Definition of Support to Research Communities in INDIGO-DataCloud

RIA-653549

Jesus Marco de Lucas (IFCA-CSIC)
marco@ifca.unican.es

EMBRC e-infrastructure working group meeting

Paris, 14th December 2016

INDIGO-DataCloud is co-founded by the Horizon 2020Framework Programme

INDIGO-DataCloud

(**IN**tegrating **D**istributed data **I**nfrastructures for **G**lobal Expl**O**itation)

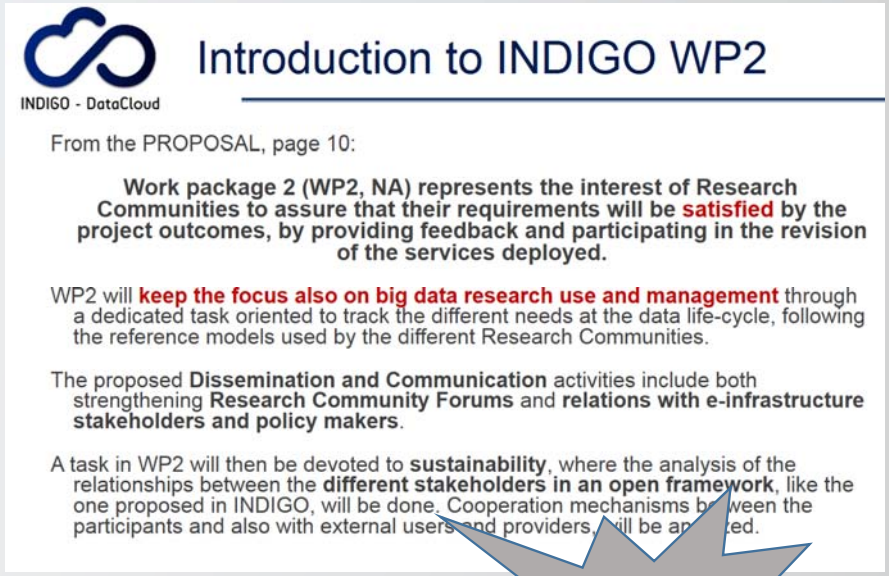
INDIGO - DataCloud

- **An H2020 project** approved in the **EINFRA-1-2014** call
  - 11.1M€, 30 months (**from April 2015 to September 2017**)
- **Who**: **26 European partners** in 11 European countries
  - Coordination by INFN (Italian National Inst. for Nuclear Physics)
  - Including developers of distributed software, industrial partners, research institutes, universities, e-infrastructures
- **What**: **develop an open source Cloud platform** for *computing and data* ("DataCloud"), tailored to science.
- **Where**: deployable on **hybrid (public or private) Cloud infrastructures**
- **For**: **multi-disciplinary scientific communities**
  - E.g. structural biology, earth science, physics, bioinformatics, cultural heritage, astrophysics, life science, climatology.
- **Why**: to answer to the technological **needs of scientists** seeking to easily and efficiently exploit distributed compute and data resources.

"Taking into account Research Requirements through Case Studies"

INDIGO - DataCloud

- Objectives and Activities
- Presentation of Research Communities
- Gathering requirements
- On Data Management
- Integration of INDIGO solutions
- **Dissemination**, Training
- **Towards Exploitation**
- WP2 deliverables, milestones
- Planning



Kick-off meeting

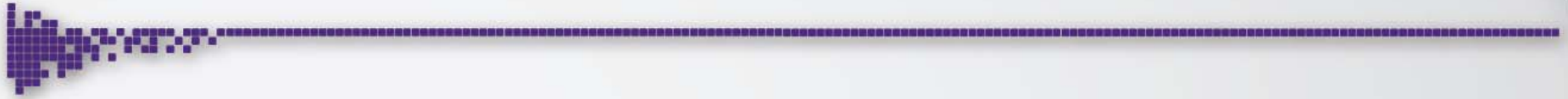
Objectives and Activities

INDIGO - DataCloud

Define the support required by Research Communities and to **test and validate the state-of-the-art services developed by INDIGO** to ensure that they will result in an increased use of production e-infrastructures in Europe, and in particular through the enhancement of services to **share, manage and process research data**.

- T2.1 Research Communities Requirements (lead by EGI.eu, UPV)
- T2.2 Defining support to Research Data (lead by INGV, CSIC)
- T2.3 Application Test and Validation (lead by U.Utrecht, CNR)
- T2.4 Dissemination towards Research Communities (lead by RBI, EGI.eu)
- T2.5 Sustainability: exploitation strategy… (lead by CSIC, EGI.eu)

Research Communities in INDIGO



SIMPLIFIED IMPACT TABLE SELECTED OBJECTIVES versus REQUESTS/ POTENTIAL IMPACT FOR COMMUNITIES O1: Development of the INDIGO Platform based on open software without restrictions on the e-Infrastructure	Life Sciences	Physical Sciences & Astronomy	Social Sciences & Humanities	Environmental Sciences
Research Communities & Initiatives , including ESFRIs	ELIXIR INSTRUCT/ WeNMR EuroBioImaging	CTA LBT WLCG	DARIAH DCH-RP	EMSO LIFEWATCH ENES
Examples of Applications	HADDOCK GROMACS AMBER GALAXY	MIDAS, IRAF, IDL, Geant4 ROOT/PROOF Geant4	Fedora Digital Libraries	Delft3D R-Studio TRUFA MATLAB
Design and development of a Platform providing advanced users and community developers a powerful and modern environment for development work. This includes programming and scripting tools, and composition of custom applications and software deployment	RELEVANT	**CRITICAL**	RELEVANT	**CRITICAL**
Developing a framework to enable the transparent execution on remote e-infrastructures of existing popular applications like MATLAB / OCTAVE, ROOT, MATHEMATICA, or R-STUDIO.	RELEVANT	**CRITICAL**	MINOR	**CRITICAL**
Provide the services and tools needed to enable a secure composition of services from multiple providers in support of scientific applications.	**CRITICAL**	**CRITICAL**	RELEVANT	RELEVANT
Develop and implement a solution that is able to deploy in a transparent	**CRITICAL**	RELEVANT	MINOR	RELEVANT

Gathering requirements

THIS WAS THE KEY TASK ALONG FIRST MONTHS, after KICK-OFF MEETING

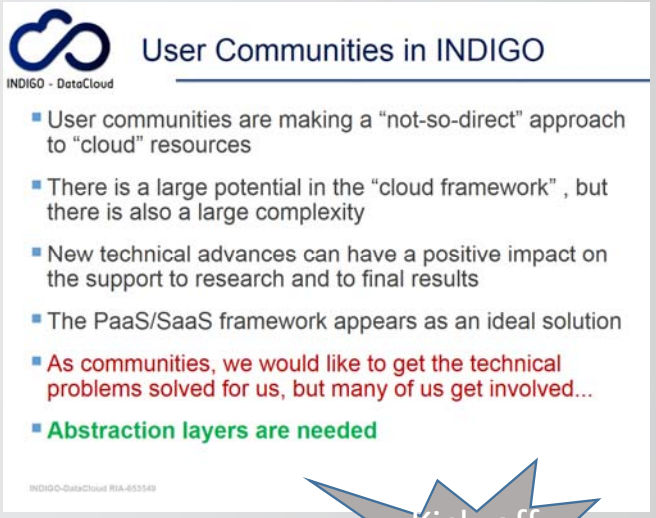
A basic problem going into the Cloud framework (impedance mismatch)

