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

National DataLake Integration

Diego Ciangottini - on behalf of Spoke3_WP4-5

INFN Istituto Nazionale di Fisica Nucleare

Spoke 3 2nd Technical Meeting

Where were we?

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

The plan

Physical level:

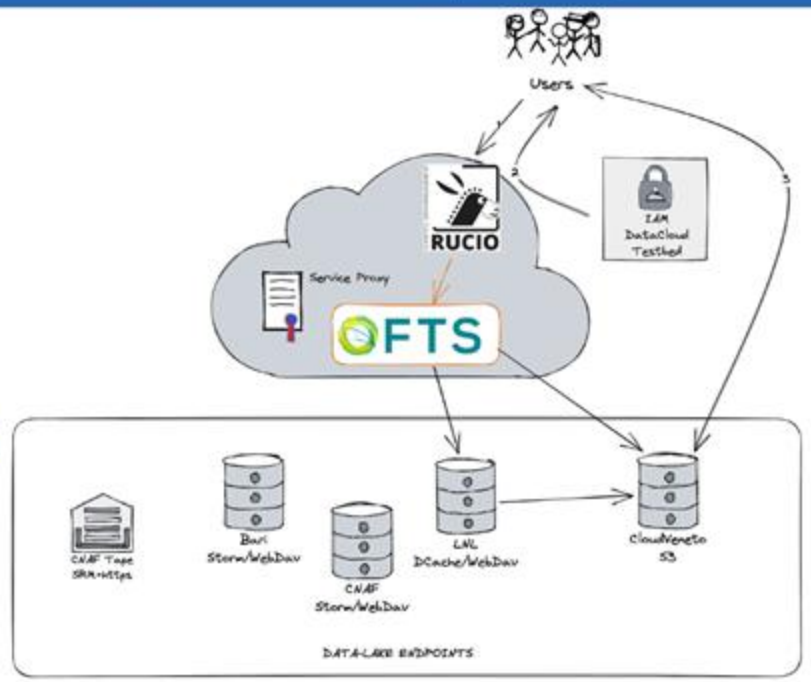
- Get initial space for a playground on a couple of storage sites
- Protocols matter: S3/object + an HTTP/WebDav might be a good first set

Physical+Logical matching (WP5 <-> INFN DataCloud):

- Setting up dev services required for managing transfers and data accessibility (FTS+RUCIO)

Interface level:

- Make Rucio interfaces with metadata databases used by archives



N.B. we acknowledge the existence of strong synergies with Spoke2 and Data-cloud initiatives, we should work in strict coordination to avoid waste of efforts.

