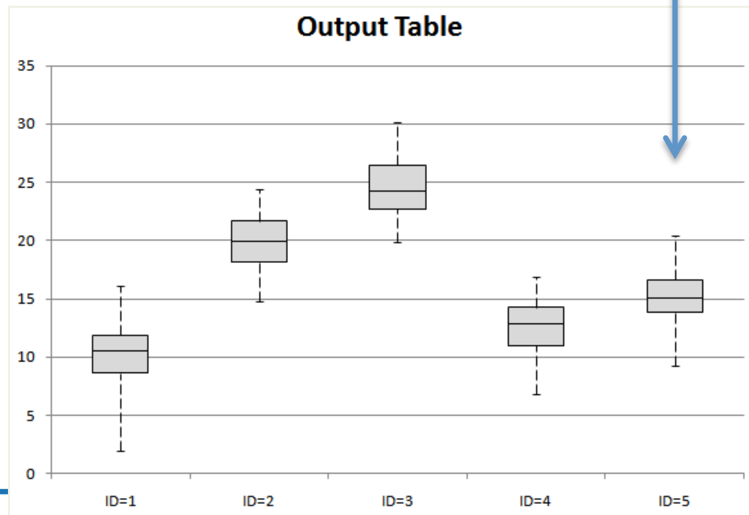
INPUTTABLE 5 tuples x 50 elements										
ID	**MEASURE**									
1	10,73	8,66	7,83	11,20	6,02	1,95	...	16,11	...	8,70
2	22,85	17,84	21,82	18,57	14,81	18,71	...	19,83	...	21,13
3	19,89	30,17	24,95	30,07	25,40	26,31	...	23,18	...	24,82
4	11,60	12,49	13,91	13,53	9,48	15,27	...	14,17	...	11,66
5	13,94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

subarray(measure, 1,18)

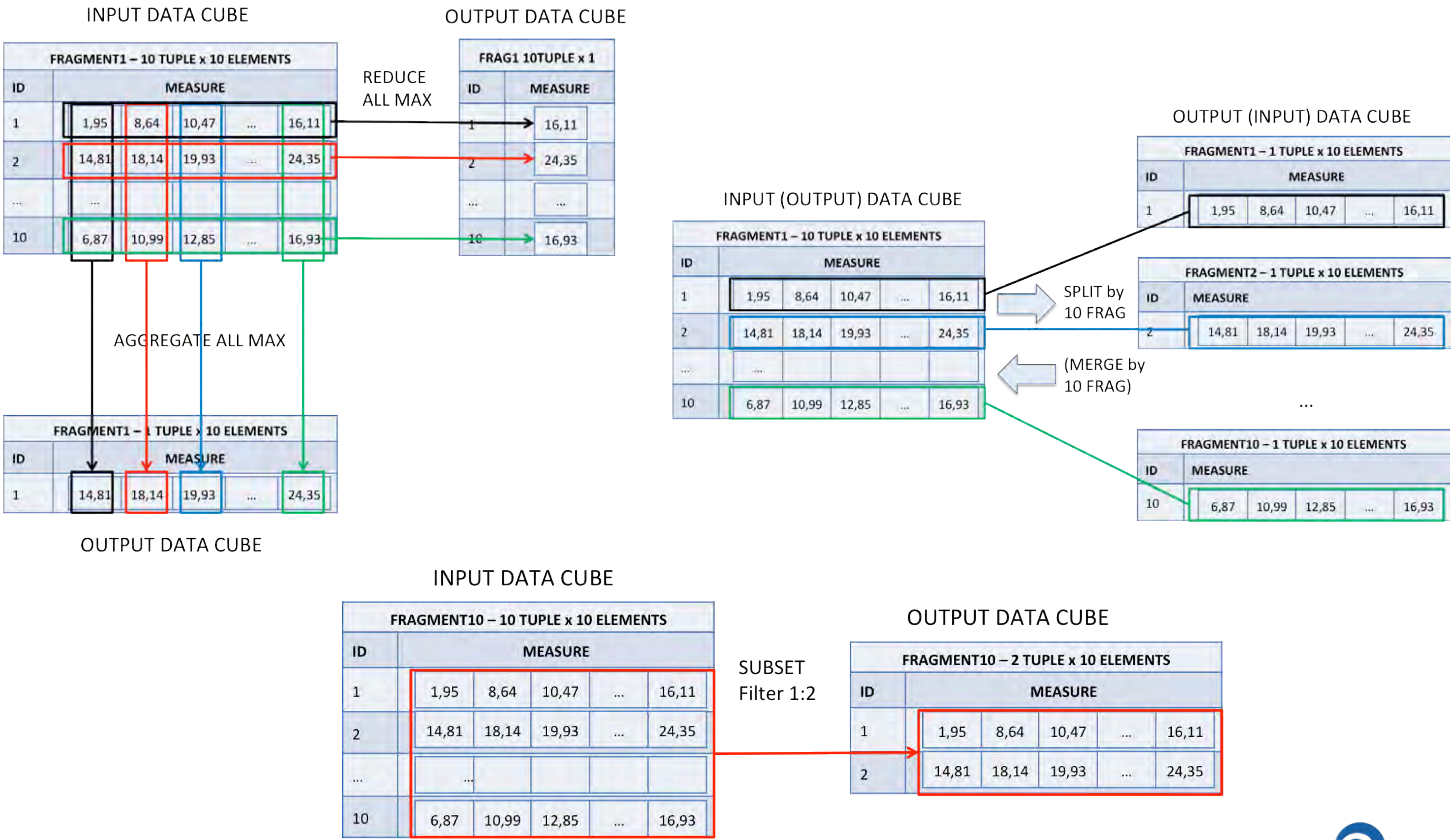
Single chunk or fragment (output)

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





Analytics operators (datacube level)



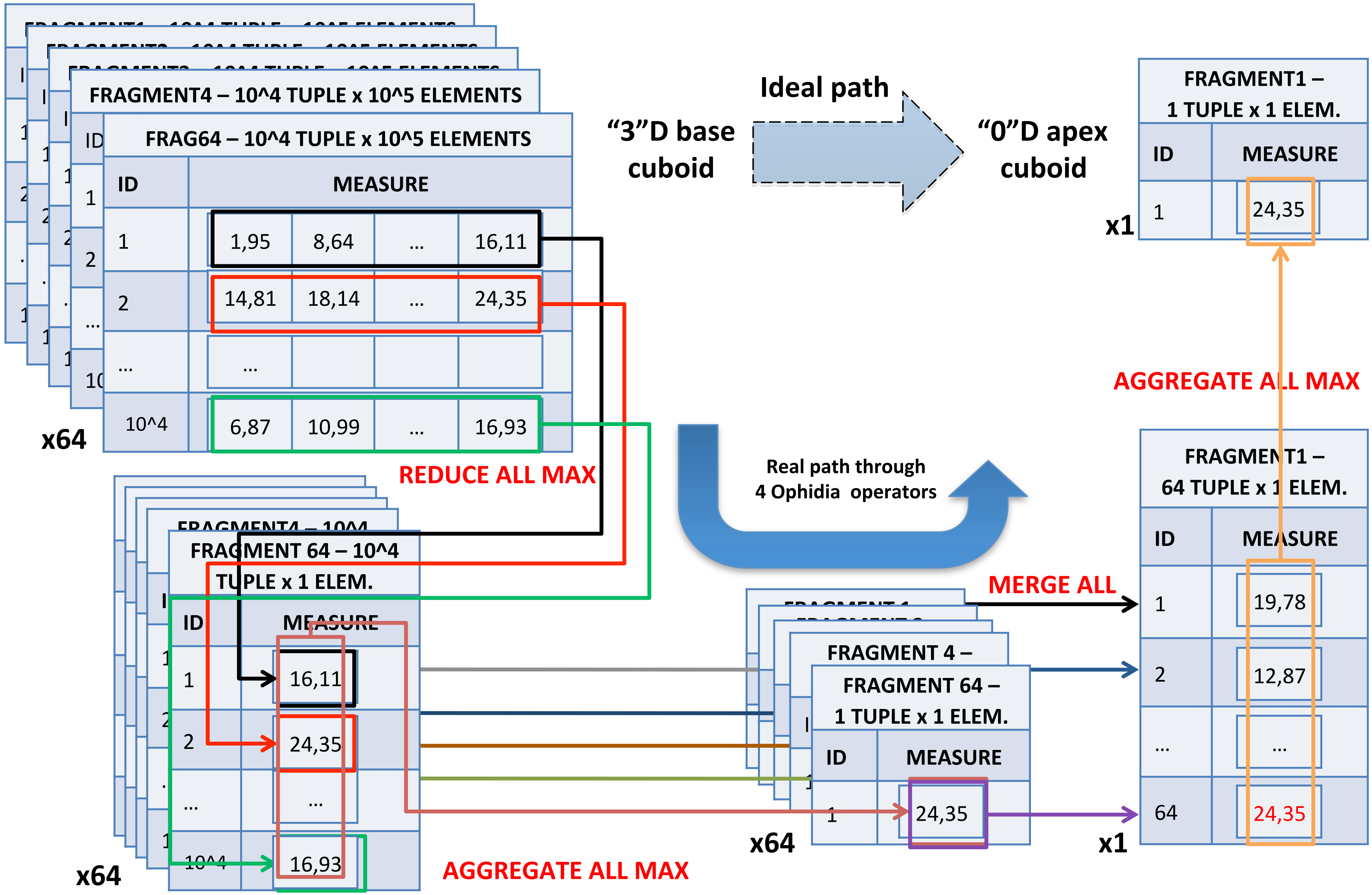
Analytics operators (datacube level)

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.

