Introduction to Cloud Computing, VRE Data Analysis Tools (JupyterLab, CCP Method Executions)

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

What is Cloud Computing

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

- USA National Institute of Standards and Technology
- Resources: networks, servers, storage, application and services
- Minimal management effort/interaction

...our definition of Cloud Computing:

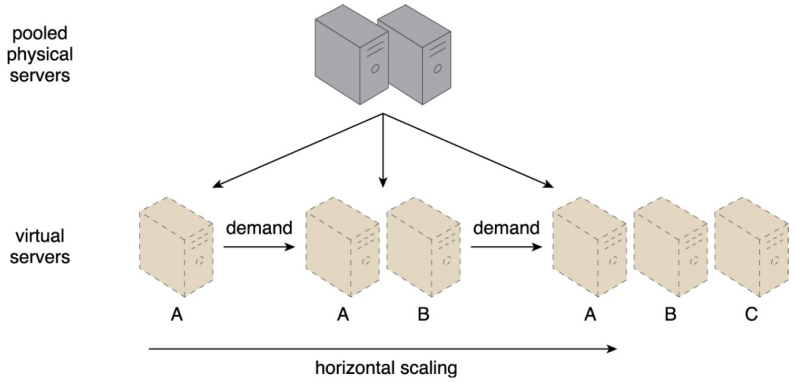
"Cloud computing allows users to run their scientific and collaborative applications in a web-based environment without the need to install software locally or worry about managing the underlying infrastructure. This approach facilitates data sharing, distributed processing, and collaboration among research groups."

The Benefits of Cloud Computing

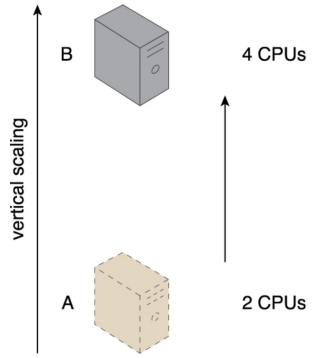
- Reduction or outright elimination of up-front IT investments, namely hardware and software purchases and ownership costs.
- On-demand access to pay-as-you-go computing resources on a short-term basis (such as processors by the hour), and the ability to release these computing resources when they are no longer needed.
- The perception of having unlimited computing resources that are available on demand, thereby reducing the need to prepare for provisioning.
- The ability to add or remove IT resources at a fine-grained level, such as modifying available storage disk space by single gigabyte increments.
- Abstraction of the infrastructure so applications are not locked into devices or locations and can be easily moved if needed.

IT-Resources Scaling

- Scaling represents the ability of the IT resource to handle increased or decreased usage demands.
- Horizontal Scaling: scaling out and scaling in

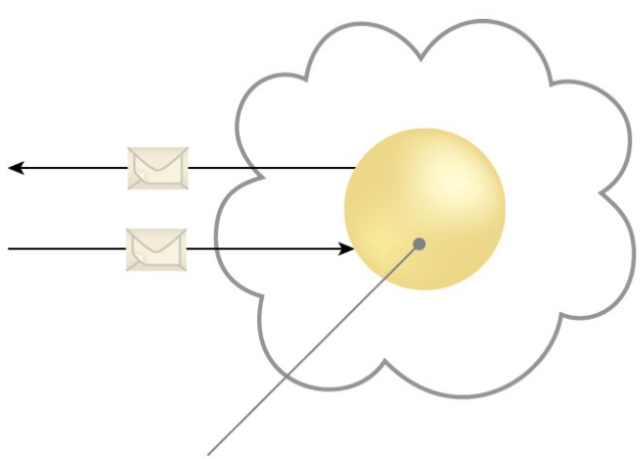


- Vertical Scaling: scaling up and scaling down

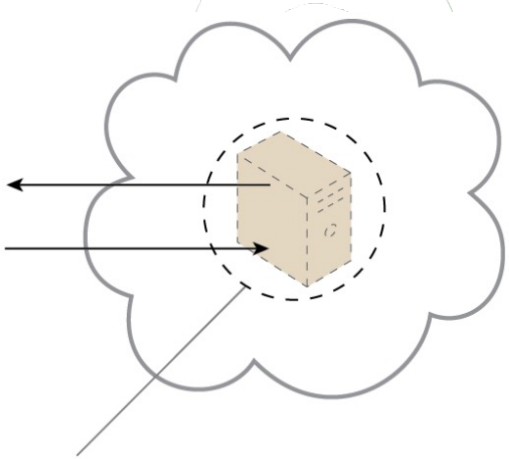


What is a Cloud Service

- A cloud service is an IT resource that is made remotely accessible via a cloud.
- It can be as a simple as a Web-based software program with a technical interface invoked via the use of a messaging protocol
- Or as a remote access point for administrative tools or larger environments and other IT resources.



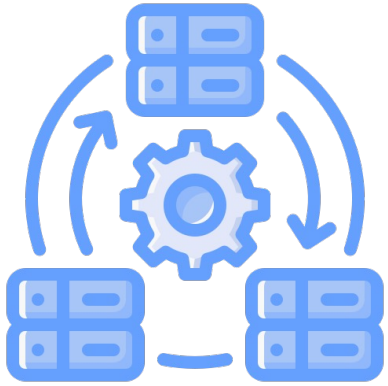
remotely accessed Web service
acting as a cloud service



remotely accessed virtual server
acting as a cloud service

Delivery Models for Cloud

- A Cloud delivery model is a specific, pre-packaged combination of IT resources offered by a cloud provider
- Three are the most common delivery models:
  - ✓ Infrastructure-as-a-Services (IaaS)
  - ✓ Platform-as-a-Services (PaaS)
  - ✓ Software-as-a-Services (SaaS)
- Function-as-a-Service (FaaS) as evolution of the IaaS and SaaS
  - ✓ Allows consumer to deploy “functions” that are executed in (Docker) containers on the cloud provider infrastructure
  - ✓ Very useful for the provider: fine-grain balancing of the resources
  - ✓ Very useful for the consumer: easy to develop little functions that can be grouped and interact directly with the cloud services



ANALYTICS Computing FRamework

Analytics Frameworks

JupyterLab

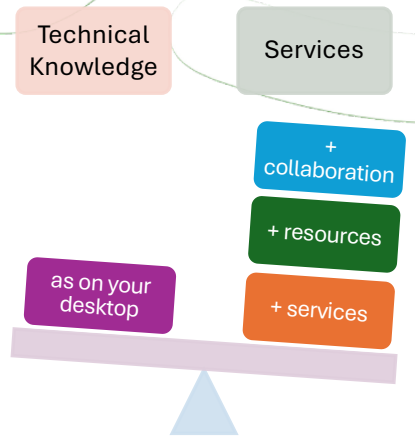
Rstudio 4+

Analytics Engines

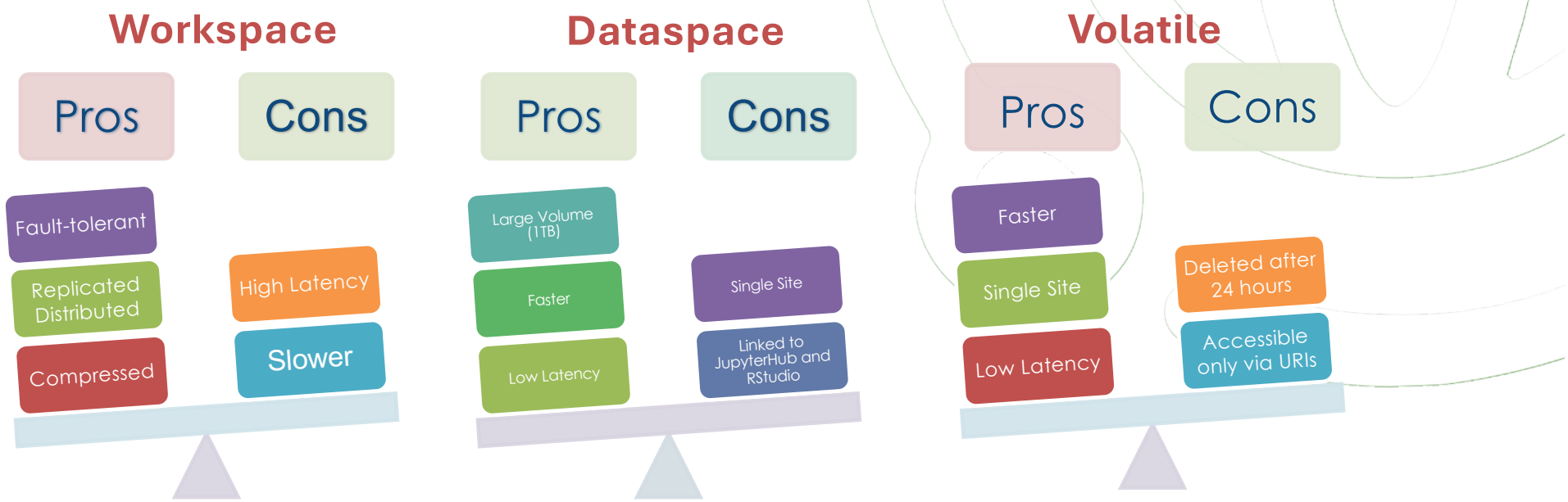
- Data Miner
- Cloud Computing Platform, CCP (the new engine)