Strategy Quick Recap

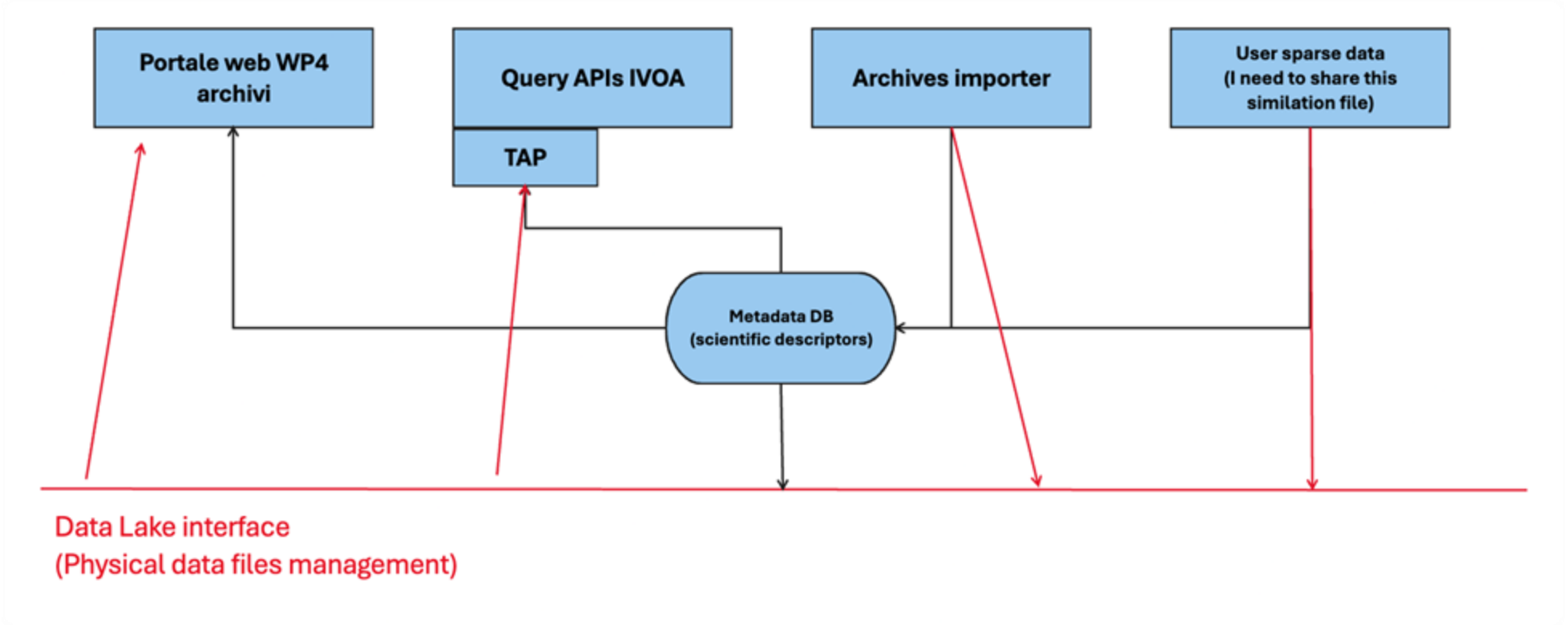
Start from INFN DataCloud experience and evolve it based on user requirements

Being particularly careful on responsibility sharing e.g.:

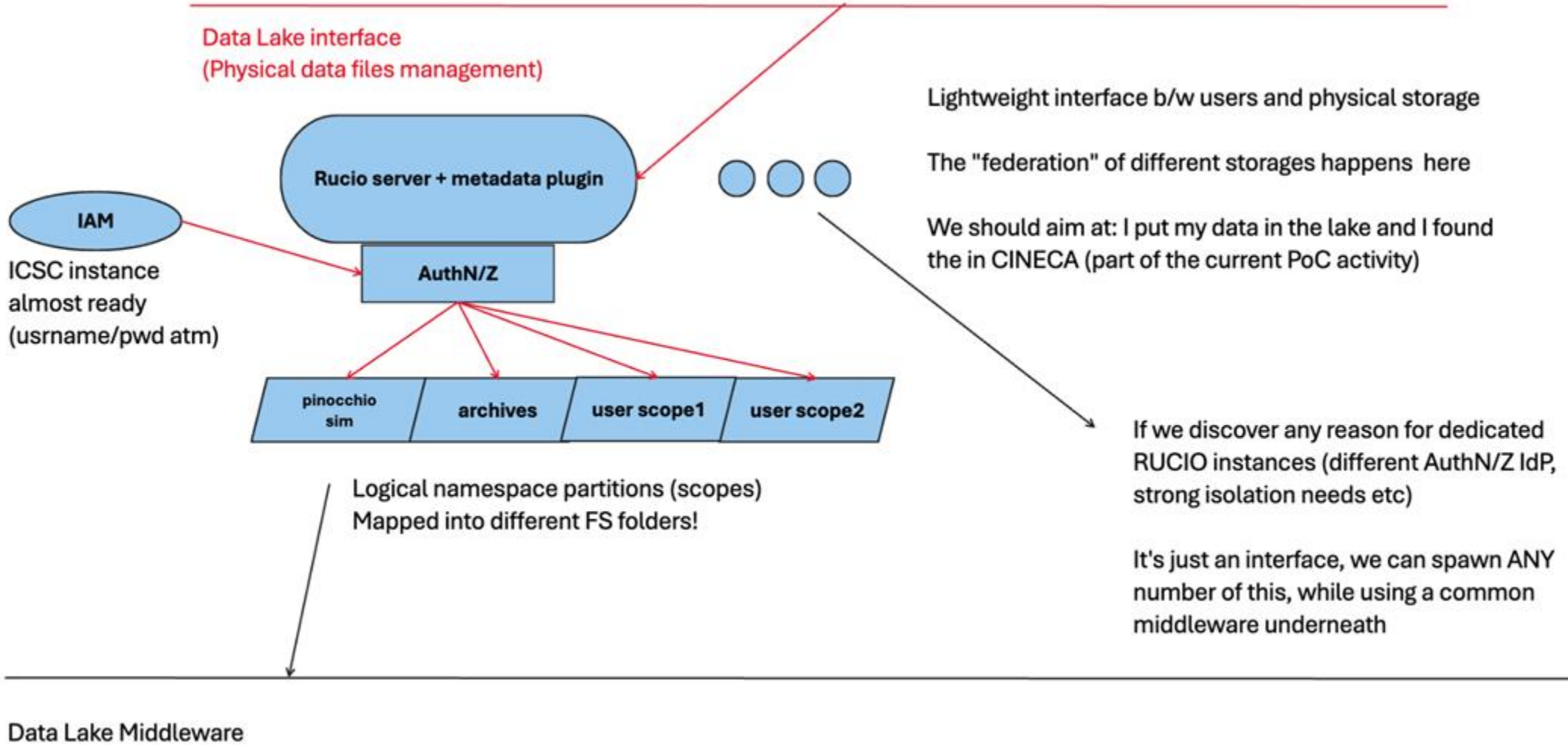
- **Spoke0 to manage data transfers and storage site** (infra central services)
- WP5 for integration of **RUCIO development services and archive databases**
- WP4 for **archives plugin integration on the frameworks**

What we have deployed is, for all these reasons, the results of a coordinated effort and discussion b/w all the stakeholders.