About 50 operators for data and metadata processing

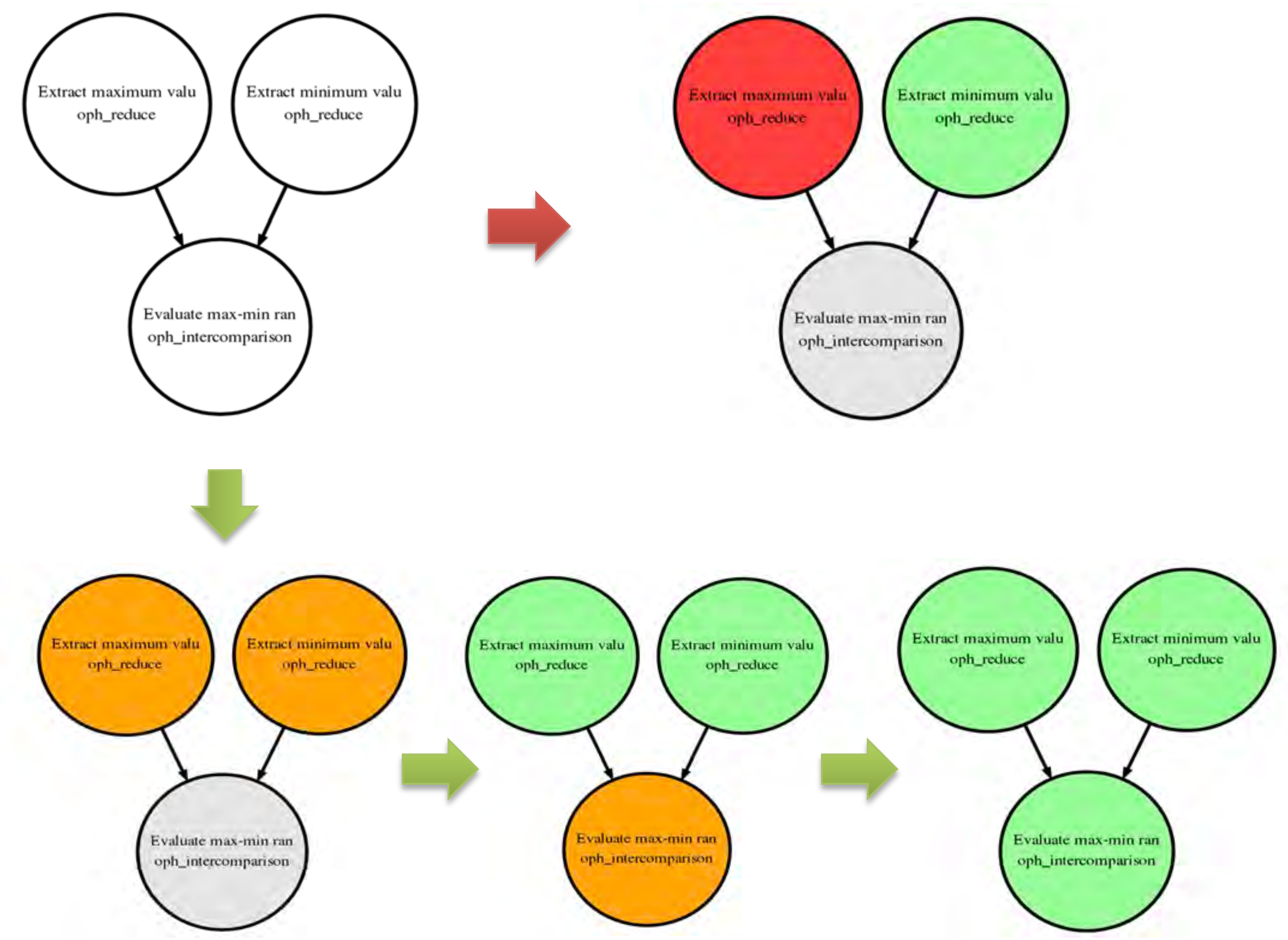
Pipelining analytics operators to reduce data



Ophidia architecture 2.0

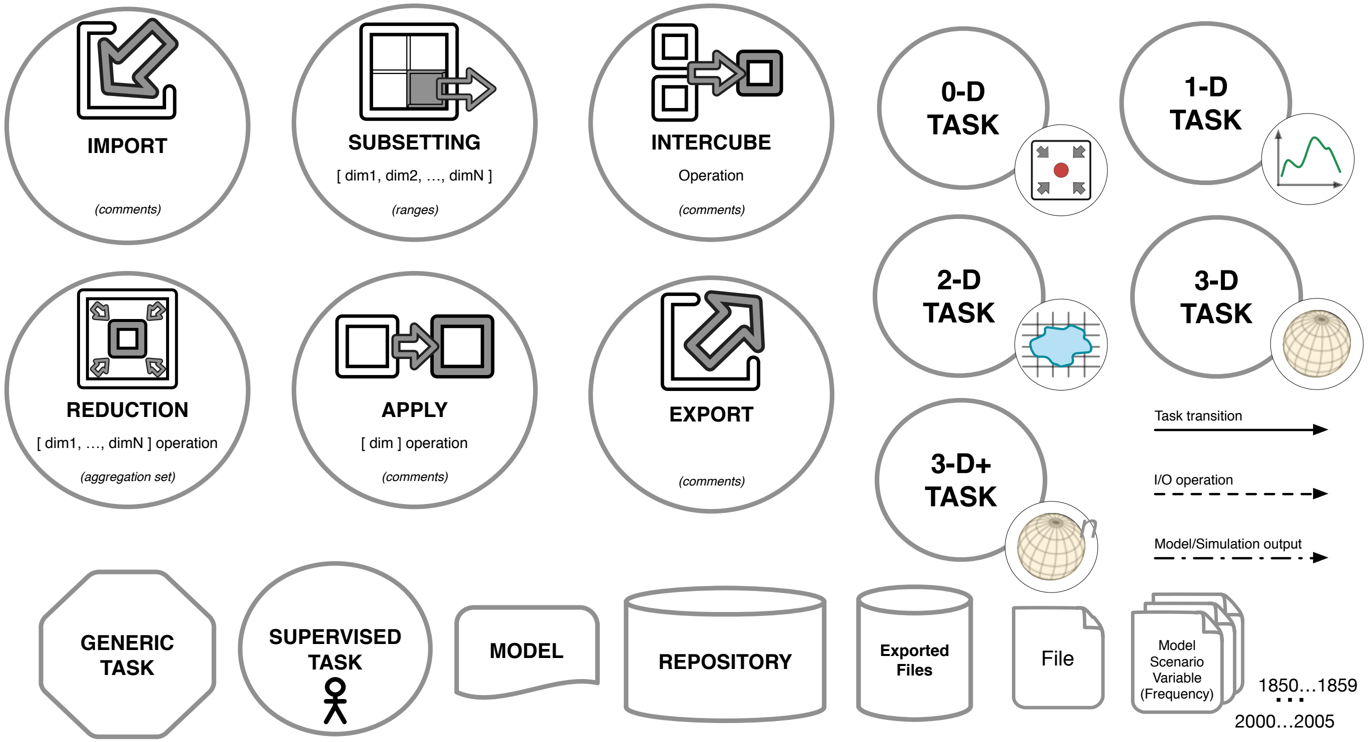
Workflows management, python applications, in-memory analytics

