DEEP Hybrid-DataCloud

Intensive computing techniques for applications needing specialised hardware

Workshop: Creating Platform-Driven E-Infrastructure Innovation On EOSC
10 July 2019

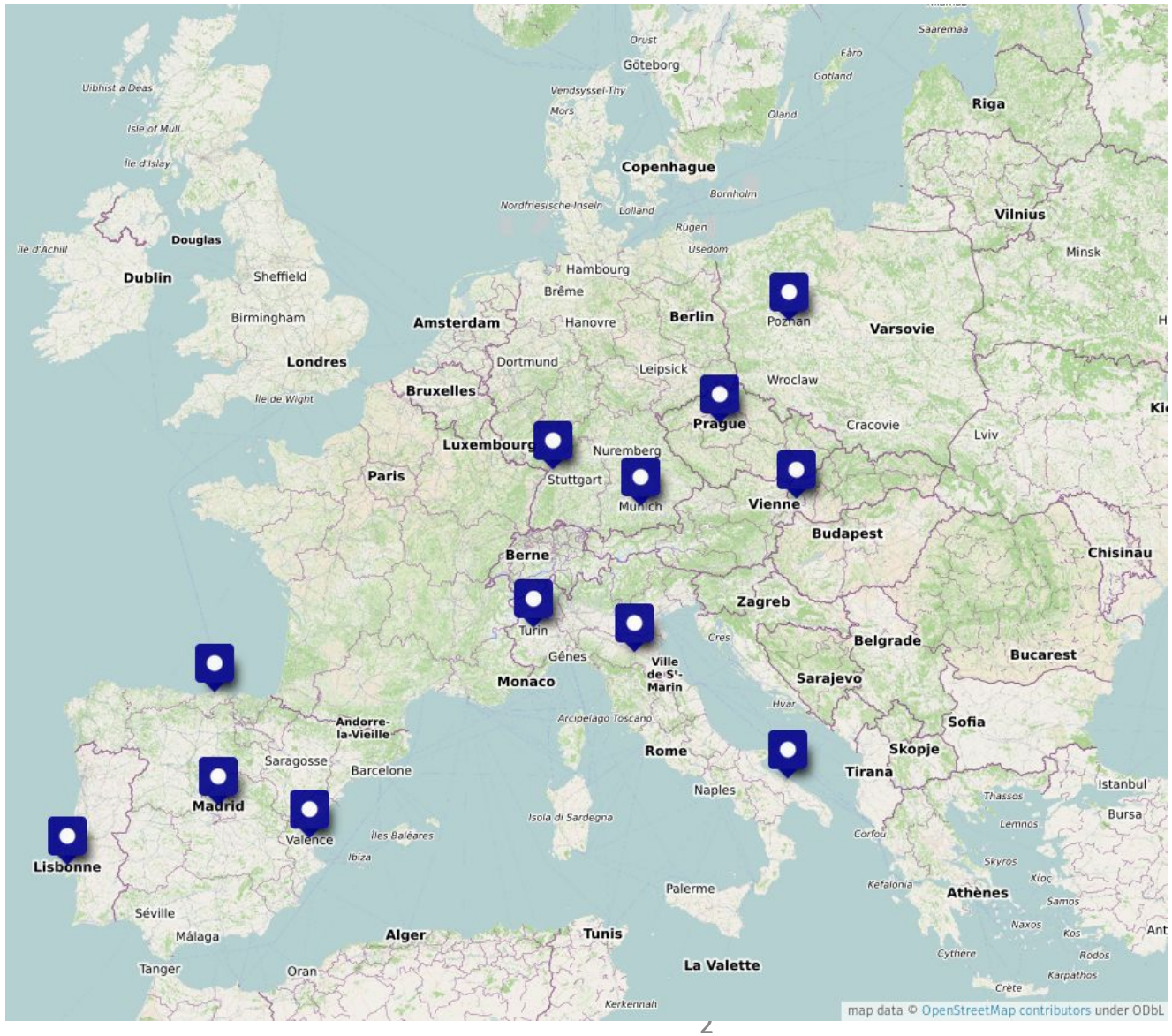
Mario David, Cristina Duma, Valentin Kozlov and Alessandro Costantini

On behalf of all the partners of DEEP

DEEP-Hybrid-DataCloud is funded by the Horizon 2020 Framework Programme of the European union under grant agreement number 777435

DEEP HybridDataCloud

- **D**esigning and **E**nabling **E**-Infrastructures for intensive data **P**rocessing in a **Hybrid DataCloud**
- Started as a spin-off project (together with eXtreme DataCloud - XDC) from INDIGO-DataCloud technologies
- H2020 project, EINFRA-21 call
- Runs November 1st 2017 – April 2020
- 9 academic partners + 1 industrial partner:
  - CSIC, LIP, INFN, PSNC, KIT, UPV, CESNET, IISAS, HMGU, Atos

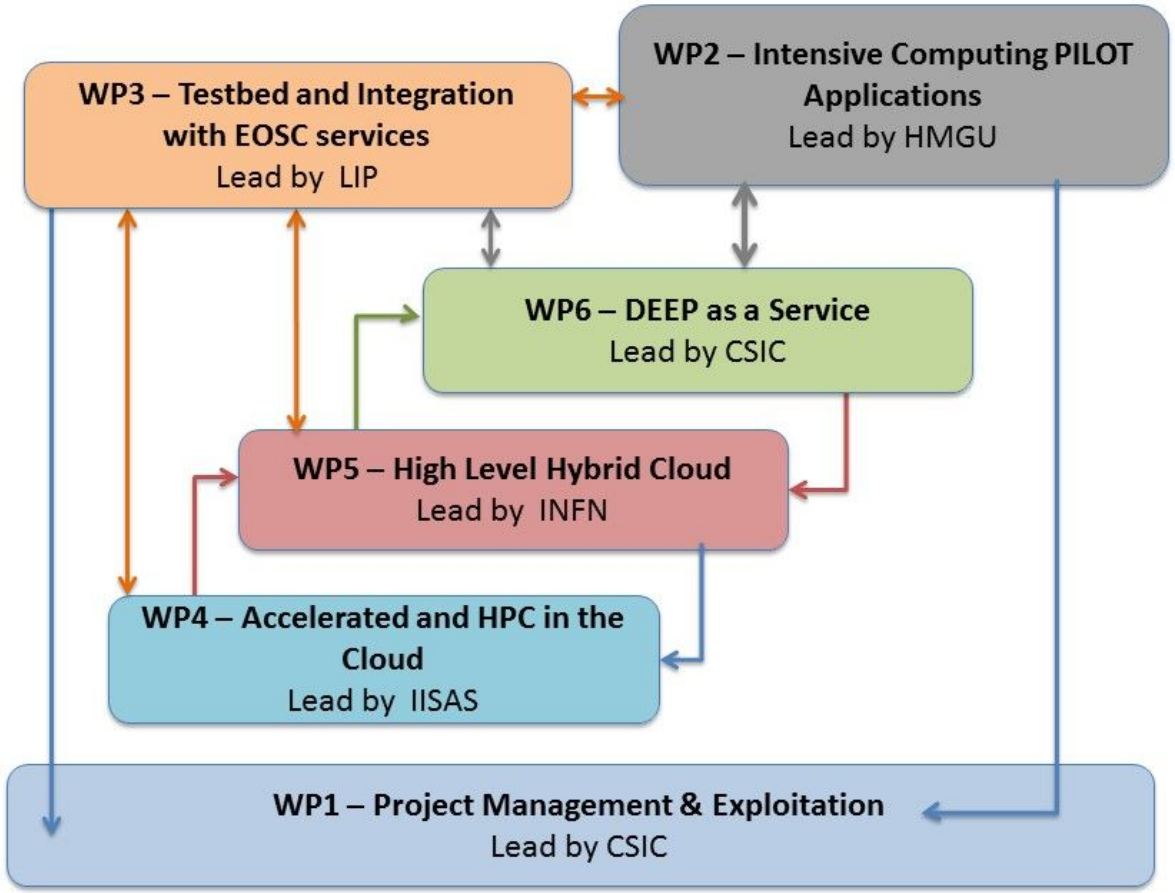


DEEP-HybridDataCloud

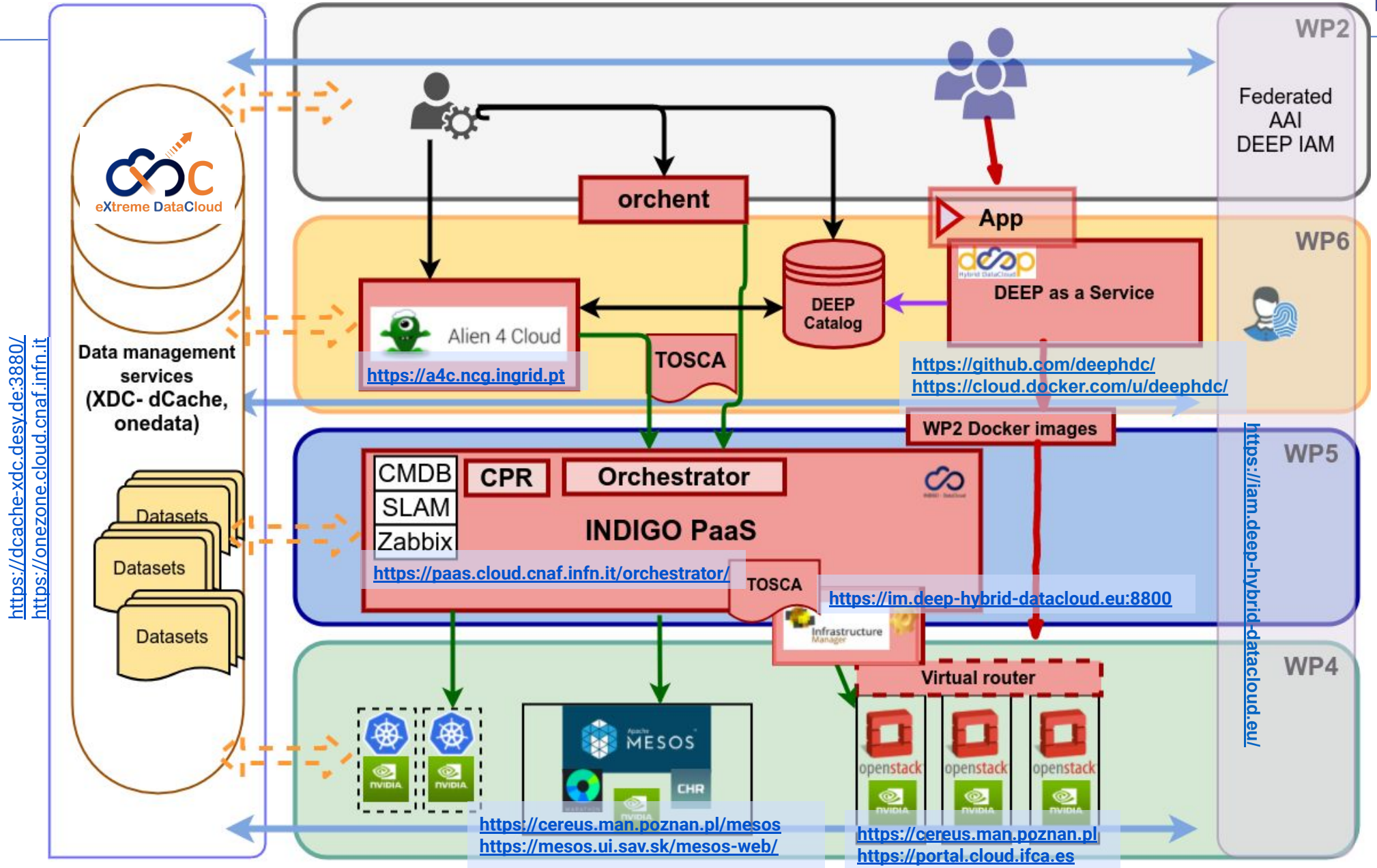
- **Goal**: prepare a **new generation** of **e-Infrastructures** that harness latest generation technologies, supporting **deep learning** and other **intensive computing techniques** to exploit **very large data sources**

- **Global objective**: **promote** the use of **intensive computing services** by different research communities and areas, and the **support by the corresponding e-Infrastructure** providers and open source projects