Communication Researchers – Developers

The components in the solution:

- Structured working documentation (D2.1, D2.4, finally D2.10)
- Agile approach: Case Studies based on User Stories
- Common supporting tools (OpenProject)
- Roles, including person in the middle!
- Champions!
- Good communication: teleconf (bi-weekly) + All Hand Meetings

(with WP3,JRA teams) (Valencia, Bari, Madrid, Amsterdam, Frascati, Catania, Krakow, Bologna)



Kick-off meeting

Case Studies

# Partner Research Community	Case Study/Application
P0 CSIC LifeWatch	**Monitoring and Modelling Algae Bloom** in a Water Reservoir Support of hydrodynamic and water quality modelling including data input-output management and visualization.
	TRUFA (Transcriptomes User-Friendly Analysis)
P1 UPV EuroBioImaging	**Medical Imaging Biobanks** The virtual Biobank integrates medical images from different sources and formats
P2 CIRMMP INSTRUCT	**Molecular dynamics simulations** Support of Molecular Dynamics simulations of macromolecules that need specific requirements in terms of computing (e.g. GPGPUs).
P3a INAF, LBT P3b INAF,LBT	**Astronomical Data Archives** Data management and analysis using different tools such as data discovery, comparison, cross matching, data mining and also workflows.
	Archive System for the Cherenkov Telescope Array (CTA) Data management, treatment and flow of data, big data archiving and processing, open data access.
P4 U. Utrecht WeNMR	**HADDOCK portal**
	DisVis allows exploring the accessible conformational space of the complex between two biomolecules defined by a few experimentally measured distances. DisVis runs in either multi-CPU mode or making use of GPGPU resources.
	PowerFit finding the optimal placement of a biomolecule into a cryo-electron miscroscopy density map by exhaustive search.
P5 CMCC ENES	**Climate models inter comparison data analysis** Linked to the Coupled Model Inter comparison Project (CMIP).

Case Studies (cont'd)

# Partner Research Community	Case Study/Application	
P6 ICCU Galleries, Libraries, Archives	**eCulture science Gateway** Digital repository collection support that allows users to upload, download digital documents and manage metadata.	
P7 EGI.eu FedCloud Community	**Chipster**	**BILS**
	READemption	**Human Brain Project**
	JAMS	**BBMRI-ERIC CC**
	HAPPI	**DARIAH CC**
	INERTIA	**EPOS CC**
	DRIHM	**Disaster Mitigation**
P8 CNR ELIXIR-ITA	**Galaxy as a Cloud service** Development a fully customizable Galaxy instance provider platform.	
P9 INGV EMSO	**MOIST- multidisciplinary oceanic information system** Data collected by the NEMO-SN1 observatory, one of the EMSO nodes used for geohazard monitoring, in proximity of Etna volcano.	
P10 RBI DARIAH	**Data Repository Platform for DARIAH** Strengthening the Use of Scientific Distributed Computing in the Arts and Humanities	

Methodology for gathering requirements

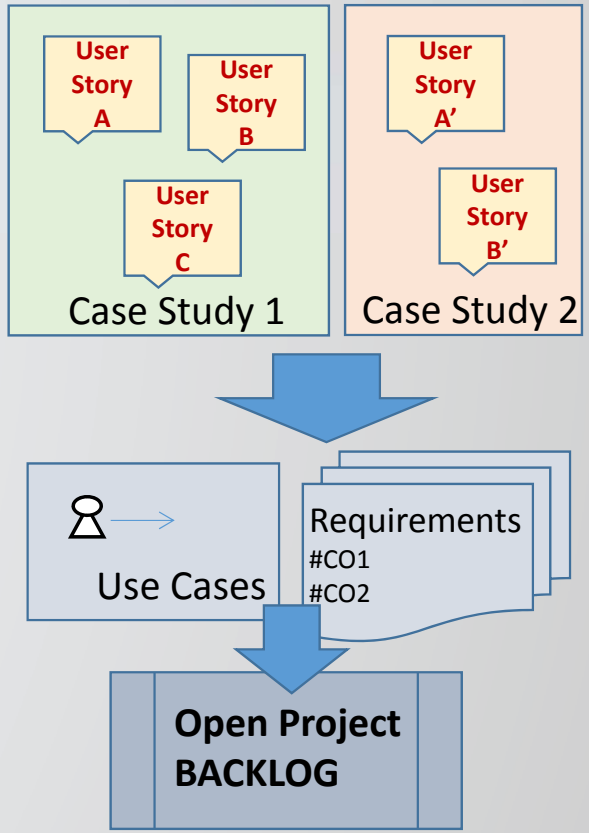
highly criticized → and refined → in several iterations

A template was **designed** to gather information from communities

- Based on **Case Studies**

  A Case Study is an implementation of a research method involving an up-close, in-depth, and detailed examination of a subject of study (the Case), as well as its related contextual conditions.

- Focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-Data Cloud solutions on it.
- A Case Study is (ideally) based on **a set of User Stories**, i.e. how the researcher describes the steps to solve each part of the problem addressed.
- User Stories are the starting point of **Use Cases**, where they are transformed into a description using software engineering terms (like the actors, scenario, preconditions, etc).
- Use Cases are useful to capture the **Requirements** that will be handled by the INDIGO software developed in JRA workpackages, and tracked by the Backlog system from the OpenProject tool.
- The template serves as a structured framework with guiding questions concerned by INDIGO development workpackages.



Methodology (cont'd)

INDIGO - DataCloud

Next: Analysis of the Annexes to identify requirements

- Performed by T2.1
- Produced large table, several entries per Case Study
  - Community, Req#, Req. Descr., Rank (Mandatory/Convenient/Optional), Current, Gaps, Solution…)

Community	Req #	Requirement	Requirement Type (Computing / Storage / PaaS	Rank (Mandatory / Convenient / Optional)	Current workflow/solution	Gaps	Proposed improvement	Potential solution for INDIGO (User community point of view)	Potential solution for INDIGO (JRA point of view)	Comments
ENES - CMCC	ENES#7	Isolation of deployments	Computing	Convenient	Currently users share the infrastructure.	Unavailable feature	Need for minimizing side-effects and Ophidia deployments are tailored to the reference data.	Deployment on containers and VMs provides the isolation.		See ENES#1, ENES#8
	ENES#8	Execution across multiple centres.	Computing / PaaS Service	Mandatory	Not provided	Unavailable feature	Interesting when exhausting resource capabilities of one deployment or when combining the processing of different data sets that are deployed on different Data Analytics infrastructures.	Task T5.3 in INDIGO deals with the geographic scheduling of workloads, however, this may not be sufficient given the interactive nature of the process. Surely changes are needed at application level and coherent global authz management could help. Metascheduling.		
		to reduce time-to-			Based on data download	Server-side approach not	It should be easy to deploy a self-configurable and auto-scalable Data	Combination of TOSCA specification, software		See ENES#1, ENES#2,

Next: identification of common requirements

- Produced single table

Methodology (cont'd)