Runtime perspective: managing analytics workflows



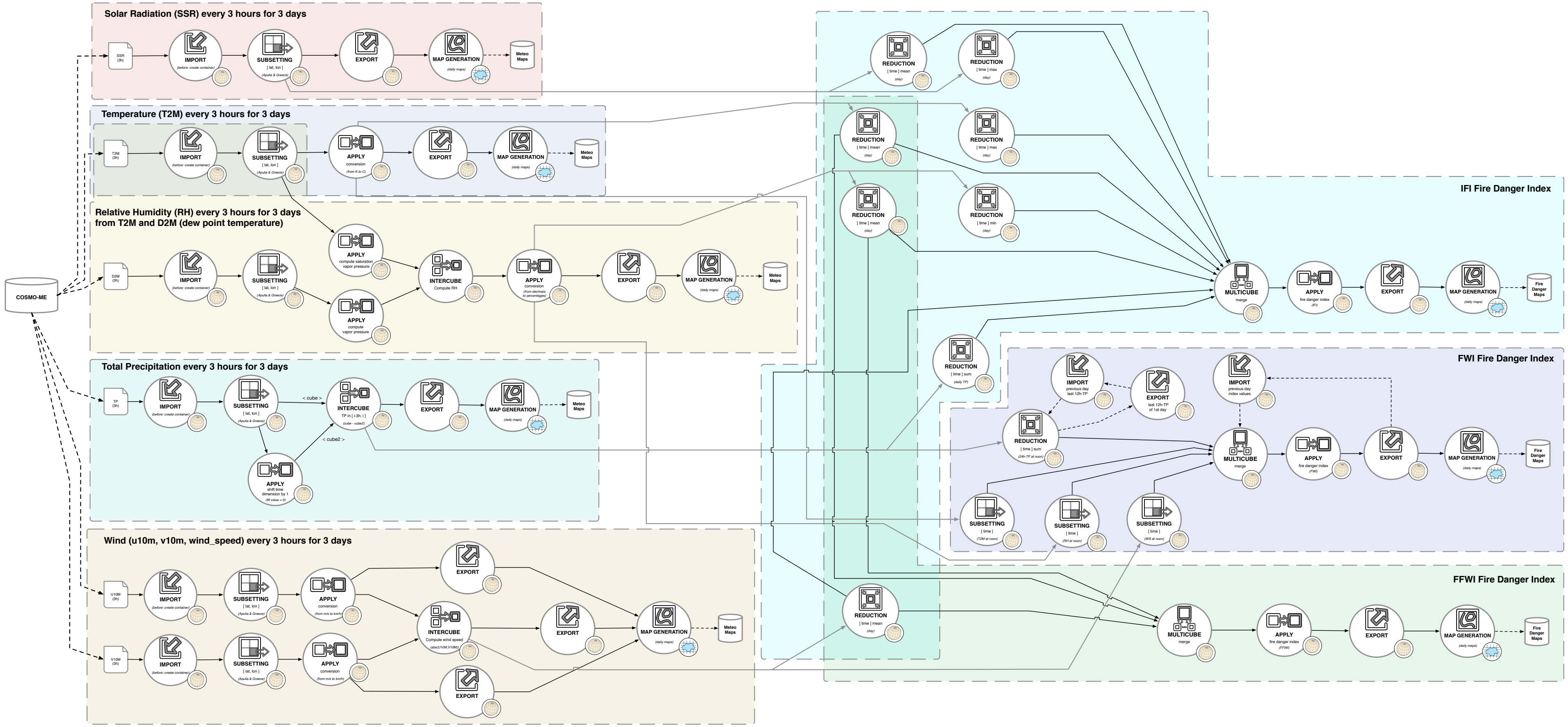
Workflows design

- ✔ A Data Analytics Workflow Modelling Language (DAWML) has been defined
- ✔ Extensible schema jointly defined with application-domain scientists
- ✔ The schema allows the definition of abstract workflows



C. Palazzo, A. Mariello, S. Fiore, A. D'Anca, D. Elia, Dean N. Williams, G. Aloisio, "A workflow-enabled big data analytics software stack for escience", HPCS 2015: 545-552

Workflow design
(case study on fire danger analysis - Italy-Greece pilot area)



See presentation by A. D'Anca (CMCC) tomorrow

Architecture 2.0

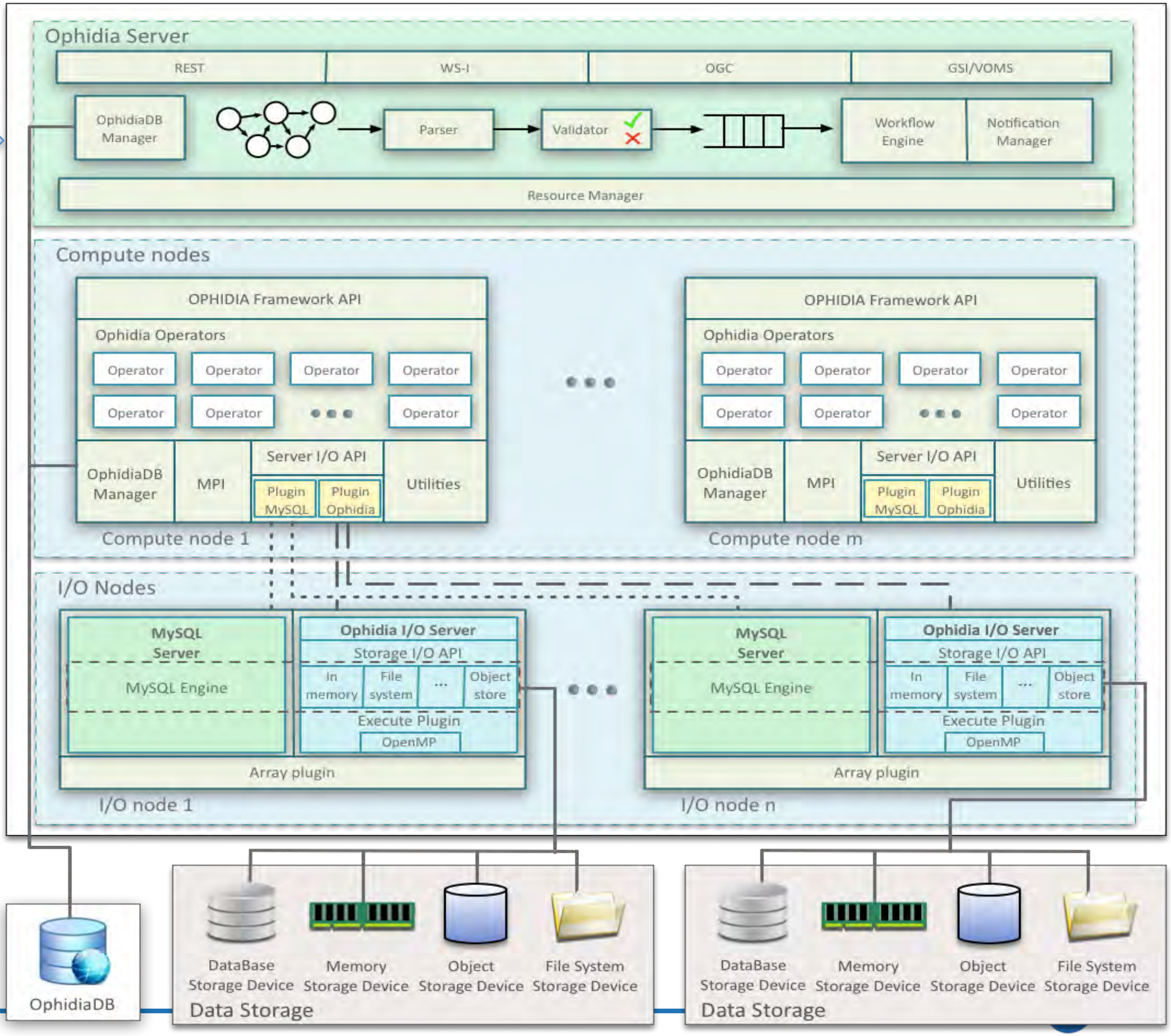
Workflow support on the server side

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



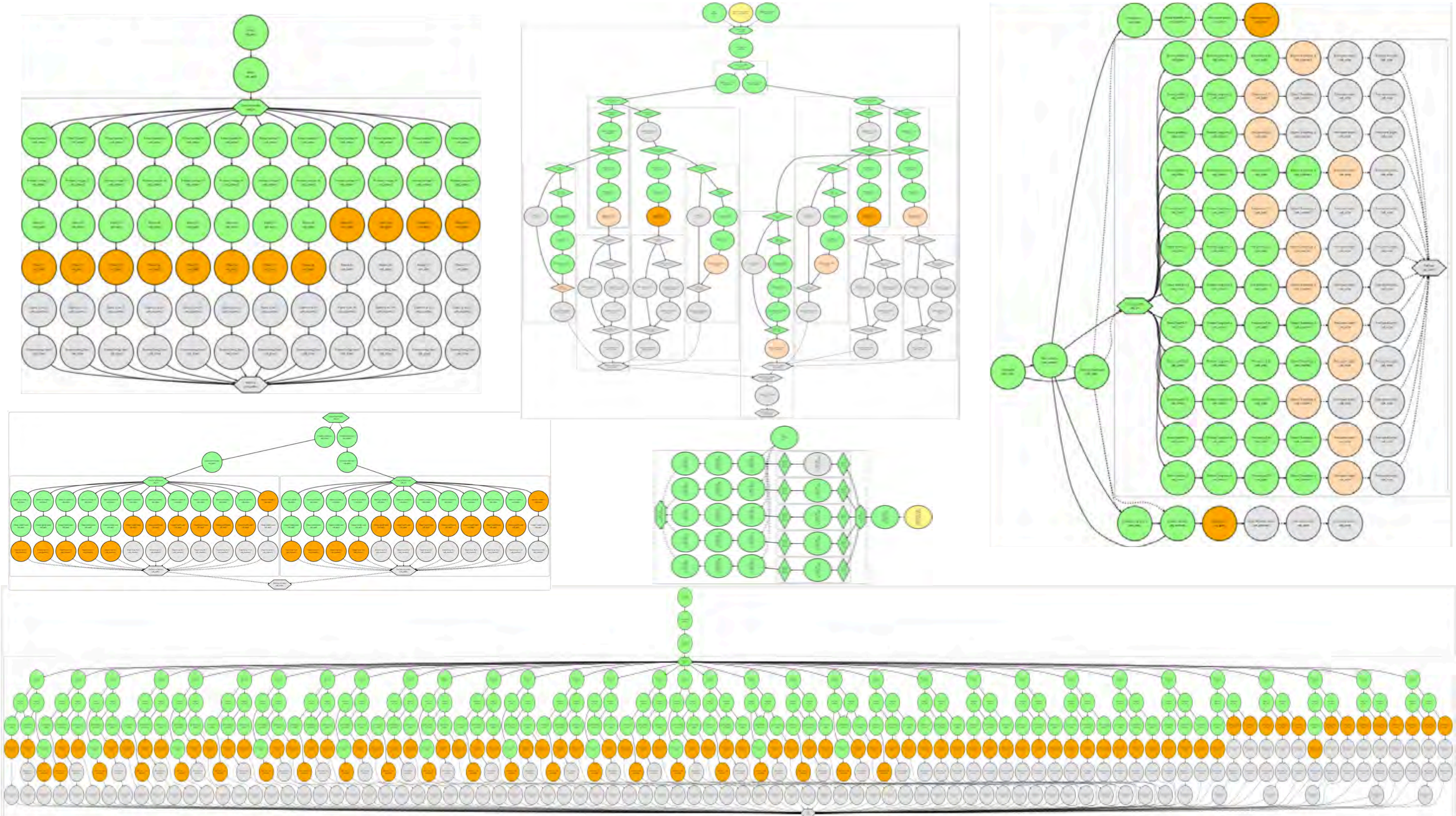
Analytics workflows support and interfaces