- **Ease and lower** the entry barrier for non-skilled scientists
  - **Transparent execution** on e-Infrastructures
  - Build **ready to use modules** and offer them through a catalog or marketplace
  - Implement common software development techniques also for scientist's applications (**DevOps**)



DEEP architecture

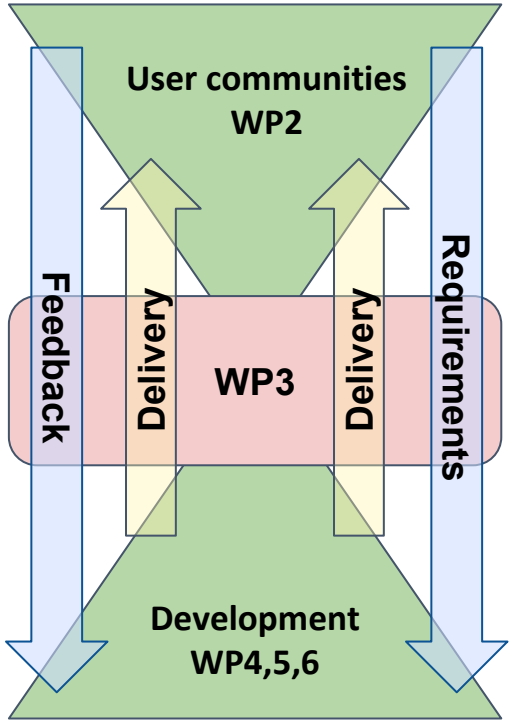


DEEP core components

Component	Workpackage
Kubernetes	WP4
Mesos/Marathon/ Chronos	WP4
OpenStack nova-lxd	WP4
udocker	WP4
PaaS Orchestrator	WP5
Orchent	WP5
Infrastructure Manager	WP5
CLUES-indigo	WP5
TOSCA types and templates	WP5, WP6
Monitoring System	WP5
CloudProviderRanker	WP5
Cloud Information Provider	WP5
INDIGO Virtual Router	WP5
Alien4Cloud - DEEP	WP6
DEEPaaS API	WP6

Documentation
Configuration
Deployment

Development & Upstream Contribution



IaaS components

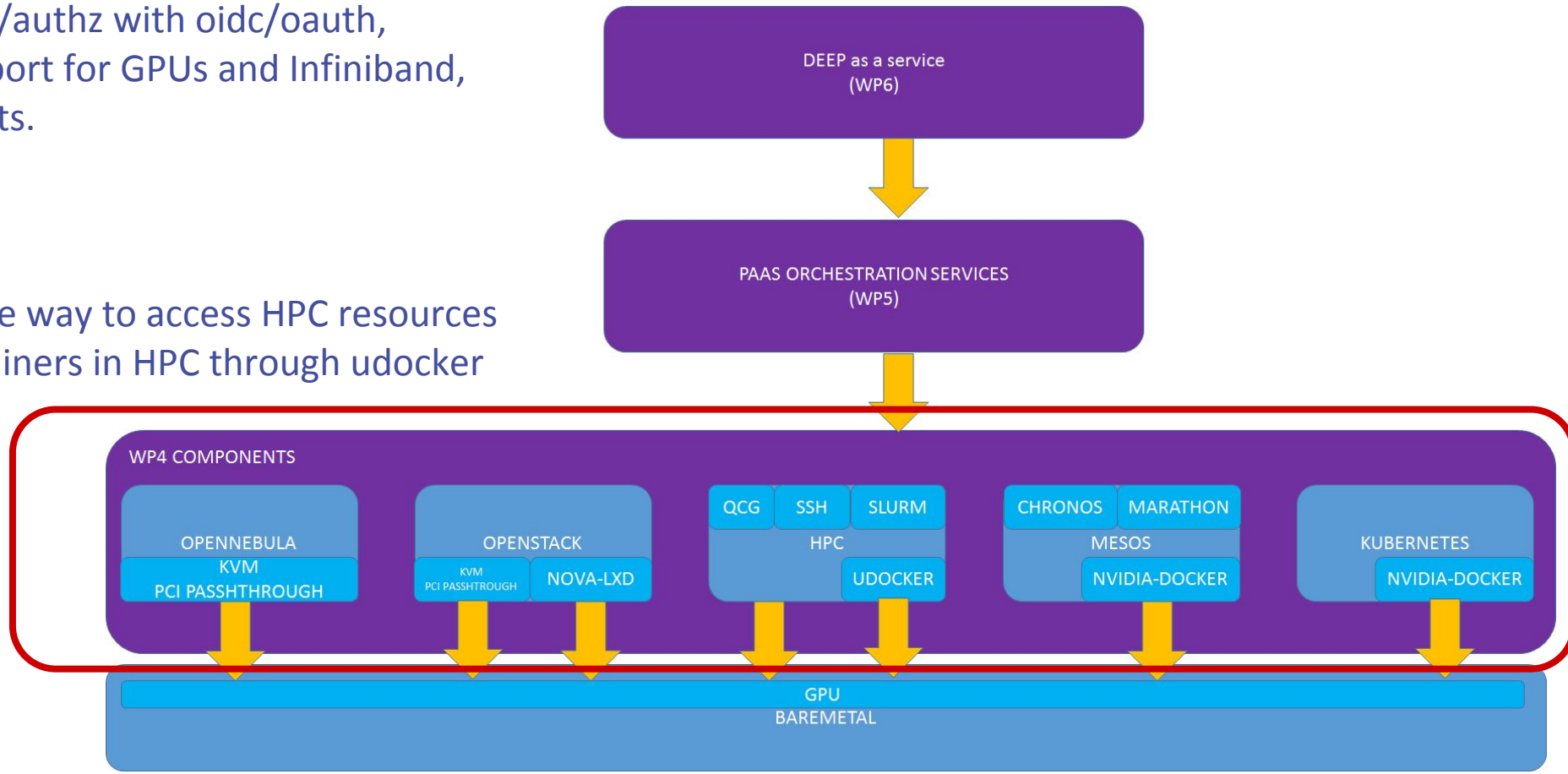
udocker: improve support of GPUs, improve support of low latency interconnects (Infiniband)

Openstack: nova-lxd - add/improve support of GPUs

Mesos/Marathon/Chronos: authn/authz with oidc/oauth, documentation/configuration support for GPUs and Infiniband, recipes for on-demand deployments.

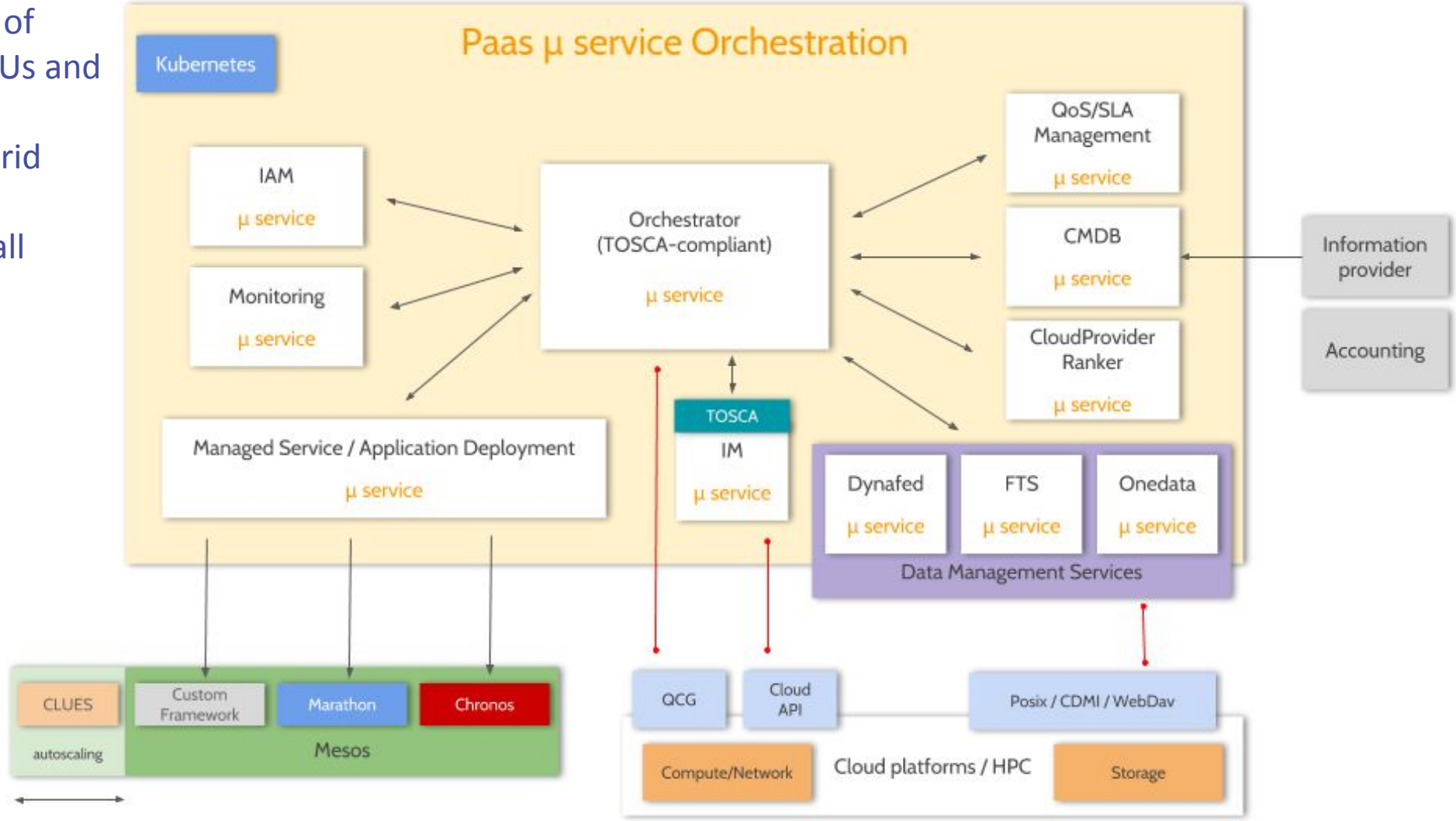
Kubernetes: same as mesos

HPC Integration Tools: Provides the way to access HPC resources from the PaaS Orchestrator - containers in HPC through udocker



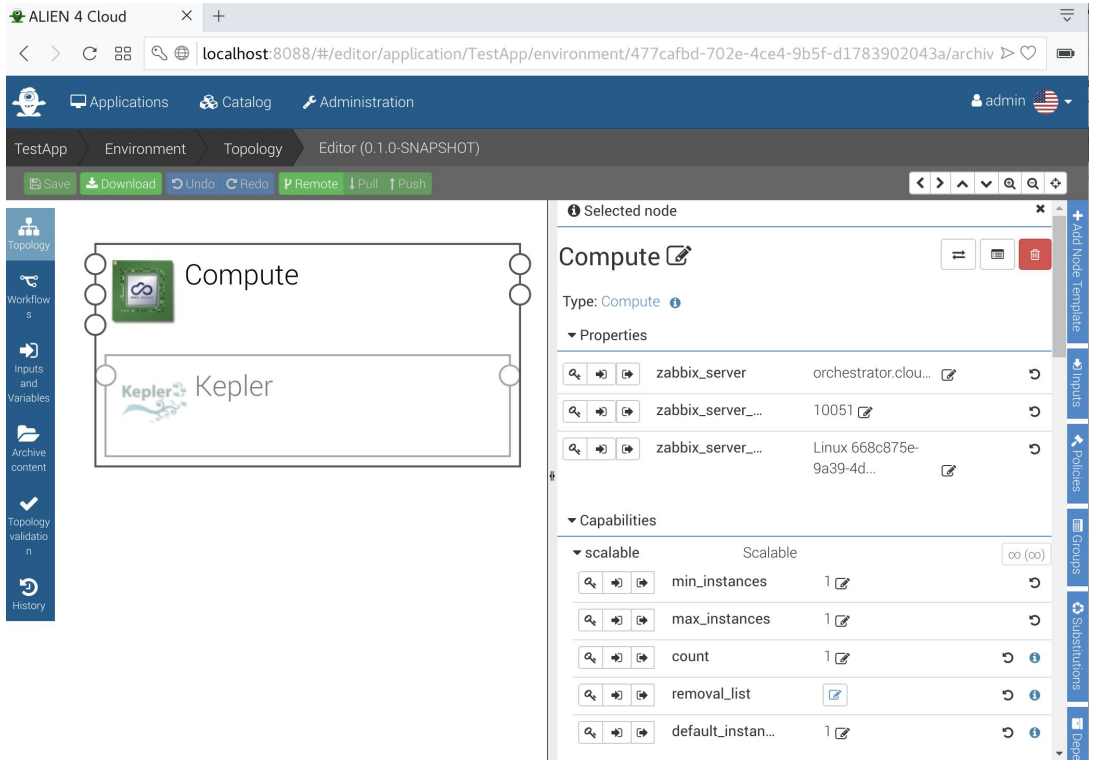
PaaS components

- Develop/Improve support of deployments requiring GPUs and Infiniband
- Better support for the hybrid multi-site deployments
- Provide Tosca recipes for all required deployments