One place for executing analysis and processes

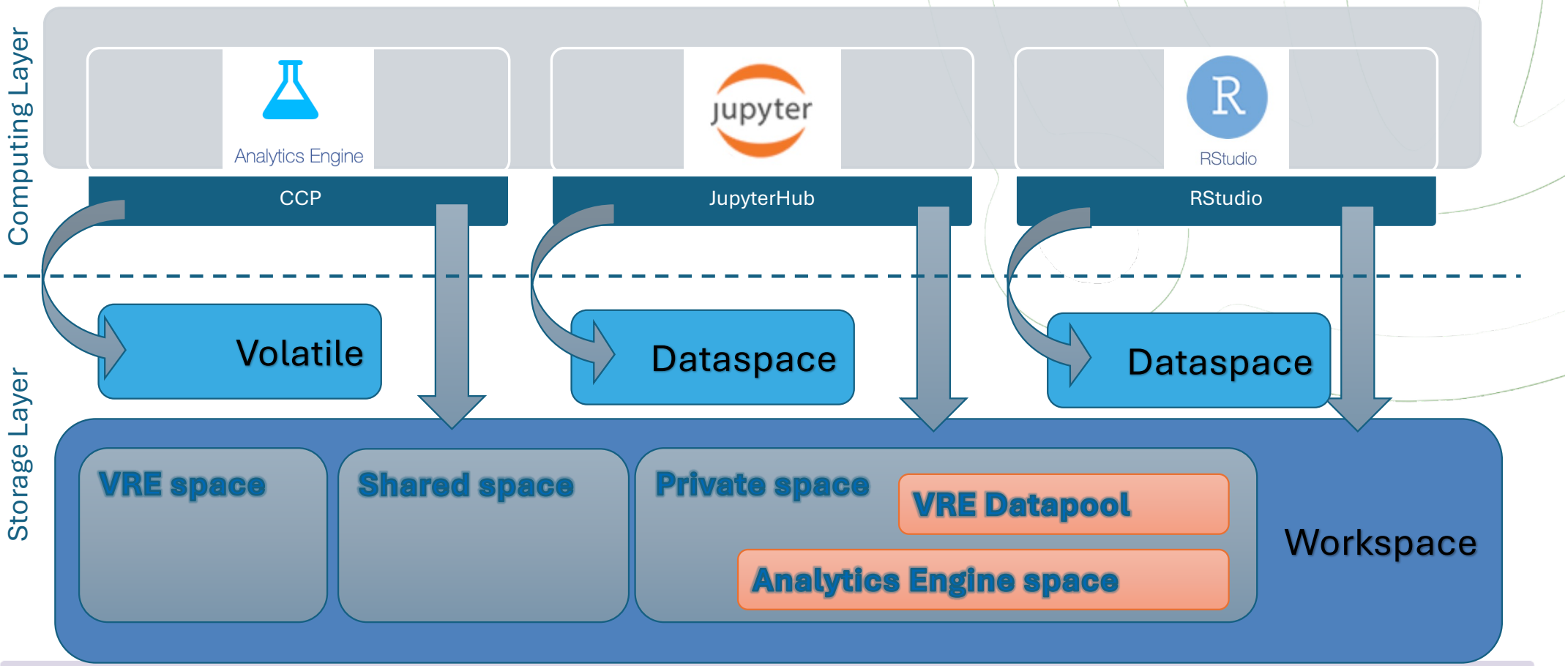
- interactive notebooks via JupyterHub and community-specific applications delivered as a Docker container extend the Analytics framework



Analytics Framework – Storage Layer



Data Analytics Options and Opportunities



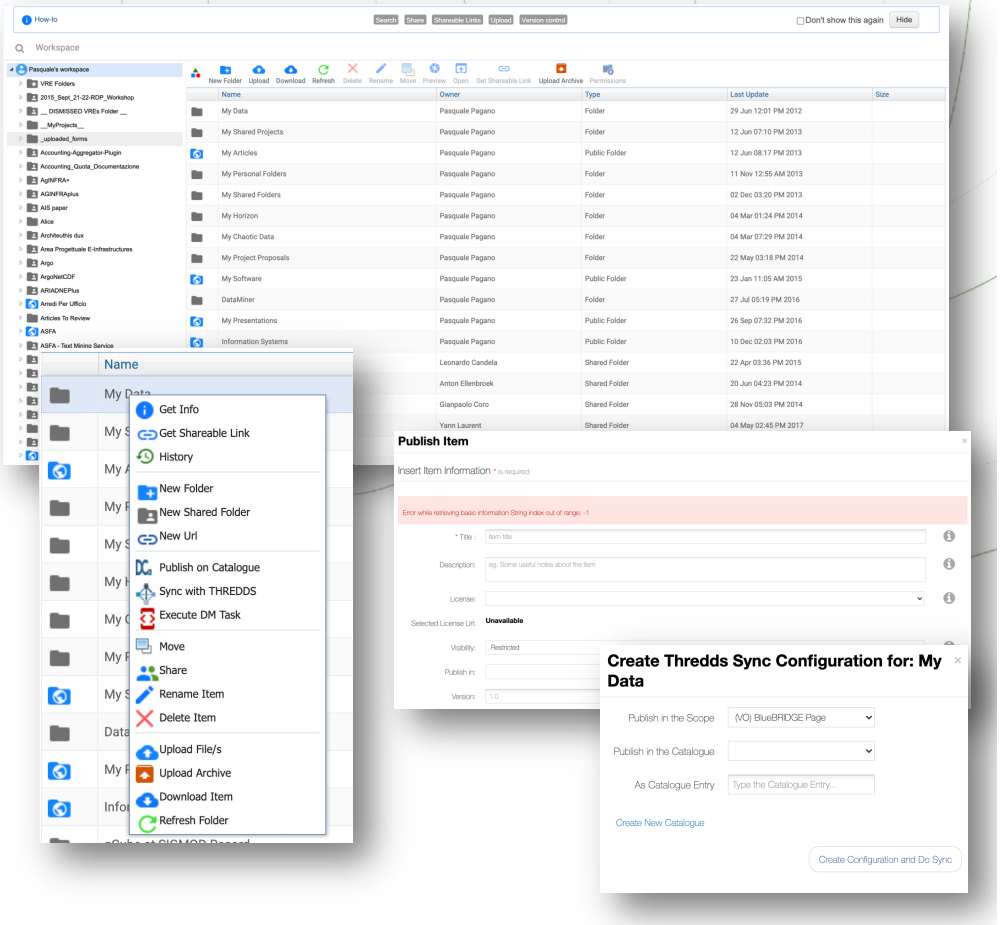
One Workspace

Just one working area across all different services

- VLab
- Rstudio
- JupyterHub notebooks
- Analytics Engine

Just one working area where to store data, script, notebooks, computational methods