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.

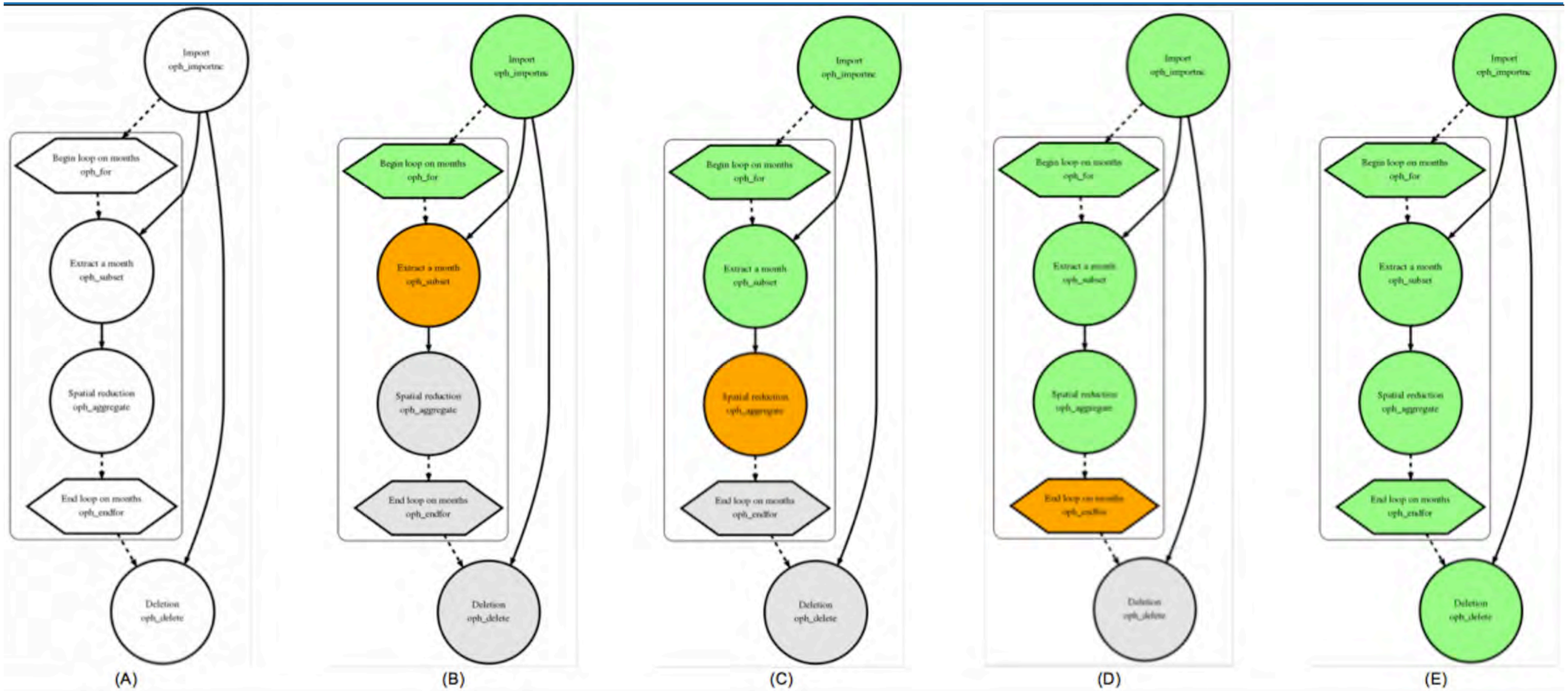
NAME	DESCRIPTION
OPH_ELSE	Start the last sub-block of a selection block "if".
OPH_ELSEIF	Start a new sub-block of a selection block "if".
OPH_ENDFOR	Close a loop "for".
OPH_ENDIF	Close a selection block "if".
OPH_FOR	Implement a loop "for".
OPH_IF	Open a "if" selection block.
OPH_INPUT	It sends commands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

Analytics workflows support and interfaces



and more…

Iterative interface: oph_for



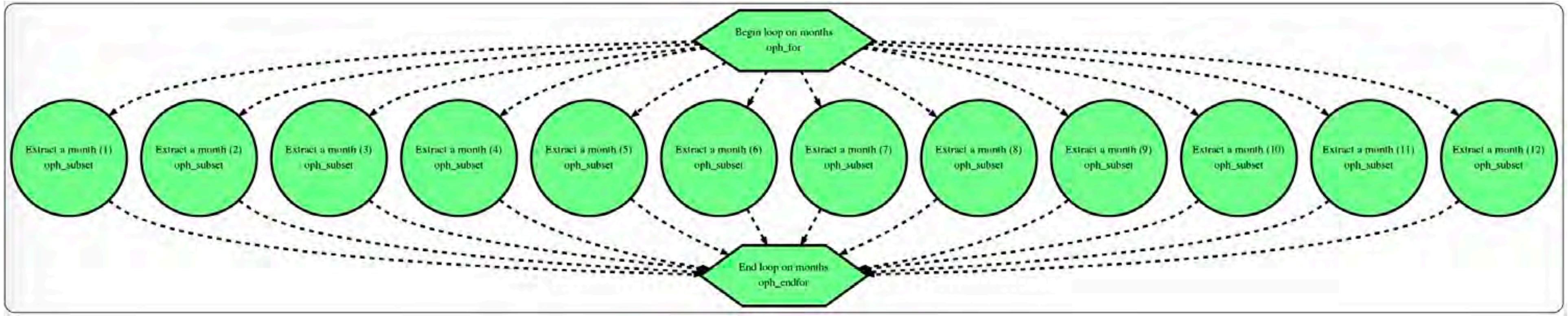
Three steps:

1. An initial task used to import data in Ophidia
2. An iterative block used to evaluate a spatial reduction over data related to the same month
3. A final task used to delete data loaded in the first task.

