Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
terabit

Data Management e Storage Federation

Verso la realizzazione del Datalake

A. Troja, M. Biasotto, D. Ciangottini, M. Delli Veneri, F. Fanzago, A. Italiano, N. Marcelli, L. Morganti, A. Rendina, M. Sgaravatto, D. Spiga, B. Spisso, S. Stalio, M. Verlato

27/05/2025

Workshop sul calcolo nell'INFN – La Biodola

Dove eravamo rimasti

WP6 si occupa da tempo di **R&D su data management** e storage federation.

L'obiettivo è verificare la fattibilità dell'adozione di un modello basato su tecnologie ben note (**RUCIO+FTS**) per la gestione dello storage presente nei centri di calcolo dell'ente.

L'implementazione di testbed di datalake, nei quali abbiamo testato e **validato con esito positivo** diverse configurazioni, ci ha permesso di verificare le funzionalità del sistema e valutarne l'impatto non solo sull'infrastruttura hardware, ma anche sui servizi e sulle operazioni.

La strada fin qui:

Storage & Data Magement, M. Sgaravatto e D. Spiga, 2022

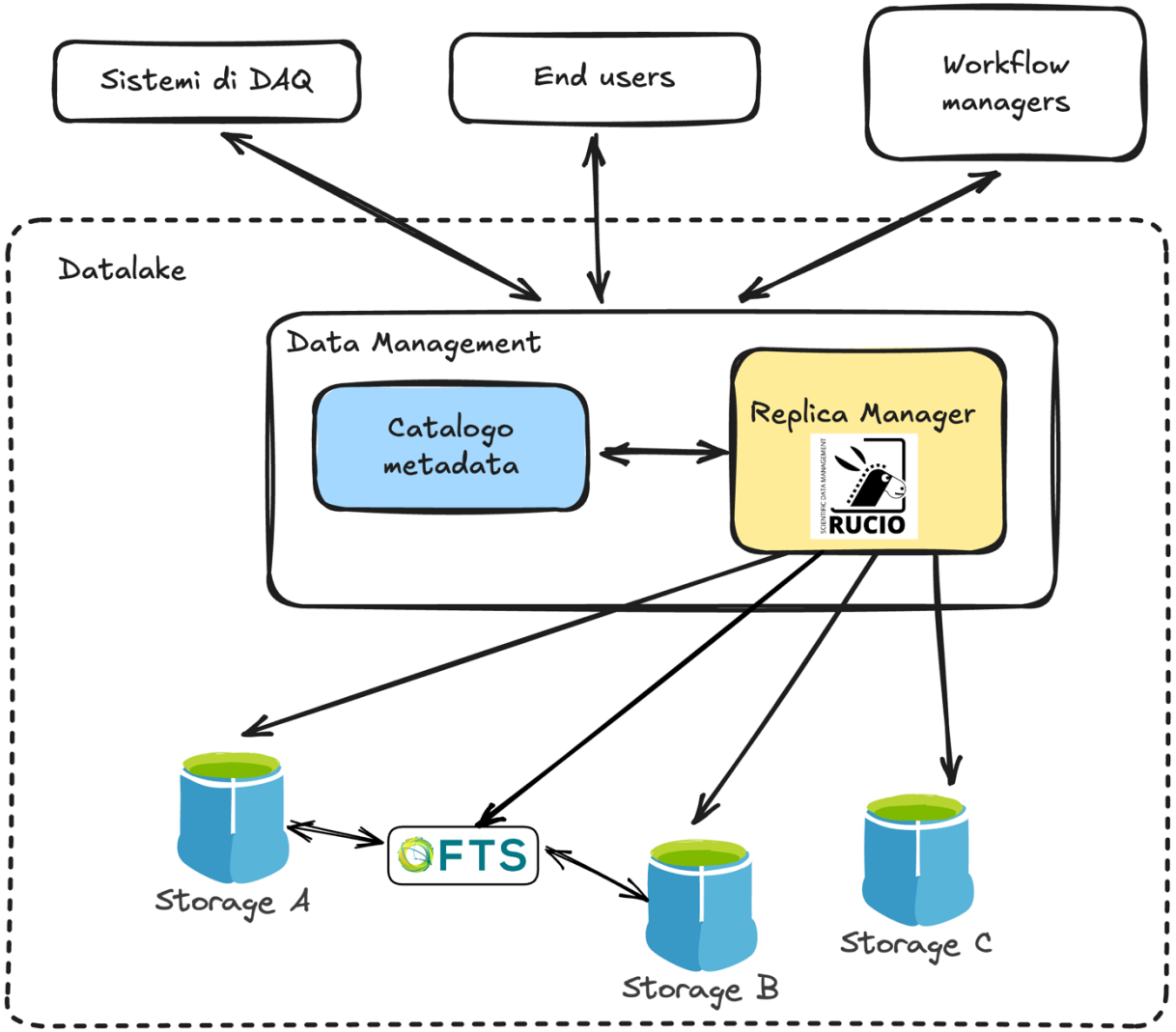
Federare lo storage distribuito nazionale, D. Ciangottini, 2023