- Private
- Shared with your colleagues
- Shared with all Vlab members



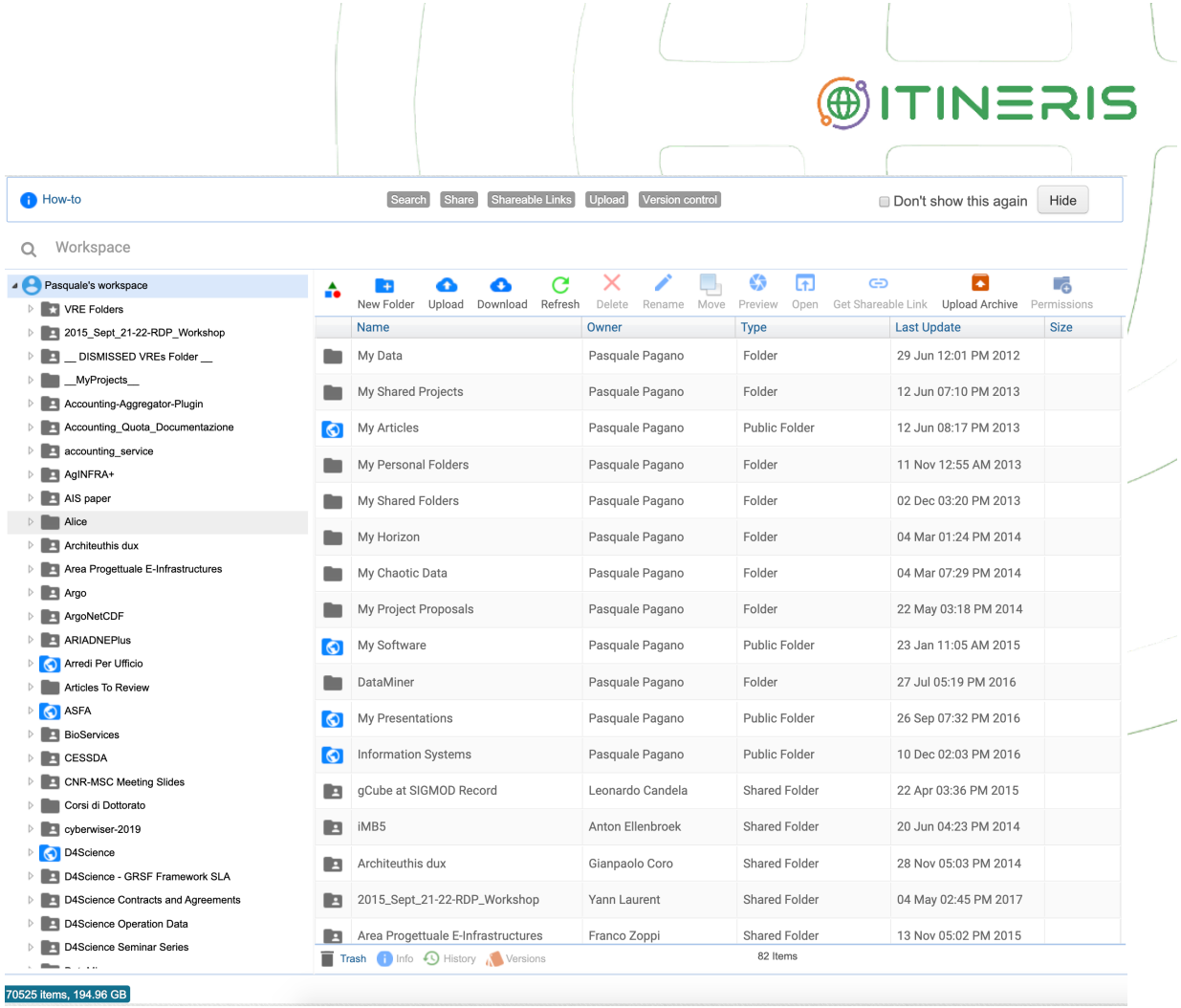
One Workspace

Resembles a **typical file system**

with **files organised in folders**,

yet it supports

files, datasets, software methods, workflows, maps, …



Workspace Folder Types

Private Folders

- Only you can access the files you upload

Shared Folders

- You decide who can access the files (Selective Sharing)

VRE Folders

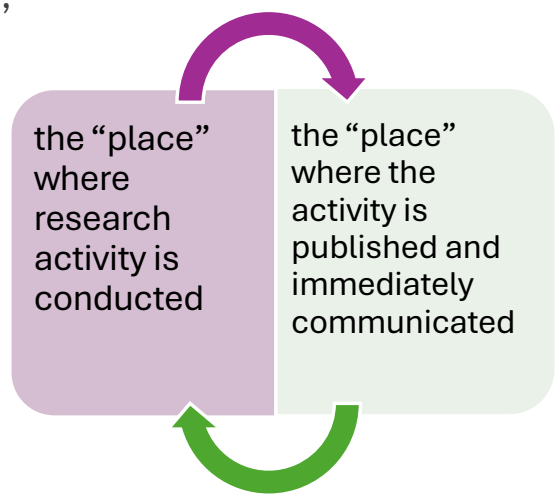
- All VRE members can access the files (Controlled Sharing)

Public Folders

- Anyone with the link can access the files

D4Science – Analytics Framework

- Offers a User-friendly ***data analytics platform*** with on-demand and scalable ***computing resources*** on the Cloud;
- Allows to ***import*** and ***execute*** methods, to ***collaborate*** by ***sharing*** methods, method executions, method inputs and results with colleagues;
- Fosters better collaboration, ***enabling teams to work together*** from anywhere with an internet connection and access the same data and applications;



How does the D4Science Analytics Framework support Open Science?

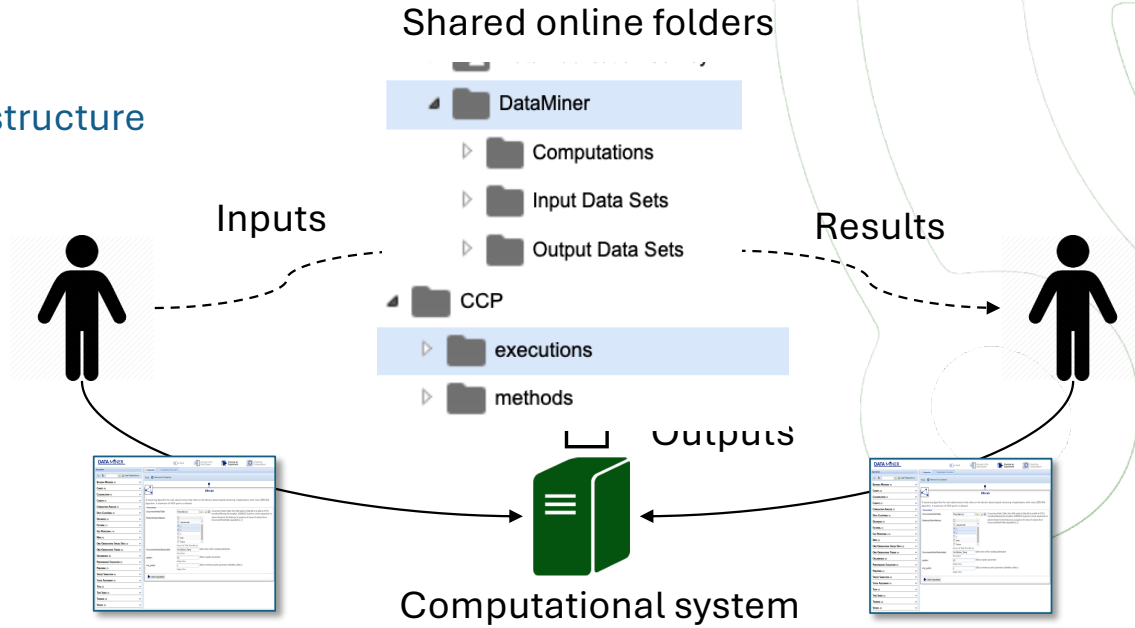
What is Open Science
An approach to the scientific process based on cooperative work and ways of disseminating knowledge, improving accessibility to and re-usability of research outputs by using digital technologies and collaborative tools.