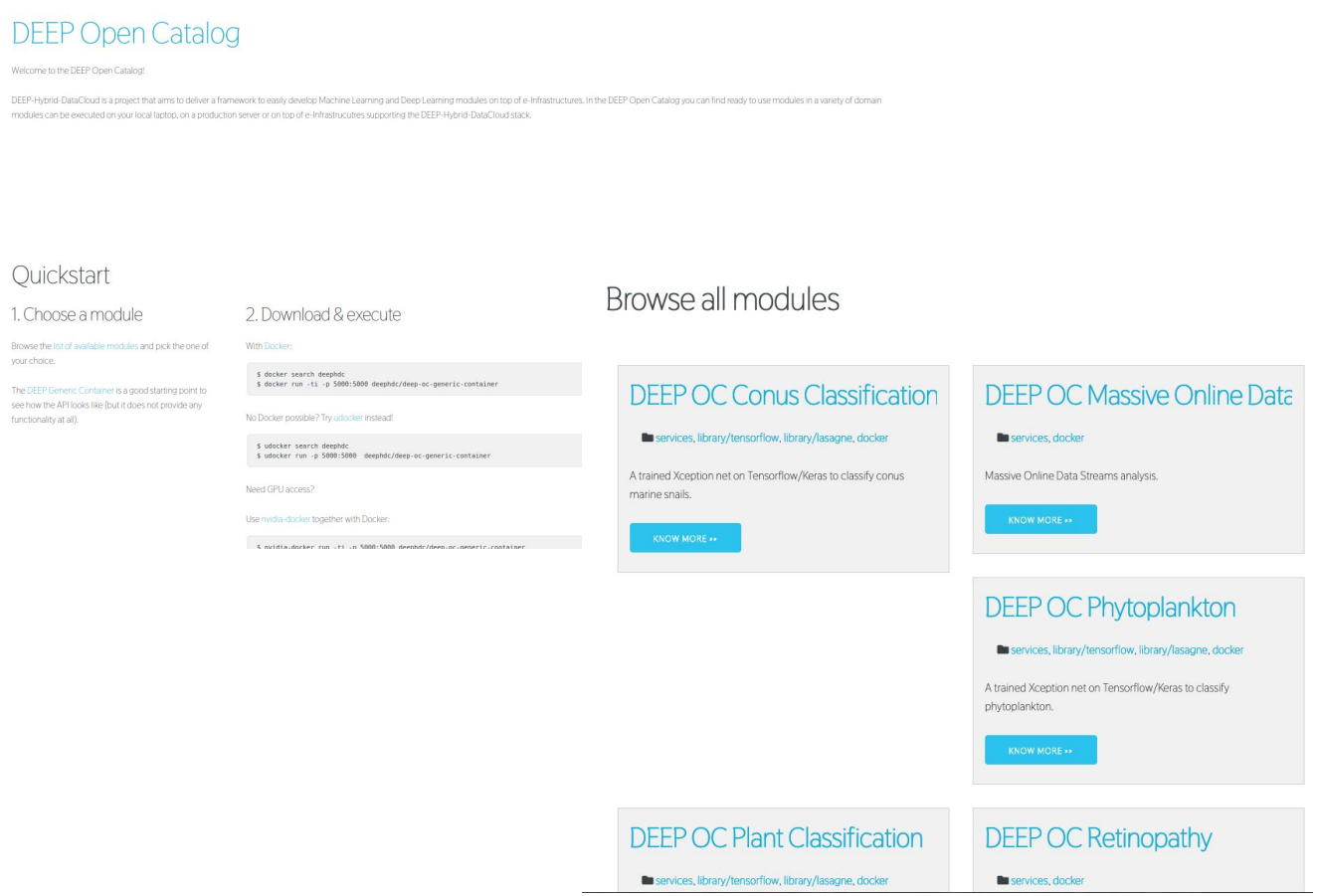
DEEPaaS components

Alien4Cloud: TOSCA template composition & deployment



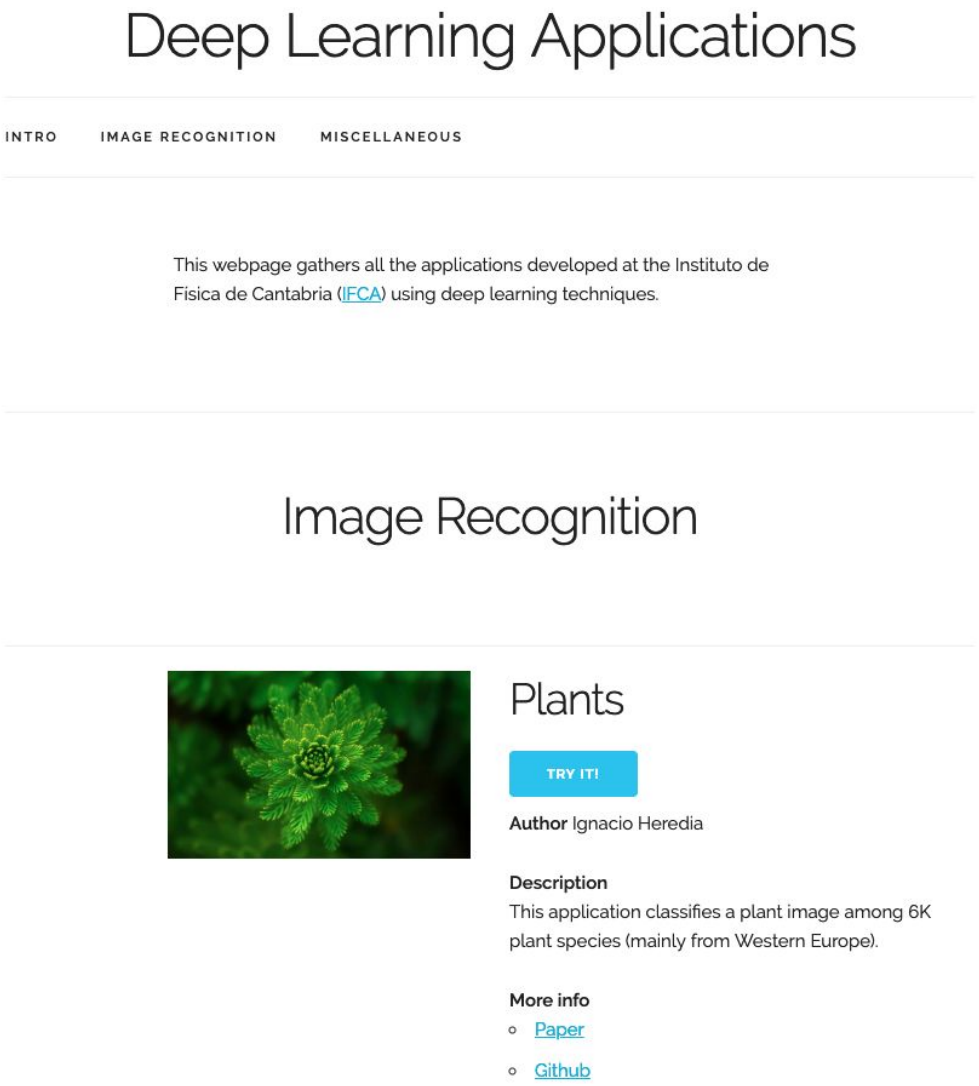
DEEP Open Catalog/Marketplace:

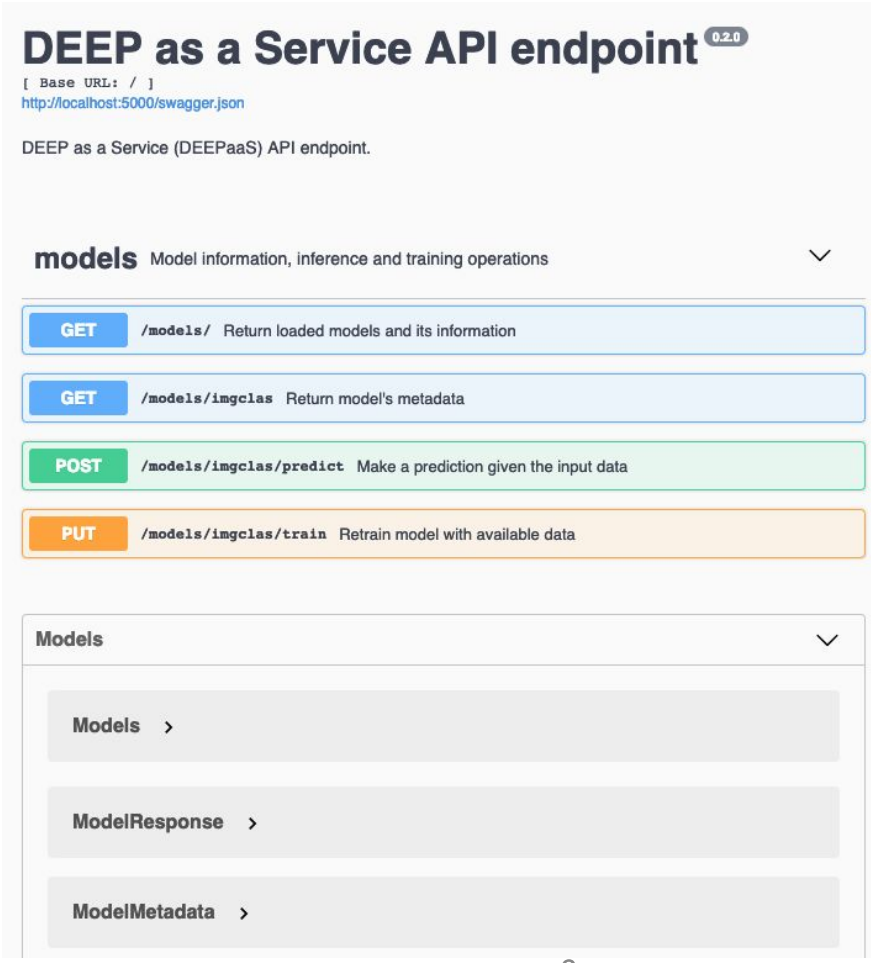
provides the universal point of entry to all services offered by DEEP.



DEEPaaS components

REST API that is focused on providing unskilled users with access to machine learning models.