Attività di R&D in corso sinergiche con i vari progetti e nell'ambito di Datacloud, M. Sgaravatto 2024

Il datalake nell'infrastruttura INFN: Applicazioni e prospettive future, A. Troja 2024

Recap: Cos'è il datalake?

- **Storage** differenti vengono federati indipendentemente dalla collocazione geografica, dalla loro implementazione, dal loro QoS (disk o tape);
- **L'utente** interagisce coi dati in maniera dichiarativa: ad esempio può dichiarare quante repliche servono per un certo file, su quanti e quali storage, per quanto tempo devono esistere;
- **I dati** possono essere organizzati gerarchicamente (dataset, container);
- Grazie alla gestione dei **metadati**, possiamo implementare funzioni di query;
- **Trasparenza**: Livello di astrazione che nasconde all'utente le eterogeneità e i dettagli implementativi.



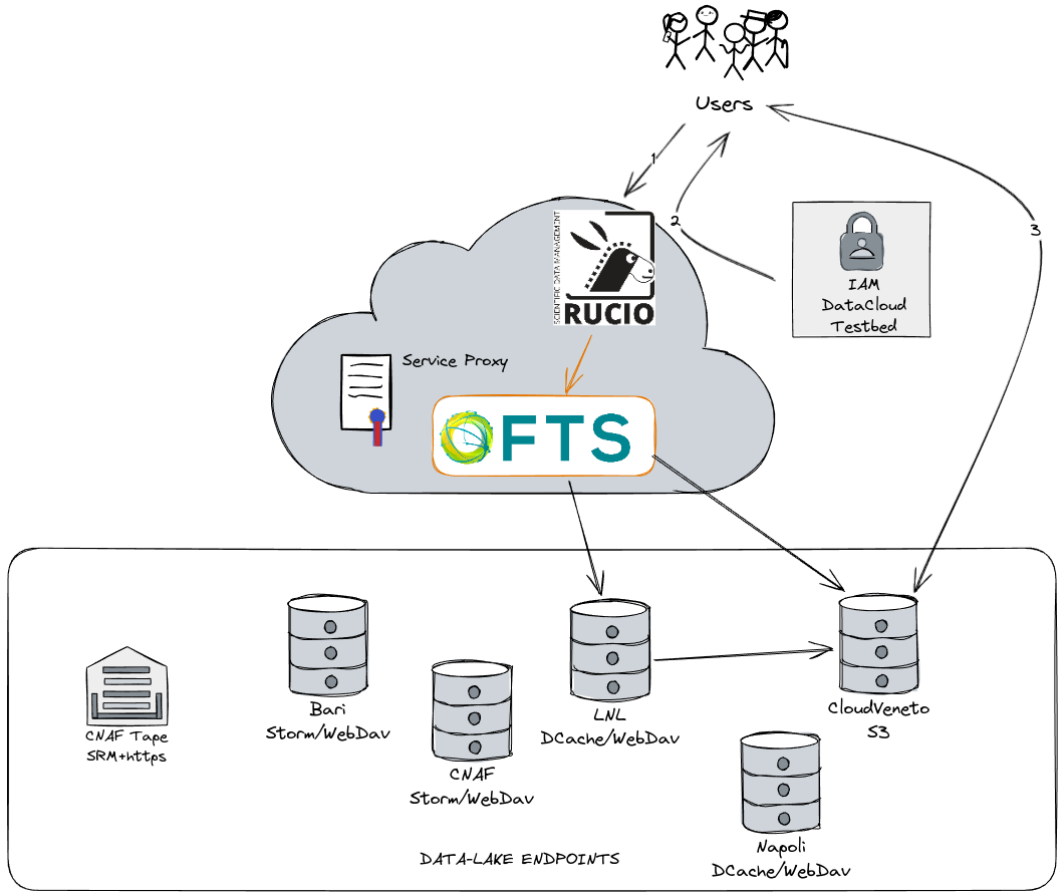
Il testbed in DataCloud WP6

Il testbed che abbiamo sviluppato integra:

- Rucio (replica manager);
- FTS in high-availability, effettua i Third Party Copy con VOMS proxy (credenziali di servizio)
- IAM (AuthN/Z via token);
- Metadata Catalog: embedded in RUCIO.
- Infrastruttura PSQL in *high availability* dove ospitare i database RUCIO.

Federati 6 storage system dell'INFN eterogenei:

- Qos (disk, tape);
- Implementazione (dCache, Storm, Ceph);
- Protocolli (WebDav, SRM, S3).



Federare lo storage distribuito nazionale, D. Ciangottini, 2023

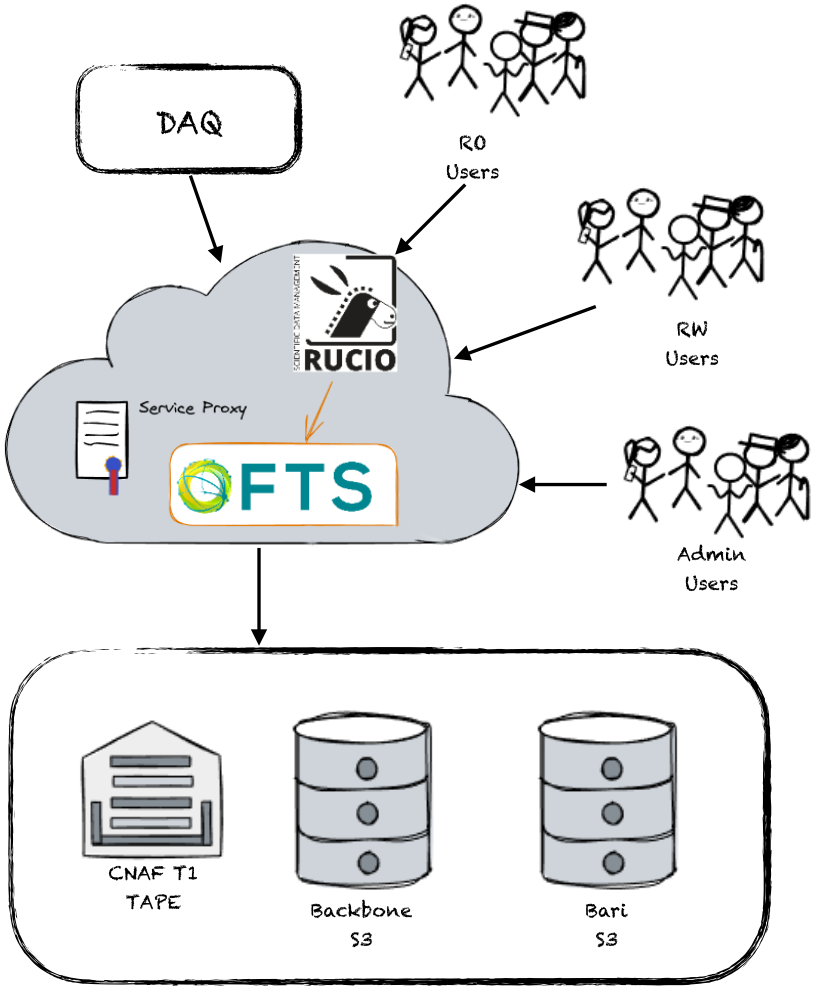
Cygno datalake

I file presenti in un buffer locale vengono inseriti nel datalake per essere processati via INFN Cloud. I dati vengono conservati nell'endpoint tape e cancellati dagli endpoint disk quando non più necessari.

Le letture da tape sono piuttosto rare, ma previste in futuro per attività di riprocessamento.