SOURCE: COMMISSION RECOMMENDATION (EU) 2018-790 of 2018

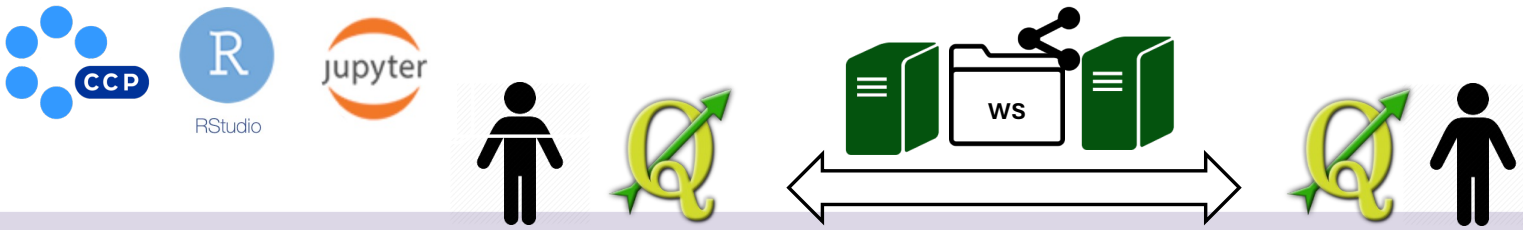
- Uses digital technologies and collaborative tools
- Enables the *repeatability* and *reproducibility* of method executions
- Enables *re-usability of methods and their outputs*
- Improves transparency and ways of disseminating knowledge

D4Science - Collaborative experiments

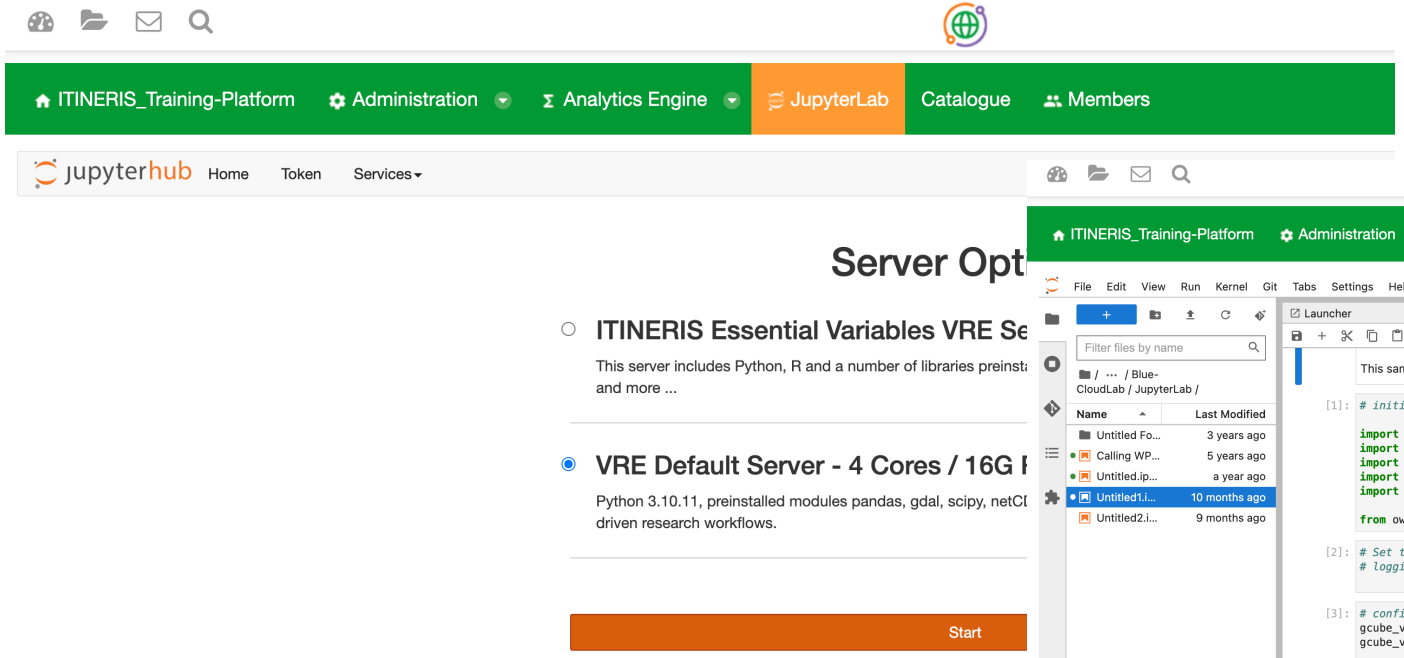
In the e-Infrastructure

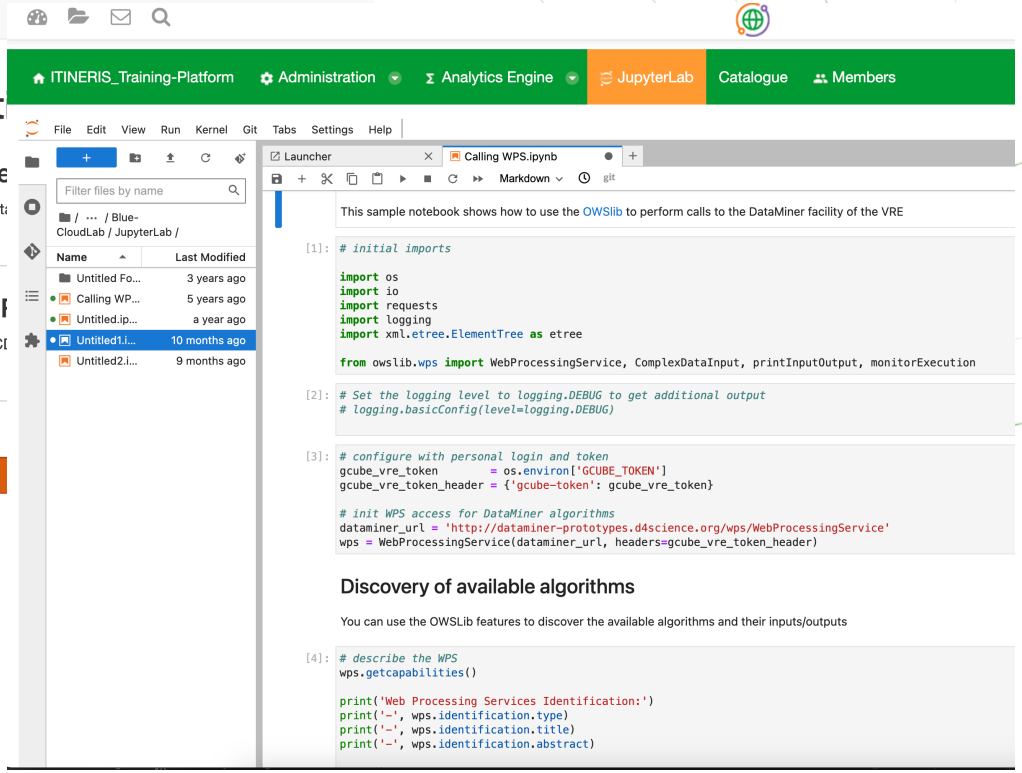


Through third party software (Via OGC Standards WPS or OGC API)



