Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

Introduction on Data Management: The Use Case

WP5 Leader: Daniele Spiga (INFN – sez. Perugia) & Elvira Rossi (Università Federico II di Napoli)

WP5 Objectives and high level view

Proposes:

- **5.1 Support of the adaptation of existing applications on the data-lake distributed infrastructure, and via innovative computational models**
- **5.2 Competence center for the design, implementation and test of computing models**

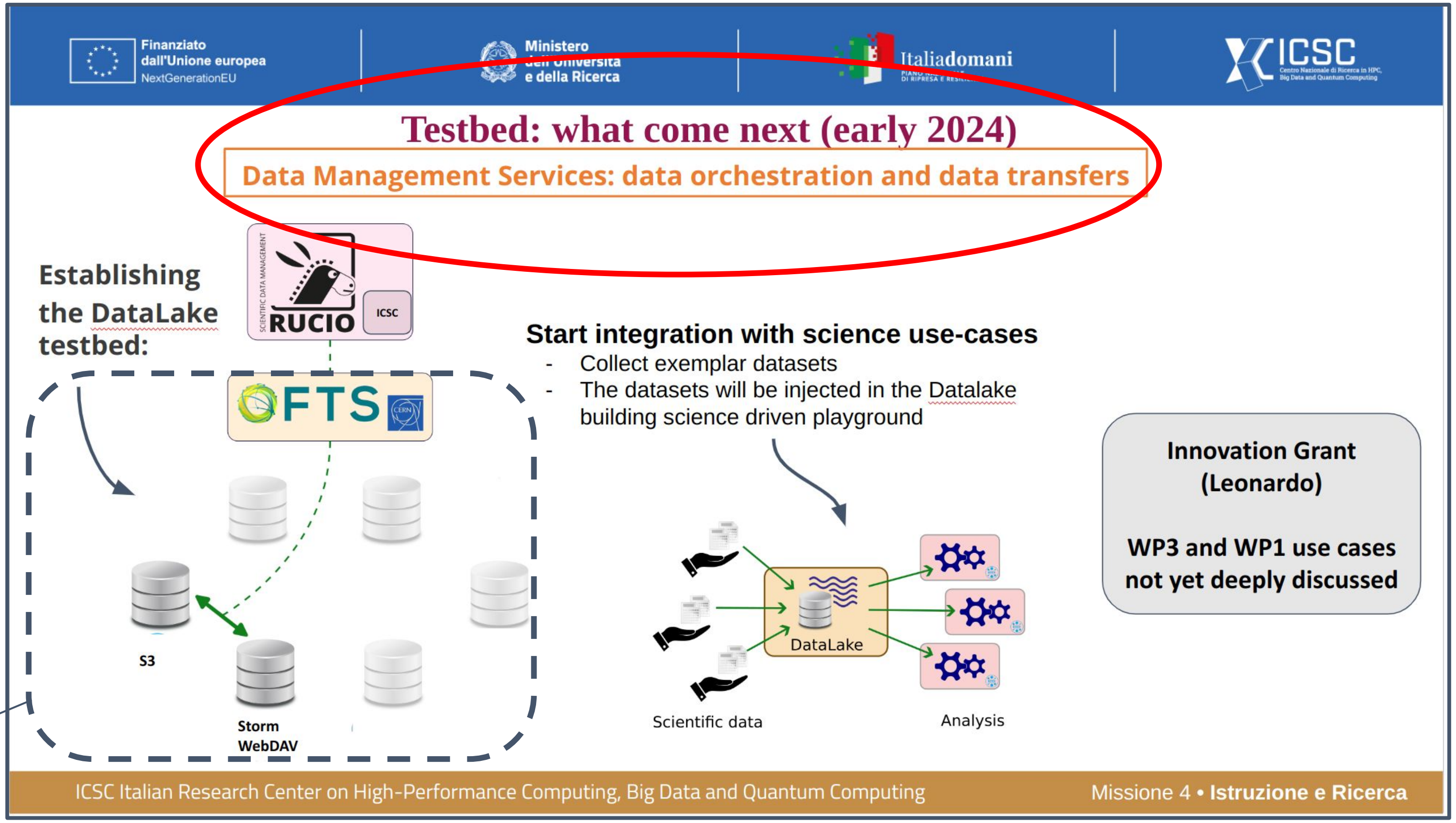
Objectives:

O5.1: Document and report best practices for integrations with the CN data lake
O5.2: Prepare tools to ease integration with the CN infrastructure
O5.3: Offer support for transitioning the computing models
O5.4: Organize training opportunities open to external users

What we already discussed and "promised"

WP5 Workplan foresees development and support for Distributed Data Management tools

- To support scientific activities within WP3/1 and Innovation Grant



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Testbed: what come next (early 2024)

Data Management Services: data orchestration and data transfers

Establishing the DataLake testbed:

Start integration with science use-cases

- Collect exemplar datasets
- The datasets will be injected in the Datalake building science driven playground

Innovation Grant (Leonardo)

WP3 and WP1 use cases not yet deeply discussed

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing Missione 4 • **Istruzione e Ricerca**

Highly Sinergic to Spoke0

Spoke2 general meeting 2023

Why a federated Data Management (DM)

To provide a consistent user experience in accessing distributed resources hosting data, through federated solutions.

To grant **interoperability between heterogeneous infrastructures** and storage solutions.

- big traditional HPC centres, Cloud, Grid and HTC possibly geographically distributed systems (== ICSC)

DM in terms of responsibilities:

- it must provide the data needed for scientific use cases applications/executables
  - What are the available data (how they are made)

- it must be able to deploy the results (output) along with the additional data needed
  - to support any post-processing steps and the final exploitation

What is the use case to solve

“The ability to store and retrieve data without caring about to know where the data actually is”

What we discuss today is necessary to satisfy this requirement.

NOTE:

- This data is assumed to be stored as **files in some file system**
- It can be also a **data objects in some object store**
- In this model we exclude data stored in specialized services, **such as a database**

However, simply storing/retrieving data is not sufficient.

What is the use case to solve (cont)

We recognize three main functionalities:

- **Dataset management and orchestration:** responsible for the locality of the files within a dataset, creating additional copies of data to satisfy demand and to remove excessive copies when under space pressure.

- **The File Transfer component:** responsible for transferring data effectively between endpoints, where the Dataset Orchestration has identified that an additional replica is Needed.

- **Storage edge service** is a component that is deployed "close to" a facility's existing storage capacity and enhances its capabilities so that it may participate within the federated data infrastructure

All this is the driver for the presented architecture (see next talk)

A workflow in practice

In summary the ultimate goal is to grant any user the possibility

- To produce data i.e. “running at CINECA Leonardo” / Running on Cloud and/or HTC at INFN
  - This might also be something like: my DAQ collects some detector acquired data - Data Producer -
- To store them in the storage (@ somewhere in the Datalake)
- Eventually to define a policy such as: make three copies of these files in all the infrastructure where I can write data
- To re-run over previously produced data (perhaps elsewhere wrt Leonardo)
- Remove unneeded copies
  - I.e. make this automagic by defining a policy accordingly

Most of this will be presented in the demo today

A final remark

The solutions we selected come from a specific domain (HEP) however they are completely HEP free

- Already adopted in other contexts at national and international level

From technical perspective all what we presented has huge interaction with the infrastructural layer. This is the reason why we decided to be in strict contact with Spoke0

Solutions presented today are in the main stream of ICSC DataLake Architecture:

- They are part of the ICSC spoke0
- They are part of the Proof Of Concept being realized for the CINECA - INFN federation
  - In turn this means that we could start discussing testbeds to support spoke2 needs in that context