Abbiamo implementato un datalake che:

- Federa due storage di tipo disco e uno di tipo tape -- Un terzo storage endpoint (ceph S3 @ CNAF) in fase di integrazione;
- Implementa policies di accesso ai dati ad hoc, più stringenti di quelle del nostro testbed, introducento un livello di autorizzazione per trasferimenti da/per tape.



DAMPE Datalake

Dopo un periodo di inattività dovuto a script non più mantenuti, grazie all'implementazione del DM Rucio, DAMPE ha ricominciato a replicare dati automaticamente verso il CNAF:

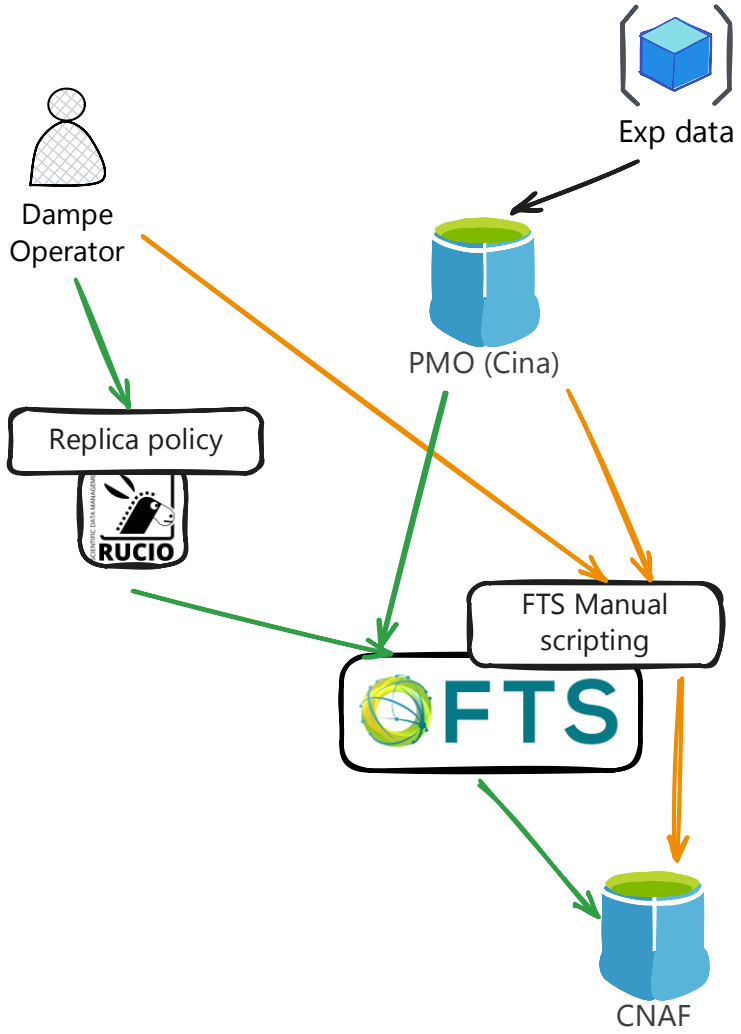
- Collegamento a bassa larghezza di banda e alto rate di errori;
- Produzione continua di dati.

Il processo è avvenuto in due step:

1. Migrazione da script manuali rsync a FTS WP6 (semi-manual);
2. CNAF e PMO inclusi nel DM, ogni file aggiunto nel dataset dell'esperimento e registrato in Rucio viene automaticamente duplicato, con verifica automatica del successo dei transfer.

Il DM DAMPE è in produzione:

- Il setup è previsto ad uso esclusivo degli operatori, ad oggi non c'è richiesta per accesso «multi utente»;
- Ci sono issue interne a storage PMO, indipendenti dal Datalake attualmente in risoluzione;
- Si vuole aggiungere un nuovo endpoint a Ginevra.