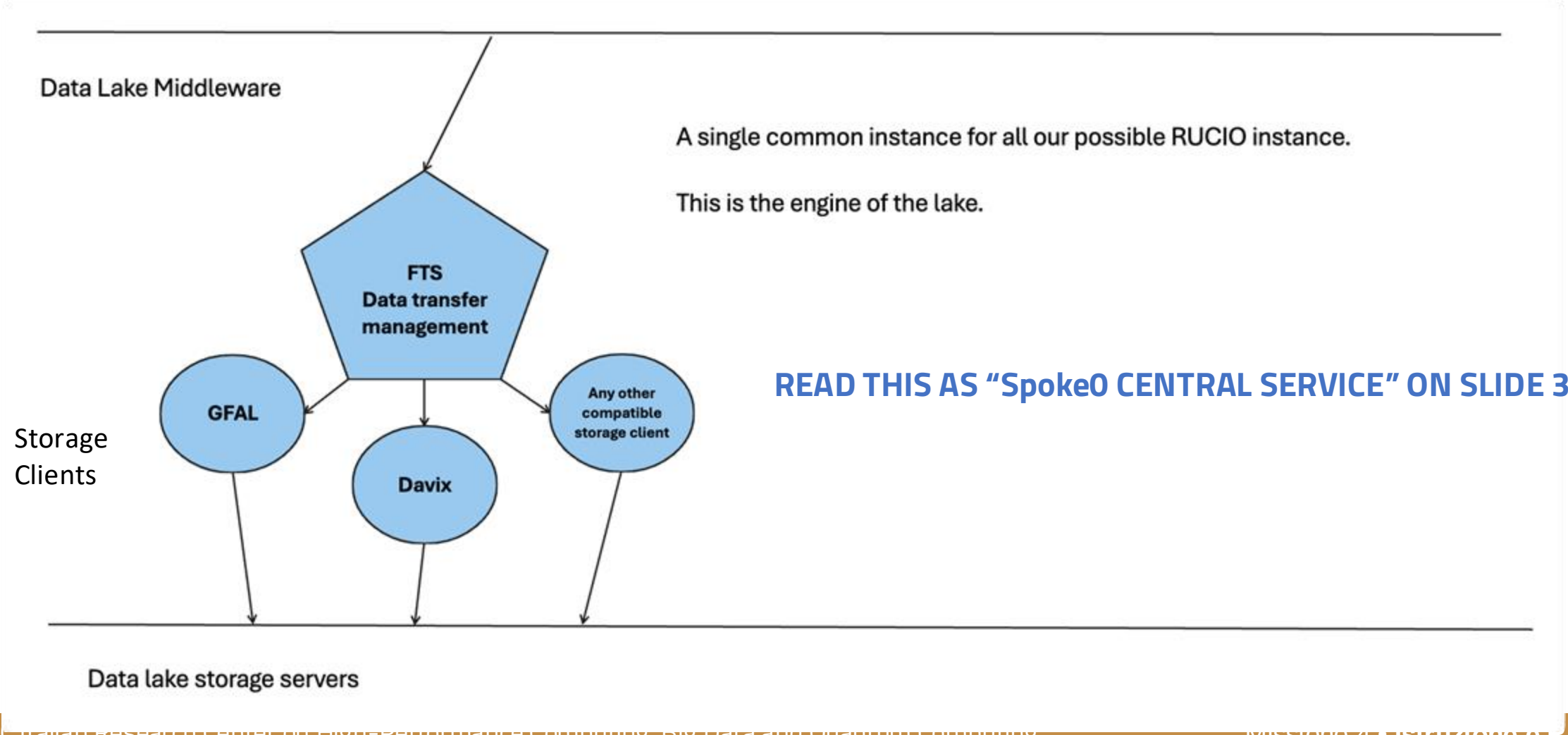
Data lake layers top to bottom: User-facing frameworks