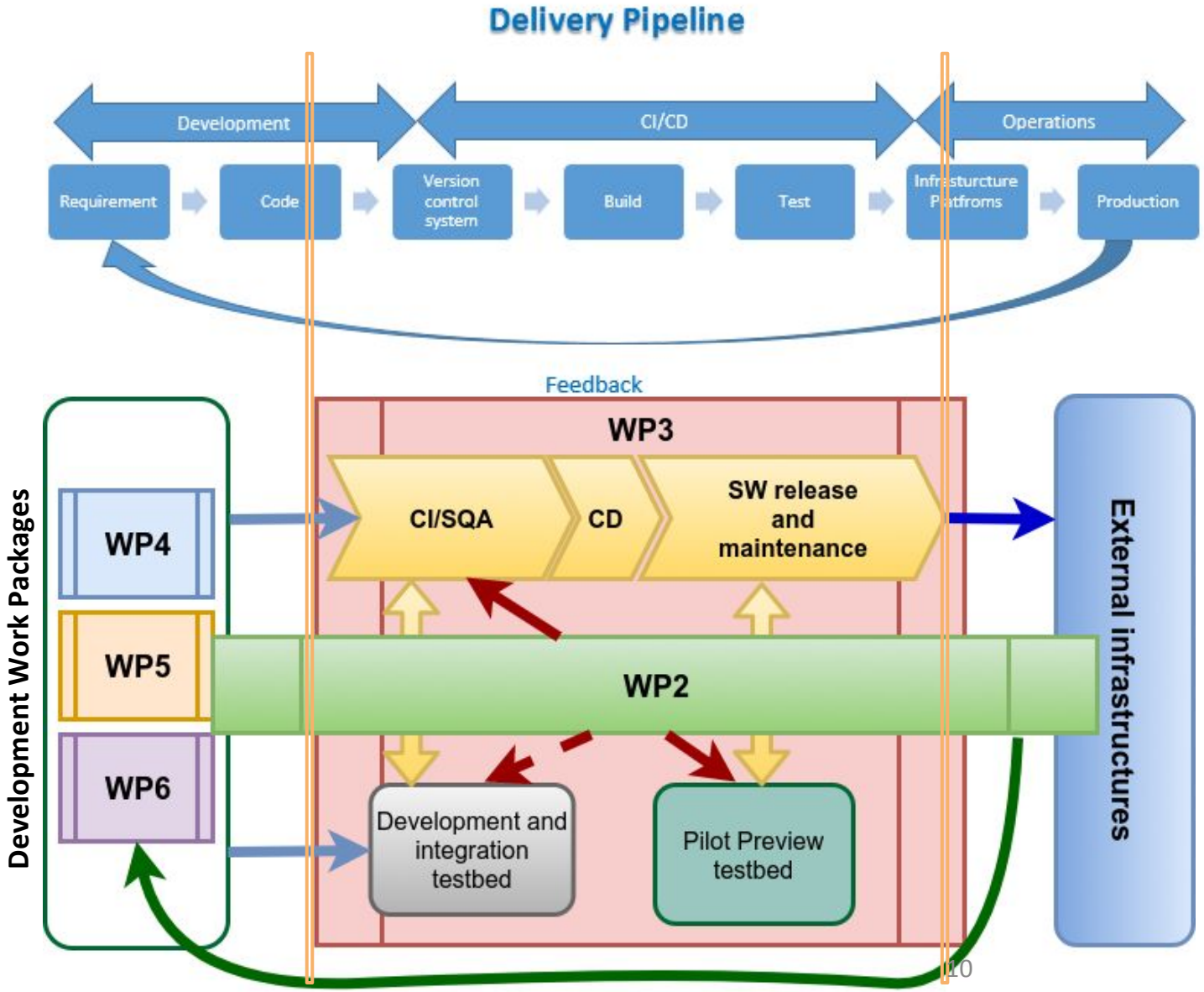


SQA, Release, Maintenance, Support and Testbeds

Hybrid DataCloud

Plan and design

Implementation

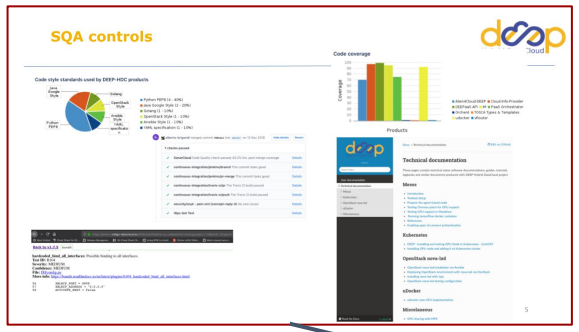


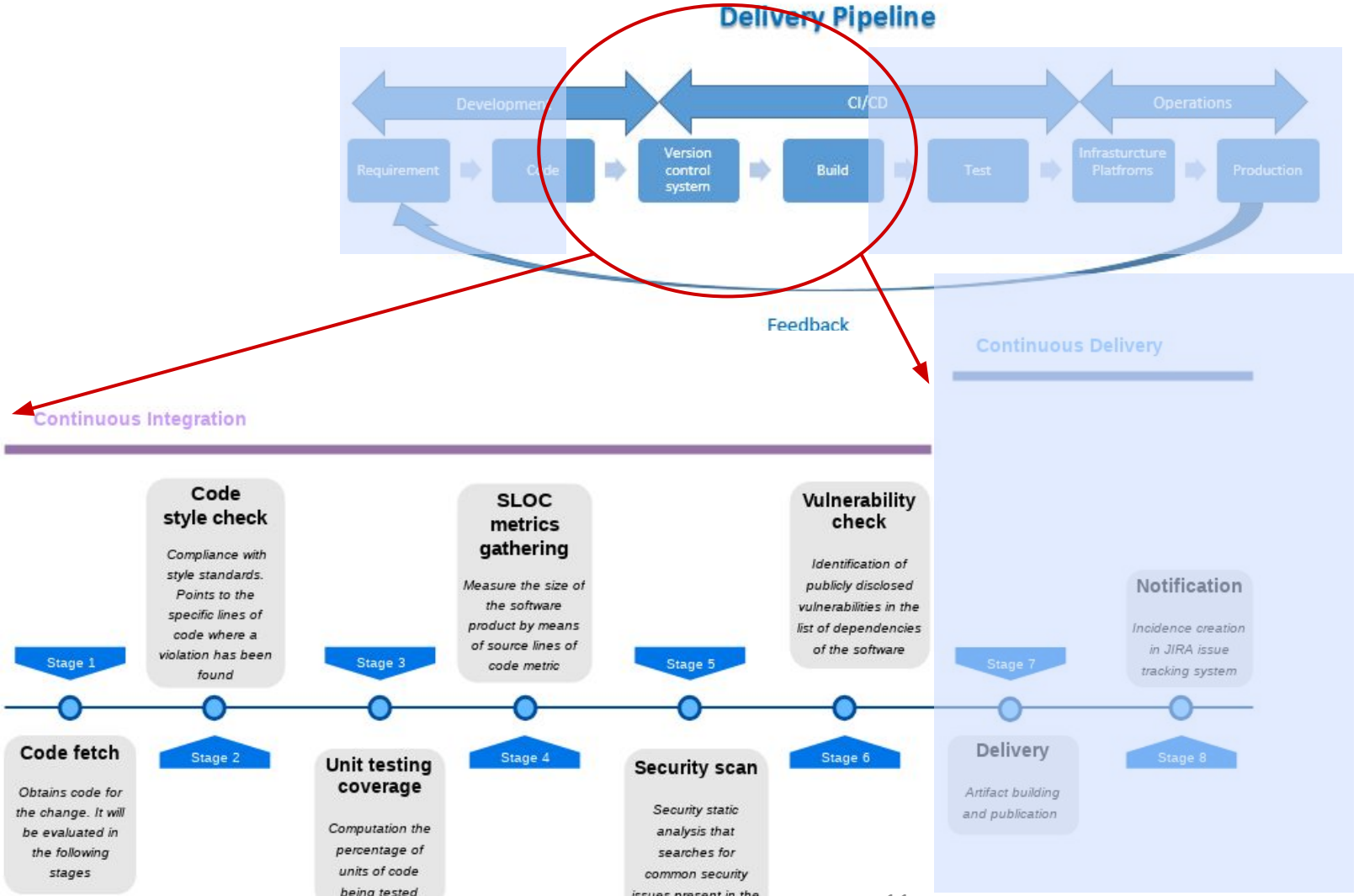
DEEP core components: CI/SQA phase

"A set of Common Software Quality Assurance Baseline Criteria for Research Projects" http://hdl.handle.net/10261/160086

Open access, open for contributions:
https://github.com/indigo-dc/sqa-baseline

1. Code fetching
2. Code style check
3. Unit testing coverage
4. SLOC metrics gathering
5. Functional and integration testing
6. Code Review
7. Documentation
8. Automated Deployment
9. Security linter/scanner
10. Vulnerability check on dependencies
11. Delivery
12. Notifications

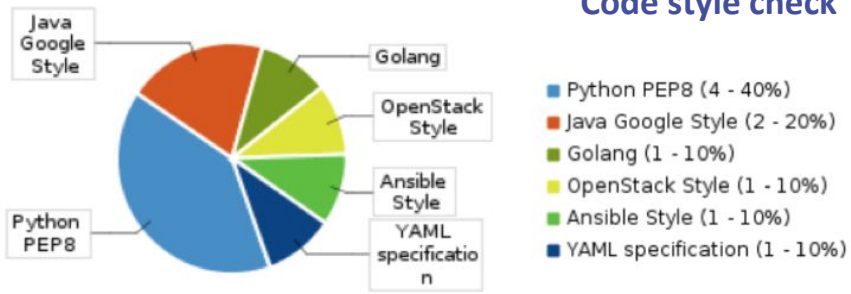




SQA controls

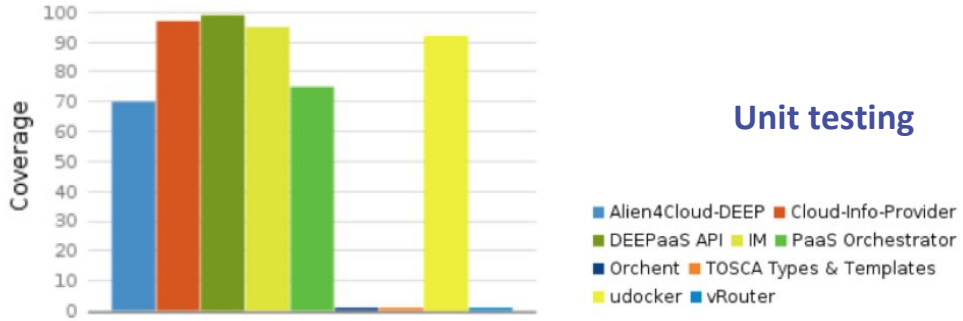
Code style standards used by DEEP-HDC products