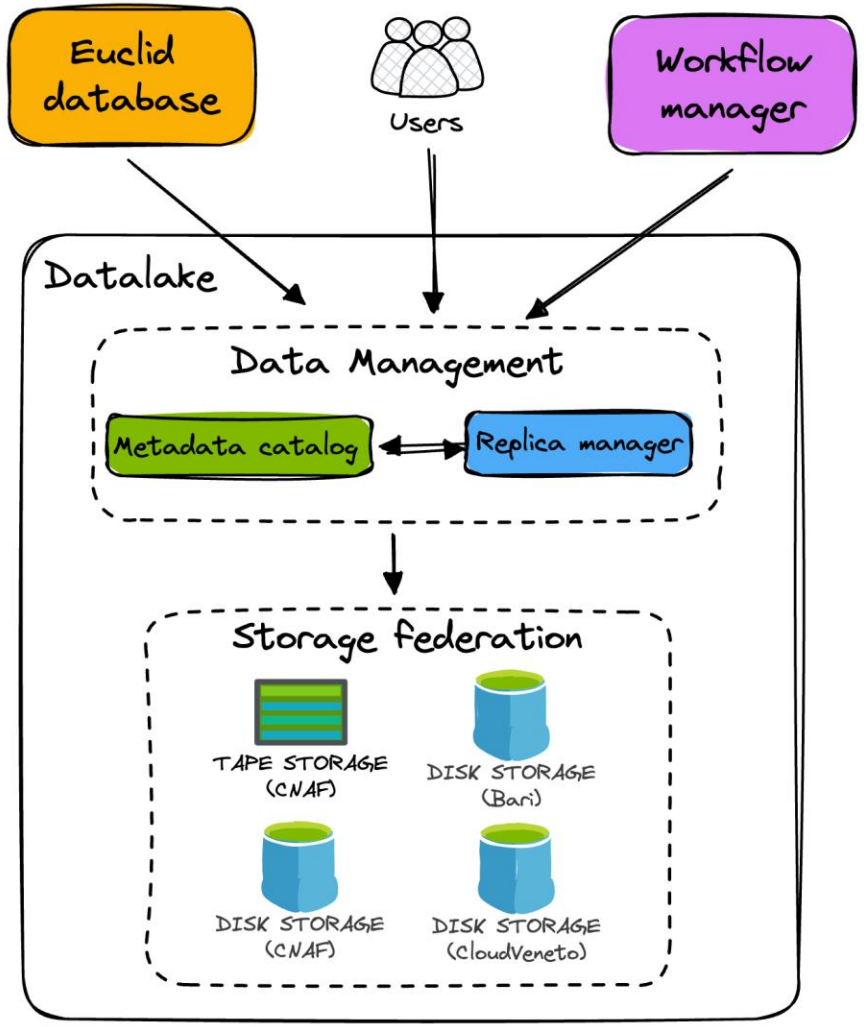
Euclid Datalake

Abbiamo sviluppato il datalake per Euclid Italia utilizzando la nuova versione 35LTS di Rucio.

Da qui a breve, metterà a disposizione alla comunità italiana dati per l'analisi delle calibrazioni dello strumento e delle sistematiche dell'esperimento, scaricandole dal database centrale.

Mentre gli storage al CNAF serviranno per la conservazione vera e propria dei dati, i trasferimenti a Bari e CloudVeneto permetteranno l'analisi vera e propria.

Degli script gestiranno i transfer tra gli storage e la vita di ogni replica fuori dal CNAF.



Onboarding di altre comunita'

Siamo in contatto con altre comunità che hanno espresso interesse a provare questi tool:

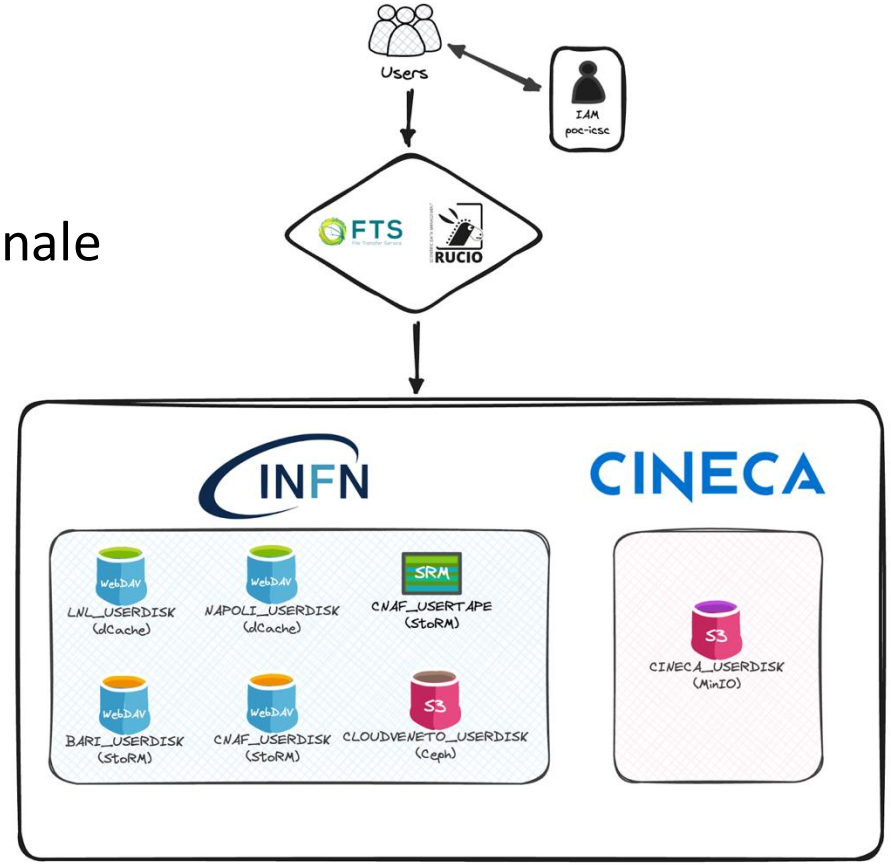
- DarkSide: seguendo la documentazione, hanno configurato una istanza RUCIO federando due storage disk. Siamo in attesa di ulteriori aggiornamenti da parte loro.

Proof of concept CINECA

L'obiettivo del PoC è dimostrare la possibilità di integrare le infrastrutture INFN e CINECA per costruire il DataLake Nazionale di ICSC/Terabit

Abbiamo quindi implementato e testato con successo un'infrastruttura testbed DM che federa storage INFN e CINECA, utilizzando un'istanza IAM e una Rucio specifiche per il PoC.

Per il futuro, sarà necessario implementare un mapping tra account Rucio e credenziali CINECA.
Consideriamo che le policy di accesso ai dati e l'user isolation saranno gestite a livello di Rucio.



Verso un datalake nazionale

Prevediamo un endpoint storm-webdav al CINECA allo scopo di montare una porzione di filesystem di Leonardo. Federando questo endpoint a Rucio, sarà possible gestire le repliche di dati con gli endpoint INFN.

- La federazione degli storage è fondamentale per la creazione di un datalake nazionale e quindi per l'integrazione delle risorse INFN-CINECA;
- Il trasferimento dati da e verso il CINECA permetterà di ottimizzare l'utilizzo delle risorse per attività di ricerca.

Non è tutto rose e fiori, le interazioni non sono banali e richiedono "tanta pazienza". Però le discussioni sono sempre piu' positive e vanno nella direzione auspicata.

Verso la produzione

Il 6 e 7 marzo del 2025, la componente WP6 dedicata allo sviluppo del data management si è riunita al CNAF per discutere i vari aspetti dell'attività di R&D e di verificarne lo stato in vista di un passaggio in produzione.

Il risultato è stata la formulazione di una proposta del modello da implementare in produzione e l'analisi dei costi collegati.

L'adozione di un sistema DM implica che:

- Gli utenti accettino di apprendere e utilizzare RUCIO per la gestione dei propri dati;
- I siti federino gli storage e li gestiscano secondo policies ben definite.

Verso la produzione - Benefici

A dispetto dei costi, i benefici dell'adozione di un sistema Rucio comprendono, tra le altre cose:

- **Modello sostenibile** dal punto delle operazioni del sistema: l'utilizzo di una tecnologia unica ben collaudata nell'ambito di altri esperimenti, semplifica le operazioni, consentendo il venir meno di un effort dedicato dentro le collaborazioni;
- Contributo alla definizione e implementazione dei **computing model** degli esperimenti;
- **Utilizzo efficiente delle risorse storage**, sincronizzando i cataloghi dei dati con quanto è effettivamente presente negli storage federati distribuiti (issue come «dark data» e «missing data» sono stati già affrontati dalle grandi collaborazioni che utilizzano Rucio, creando strumenti utilizzabili da tutti;
- Possibilità di avere **un monitoring centralizzato**;
- Astraendo il livello fisico da quello logico, si **semplifica la gestione delle risorse**, ad esempio nei processi di assegnazione del pledge.