INDIGO - DataCloud

Req #	Requirement	Requirement	Rank	Proposed improvement	Potential solution for INDIGO (User	Pot	EuB	Lif	ELI	Had	CIR	Fe	DA	INA	CMC	CT	AL	ING
CO#1	Deployment of Interface SaaS	Computing / PaaS	Mandatory	A mechanism to facilitate the deployment of a customised Haddock	The portal could be instantiated by means of a set of containers and/or specific base		M	C	M	M	M			C	C	M		
CO#2	Deployment of Customized computing back-ends as batch queues	Computing / PaaS	Mandatory	Each instance may have an independent software configuration,	A devops tool integrated with the deployment service to install and configure		M	M	M	M	M	C		C	M	C	M	
CO#3	Deployment of user-specific software	Computing / PaaS	Mandatory	Manual installation may be cumbersome for large-scale application involving many computing	Ability of a user to easily construct a software installation and configuration specification (e.g. TOSCA) for their own		M		M						C			
CO#4	Automatic elasticity of computing batch queues	Computing / PaaS	Mandatory	When moving to the cloud, users should be provided with the exact	Monitoring services may be integrated with the deployment, which will trigger the		M	M	M	M	M			C	M		M	
CO#5	Terminal access to the resources.	Computing / PaaS service	Mandatory	This feature must be linked to the AAI	This will require ssh ports to be open and direct access to the VMs. The massive		M		M	M	M				M			
CO#6	Privileged access	Computing / PaaS service	Mandatory	This feature must be linked to the AAI	A single special user in the "sudo" group.		C		M	M	M				M			
CO#7	Execution of workflows	Computing / PaaS	Mandatory	Processing done on the cloud where the outputs of the processing are	Workflow engine can be deployed as any other application. Back-end could be a		M		C	C		O			M	M		
CO#8	Provenance information	Computing / PaaS Service	Convenient	Very important for revision of papers and project proposals.	Repository of data and software that could be deployed or inspected on demand.		C											
CO#9	Cloud bursting	Computing / PaaS Service	Mandatory	Supplementing the computing capacity with special instances	Automatic contextualization and configuration will enhance the		C	C	C	M					M			
CO#10	Data-aware scheduling	Computing / PaaS Service	Convenient	Currently storage and computing are highly coupled.	This will affect the scheduling. Moving computing to data. Maybe the use of				C			C			M			
CO#11	Provisioning of efficient Big Data Analysis solutions exploiting server-side and declarative approaches	Computing / Storage / PaaS Service	Mandatory		Currently it uses a hierarchical set of databases that are coordinated through distributed memory parallel computing										M			
CO#12	Execution across multiple centres.	Computing / PaaS Service	Mandatory	Interesting when exhausting resource capabilities of one deployment or when	Task T5.3 in INDIGO deals with the geographic scheduling of workloads,										M			
CO#13	On-line processing of data	PaaS	Mandatory	Special management of post-processing jobs that could be sent to	Despite that this may look similar to any other processing, two aspects need to be		C	M			M			M	M		M	C
CO#14	Special hw configuration - MPI, multicore, GPGPU	Compute / PaaS	Mandatory	More flexibility in the way the requirements are defined and the	Three main issues must be analysed here (not all for the User Cases selected)ç: 1) The		C	C			M				M			

Req #	Requirement	Requirement	Rank	Proposed improvement	Potential solution for INDIGO (User community	Pot	EuB	Lif	ELI	Had	CIR	Fe	DA	INA	CMC	CT	AL	ING
SO#1	Shared storage accessible like a POSIX filesystem	Storage / PaaS Service	Mandatory	Limited storage and no scalability	Data volumes that can be mounted (R/W) on multiple VMs using an efficient protocol. Block-based storage will offer a		M	M	M	M				M			M	
SO#2	Persistent data storage	Storage	Mandatory		Disk storage in the VMs must be persistent even if the VM is undeployed, and only removed if explicitly requested.		M	M	M		M		M	M	M	M	M	M
SO#3	Long-term availability of results	Storage	Mandatory	External, long-term, self-maintained storage.	Interoperability with other infrastructures.		C				M		M					
SO#4	Local user storage	Storage / PaaS Service	Mandatory	Separate individual volumes will increase scalability and privacy.	Individual storages deployed as R&W volumes.		M	M	M					M		C		
SO#5	Availability of reference data	Storage / PaaS Service	Mandatory		A shared, read-only volume should be available with all the reference data.				M	M	M		M	M		M		
SO#6	Interoperability with IS-ENES/ESGF	Storage / PaaS Service	Mandatory	No improvement, keeping this feature	Basic data access functionality through ESGF protocols (HTTP, OPeNDAP) associated to metadata catalogues (Thredds); User authentication based on OpenID federation; and Solr search and discovery service.										M			
SO#7	Metadata management / Database as a Service	Storage / PaaS Service	Convenient		Metadata services as part of the storage services		C							C	C	C	M	M
SO#8	Share data capabilities	Storage/IaaS service	Convenient	Block storage with added NFS-like capability of multiple access.	One Data Storage solution							M			C			C
SO#9	Data replication	PaaS	Mandatory	Hide the data topology to the user, data federation, data replication capabilities	OneData used to federate community repositories, and allow an easy access to the datasets, and to replicate the data where necessary, based on community parameters.							M		M				
SO#10	Distributed storage	Storage / PaaS service	Mandatory	Cloud or grid based solutions have not proven to be efficient yet.	Cloud back-end will facilitate the deployment on a wider range of infrastructures.								M	M	M	M		M
SO#11	Dropbox-like storage	Storage / PaaS service	Convenient	Facilitate interaction with users in uploading and downloading files	Client tools for accessing storage from desktop systems.			C						C			M	

Req #	Requirement	Requirement	Rank	Proposed improvement	Potential solution for INDIGO (User community point of	Pot	EuB	Lif	ELI	Had	CIR	Fe	DA	INA	CMC	CT	AL	ING
SO#9	Data replication	PaaS	Mandatory	Hide the data topology to the user, data federation, data replication capabilities	OneData used to federate community repositories, and allow an easy access to the datasets, and to replicate the data where necessary, based on community parameters.							M		M				
SO#10	Distributed storage	Storage / PaaS service	Mandatory	Cloud or grid based solutions have not proven to be efficient yet.	Cloud back-end will facilitate the deployment on a wider range of infrastructures.								M	M	M	M		M
SO#11	Dropbox-like storage	Storage / PaaS service	Convenient	Facilitate interaction with users in uploading and downloading files	Client tools for accessing storage from desktop systems.			C						C			M	
PL#1	Global-level AAI	PaaS	Mandatory	Centralized mechanism to define general authorisation policies will give scalability and a coherent mechanism.	Use a repository of credentials and authorisation tokens that could provide a coherent global mechanism. Use of a centralised credential system and the management of users and tenants, such as OpenStack Keystone.		M	M	M	M	M		M	M	M	M		M
PL#2	On-line access to data	Computing / Storage / PaaS	Mandatory	Interactive access to the VMs to avoid downloading huge amounts of data for consolidated inspection of results	This will have to main impacts. If VMs are not be provided of public IPs, reverse tunnelling or any other solution must ensure that the ports used in the interactive access are provided (VNC-like). In any case, firewall rules must enable this kind of traffic.		C							M	M		M	M
PL#3	Network configuration	IaaS	Optional	Extend current standard interfaces to support network configuration, such as VPN-aaS, Firewall-aaS								O						
PL#4	Monitoring and operation	PaaS	Convenient	Keep functionality	Monitoring of resources is competence of the infrastructure provider. Monitoring of the services need to be analysed.					C	C							