Data lake layers top to bottom: DM interface



Data lake layers top to bottom: DM middleware

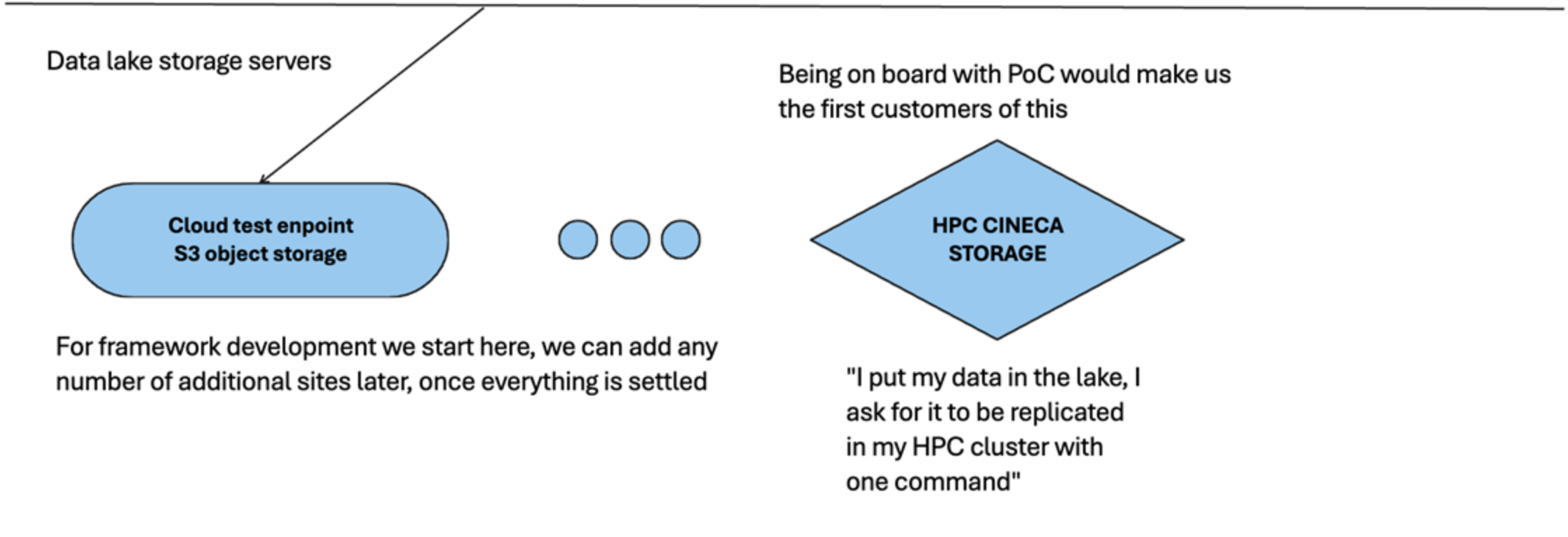


A single common instance for all our possible RUCIO instance.

This is the engine of the lake.

READ THIS AS "SpokeO CENTRAL SERVICE" ON SLIDE 3

Data lake layers top to bottom: DM middleware



What is there already?

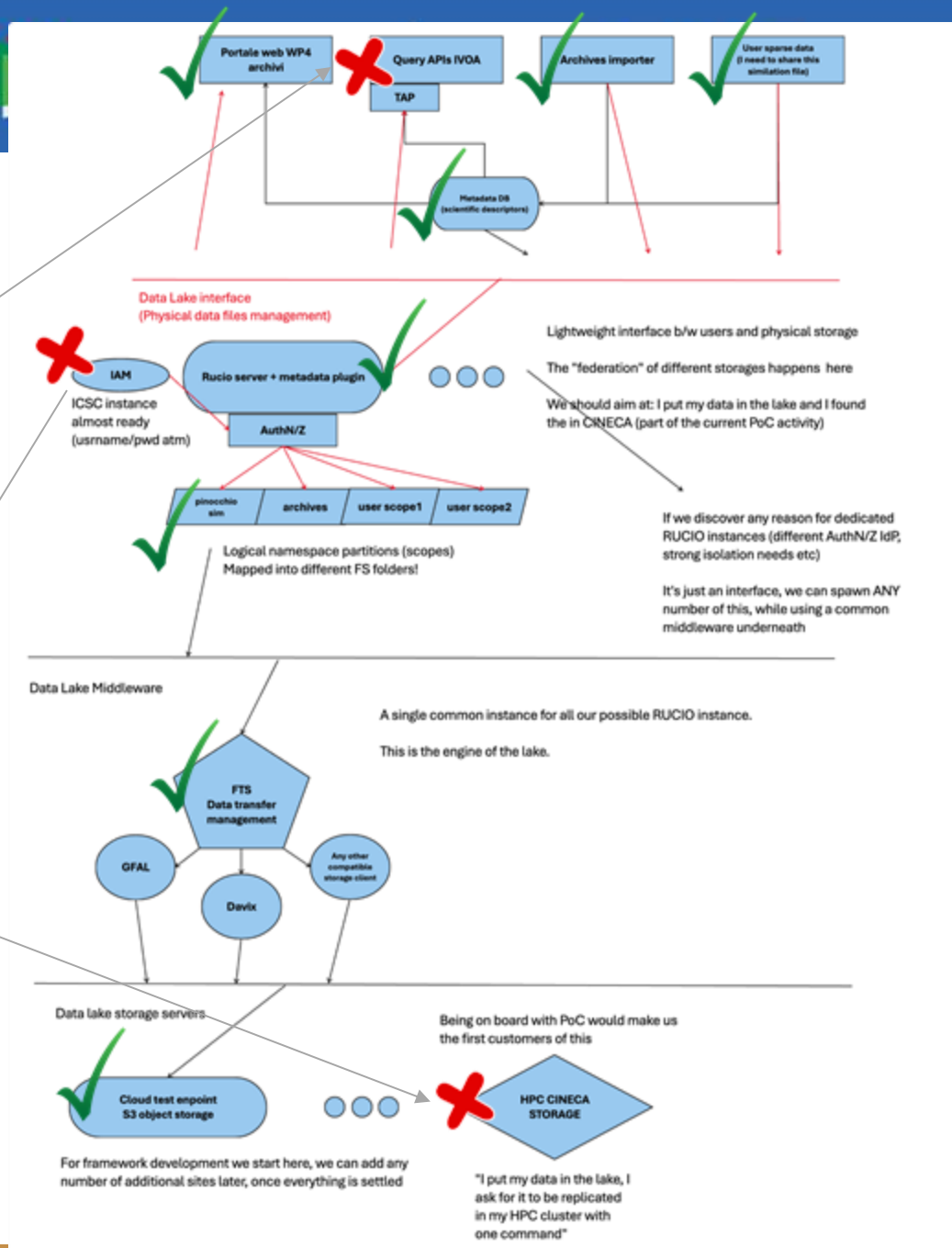
We have multiple e2e paths validated!