Code style check

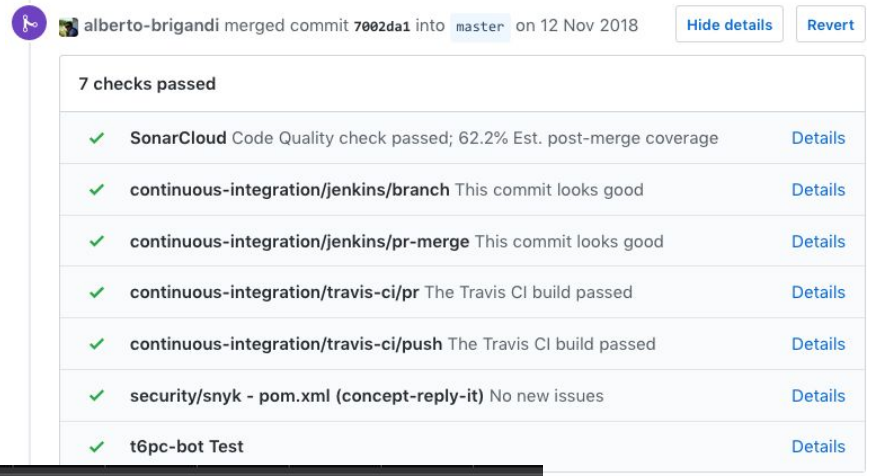


Code coverage

Unit testing

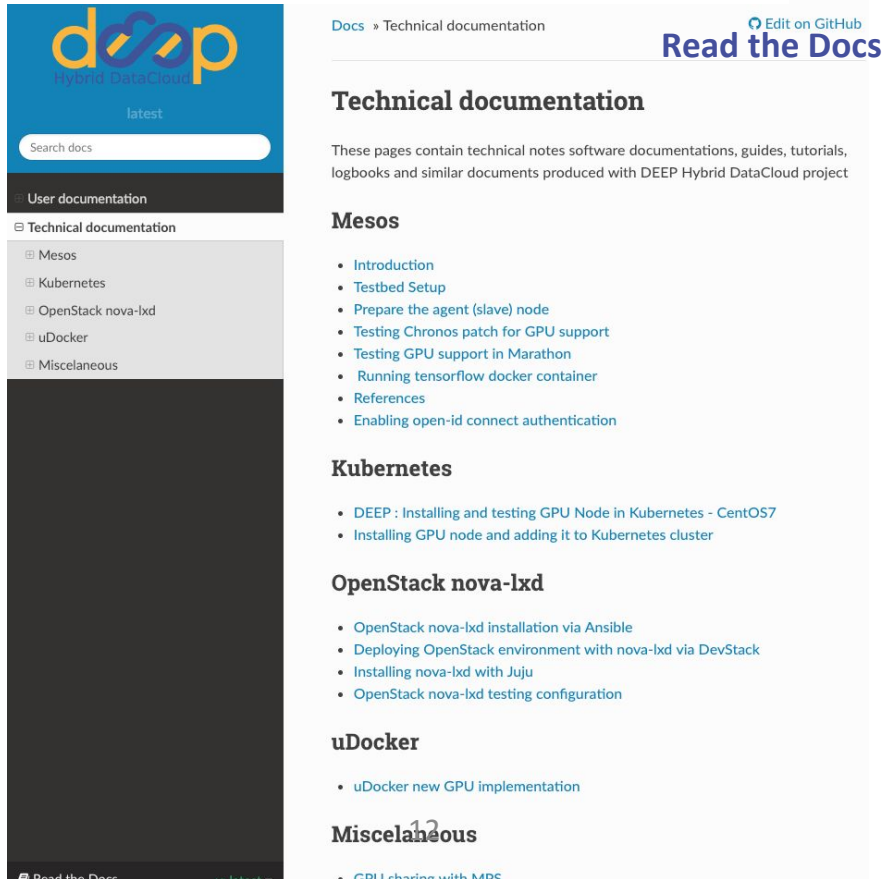


github PR



Bandit (Python): security check

Read the Docs



DEEP core components: CD

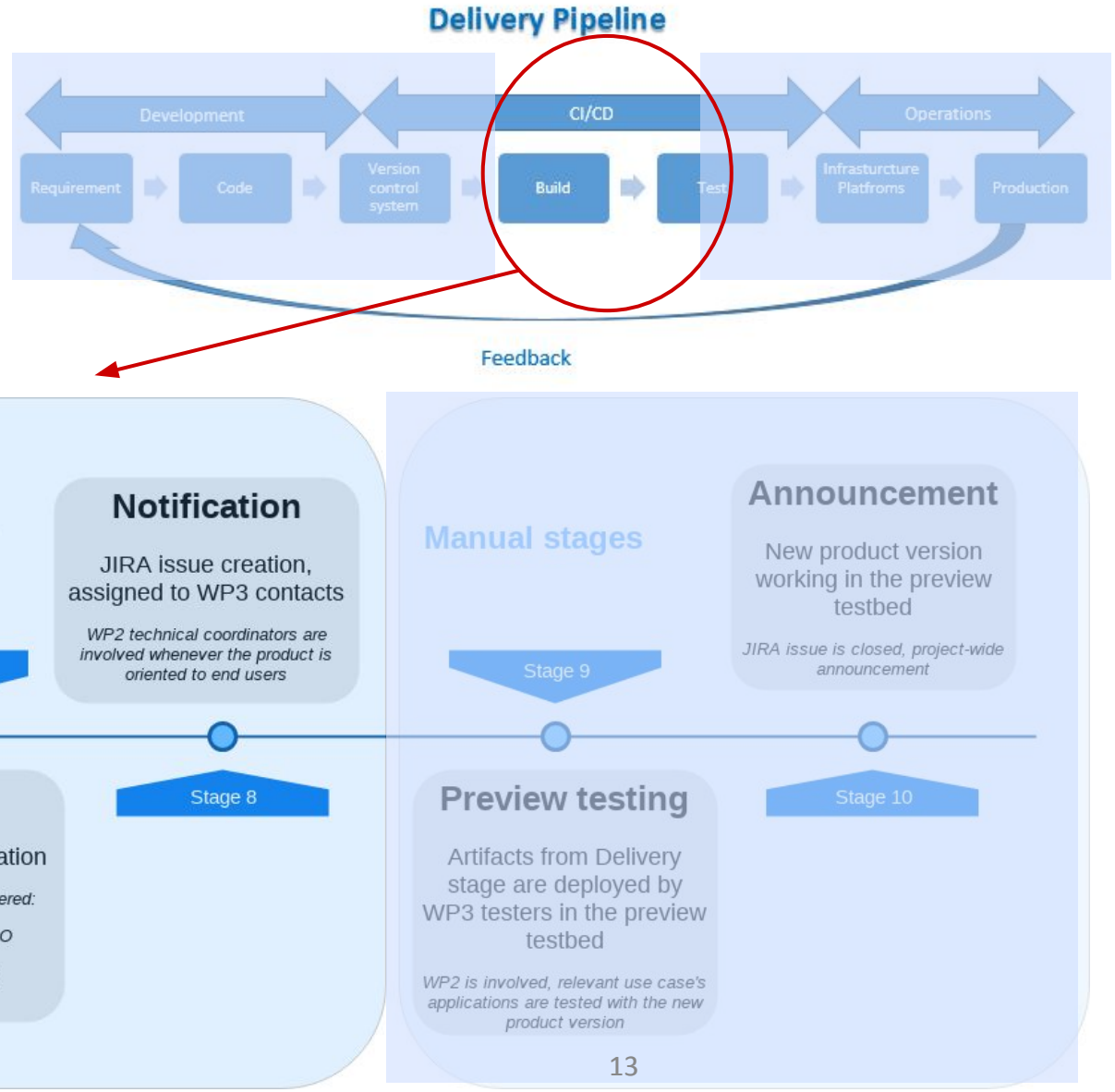
Delivery: **Automatic build**

- RPMs/DEBs.
- Docker images.
- Python packages.

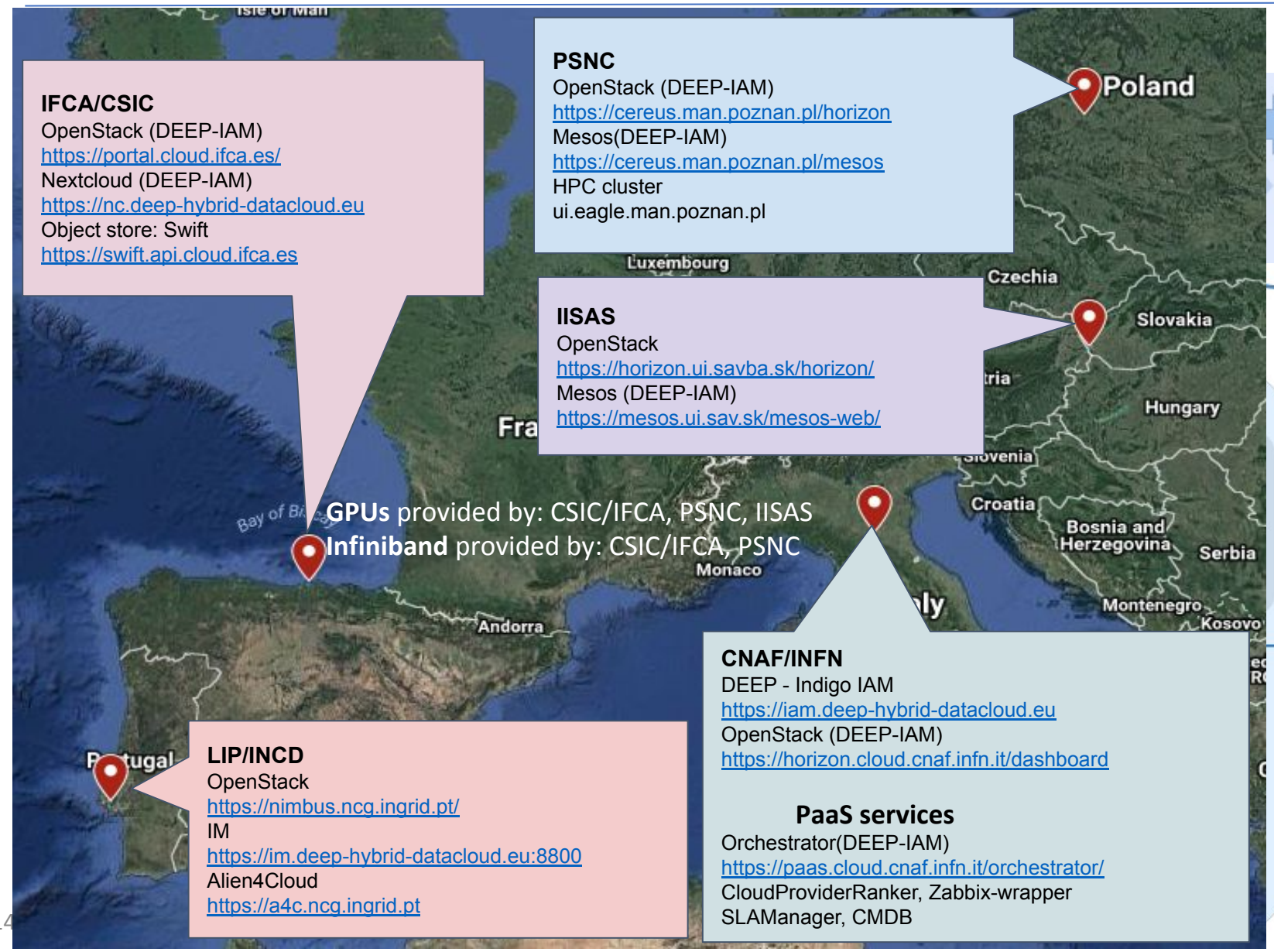
DEEPaaS component → PyPI automatically.

Notification: **Immediate availability** of artefacts:

- **Automatic creation** of JIRA issue.
- Component version and artefact location.
- **Notification** sent to WP3 for **testing**.



DEEP core components: Deployment Pilot Preview



IFCA/CSIC
OpenStack (DEEP-IAM)
https://portal.cloud.ifca.es/
Nextcloud (DEEP-IAM)
https://nc.deep-hybrid-datacloud.eu
Object store: Swift
https://swift.api.cloud.ifca.es

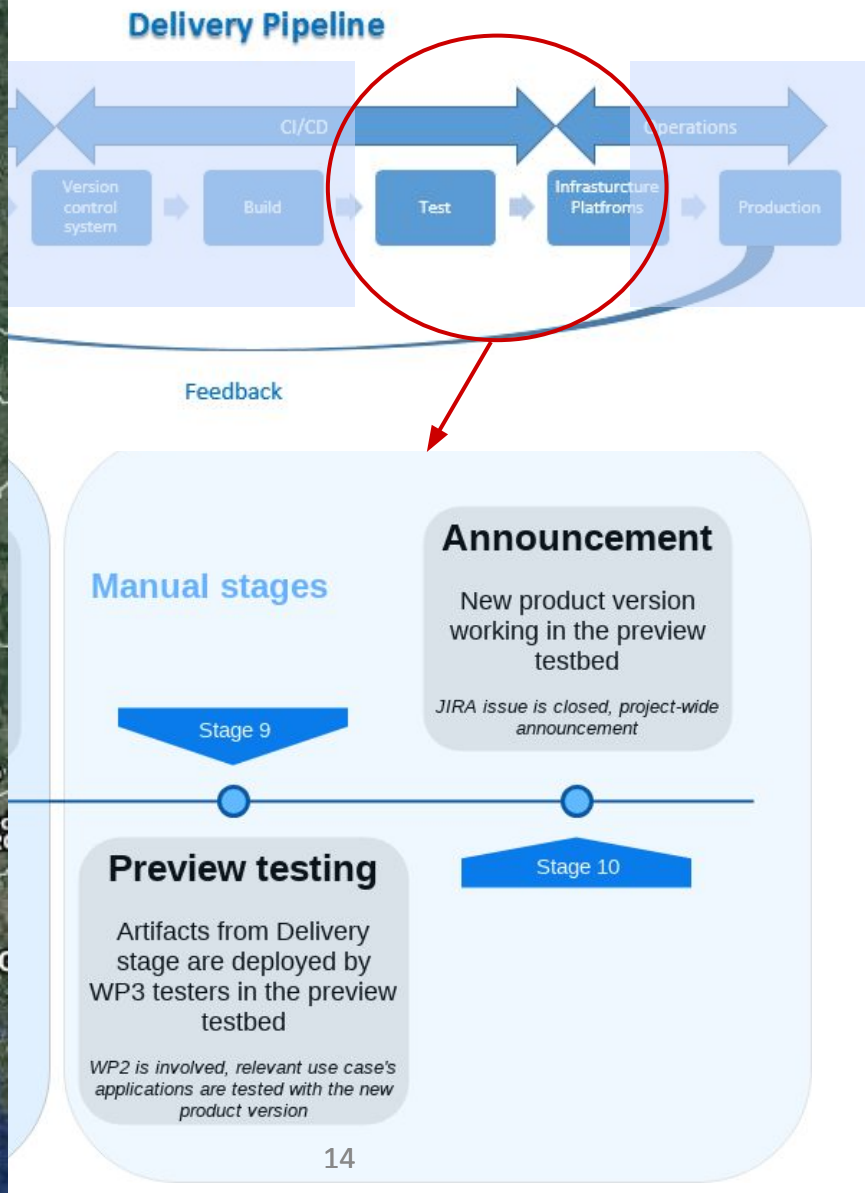
PSNC
OpenStack (DEEP-IAM)
https://cereus.man.poznan.pl/horizon
Mesos(DEEP-IAM)
https://cereus.man.poznan.pl/mesos
HPC cluster
ui.eagle.man.poznan.pl