INDIGO - DataCloud

Requirements and mapping to services

Case Study	Communities' Specific Requirements	Common Requirements	Requested INDIGO service components	In the 1st Release	In Future Releases
P0_1: Monitoring and Modelling Algae Bloom in a Water Reservoir	LWAB#1: Model processing (hydrodynamic, water quality)	CO#2, CO#4, SO#1,SO#2	Orchestrator, Mesos/Chronos, IAM, Zabbix, OneData	Orchestrator, Mesos/Chronos, IAM, Zabbix Server, OneData	-
	LWAB#2: Distributed storage (Dropbox like)	SO#11	OneData	OneData	-
	LWAB#3: Online post processing	CO#13, SO#1, PL#2	OneData, Ophidia	OneData, Ophidia	-
	LWAB#4: Data & Metadata Management	SO#7	OneData – Metadata	-	OneData - Metadata
P1: Medical Imaging Biobanks	EB#1: Persistent (but medium-term) data storage volumes with standard POSIX file Access	SO#1, SO#2 SO#5, SO#6	OneData	OneData	-
	EB#2: ACL in the access to data	SO#2, SO#4, SO#8, SO#10	OneData, IAM	OneData, IAM	-
	EB#3: Execution of data-driven and computing-intensive workflows	CO#2, CO#7, CO#9, CO#10, SO#6	Future Gateway	Future Gateway	-
	EB#4: Availability of customised software	CO#2,CO#3, CO#14	TOSCA recipes, IM, OneData Client, Mesos	TOSCA recipes, IM, OneData Client, Mesos	-
	EB#5: Deployment of own software	CO#2, CO#3, CO14,	TOSCA	TOSCA recipes,	-

On Data Ingestion and Data Management…

- Align the initial vision of the different research communities with the current advances and recommendati… consider Research Data Alliance (RDA, https://www.rd-alliance.org/).
- The exploitation of INDIGO-**Data**Cloud solutions requires a careful consider… along the full data life cycle. Data collection, storage, processing, analytics on b… …ese …nt activities and many others, like for example related simulations, benefit of a well defined plan…ing to exploit the possibilities offered by the Cloud framework.
- The work started with the initial implementation of a Data Management Plan (DMP).
- But…the answers, showed the need for further work to inform the different Research Communities of the current recommendations on data management, the need to carefully take them into account, and to further detail those data management needs as requirements to INDIGO JRA.
- Fortunately, INDIGO JRA teams are already aware of the fact that solutions are required in this context, as they actively participate in the RDA and similar efforts.
- So the first joint discussion among the Champions from the Research Communities and the JRA developers was much productive, and as a result most of the initial requirements have been already satisfied in the INDIGO Midnight Blue release…. BUT

Submitted 6 proposals to RDA-Open Call for collaboration

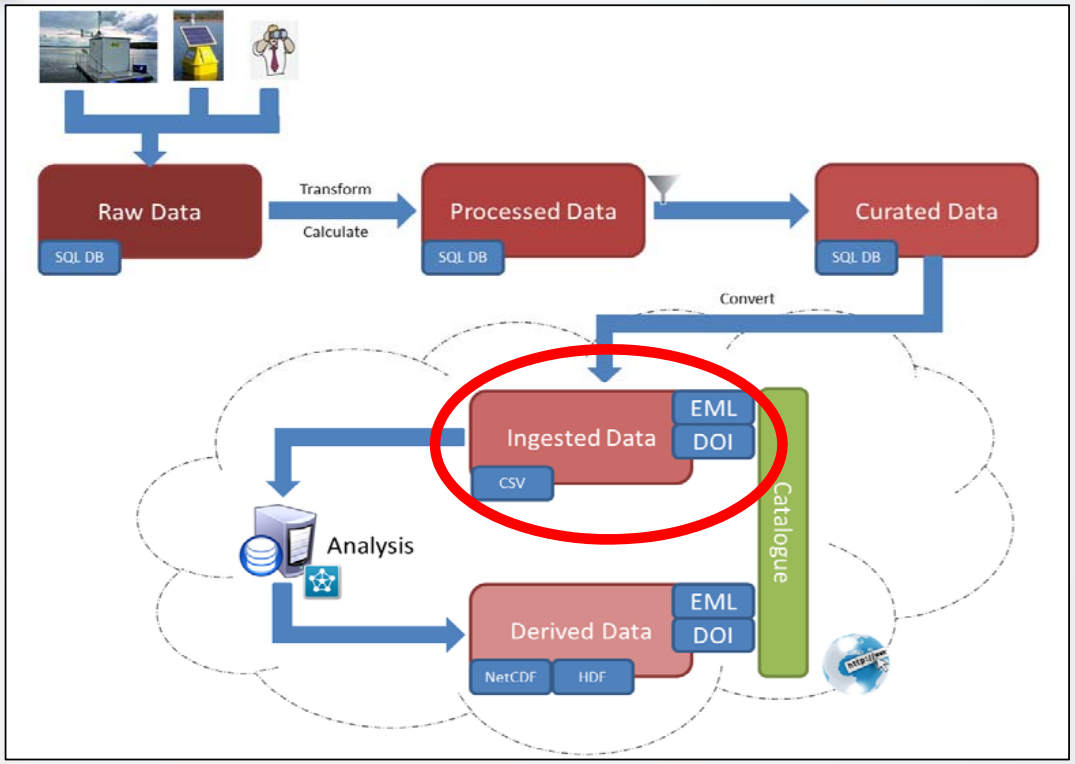
…along the Data Life Cycle in the Cloud

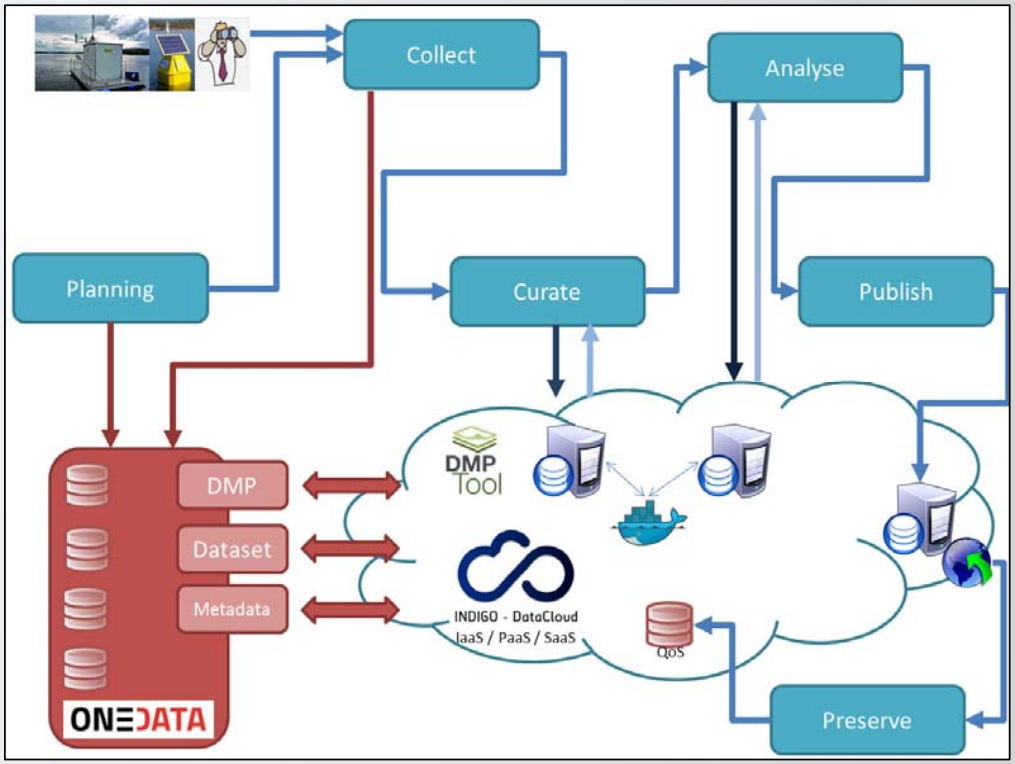
INDIGO - DataCloud

… to fully exploit the potential of a Cloud for Data, we need more detailed plans for data management, and in particular to address Big Data challenges.