JupiterLab platform benefits

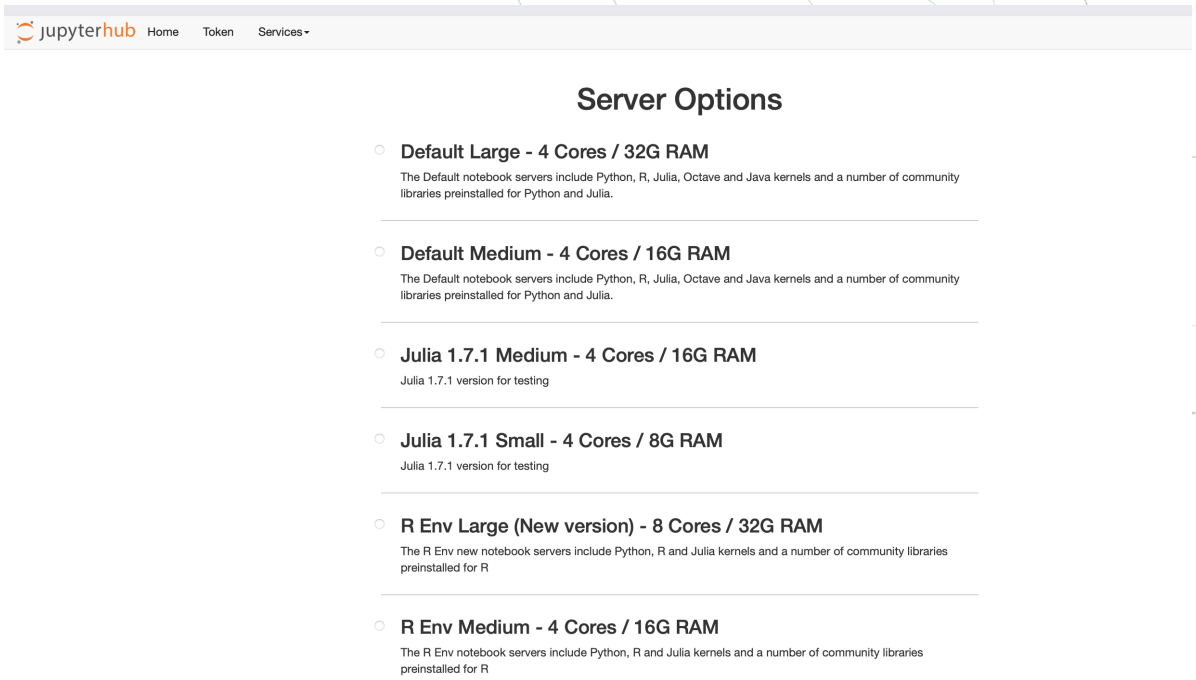




JupiterLab platform benefits

- The Server Options can be customised differently for each Vlab / WB

  - Server image and required libraries
  - No. of CPU Cores
  - RAM Memory avail.



Rstudio platform benefits

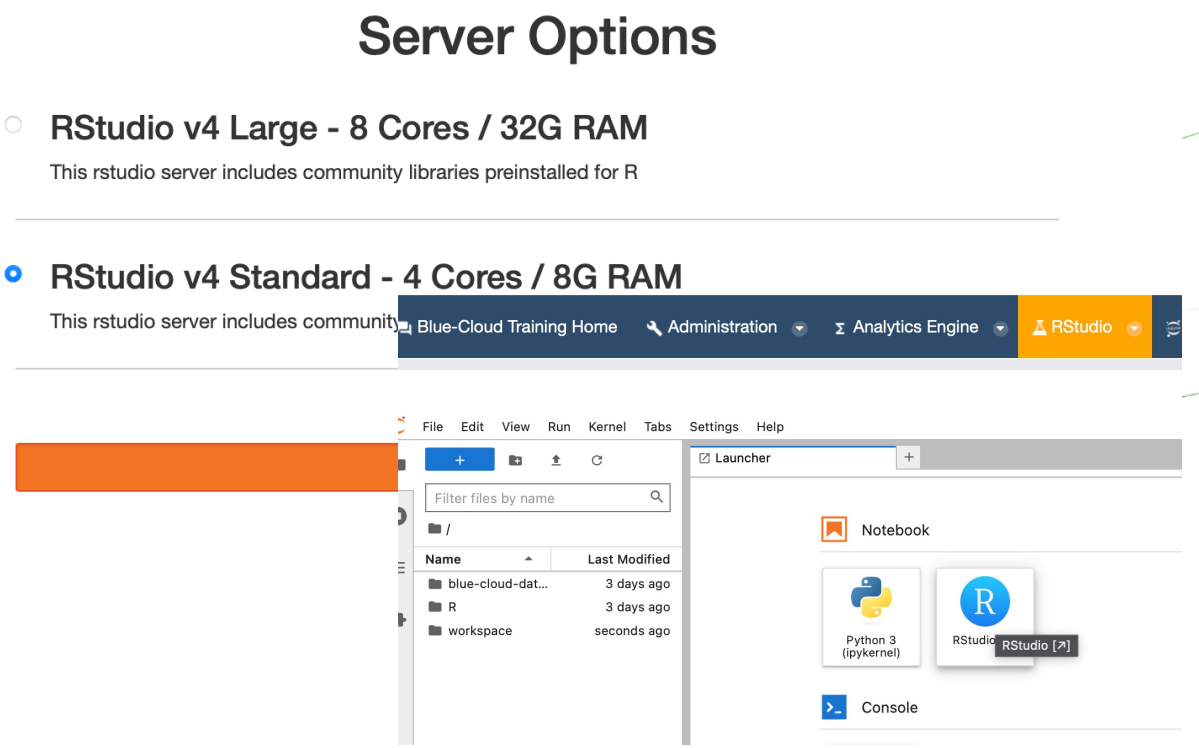
RStudio

- RStudio server **no longer shared,** dedicated instance spawned on demand
- Each VLab can define its RStudio servers configuration, i.e. select packages

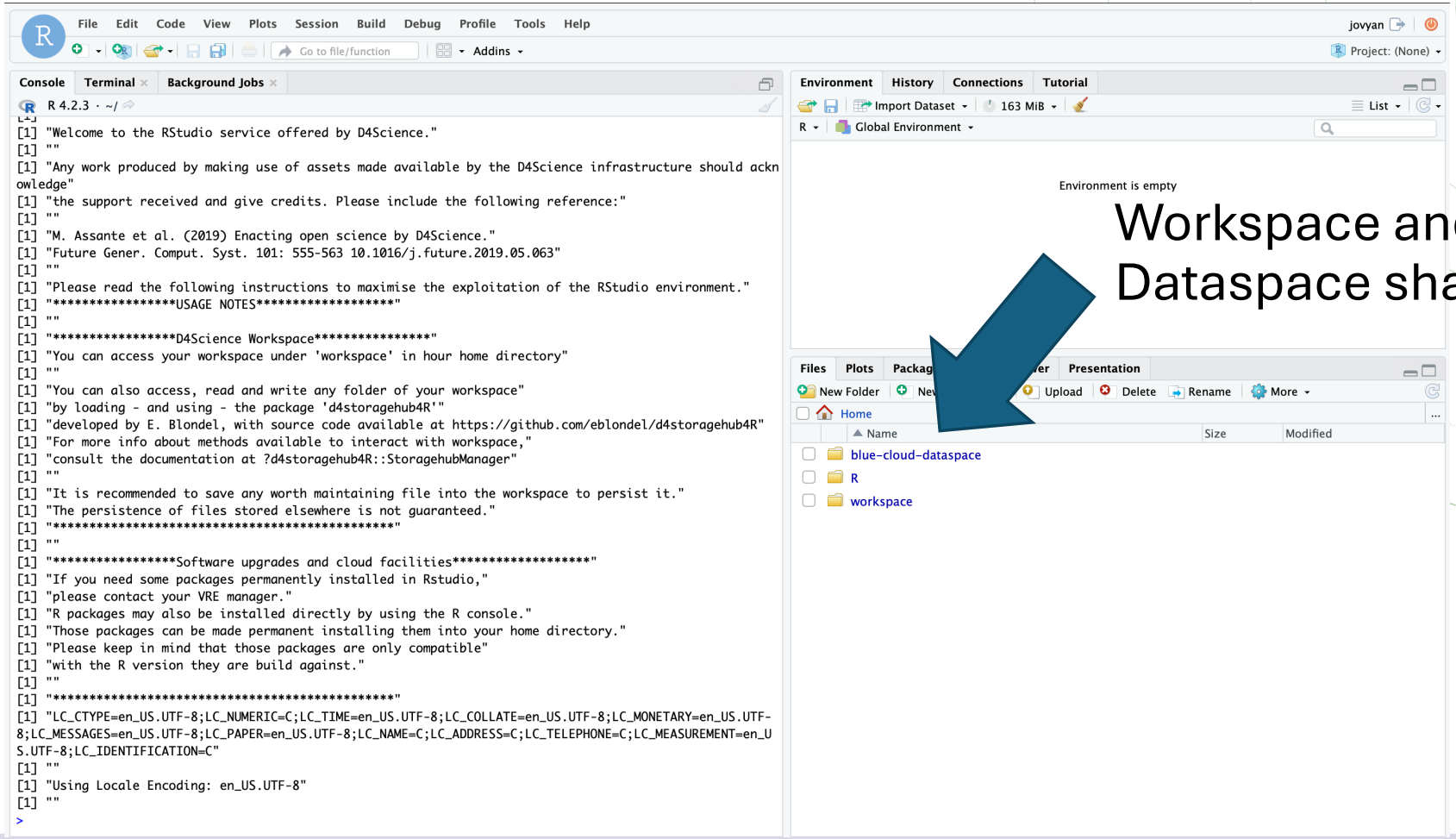
Rstudio platform benefits

- The Server Options can be customised differently for each Vlab / WB

  - Server image and required librar
  - No. of CPU Cores
  - RAM Memory avail.