IISAS
OpenStack
https://horizon.ui.savba.sk/horizon/
Mesos (DEEP-IAM)
https://mesos.ui.sav.sk/mesos-web/

LIP/INCD
OpenStack
https://nimbus.ncg.ingrid.pt/
IM
https://im.deep-hybrid-datacloud.eu:8800
Alien4Cloud
https://a4c.ncg.ingrid.pt

CNAF/INFN
DEEP - Indigo IAM
https://iam.deep-hybrid-datacloud.eu
OpenStack (DEEP-IAM)
https://horizon.cloud.cnaf.infn.it/dashboard

PaaS services
Orchestrator(DEEP-IAM)
https://paas.cloud.cnaf.infn.it/orchestrator/
CloudProviderRanker, Zabbix-wrapper
SLAManager, CMDB

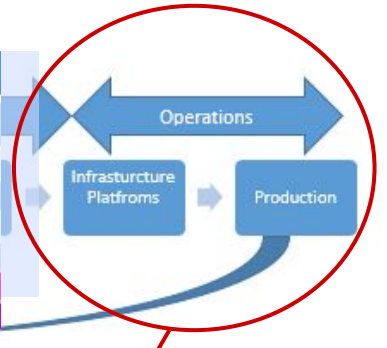


14

DEEP core components: Software releases

Release, Maintenance & support schedule

	Update Type	2018		2019												2020			
		Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr
DEEP-1	Full																		
	Standard																		
	Security																		
DEEP-2	Full																		
	Standard																		
	Security																		

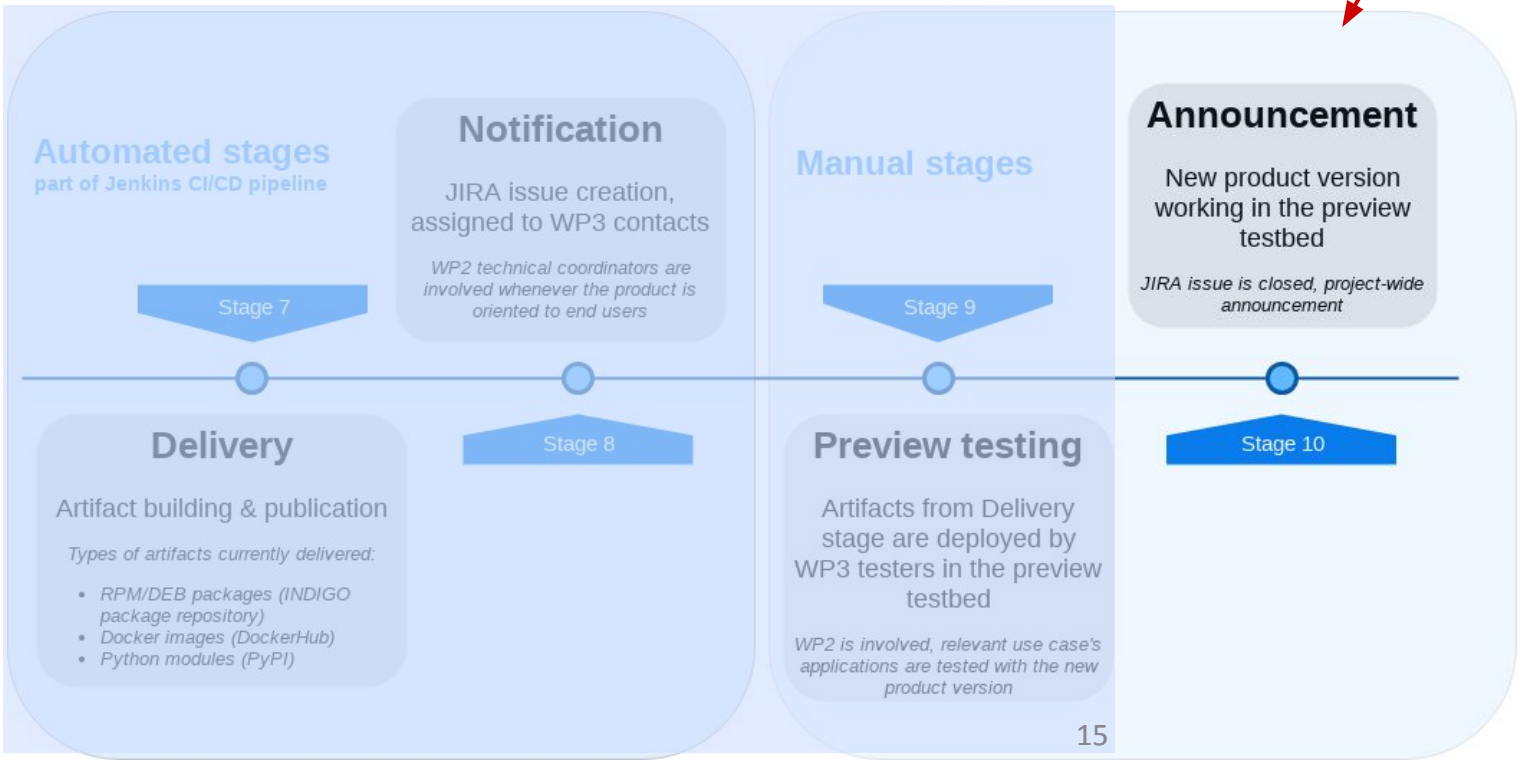


Time-based releases:

2 major releases planned during the project lifetime.

Define periods for **full** and **standard** maintenance, **security** updates and **end-of-life**.

Updates of the components and services are provided as soon as new versions are announced and the SQA pipelines and controls finish successfully.



DEEP-1/Genesis release

DEEP - 1 (Genesis) QC reports

Product	VCS		License	CodeStyle	Unit Testing (%)	Functional Testing	Integration Testing	Docs	Code Review	Automated Deployment	Security	Artefacts
	code	tag										
Alien4Cloud-"plugin"	✓	1.1-r0	✓	✓	✓ = 70 (lowest:55)	★	-	✓	✓	★ = manual	✓	docker image: ✓
Cloud-Info-Provider-DEEP	✓	0.10.4	✓	✓	✓ = 97 (lowest: 86)	✓	-	✓	✓	★ = manual	✓	rpm & debs: ✓
DEEPaaS API	✓	0.1.0	✓	✓	✓ = 99 (lowest 95)	✓	✓	✓	✓	✓	✓	docker image: ✓
IM	✓	1.7.5	✓	✓	✓ = 95 (lowest 75)	✓	✓	✓	✓	✓	★	docker image: ✓ rpm & debs: ✓
PaaS Orchestrator	✓	2.1.1 -FINAL	✓	✓	✓ = 75 (lowest 71)	✓	✓	✓	✓	★ = manual	✓	docker image: ✓
Orchent	✓	1.2.2	✓	★	★	✓	✓	✓	✓	✓	✓	rpm & debs: ✓
TOSCA Types & Templates	✓	3.0.0	✓	N/A	N/A	★	★	★	✓	✓	-	tarballs: ✓
udocker	✓	1.1.3	✓	✓	✓ = 92 (lowest 62)	✓	✓	✓	✓	✓	✓	tarballs: ✓
vRouter	✓	DEEPv1	✓	-						✓		Ansible playbook

Upstream contributions

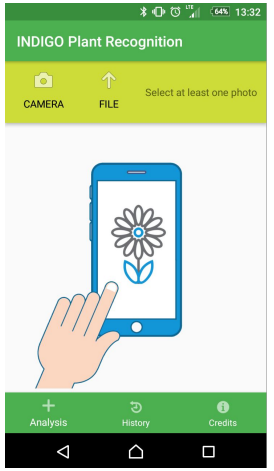
Upstream Code	Link	Partner
Tosca-parser	https://github.com/openstack/tosca-parser/commit/3af43cb9a88863ccb7f8991ca379163fe23f33e3 https://github.com/openstack/tosca-parser/commit/23cd991e884b67d41378a69ee30ef38b5d760e68 https://github.com/openstack/tosca-parser/commit/c08022d0b71ca7936e084bc40805af0b89b724ae	UPV
Apache-libcloud	https://github.com/apache/libcloud/pull/1215 https://github.com/apache/libcloud/pull/1242 https://github.com/apache/libcloud/pull/1269	UPV
Apache OpenWhisk	https://github.com/apache/incubator-openwhisk-devtools/pull/165 https://github.com/apache/incubator-openwhisk-devtools/pull/162 https://github.com/apache/incubator-openwhisk-devtools/pull/161	CSIC
cloud-info-provider	https://github.com/EGI-Foundation/cloud-info-provider/pull/137 https://github.com/EGI-Foundation/cloud-info-provider/pull/126 https://github.com/EGI-Foundation/cloud-info-provider/pull/119	CSIC

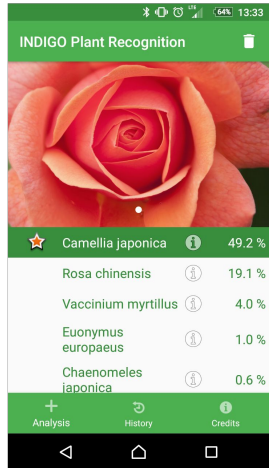
User communities and applications

Citizen Science: Plant classification Image Classification

For training and testing image classifiers (CNNs; TensorFlow).
From this model the following services are derived:

- Plants (dataset: up to 1 TB)
- Conus marine snails
- Seeds
- Phytoplankton

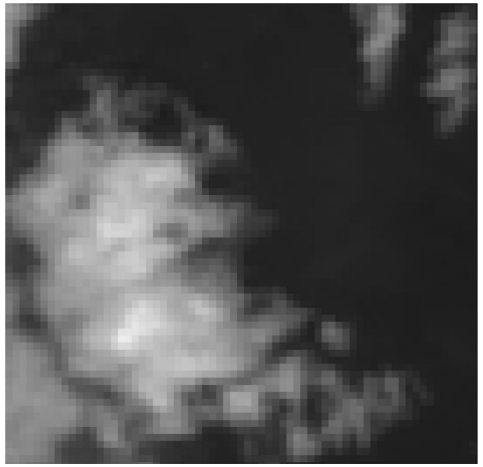


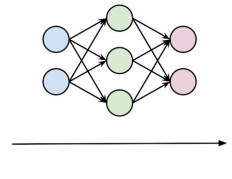


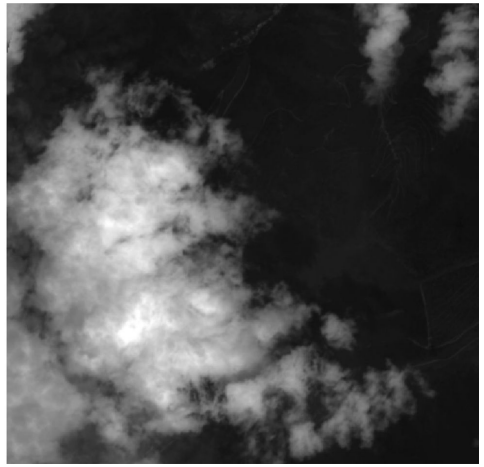
Earth Observation: Satellite Imagery

Explore application of **Machine Learning** for satellite imagery (e.g. remote object detection, terrain segmentation, meteorological prediction).