- Deliverable D2.11, provides the background required to help the Research Communities to develop these more detailed plans **taking as a reference stage data ingestion, defined as the point in the data life cycle when the data is prepared for re-use, including also potential external users.**
- The deliverable analyses in detail a Case Study for the Algae Bloom prediction, as it includes most of the features of interest, and it is already quite advanced regarding the use of INDIGO services related to data management, in particular solutions like OneData or the potential interest of QoS in storage for preservation, but also others more subtle, like organizing parameter scan simulations using Cloud instances.

Ingested Data in the Life Cycle scheme

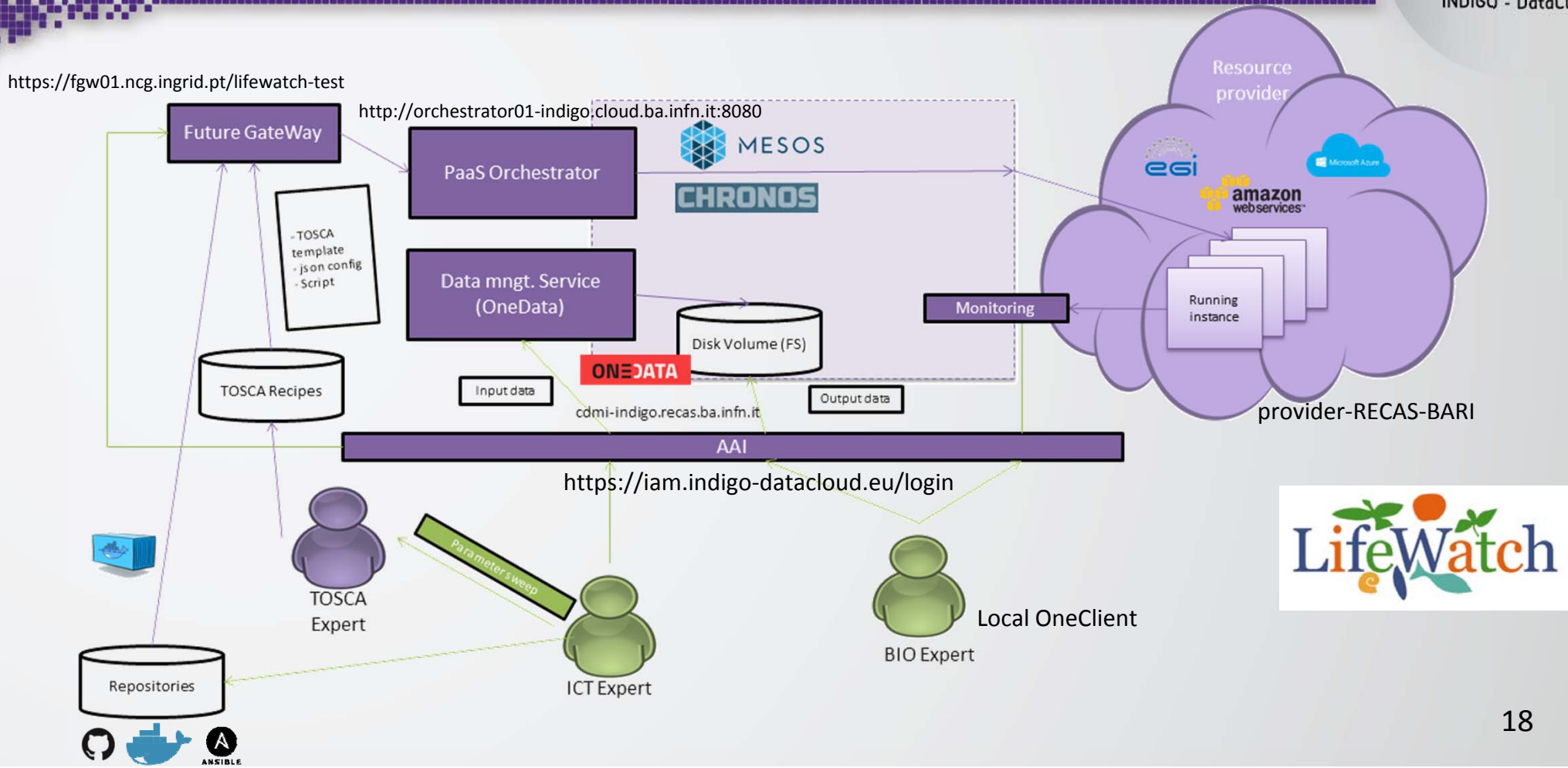




Case Study: Algae Bloom in a Water Reservoir

- **Research Community: LifeWatch (ESFRI)**
  - Topic/Area: Biodiversity & Ecosystem research
- **Objective of the Case Study:**
  - **Monitor the evolution of the potential eutrophication of a Water Reservoir** including the Data Life Cycle management. Hydrodynamic and Water Quality models for forecasting.
  - Schedule: first version of model running by the end of the year. Prototype of Data Life Cycle Management by the second quarter of 2017. In production by third quarter 2017.
- **Innovation challenge:**
  - Different components at different Data Life Cycle stages.
  - Each Model test requires ~20GB and potentially o(10^{2}-10^{4}) (multi-parametric)
- **Teams involved:** IFCA/CSIC Team + Ecohydros **(SME)** Team (consulting).
- **Final user community**: Researchers (LifeWatch Community), Water management authorities, ICT Groups, Limnology groups.
- **Impact**:
  - Pro-active management actions on water reservoirs , including new policies.
  - Definition of monitoring instrumentation and parameters to be under control.

Solution Developed



INDIGO added value

- **Scalable (storage and computing) resources in the cloud to perform o(10^2-10^4) tests…**
- **…and share directly within the community**
- **User Friendly interface to use cloud resources:**
  - Final users only need to fill a form to submit a new simulation, avoiding the script edition or direct contact with the infrastructure (Supercomputer, Grid, Cloud) (very helpful for non IT experts).
  - First time we use a flexible and “universal” user authentication (quite relevant to collaborate with SMEs also)
  - Transparent access to shared large storage (OneData)

ELIXIR-ITALY Case Study: Galaxy workflow

Galaxy is a workflow manager adopted in many life science research environments in order to facilitate the interaction with bioinformatics tools and the handling of large quantities of biological data.