Rstudio example



Analytics Engines

- Data Miner
- Cloud Computing Platform, CCP

Cloud Computing Platform Vision

3 years of work but decade long experience (D4Science)

- Presentations, interviews, minutes and write-ups

Stems from a use case based analysis …

- 10 abstract use cases
- A sound logical and architectural model

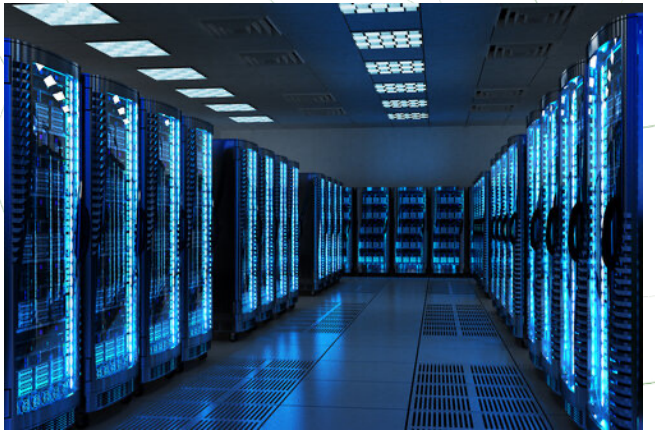
Developed in agility with early prototypes

Embraces heterogeneity

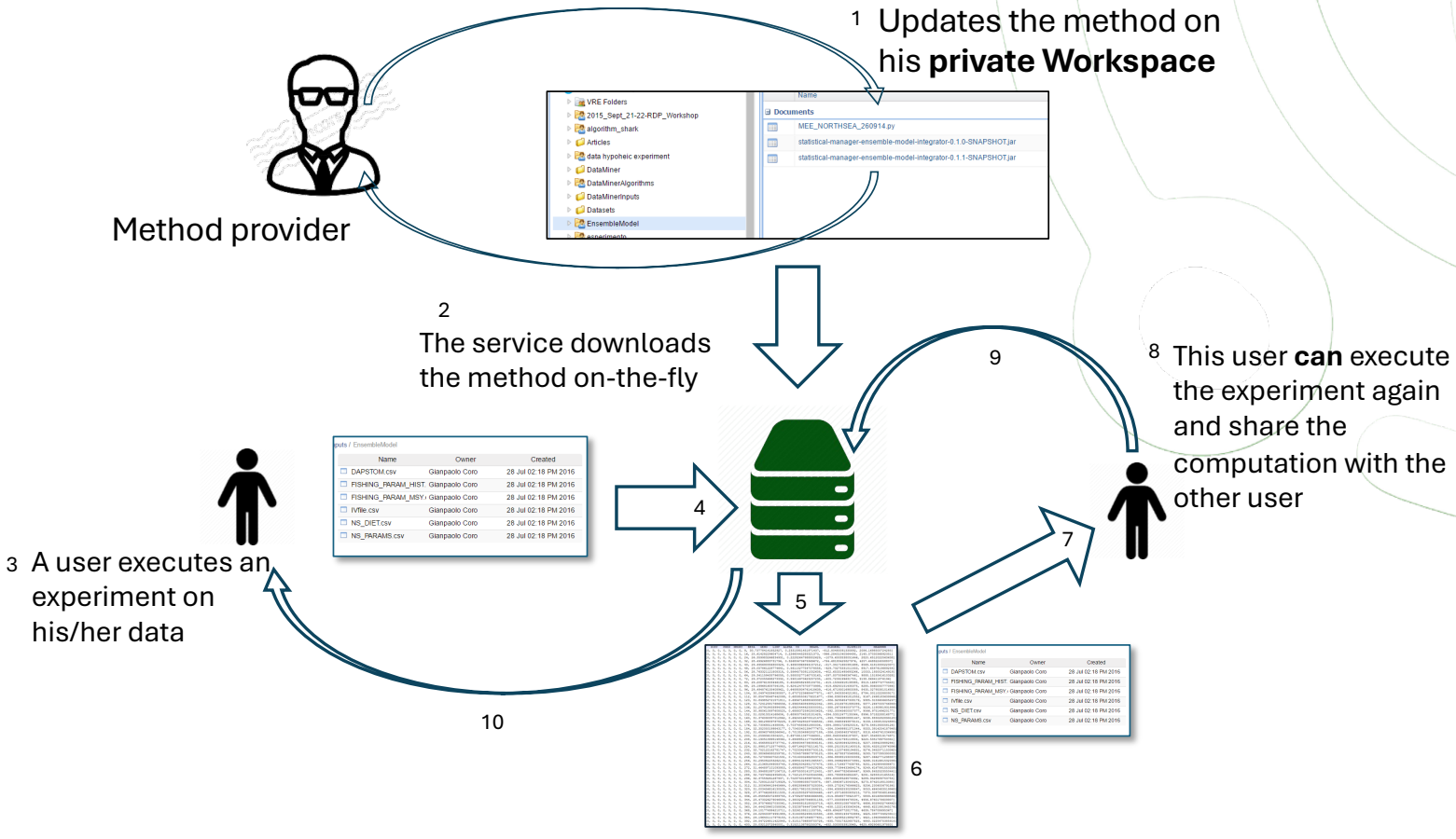
- supporting as many as possible different libraries, languages, tools, interfaces, services
- keep being manageable

CCP Main Features

- Container based – isolated runtime
- Multiple execution infras on the same Vlab, selectable by users
- Per-method Web UI and REST API access via Standard
- Method import support via a dedicated tool supporting any prog. language
- "*out-of-the-box*" methods **as-a-Service** and automatic code generation for Python, R, Julia, IPy notebooks
- Provenance management



CCP Security and privacy



A note on CCP provenance management

- In our context is ***the lineage and processing history of a research data product, and the record of the processes that led to it;***

- Provenance plays an important role in Data Science;
  - Experiments are often developed iteratively involving multiple executions with different versions of data sources, accessing multiple applications and infrastructure services;

- CCP adopts the ***the PROV Ontology (PROV-O***) expresses the PROV Data Model using the OWL2 Web Ontology Language (OWL2).

"Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."

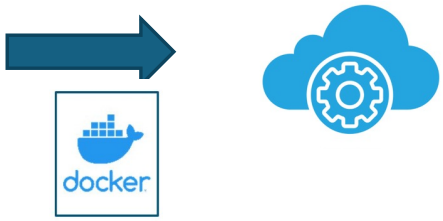
W3C Recommendation https://www.w3.org/TR/prov-dm/

```
<prov:document
    xmlns:prov="http://www.w3.org/ns/prov#"
    xmlns:d4s="http://d4science.org/#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <prov:activity id="activity-f07872aa-b783-4b4e-8512-95545aec7a94">
        <prov:startTime>2024-03-26T07:53:17.376Z</prov:startTime>
        <prov:endTime>2024-03-26T07:56:03.306Z</prov:endTime>
        <prov:type xsi:type="xsd:QName">d4s:computation</prov:type>
        <prov:softwareAgent prov:id="d4s:sobigdata.d4science.org"/>
        <prov:person prov:id="d4s:andrea.rossi"/>
        <prov:entity prov:id="d4s:operator_name">
            <prov:value xsi:type="xsd:string">SimpleImageClassifier</prov:value>
        </prov:entity>
        <prov:entity prov:id="d4s:operator_version">
            <prov:value xsi:type="xsd:string">1.0.1</prov:value>
        </prov:entity>
        <prov:entity prov:id="d4s:operator_description">
            <prov:value xsi:type="xsd:string">A simple image classifier with parametrizable url to input picture compatibl
        </prov:entity>
        <prov:entity prov:id="d4s:VRE">
```