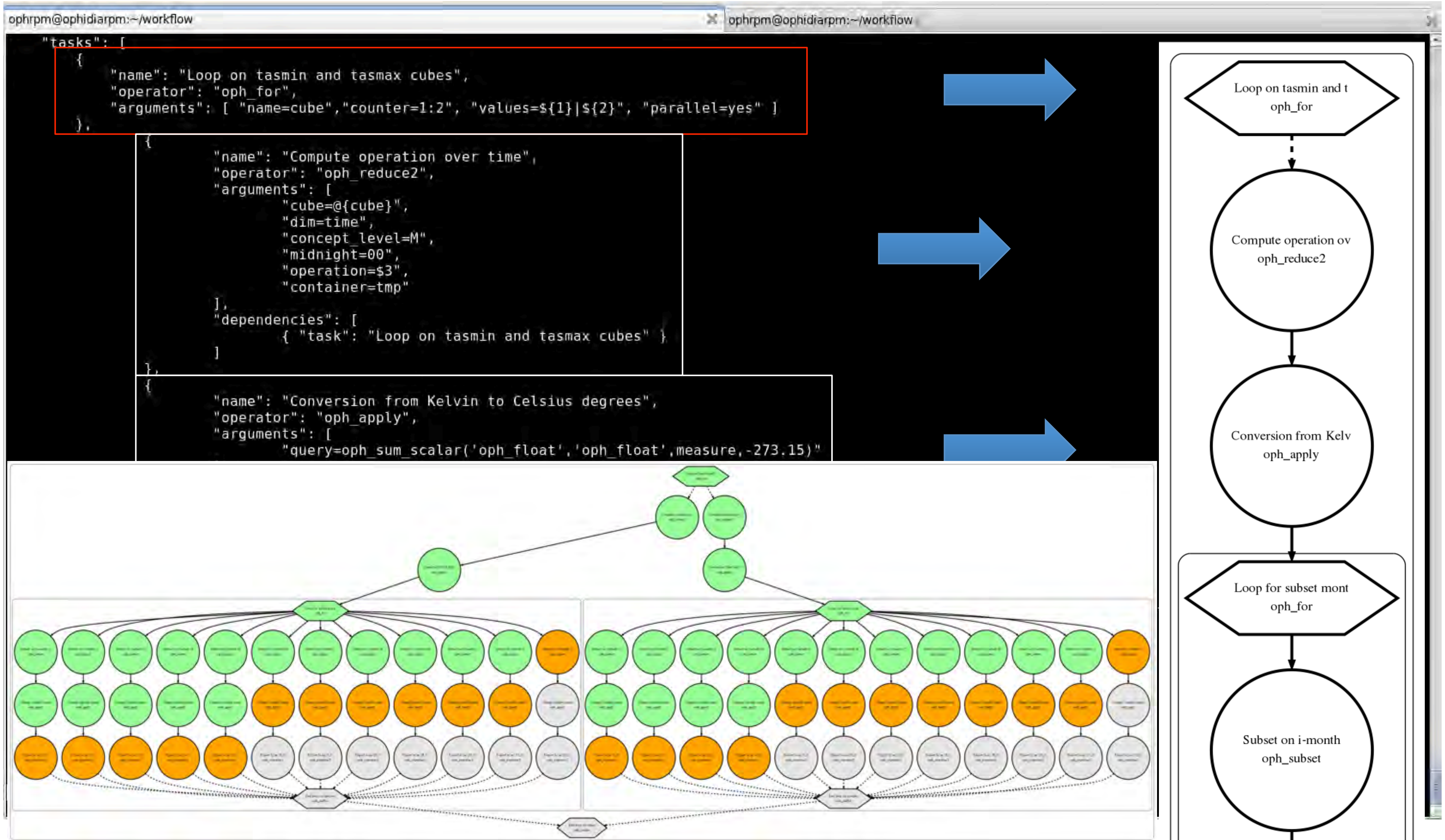
Parallel interface: oph_for, parallel=yes

Operator **oph_for** (*parallel*)

```
{
    "name": "Begin loop on months",
    "operator": "oph_for",
    "arguments":
    [
        "parallel=yes",
        "name=index",
        "counter=1:12",
        "values=Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec"
    ]
}
```



Parallel interface: nested for loops

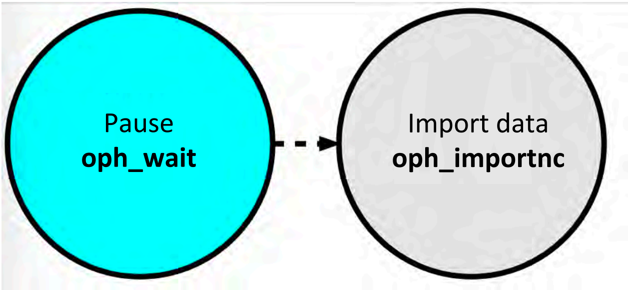


***Youtube video**: https://www.youtube.com/watch?v=PTZkw60YCNU*

Parallel interface: nested for loops (execution)

```
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]

[Response]:
Workflow Status
---------------

OPH_STATUS_COMPLETED

Workflow Progress
-----------------

+===========================+========================+
| NUMBER OF COMPLETED TASKS | TOTAL NUMBER OF TASKS |
+===========================+========================+
| 82                        | 82                     |
+===========================+========================+

Workflow Task List
------------------

+=========================================================+=======================+========+=======+===========+===========================+=====+=============+
| OPH JOB ID                                              | SESSION CODE          | WORKFL | MARKE | PARENT MA | TASK NAME                 | TYP | EXIT STATUS |
|                                                         |                       | OW ID  | R ID  | RKER ID   |                           | E   |             |
+=========================================================+=======================+========+=======+===========+===========================+=====+=============+
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3145  | 3144      | Loop on tasmin and tasmax | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3145              | 1449455166146380      |        |       |           |  cubes                    | PLE | COMPLETED   |
|---------------------------------------------------------|-----------------------|--------|-------|-----------|---------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3146  | 3144      | Compute operation over ti | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3146              | 1449455166146380      |        |       |           | me (1)                    | PLE | COMPLETED   |
|---------------------------------------------------------|-----------------------|--------|-------|-----------|---------------------------|-----|-------------|
| http://193.204.199.174/ophidia/sessions/37669923831130 | 37669923831130223251 | 247    | 3147  | 3144      | Compute operation over ti | SIM | OPH_STATUS_ |
| 2232511449455166146380/experiment?247#3147              | 1449455166146380      |        |       |           | me (2)                    | PLE | COMPLETED   |
|---------------------------------------------------------|-----------------------|--------|-------|-----------|---------------------------|-----|-------------|
| 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   |
|---------------------------------------------------------|-----------------------|--------|-------|-----------|---------------------------|-----|-------------|
```

Operator oph_wait (type=clock)

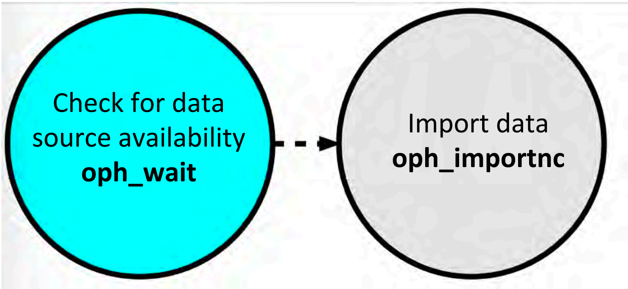


Operator **oph_wait** (*clock*)

Import data

```
{
        "name": "ExampleW1",
        "author": "Foo",
        "abstract": "Simple workflow with a pause",
        "exec_mode": "sync",
        "ncores": "1",
        "tasks":
        [
                {
                        "name": "Pause",
                        "operator": "oph_wait",
                        "arguments": [ "timeout=10" ]
                },
                {
                        "name": "Import data",
                        "operator": "oph_importnc",
                        "arguments":
                        [
                                "measure=measure",
                                "src_path=source.nc"
                        ],
                        "dependencies":
                        [
                                { "task": "Pause" }
                        ]
                }
        ]
}
```

Operator oph_wait (type=file)



Operator **oph_wait** (*file*)

Import data

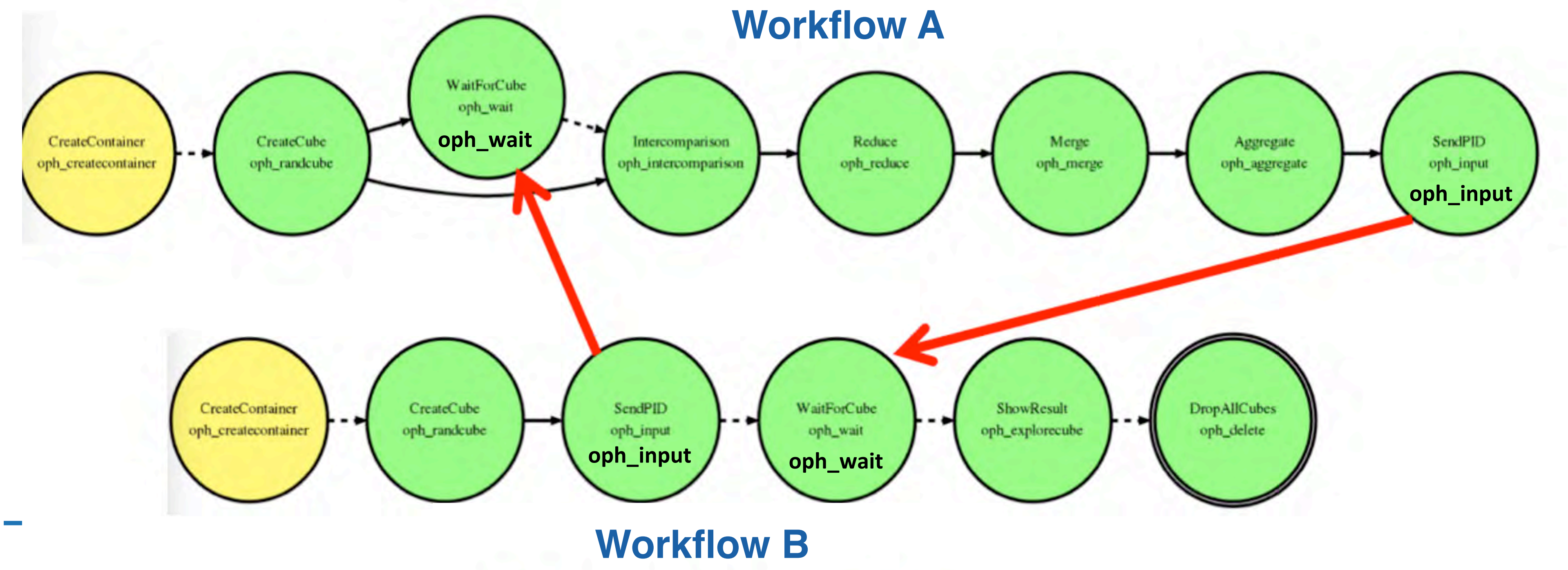
```
{
        "name": "ExampleW2",
        "author": "Foo",
        "abstract": "Simple workflow with source checking",
        "exec_mode": "sync",
        "ncores": "1",
        "tasks":
        [
                {
                        "name": "Check for data source availability",
                        "operator": "oph_wait",
                        "arguments":
                        [
                                "type=file",
                                "filename=http://localhost/source.nc",
                                "timeout=5"
                        ]
                },
                {
                        "name": "Import data",
                        "operator": "oph_importnc",
                        "arguments":
                        [
                                "measure=measure",
                                "src_path=http://localhost/source.nc"
                        ],
                        "dependencies":
                        [
                                { "task": "Check for data source availability" }
                        ]
                }
        ]
}
```