Verso la produzione – Modelli di deployment

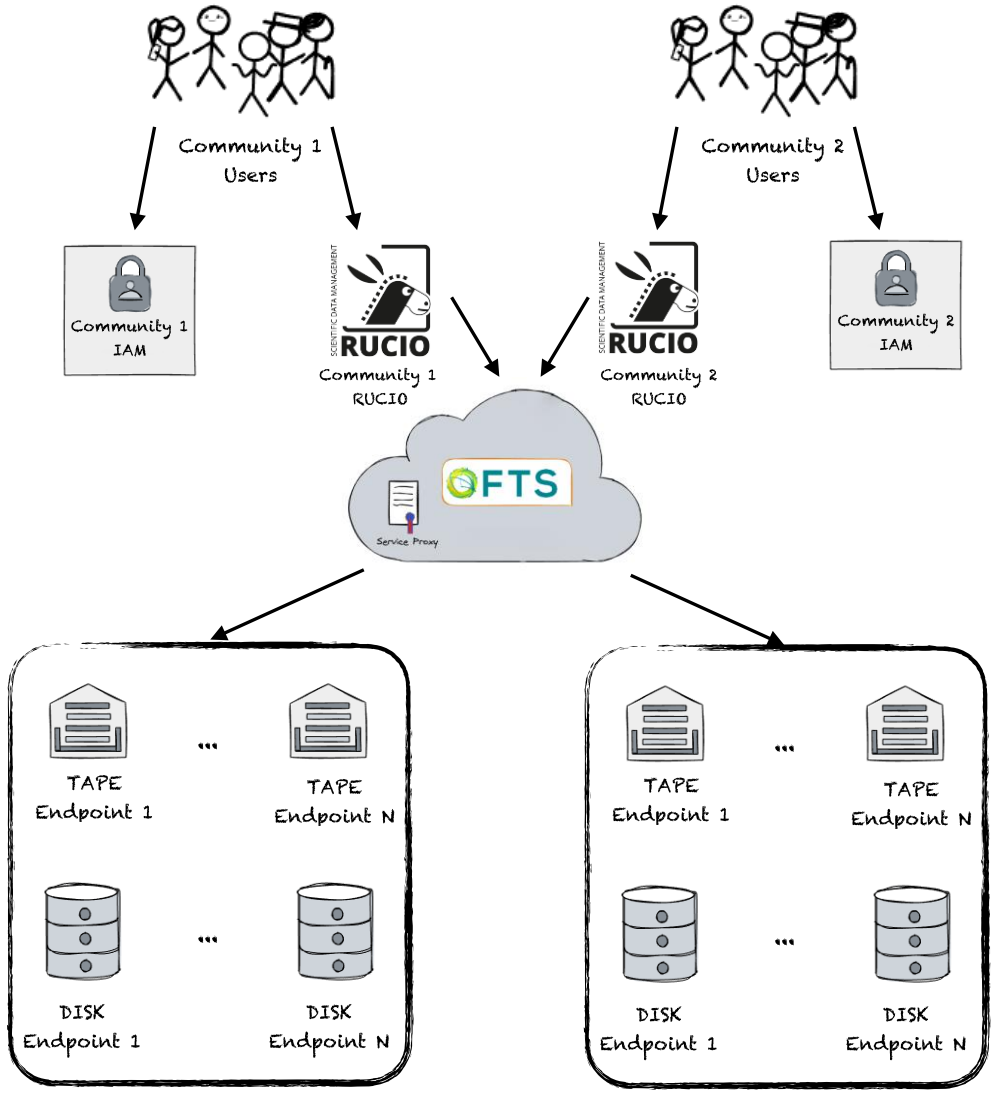
Pensiamo debbano essere previsti due modelli di deployment di un sistema di data management:

- Rucio gestito da una **collaborazione scientifica**: Il DM sarebbe a servizio di un gruppo/esperimento afferente ad una CSN, per cui esiste un pledge, o anche gruppi di utenti in collaborazione con quote storage;
- Rucio **«catch-all» nazionale**: il DM sarebbe a servizio di «utenti semplici» per i quali non esiste un pledge, o piccole comunità non strutturate che fanno uso dell'infrastruttura distribuita (ad esempio, gruppi di teorici).

DM di comunità

Ad uso di una comunità di utenti, deployato su un cluster K8s centralizzato gestito da Operations.