A few missing but ongoing (more complicated interfaces to be adapted)

Some feature are not fully on Spoke3 ballpark, but we are onboard with the ongoing PoC

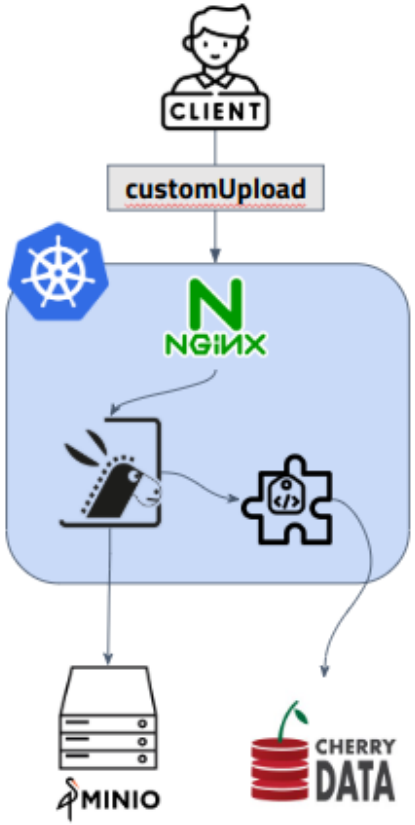
EVERYTHING atm is hosted on "in-kind" cloud resources at INFN!

RAC resource arrived last week, we are going to REPLICATE the whole setup in a more robust production-ready manner



Interoperability

- **Interoperable Data Lake (IDL) activity** is a practical example of how you can treat each of the step in the presented chain as a set of **building blocks. (again strong synergy to converge over a common set of these blocks)**
- You can **compose the service that you need** just with a few configuration changes, while at the core **the system still work the same.**
- The plugin mechanism for the metadata guarantees the extension toward **any kind of database.**



Next challanges

- **AuthN/Z schema and case studies**
  We will start immediately (already started for what we can) to play with ICSC provided IAM instance.

  The scope is to understand the shortcomings and where/when/if we will need to tweak our system to be compliant with the use case needs

- **HPC integration**
  Cross activity with Spoke0 and Spoke2. This is a huge step toward implementing the federation of data b/w cloud and HPC clusters. Currently happening in the context of Spoke0 CINECA PoC

- **ICSC resources**
  Last week we started the migration. It will not be a 1to1 migration, since we are taking the occasion to instantiate a more mature operational system (it might take roughly a month to be on feature parity with the current setup)

- **Integration with WP5 scientific hub**
  see later in Matteo's talk