ELIXIR-ITALY, the Italian node of ELIXIR, is developing a fully customizable Galaxy instance provider platform founded on the technologies elaborated within the INDIGO-DataCloud project framework. The goal is to provide, through an easy setup procedure, an on-demand workspace ready to be used by life scientists and bioinformaticians.

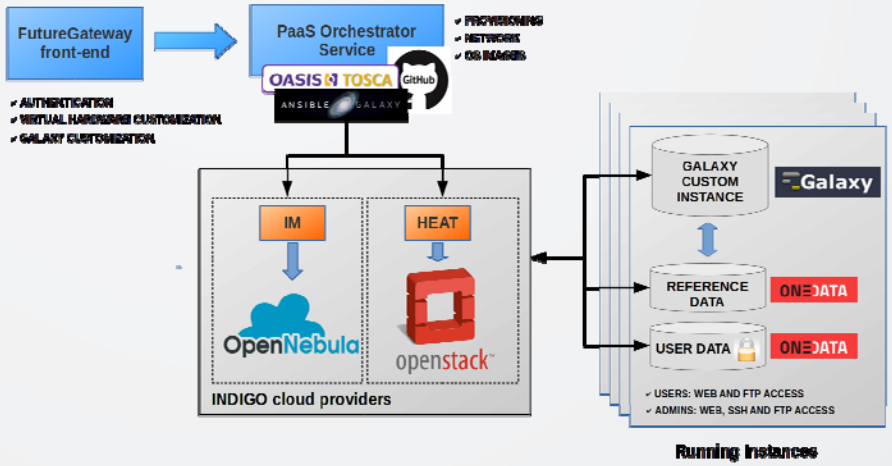
Key features:

- Isolated Galaxy instances;
- Galaxy customization;
- Virtual hardware customization.

Specific Requirement	Generic Requirement	Service component
Galaxy instance deployment Galaxy software customization	CO#1,CO#3 Employment of Interface SaaS	FutureGateway, Orchestrator, TOSCA, IM
Instance Isolation File-system like storage Persistent storage	SO#1, SO#2,SO#4, SO#8 Persistent data storage	OneData and/or IaaS local block storage
Galaxy instance access	PL#2, PL3 On-line access to data Network configuration	FutureGateway, Orchestrator, SSH
Cloud bursting facilities Automatic elasticity	CO# 4, CO#9, Automatic elasticity	CLUES

ELIXIR-ITALY Case Study

Each instance will be tailored to the specific user needs by the users themselves through a web interface that allows the selection among different sets of tools and different virtual hardware setups. Once deployed each Galaxy instance will be fully customizable with tools and reference data and running in an insulated environment, thus providing a suitable platform for research, training and also clinical scenarios involving sensible data.



Galaxy cloud service architecture: the prototype is based on the coordination of separated components, provided by the INDIGO e-infrastructures.

INDIGO Services

- **FutureGateway Portal:** the web front-end is designed to grant user friendly access to this service, allowing for an easy configuration and launch of Galaxy instances.
- **Orchestrator:** the INDIGO Orchestrator Service, based on the TOSCA orchestration language automatically setup Galaxy instances with all their required components deployed and configured using the Ansible role indigo-dc.galaxycloud. It is currently hosted on the INDIGO github repository and installed through Ansible-Galaxy (ansible-galaxy.com). The role supports both Virtual Machines and Docker containers.
- **OneData:** persistent storage is needed to store users and reference data and to install and run new (custom) tools and workflows. The users' data access rights will be controlled through the OneData INDIGO component.
- **Elastic cluster support**: It is provided by integrating SLURM within the Ansible role.

Climate Model Intercomparison Data Analysis case study (ENES)

- The proposed case study is directly connected to the **Coupled Model Intercomparison Project** (CMIP) and to the **Earth System Grid Federation** (ESGF) infrastructure
- CMIP* experiments provide input for multi-model analytics experiments (e.g. trend analysis, climate change signal analysis)
- Key challenges:
  - **Data distribution** is inherent in the infrastructure
  - Research community infrastructure is mainly for **data sharing**
  - **Data download is a big barrier** for end-users (download can take from several days to weeks!)
  - **Data analysis** is mainly performed using **client-side** and **sequential approaches**
  - The complexity of the data analysis needs more **robust end-to-end support**
  - **Scientific data formats** (e.g. NetCDF) needs to be properly managed

Added value of the INDIGO solution for ENES

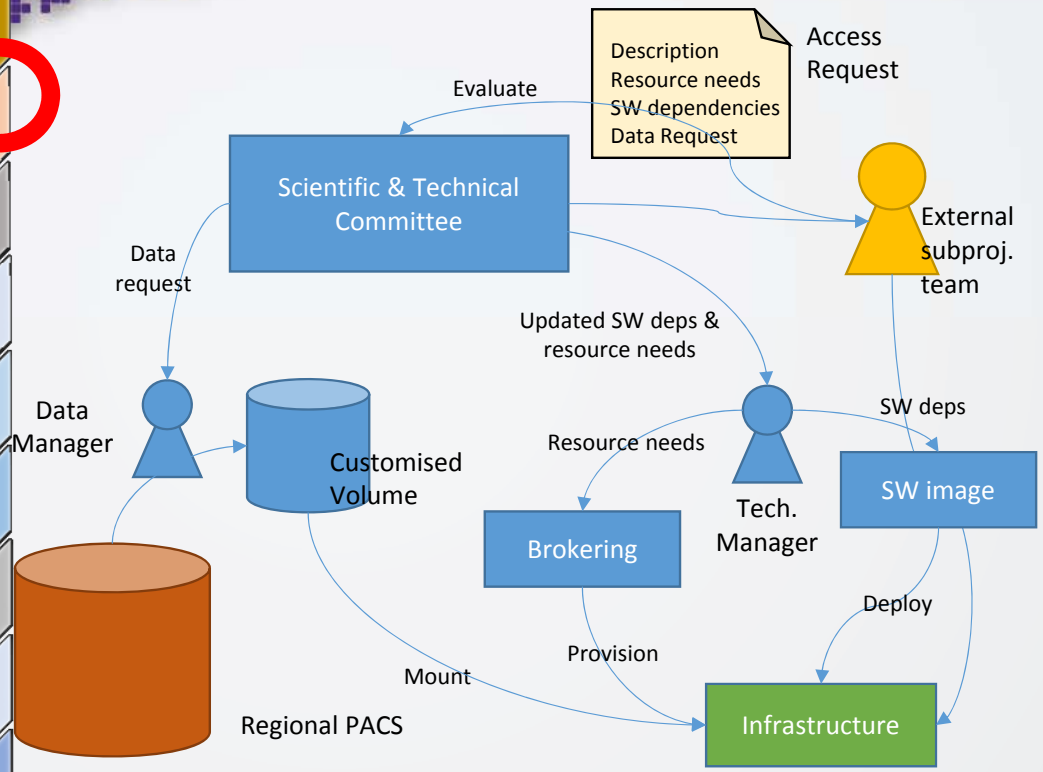
The solution implemented in INDIGO:

- implements a different **paradigm** (from **client-** to **server-side**)
- intrinsically **reduces data movement**
- makes **lightweight** the end-user setup
- fosters **re-usability** (of data, final/intermediate products, workflows, sessions, etc.)
- **complements**, **extends** and **interoperates** with the **ESGF** stack
- provides a new "**tool**" for scientists to run multi-model experiments

And

It drastically reduces the time to the solution!