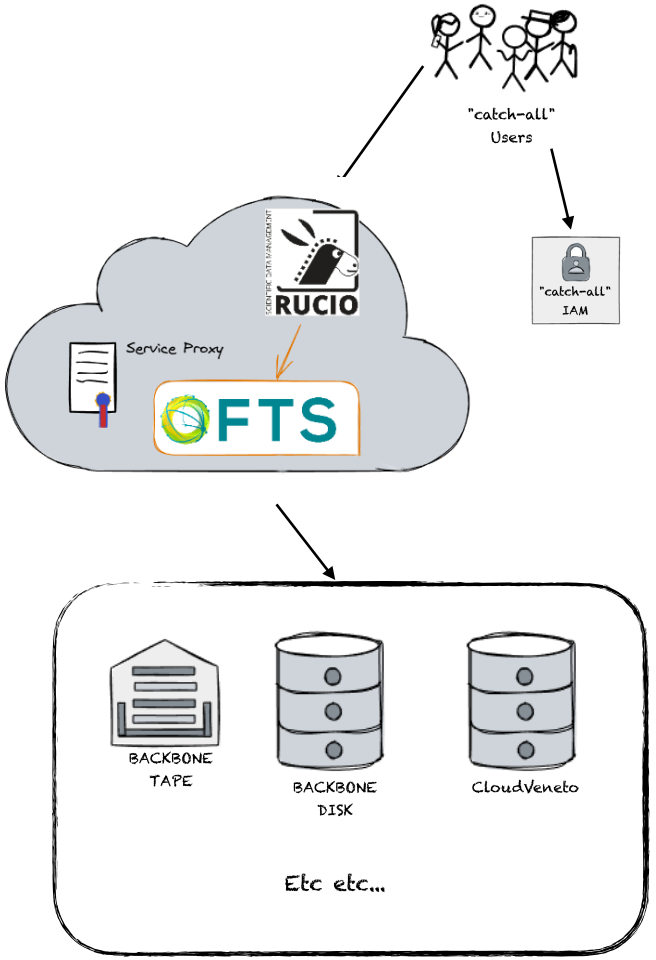
- Istanza RUCIO di comunità, gestita da essa;
- Integrato con uno IAM che gestisce gli utenti e le policies della comunità;
- FTS multi-VO (gestito centralmente);
- Integrazione degli storage *pledged* della comunità;
- Database gestito centralmente;
- DataCloud si fa carico del supporto.



DM «catch-all»

Istanza di DM messa a disposizione di utenti «singoli».
Esattamente come per il DM di comunità, i servizi ancillari (cluster K8s, database, FTS), sono gestiti centralmente dal team Operations.

Permette ad un utente singolo di utilizzare il DM anche per dati «personali», con una singola interfaccia a fronte di diversi backend storage --> semplificazione.



- Istanza di RUCIO gestita centralmente;
- IAM «catch-all» gestito centralmente;
- Federazione degli storage dei vari centri di calcolo dell'INFN.

Alcuni prossimi passi (sempre loro)

- Onboarding di altre comunità interessate e ricerca di utenti interessati «personalmente»;
- DM nazionale «catch-all»
- Integrazione delle evoluzioni architetturali:
  - Storm REST API in sostituzione di SRM anche per il tape;
  - Uso di token al posto di VOMS proxy anche per i TPC fatti da FTS;
  - Etc …
- Integrazione con *metadata catalog* esterni
  - Sinergie con ICSC spoke3

Osservazioni finali

In base all'attività svolta, possiamo affermare che il sistema non presenta criticità implementative.

Il sistema presentato è uno dei quattro elementi portanti dell'infrastruttura datalake di ICSC/Terabit ed è un elemento portante delle attività di Spoke2, in particolare il progetto IDL con Leonardo (innovation grant).

Al fine di passare ad un'eventuale fase di produzione, riteniamo importante un'analisi/discussione. I temi d'affrontare, tra gli altri:

- Valore aggiunto nell'assegnare una quota storage ad «utenti semplici» su storage distribuito?
- Impegno non trascurabile di Operations nei confronti delle CSN mid/long term;
- Necessità di pianificazione «periodica», soprattutto in fase iniziale.

Grazie per l'attenzione

antonino.troja@pd.infn.it