Interactive workflow and inter-wf dependencies
operator oph_wait (type=input) and oph_input

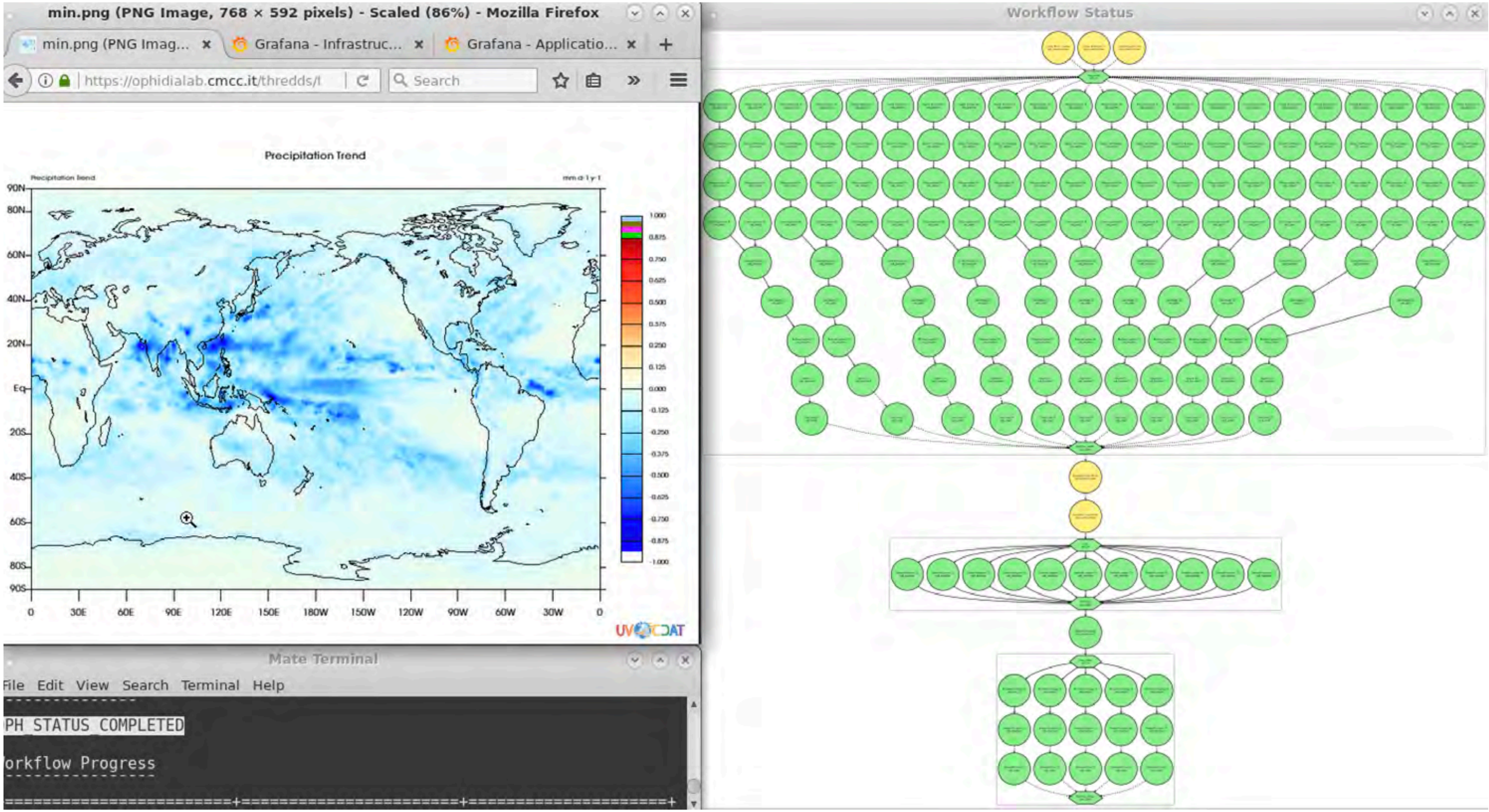
oph_wait and **oph_input** interfaces can be effectively used to setup **interactive** tasks, namely tasks that can be completed only after the user sends additional data

Besides interactive workflows, the operators wait and input can also **enable inter-wf dependencies**

It is also possible that **more than two workflows** interact in this way.



Centralized multi-model analytics experiments (CMIP/ESGF context)



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

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

See presentation by A. D'Anca (CMCC) tomorrow

Distributed analytics experiments
(Two-level workflow strategy)

2016 IEEE International Conference on Big Data (Big Data)

Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the Earth System Grid Federation eco-system

S. Fiore[1], M. Plóciennik[2], C. Doutriaux[3], C. Palazzo[1], J. Boutte[3], T. Żok[2], D. Elia[1], M. Owsiak[2], A. D'Anca[1], Z. Shaheen[3], R. Bruno[4], M Fargetta[4], M. Caballer[5], G. Moltó[5], I. Blanquer[5], R. Barbera[4,6], M. David[7], G. Donvito[4], D. N. Williams[3], V. Anantharaj[8], D. Salomoni[4], and G. Aloisio[1,9]

[1]Euro-Mediterranean Center on Climate Change Foundation (CMCC), Italy
[2]Poznan Supercomputing and Networking Center (PSNC), Poland
[3]Lawrence Livermore National Laboratory (LLNL), California, USA
[4]Italian National Institute of Nuclear Physics (INFN), Italy
[5]Universitat Politècnica de València (UPV), Spain
[6] University of Catania, Italy
[7]Laboratório de Instrumentação e Fisica Experimental de Particulas (LIP), Portugal
[8]Oak Ridge National Laboratory (ORNL), Tennessee, USA
[9]University of Salento, Italy

Abstract—A case study on *climate models intercomparison data analysis* addressing several classes of multi-model experiments is being implemented in the context of the EU H2020 INDIGO-DataCloud project. Such experiments require the availability of large amount of data (multi-terabyte order) related to the output of several climate models simulations as well as the exploitation of scientific data management tools for large-scale data analytics. More specifically, the paper discusses in detail a use case on precipitation trend analysis in terms of requirements, architectural design solution, and infrastructural implementation. The experiment has been tested and validated on CMIP5 datasets, in the context of a large scale distributed testbed across EU and US involving three ESGF sites (LLNL, ORNL, and CMCC) and one central orchestrator site (PSNC).

Keywords-big analytics, workflow management, cloud computing, ESGF, INDIGO-DataCloud.

I. Introduction

The increased models resolution in the development of comprehensive Earth System Models is rapidly leading to very large climate simulations output that pose significant scientific data management challenges in terms of data sharing, processing, analysis, visualization, preservation, curation, and archiving [1-3].

In this domain, large scale global experiments for climate model intercomparison (CMIP) have led to the development of the Earth System Grid Federation (ESGF [4-5]), a federated data infrastructure involving a large set of data providers/modelling centers around the globe, which includes the European contribution - regarding the ENES [6] community – through the IS-ENES project.

From an infrastructural standpoint, ESGF provides a production-level support for search & discovery, browsing and access to climate simulation data and observational data products. ESGF has been serving the Coupled Model Intercomparison Project Phase 5 (CMIP5) experiment, providing access to 2.5PB of data for the Intergovernmental Panel on Climate Change (IPCC) [7] Assessment Reports 5 [8], based on consistent metadata catalogues. More precisely, the Coupled Model Intercomparison Project (CMIP) has been established by the Working Group on Coupled Modelling [9] (WGCM) under the World Climate Research Programme [10] (WCRP).

It provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation and data access. This framework enables a diverse community of scientists to analyse General Circulation Models (GCMs) in a systematic fashion, a process that serves to facilitate models improvement.