Medical Imaging Biobanks

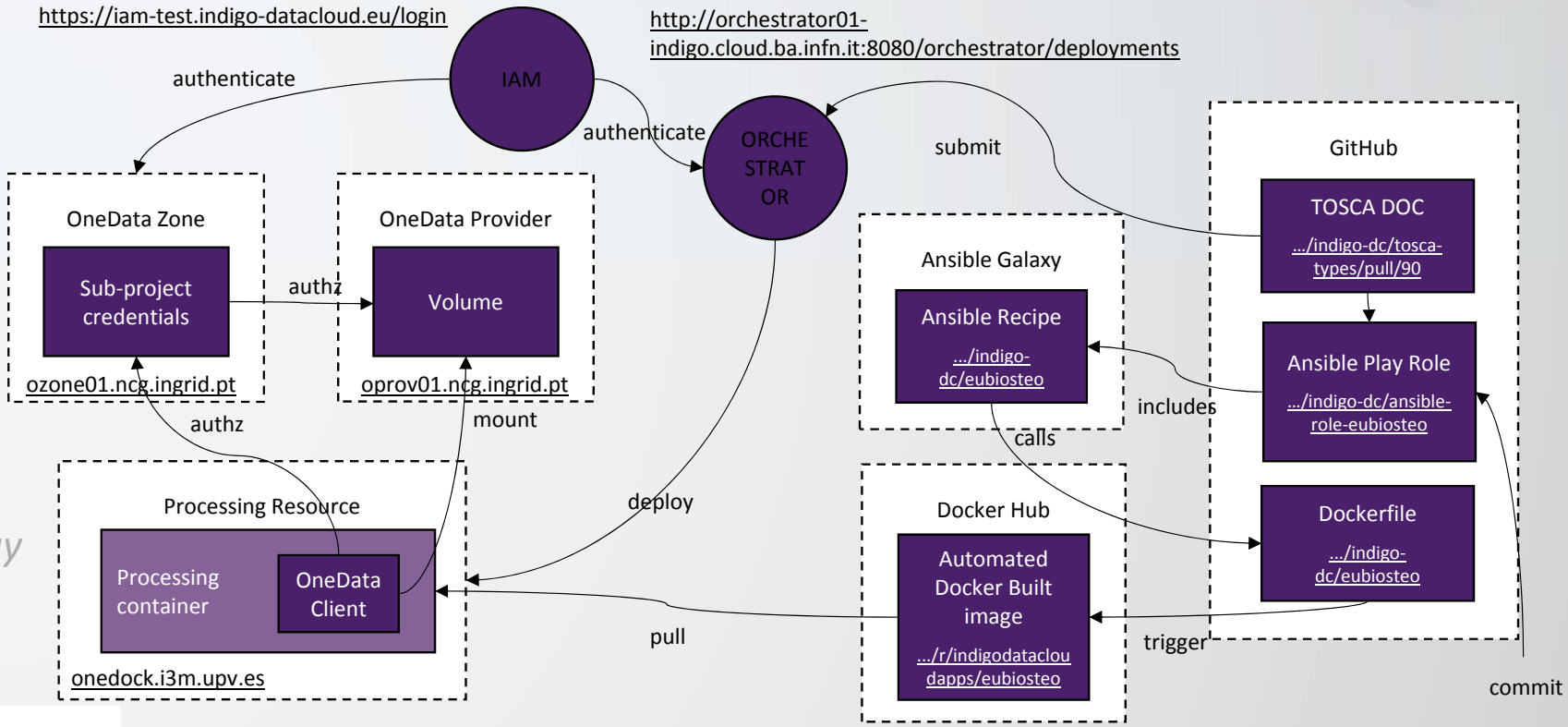


http://ceib.san.gva.es/bimcv

- BIMCV (an EuroBioImaging ESFRI node) manages a population database from an area of 5 Million people
- BIMCV will receive applications for projects
  - E.g. Training a set of models for the automatic segmentation of bone tissues in osteoporotic women of an age above 70, including sound control subjects.
- BIMCV needs a system to setup virtual infrastructures for Medical Imaging Biomarkers projects
  - BIMCV will provide processing tools and pipelines and will allow the use of any third party tool
- INDIGO-DC offers BIMCV
  - The capability of creating secured shared remote volumes accessible through POSIX.
  - The definition of complex applications involving elastic batch queues, graphical interfaces compatible with on-premise, research and public clouds.
  - High and convenient customizability of software for both containers and VMs.
  - Single sign-on, integrated deployment, QoS.

Medical Imaging Biobanks – INDIGO Components

- IAM
- IM-TOSCA
- ONEDATA
- Orchestrator
- ONEDock
- *CLUES*
- *QoS*
- *FutureGateway*



EGI Case Study: Three generic User Stories

- EGI: International standard-based federation of Resources Providers
  - **A diverse variety of User Communities (from Earth Observation to Humanities)**
  - Common needs across multiple communities
- Three stories covering common needs:

1- Creating Virtual Machines sharing common datasets on multiple heterogeneous and distant sites

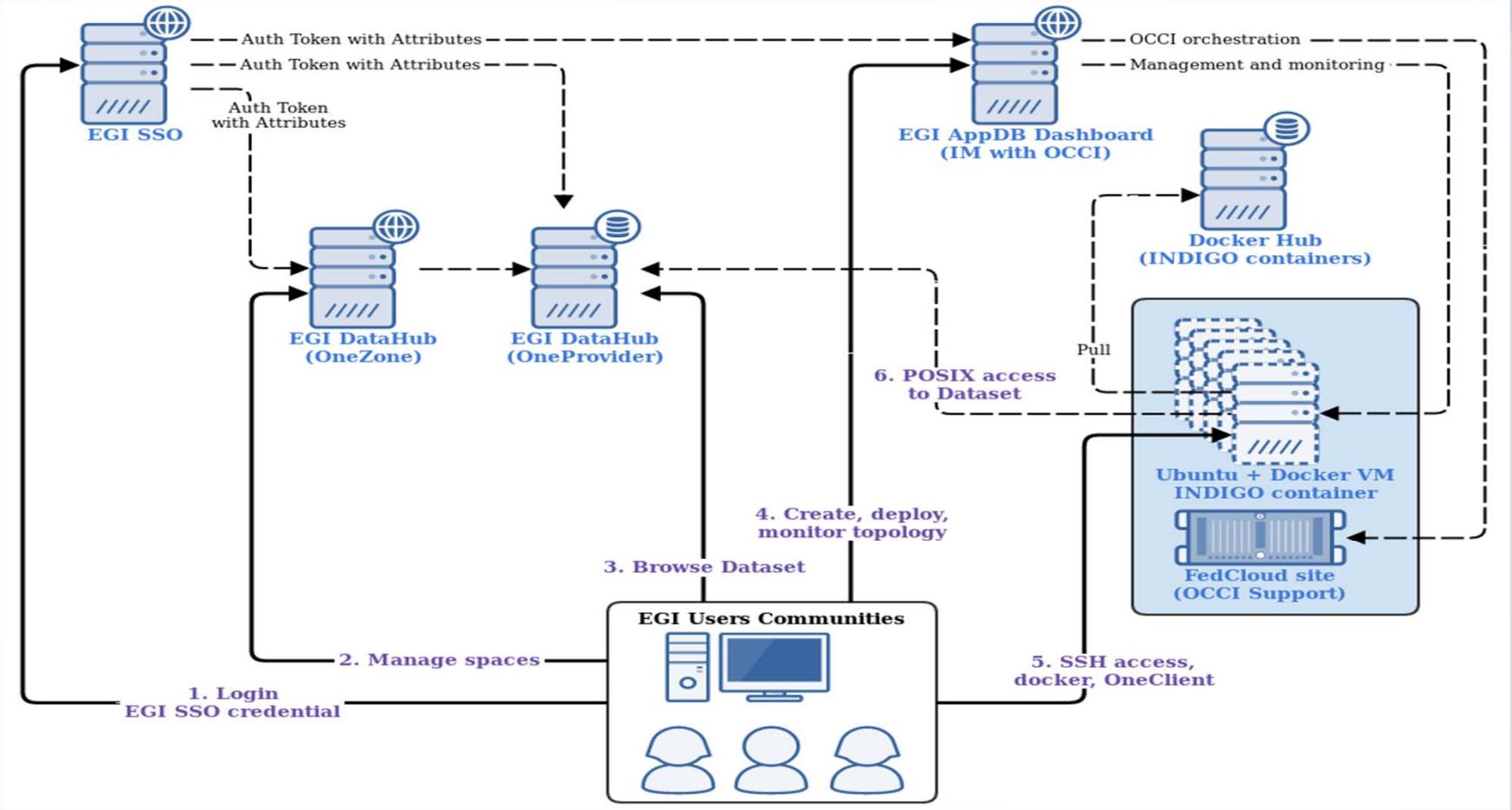
2- Running an application from a docker container accessing remote storage sites via POSIX