Currently being developed is **super-resolution service** to upscale low resolution bands to high resolution with deep learning (e.g. DSen2; TensorFlow. Dataset: ca. 1 TB)







User communities and applications

Hybrid DataCloud

Biological and Medical Science: Retinopathy

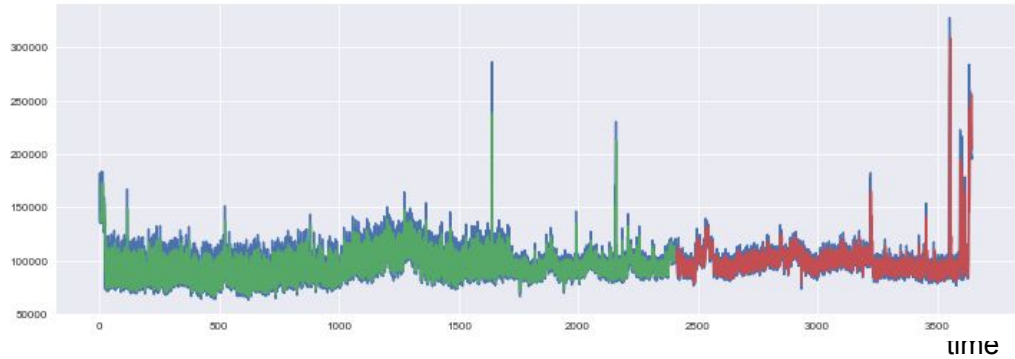
Diabetic retinopathy is a fast-growing cause of blindness worldwide. The use-case focuses on a deep learning approach (CNNs; TensorFlow) to automated classification of retinopathy based on color fundus retinal photography images (DR=0 (=healthy) .. 4 (blind)).
Dataset: ca. 100 GB

DR=0

DR=4

Computing Security:
Massive Online Data Streams: Online analysis of data streams

Intrusion detection systems: provide an architecture able to analyze massive on-line data streams, also with historical records, in order to generate alerts in real-time. Based on proactive time-series prediction adopting artificial neural networks (e.g. LSTM, GRU; TensorFlow). Dataset: 100 GB currently, then up to 2 TB /day

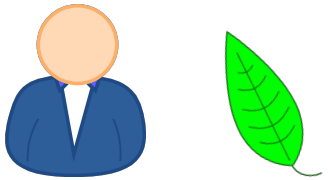


dataset, prediction (train), prediction (test)

Physics: Post-processing

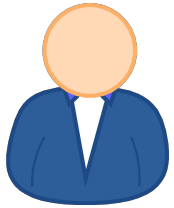
Of HPC simulations (Lattice QCD): analysis of a large number of configurations for Lattice QCD simulation. Move the configurations to long-term storage, perform checks and metadata operations. Requirements: Infiniband, data of 1 TB

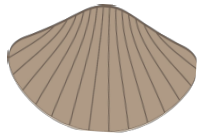
DEEP serves different users' profiles



Unskilled* **Users**: want to use a trained deep neural network for prediction/classification of their own data
→ no expert knowledge of ML
→ no access to high-level computing resources

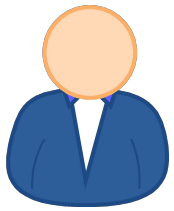
docker
DEEPaaS API

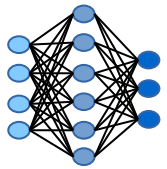




Intermediate* **Users**: want to use a trained deep neural network and adapt it for solving their problem (transfer learning)
→ some knowledge of ML
→ access to limited computing resources

eXtreme DataCloud
docker
ONEDATA
dCache
DEEPaaS API



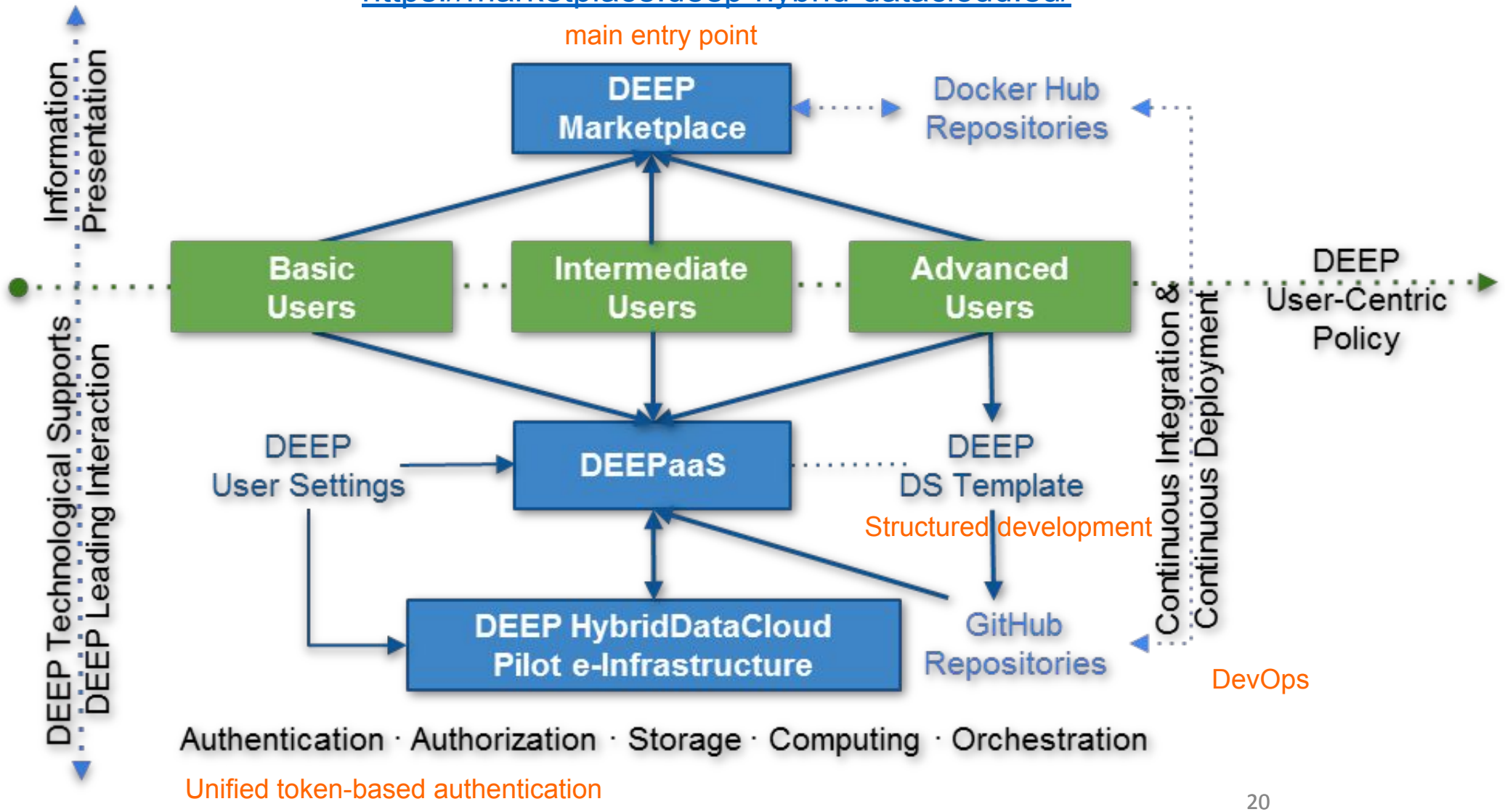


Advanced* **Users**: want to develop their own deep neural network (with special requirements as e.g. data privacy)
→ expert knowledge of ML
→ need access to high-level computing resources

eXtreme DataCloud
docker
ONEDATA
dCache
COOKIECUTTER
DEEPaaS API

(*) - knowledge level

DEEP from a user point of view



DEEP CI/CD for user applications

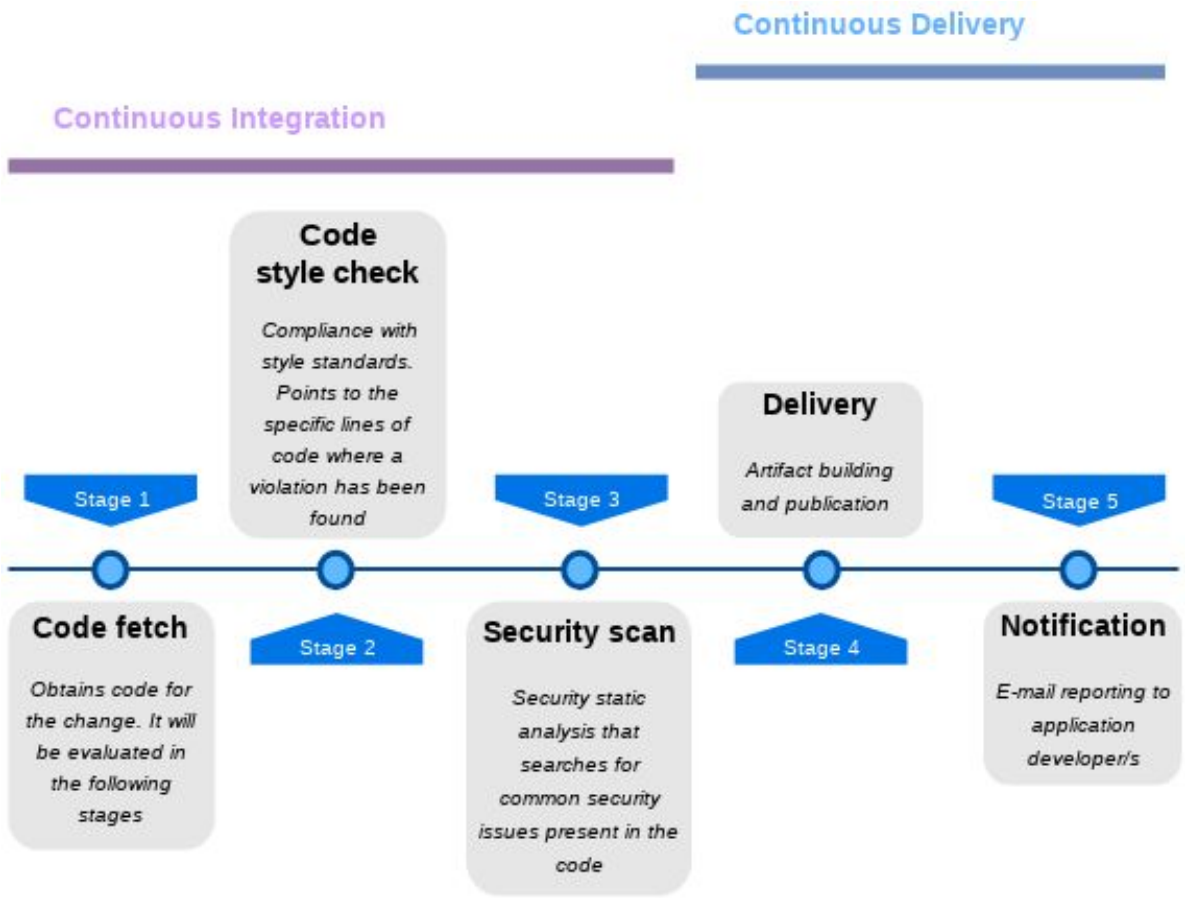
Development and **integration** in **cloud** resources.