CMIP5 has promoted a standard set of model simulations in order to:

- evaluate how realistic the models are in simulating the recent past;
- provide projections of future climate change on two time scales, near term (out to about 2035) and long term (out to 2100 and beyond); and
- understand some of the factors responsible for differences in model projections, including quantifying some key feedbacks such as those involving clouds and the carbon cycle.

In such a context, running a multi-model data analysis experiment is very challenging, as it requires the availability of large amount of data (multi-terabyte order) related to multiple climate models simulations as well as scientific data management tools for large-scale data analytics.

The remainder of this work is organized as it follows. Section II provides the current workflow for the multi-model climate data analysis in the CMIP context, whereas Section III presents the paradigm shift needed to address such large-

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

- *A first experiment across sites was demonstrated at the INDIGO Review, November 2016 in Bologna*
- *Strong synergy with the ESGF Compute Working Team*
- *International collaboration across the Atlantic*

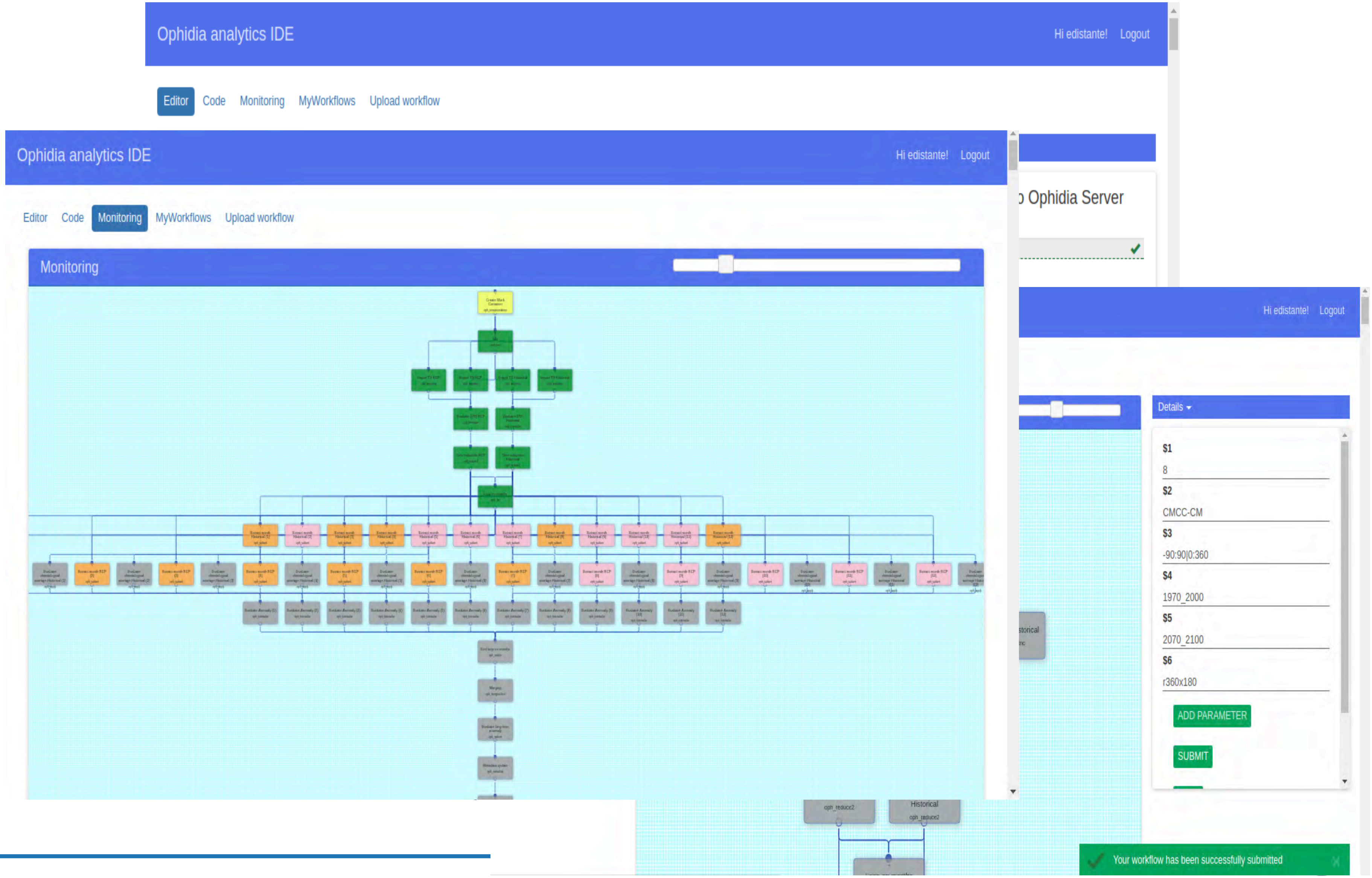
S. Fiore, M. Plóciennik, et al.: Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the Earth System Grid Federation eco-system. BigData 2016: 2911-2918

See next presentation by M. Plociennik (PSNC)

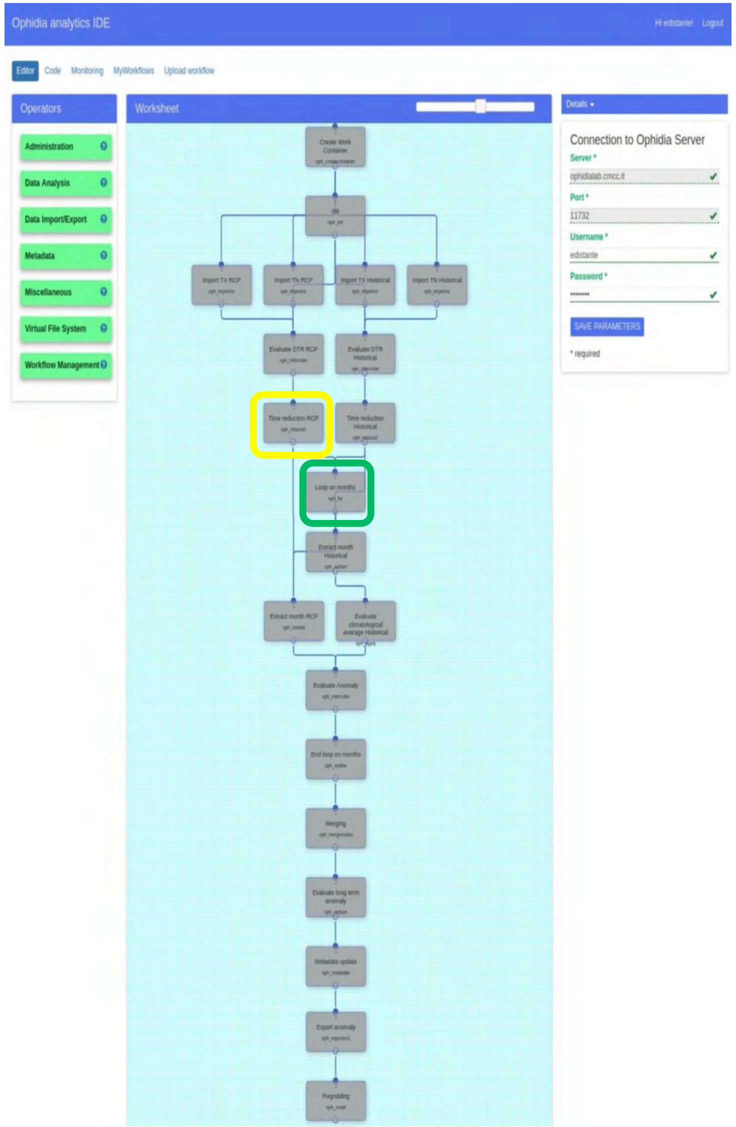
Looking forward

Workflow IDE

The analytics workflow IDE



Easy and automated generation of JSON code




```
{
        "name": "dtr anomaly",
        "author": "CMCC",
        "abstract": "This workflow computes the anomaly of DTR (Diurnal Temperature Range) index with respect to past values. It works on two input files (tasmin/tasmax variable): $1 is ncores, $2 is the model, $3 is spatial filter (lat|lon ranges), $4 is the first time filter (historical), $5 is the second time filter (scenario), $6 is the grid of output map using the format r<lon>x<lat> (e.g. r360x180), i.e. a global regular lon/lat grid (this parameter is optional and by default the lon/lat grid of input file is adopted).",
        "exec_mode": "sync",
        "cwd": "/",
        "ncores": "${1}",
        "on_exit": "oph_delete",
        "tasks":
        [
                {
                        "name": "Init",
                        "operator": "oph_set",
                        "arguments": [
                                "key=missingvalue",
                                "value=1.e+20"
                        ]
                },
                {
                        "name": "Create Work Contai
                        "operator": "oph_createcont
                        "arguments": [
                                "container=dtr",
                                "dim=lat|lon|time",
                                "dim_type=double|do
                                "hierarchy=oph_base
                                "compressed=no",
                                "ncores=1",
                                "base_time=1850-01-
                                "calendar=standard"
                                "units=s"
                        ],
                        "on_error": "skip"
                },
                {
                        "name": "Import TX RCP",
                        "operator": "oph_importnc",
                        "arguments":
                        [
                                "src_path=/data/cmi
                                "measure=tasmax",
                                "base_time=1850-01-
                                "units=d",
```