3- Using EGI Single Sign On (SSO) credential to access any service of the INDIGO-DataCloud platform (according to access rights)

One prototype, lots of technology!

- Using EGI SSO account, browse a dataset in **EGI DataHub**
  - **INDIGO contribution: OneZone and OneProvider**
- Using EGI SSO and AppDB VMops Dashboard to instantiate, manage and monitor a Virtual Machine (VM) with Docker support on an EGI cloud site
  - **INDIGO contribution: Infrastructure Manager (IM)**
  - **OCCI used in IM and FedCloud sites**
- Using ssh to access the VM with the provided secure key
  - **INDIGO contribution: Infrastructure Manager**
- Running a Docker container inside the VM to access the dataset
  - **INDIGO contribution: OneClient, OneProvider and INDIGO containers**

Overview


Auth Token with Attributes
Auth Token with Attributes
Auth Token with Attributes
EGI SSO
OCCI orchestration
Management and monitoring
EGI AppDB Dashboard
(IM with OCCI)
Docker Hub
(INDIGO containers)
EGI DataHub
(OneZone)
EGI DataHub
(OneProvider)
Pull
6. POSIX access
to Dataset
Ubuntu + Docker VM
INDIGO container
FedCloud site
(OCCI Support)
4. Create, deploy,
monitor topology
3. Browse Dataset
EGI Users Communities
2. Manage spaces
1. Login
EGI SSO credential
5. SSH access,
docker, OneClient

Benefits gained from INDIGO solutions

- Infrastructure Manager
  - Transparent management of VM lifecyle from AppDB
  - Automatic mounting of EGI block storage inside the VM
  - Harmonized access to different cloud middleware (abstraction layer)
  - Possibility to orchestrate multiple VMs (under study)
- OCCI: ooi (OpenStack), NOW (OpenNebula) and rOCCI
  - Automated, precise network configuration (planned integration)
- OCCI: rOCCI-server
  - Federating commercial providers (Amazon AWS resources) into EGI (future)
- Onedata
  - Easy sharing of datasets and produced outputs
  - Access from multiple client (web, VM, containers)
  - POSIX-like file access

DisVis and Powerfit & INDIGO DataCloud

INDIGO - DataCloud

Docker containers for multiple CPUs and GPGPU (also used for grid)

https://hub.docker.com/u/indigodatacloudapps/

Access to VM: SSH

Pull Docker images for DisVis and PowerFit

User

Grid certificate

IM: Infrastructure Manager

Selection of base image with Docker in it

Applications Database
Supporting egi

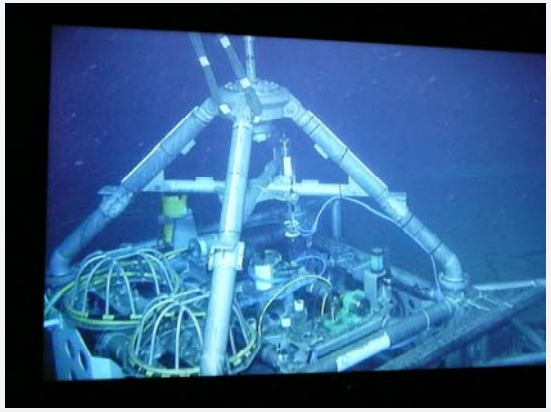
https://appdb.egi.eu/

Future work:

- Integration of OneData for input/output
- A VM with visualization tools – Chimera, PyMOL

MOIST, a case study applied to one of the EMSO nodes

*This case study is a pilot experience used to describe some of the activities performed by **INGV** within the European Research Infrastructure **EMSO** (European Multidisciplinary Seafloor and water column Observatory)*



We consider data collected by the NEMO-SN1 observatory, one of the EMSO nodes used for geohazard monitoring (Western Ionian Sea, Italy).

Areas of particular interest

Area	Present	Requirements
Data storage	Single local server, External server (if needed)	**SO#10, SO#7:** Distributed data storage Metadata storage
Data fruition	Only internal storage access Web site- only some data published	**SO#1, SO#4**: Efficient data retrieval
Data analysis	Usually performed on PC No web tools available Low Computing Resources	**CO#13, PL#2**: Online data processing
Data access management	Typical Linux-based user and group management Access by ssh and samba	**PL#1**: Tracking user access and authorization

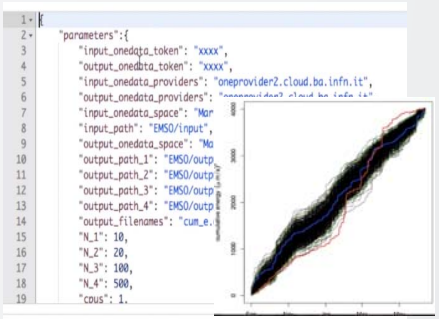
Current work and future steps

Global Data Access

Testing ONEDATA as single interface between different servers at INGV and partners for data transfer and sharing of scientific work

PaaS-Orchestator

Executed R job on PaaS (INFN-Bari) with a docker container described by TOSCA template (left panel) through INDIGO Orchestrator and Apache Mesos



Graphical output (right panel): SDE energy cumulate curve (red) are compared to random process cumulate curves (black)

Big Data Analytics

Implementing import/export functions of seismological data files in SAC (Seismic Analysis Code) format into Ophidia framework. Extension to other seafloor observatory data types.

Multiparametric data analysis based on **Ophidia**

INDIGO-IAM

Identity and Access Management

Thank you

INDIGO - DataCloud

https://www.indigo-datacloud.eu

Better Software for Better Science.