Jenkins pipeline for user applications:

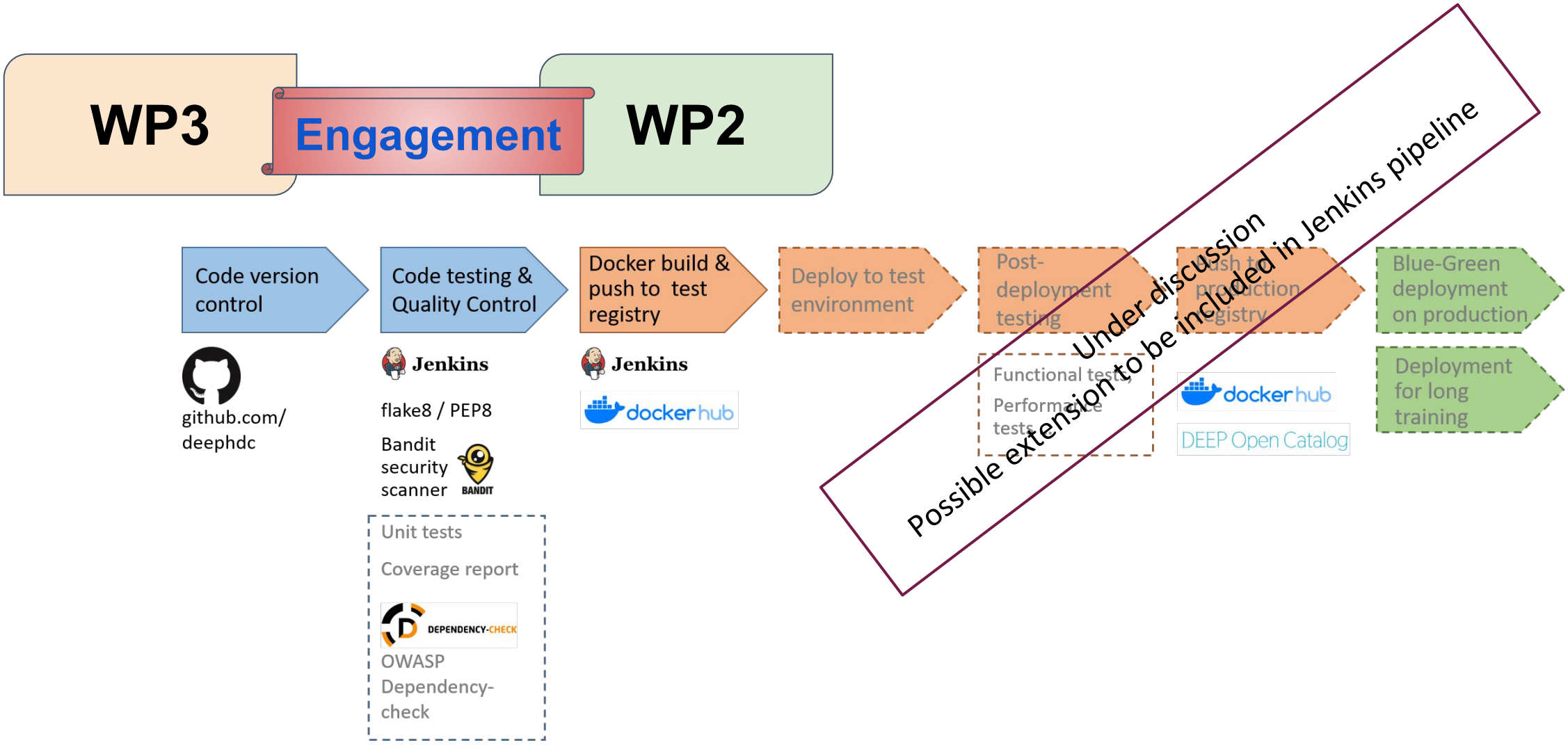
CI/SQA: **Code style**, **security scan**.

CD:

Immediate availability of application.
Automatic building (Docker images).
Automatic publishing (Docker Hub).
Notification (email to developers).



DEEP user applications



Achievements: users perspective

- **Encapsulation and isolation** of different environments (using container)
  - Enable operativity in different infrastructure
- **DEEPaaS** as an entry points with **flexible design**
  - Allow different training arguments for each use-case
- DEEP DS **template** and DEEPaaS **API**
  - ready-made template that facilitates standardisation and (semi)automatic creation of necessary files and codes

Achievements: users perspective

- DEEP **OC software automation** DevOps: CI/CD pipelines for user applications
- DEEP Leading Interaction
  - A good way to provide technical support for users as well as accompanying "HowTo" documentation with details
- Udocker extended support
  - Older Linux kernels are also supported and does not require root privileges
  - Functionalities were critically valuable in test running docker images at an old local cluster.
- Orchestrator Dashboard: WebUI where users can submit TOSCA templates to the orchestrator

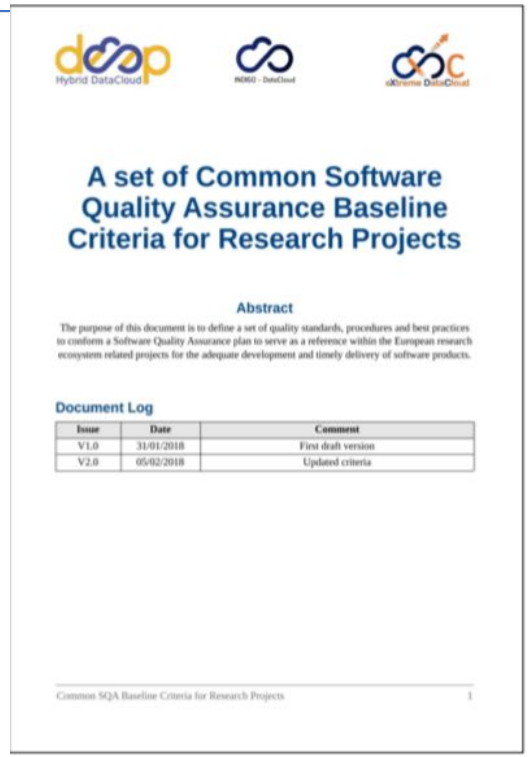
Use cases status

	Problem	Goal	DEEP services	Status
Plant Classification	Automatically identify plant species from images using Deep Learning	Perform image classification on different datasets by performing the so called transfer learning	Tensorflow, Keras via DEEP PaaS Orchestrator, OIDC-agent	Deployed through orchestrator and oidc-agent in the DEEP testbed
Satellite Imagery	Explore possible applications of machine learning techniques to satellite imagery from different sources	Support a super -resolution service to upscale low resolution bands to high resolution with Deep Neural Networks	Tensorflow, Keras, Pytorch via DEEP Paas Orchestrator, OIDC-agent, Alien4Cloud	Only DEEPaas API available Will be implemented in the DEEP testbed via A4C
Retinopathy	Automated classification of retinopathy based on color fundus retinal photography images	Perform deep learning approach for image classification	Tensorflow via DEEP PaaS Orchestrator, OIDC-agent	Deployed through orchestrator and OIDC-agent in the DEEP testbed
MODS	Intrusion detection systems: provide an architecture able to analyze massive on-line data streams, also with historical records, in order	Generate alerts in real-time using ML and DEEP learning approaches.	Tensorflow, Keras via DEEP PaaS Orchestrator, OIDC-agent, Alien4Cloud	Deployed through orchestrator and OIDC-agent in the DEEP testbed. To be implemented via A4C

DEEP-HybridDataCloud highlights

DEEP vision & work on Software Quality Assurance

- **Vision:**
  - We support the HLEG vision on **delivering quality software for the EOSC**
  - We produced ***"A set of common software quality assurance baseline criteria for research projects"***
    - http://hdl.handle.net/10261/160086
    - Done together with the **eXtreme DataCloud** and **INDIGO** projects
    - On 2019: **Open** for collaborations, collaborative document: https://indigo-dc.github.io/sqa-baseline/
  - **Objective**: ***align baseline criteria within different projects***
- **Work**:
  - Current **EOSC synergies** (eXtreme-DataCloud)
    - SQA baseline
    - **Automation**: continuous integration and delivery for core products
      - Common library for CI/CD pipeline functionalities
    - **Agile** software development: jump-started from WP2 requirements
  - **DEEP goes beyond**: automation techniques supporting user communities
    - **Continuous integration** and **delivery** pipelines in place: Docker Hub images re-creation triggered by changes in i) DEEPaaS and ii) application itself
    - Initial **continuous deployment prototype**: readiness/provision of training and inference as a service
    - Rendering and generation of the **marketplace portal:** leveraging (JSON) schema- validated metadata descriptions


A set of Common Software Quality Assurance Baseline Criteria for Research Projects

Abstract

The purpose of this document is to define a set of quality standards, procedures and best practices to conform a Software Quality Assurance plan to serve as a reference within the European research ecosystem related projects for the adequate development and timely delivery of software products.

Document Log

Issue	Date	Comment
V1.0	31/01/2018	First draft version
V2.0	05/02/2018	Updated criteria

Common SQA Baseline Criteria for Research Projects 1

DEEP-Genesis: 1st platform and release

- First software release and prototype platform released **Jan. 2018**
- More than 12 software components, 4 different services, several upstream contributions, more than 10 models in marketplace
- DEEP-Genesis: **initial service catalog**
  - All services are OIDC-ready and follow the AARC and AARC2 blueprint recommendations

Service	Functionalities	Preview endpoint
Visual application topology composition and deployment	• Graphical composition of complex application topologies • Deployment through PaaS orchestrator	https://a4c.ncg.ingrid.pt
ML/DL training facility as a service	• Provide continuous training and retraining of developed models	On demand
DEEP as a Service	• Deployment of DEEP Open Catalog components as server-less functions • Provide inference/prediction endpoints	(beta, internal preview only) https://deep.cloud.ifca.es/
DEEP Open Catalog	• Ready-to-use machine learning and deep learning applications, including: ➢ Machine learning frameworks + JupyterLab ➢ ML/DL ready to use models ➢ BigData analytic tools	https://marketplace.deep-hybrid-datacloud.eu

Collaboration with ongoing initiatives

- Collaboration with **EINFRA-21** projects:
  - **eXtreme-DataCloud**: Integration of data management solutions (XDC) and computing solutions (DEEP), exploiting event driven executions. Work on software quality.
  - **DARE**: Provide ML/DL services to integrate into workflows
- Collaboration with **other** initiatives:
  - **EOSC-Hub**: integration of developments into production tools (cloud-info-provider, TOSCA-Parser).
  - **EGI.eu**: Improved support for accelerators in Cloud resources
  - Developments merged upstream
- Collaborations with **external** user communities:
  - Royal Botanical Garden of Madrid, LifeWatch ERIC, Mouse Motor Lab (Rowland Institute Harvard), Centre for Automatic and Robotics (CSIC)

Selected DEEP early results

- G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. López García, I. Heredia, P. Malík, and L. Hluchý. "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey". In: Artificial Intelligence Review (Jan. 2019). ISSN: 1573-7462. DOI: 10.1007/s10462-018-09679-z
- User communities publications:
  - N. Tran, T. Nguyen, B.M. Nguyen, G. Nguyen. "A multivariate fuzzy time series resource forecast model for clouds using LSTM and data correlation analysis". Procedia Computer Science, Elsevier, 2018, Volume 126, pp. 636-645, ISSN 1877-0509. DOI 10.1016/j.procs.2018.07.298
  - G. Nguyen, B.M. Nguyen, D. Tran, L. Hluchý. "A heuristics approach to mine behavioural data logs in mobile malware detection system". Data & Knowledge Engineering, Elsevier, 2018, Volume 115, pp. 129-151, ISSN 0169-023X, DOI 10.1016/j.datak.2018.03.002
  - B. M. Nguyen, H. Phan, D. Q. Ha, G. Nguyen. "An Information-centric Approach for Slice Monitoring from Edge Devices to Clouds", Procedia Computer Science Volume 130, 2018, Pages 326-335. DOI: 10.1016/j.procs.2018.04.046
- Published articles by user communities not in the project, exploiting DEEP-HybridDataCloud software components:
  - I. Heredia Cacha. Application of a Convolutional Neural Network for image classification to the analysis of collisions in High Energy Physics. CHEP 2018 Conference, Sofia, Bulgaria. Oral Contribution
  - L. Lloret; I. Heredia; F. Aguilar; E. Debusschere; K. Deneudt; F. Hernández. Convolutional Neural Networks for Phytoplankton identification and classification. Biodiversity Information Science and Standards. 2018. Oral Contribution
  - F. Pando; I. Heredia; C. Aedo; M. Velayos; L. Lloret; J. Calvo. Deep learning for weed identification based on seed images. Biodiversity Information Science and Standards. 2018. Oral Contribution

Thank You

Next steps: users perspective

- Serverless framework planned to be developed
  - based on OpenWisk platform
- Further development of FLAAT:
  - **FLA**sk support for handling OIDC **A**ccess **T**okens
- Extend the CI/CD pipeline:
  - Include deployment, testing of produced application Docker images from its DockerHub repository.
- General improvements:
  - Improve documentation based on feedback from users.
  - Perform training actions: through dedicated video conferences and webinars
  - Individual support through 1-to-1 TeleConferences
- Development Docker Image (DDI)