OPH_REDUCE2

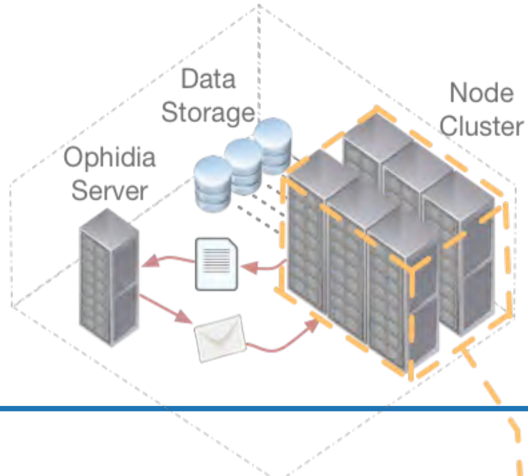
```
        "name": "Time reduction RCP",
        "operator": "oph_reduce2",
        "arguments":
        [
                "operation=avg",
                "dim=time",
                "concept_level=M",
                "description=Monthly Averages RCP",
                "missingvalue=@{missingvalue}"
        ],
        "dependencies": [
                { "task": "Evaluate DTR RCP", "type": "single" }
        ]
},
{
        "name": "Time reduction Historical",
        "operator": "oph_reduce2",
        "arguments":
        [
                "operation=avg",
                "dim=time",
                "concept_level=M",
                "description=Monthly Averages Historical",
                "missingvalue=@{missingvalue}"
        ],
        "dependencies": [
                { "task": "Evaluate DTR Historical", "type": "single" }
        ]
},
{
        "name": "Loop on months",
        "operator": "oph_for",
        "arguments": [
                "key=month",
                "counter=1:12",
                "parallel=yes"
        ],
        "dependencies": [
                { "task": "Time reduction RCP" },
                { "task": "Time reduction Historical" }
        ]
},
```

OPH_FOR

Deployment of Ophidia on HPC clusters and path towards large scale workflows

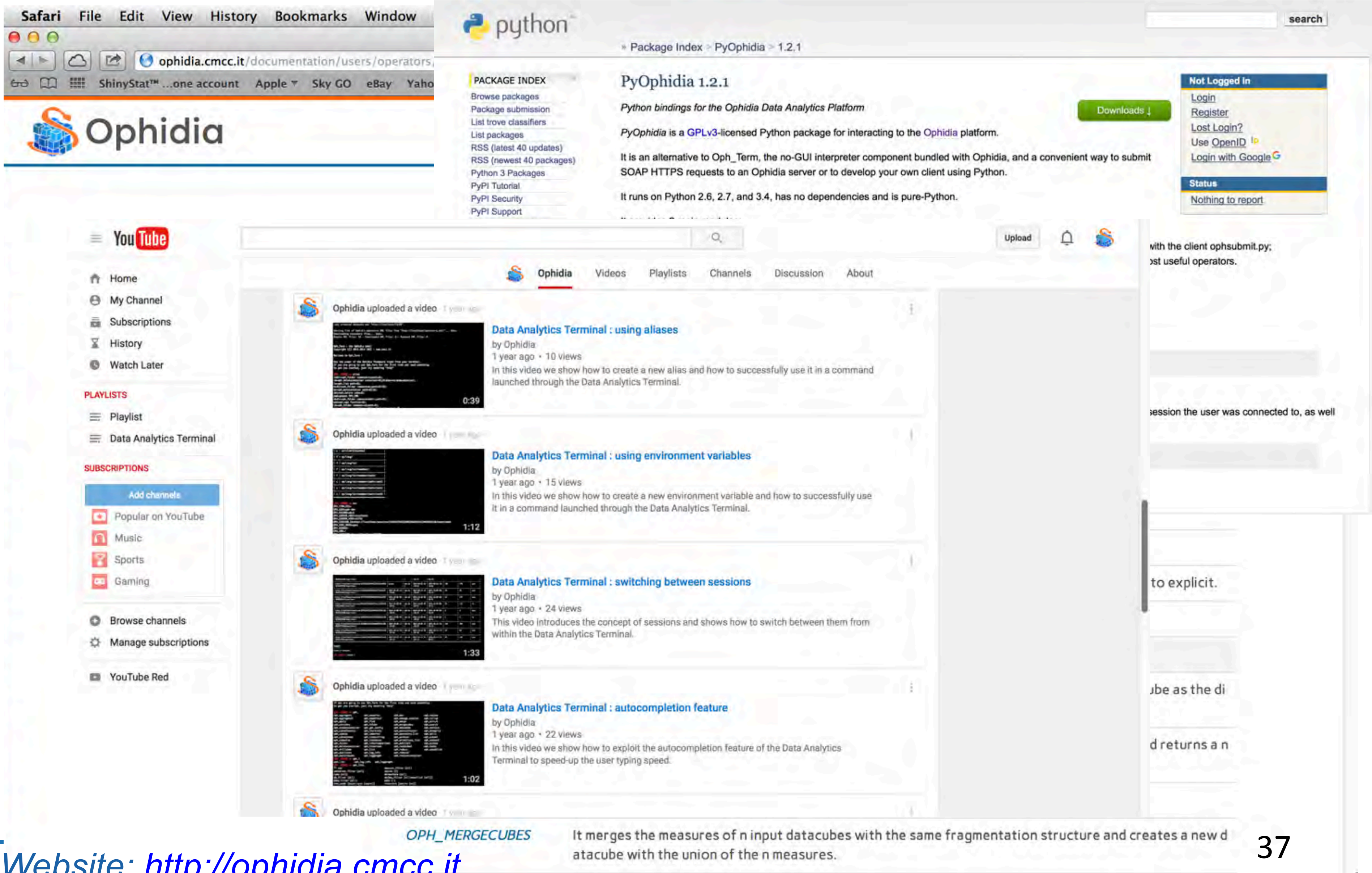
So far we had a dedicated cluster hosting Ophidia permanently for data analytics purposes/experiments.

- ✔ *Our latest development relates to **automated deploy** of Ophidia on the Athena HPC Cluster at the CMCC SuperComputing Centr*
- ✔ *A user can submit a **deploy task** on the cluster and get his/her own instance running*
- ✔ *From the Ophidia terminal he/she can start running an analytics session, **run workflows**, produce output and store them persistently on the storage*
- ✔ *After that h/shee can undeploy the cluster with a separate **undeploy task** and release the resources*
- ✔ *Such feature is still in **alpha stage** and we'll released once further tested in production by our scientists*
- ✔ *Ongoing: to replace the available software components with containers*



Useful resources and final remarks

Ophidia documentation and social/multimedia content



Website: http://ophidia.cmcc.it

Conclusions

- ✔ ***ECAS*** *represents the* ***community evolution*** *of Ophidia*
  - ✔ *One of the thematic service in the context of the* ***EOSC-hub***
- ✔ ***OLAP approach*** *for big data – multidimensional data model*
- ✔ *It provides access via* ***CLI*** *(end-users) and* ***API*** *(devel users)*
- ✔ *Programmatic access via* ***C*** *and* ***Python APIs***
- ✔ ***Workflow support*** *to build very complex analysis experiments*
  - ✔ *Design language*
  - ✔ *Conditional and loop statements*
  - ✔ *Parallelism support*
  - ✔ *Interactive & batch*
  - ✔ *Inter-wf dependencies*
  - ✔ *CLI and web- based tools*

Publications

[11] 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

[10] 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.

[9] 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.

[8] M. Plociennik, S. Fiore, G. Donvito, M. Owsiak, M. Fargetta, R. Barbera, R. Bruno, E. Giorgio, D. N. Williams, and G. Aloisio, "Two-level Dynamic Workflow Orchestration in the INDIGO DataCloud for Large-scale, Climate Change Data Analytics Experiments", International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA. Procedia Computer Science, vol. 80, 2016, pp. 722-733

[7] 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

[6] 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

[5] S. Fiore, M. Mancini, D. Elia, P. Nassisi, F. V. Brasileiro, I. Blanquer, I. A. A. Rufino, A.C. Seijmonsbergen, C. O. Galvao, V. P. Canhos, A. Mariello, C. Palazzo, A. Nuzzo, A. D'Anca, G. Aloisio, "Big data analytics for climate change and biodiversity in the EUBrazilCC federated cloud infrastructure", Workshop on Analytics Platforms for the Cloud, In Proceedings of the 12th ACM International Conference on Computing Frontiers (CF '15), May 18th, 2015, Ischia, Italy. Article 52, 8 pages.

[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).

Thanks

Do you want to join?

ECAS/Ophidia is an open source effort

Feel free to get in touch with us

sandro.fiore@cmcc.it

http://ophidia.cmcc.it @OphidiaBigData www.youtube.com/user/OphidiaBigData