CCP Methods, editor, and execution

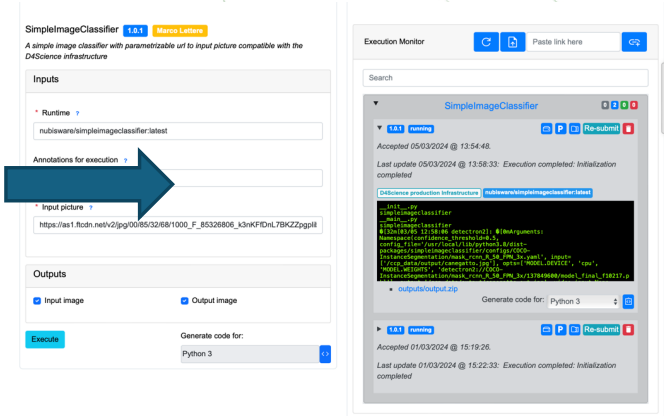
Method:
Script, Algorithm
Software

Importing phase



METHOD EDITOR

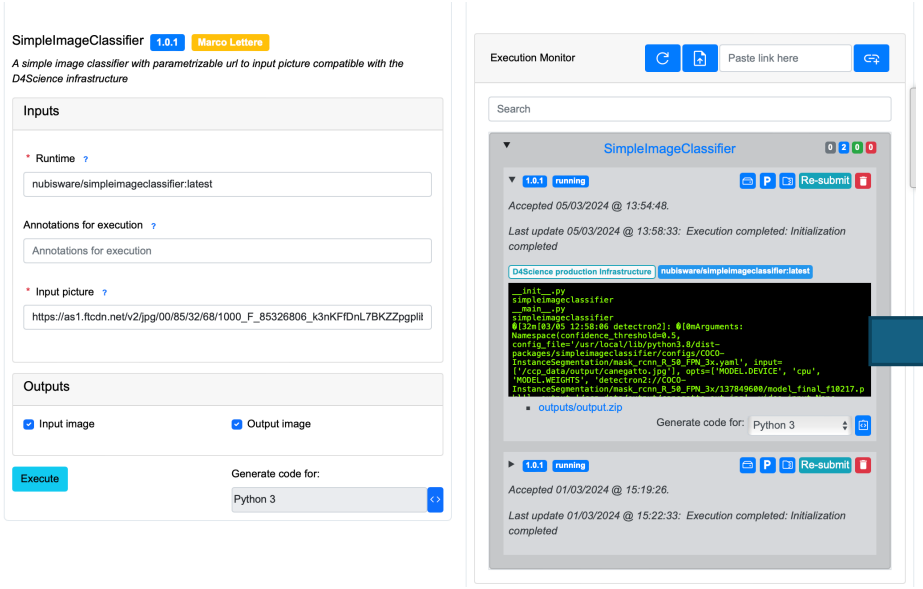
Resulting Web UI &Web API

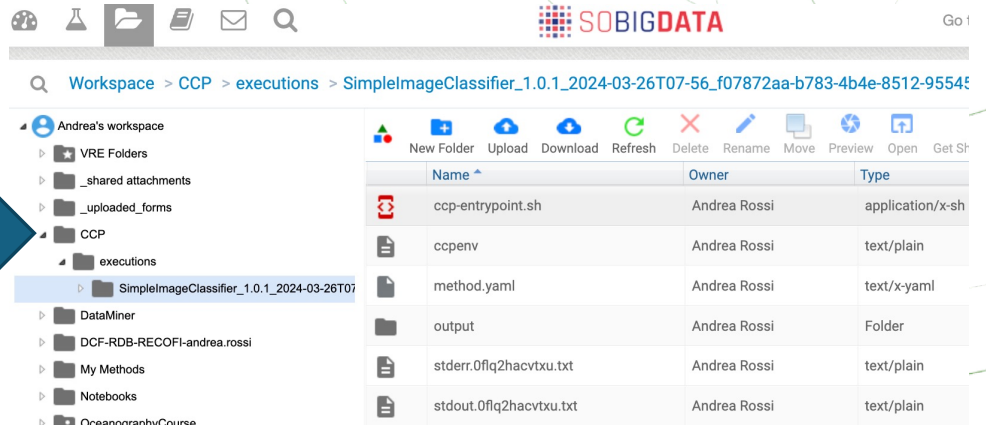


METHOD EXECUTIONS

CCP Methods, editor, and execution in the workspace

Web UI / Web API

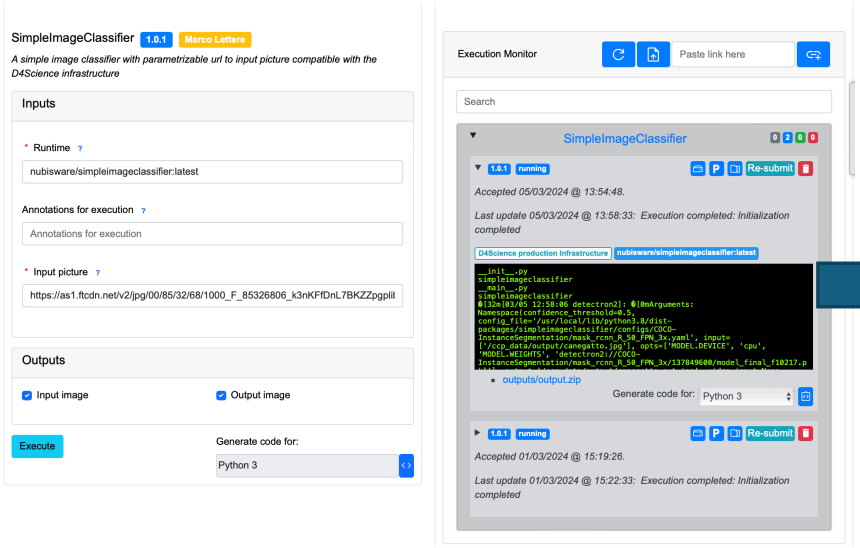


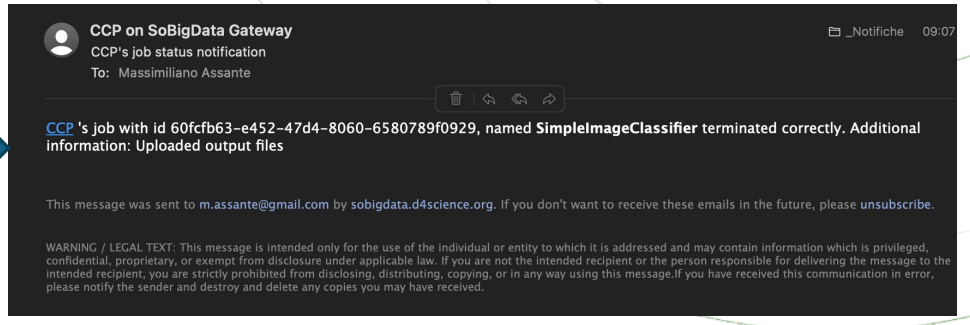


METHOD EXECUTIONS

CCP Methods, editor, and execution in your email inbox

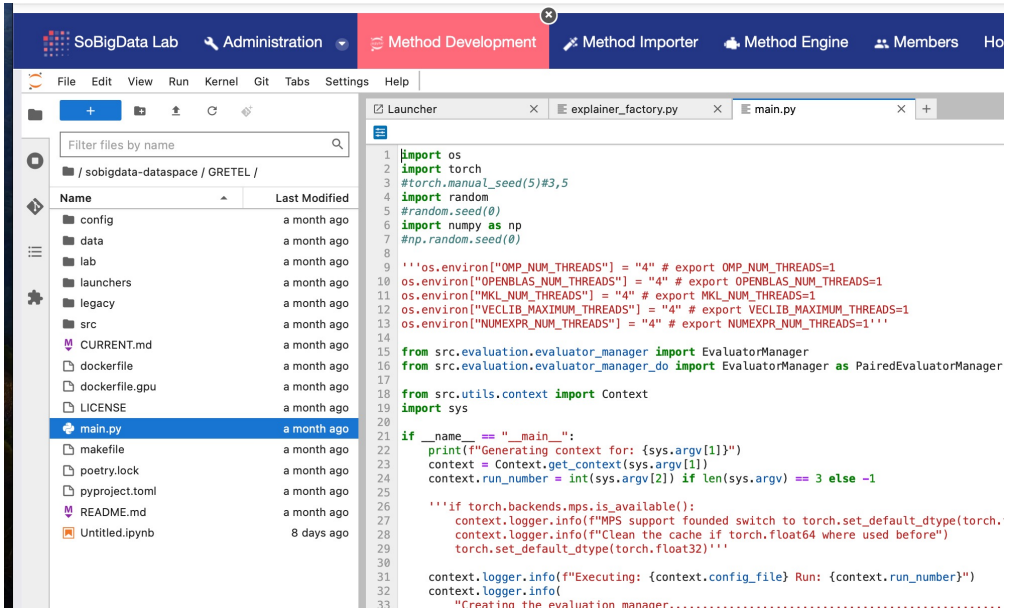
Web UI / Web API

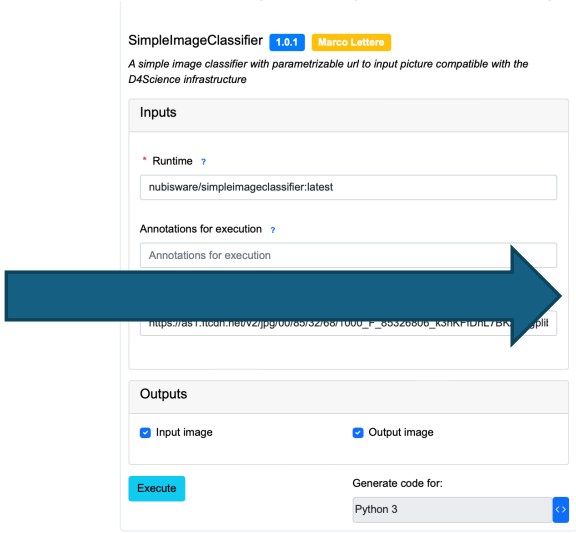


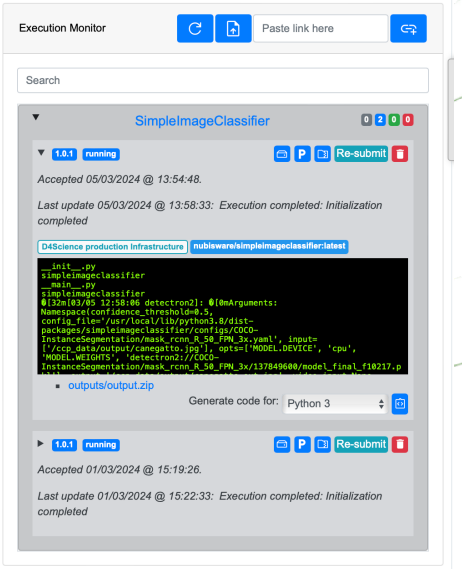


METHOD EXECUTIONS

CCP Methods, editor, and execution programmatically

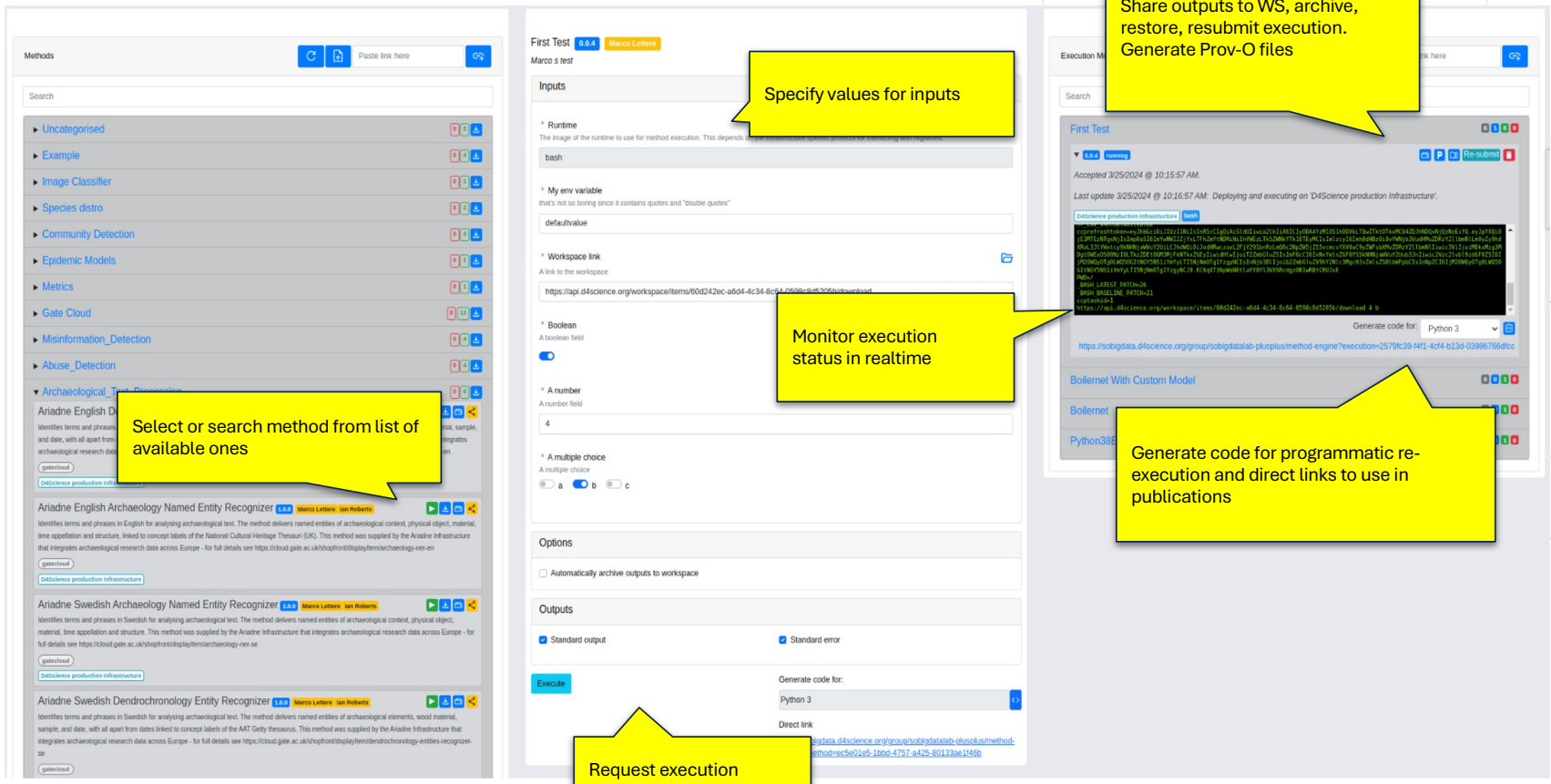






METHOD EXECUTIONS

Method Execution



Method execution: few details

- D4Science infrastructure is based on Docker containers.
- Any publicly availabe Docker registry
  - Including an experimental D4Science specific registry
- Environmental variables are passed by default to your Method
  - ease authentication when interacting with D4Science services (**ccpiamurl**, **ccpclientid**, **ccpcontext**, **ccprefreshtoken**)
  - **ccptaskname** and **ccptaskid**
- Special folder **/ccp_data** which is zipped and uploaded to the execution repository
  - The size is limited so DON'T upload huge files.
  - Use workspace or other storage facilities for large files.
  - Some meta outputs are kept in the outputs for debugging purposes
- Please consider good housekeeping :-)
  - Delete useless executions. They are considered volatile.
  - Archive to your workspace executions or execution outputs that have a value.
  - Backup your Methods to the workspace

ITINERIS

Let’s play with CCP now!

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell’Università e della Ricerca

Italiadomani