Collaborative Ground Segment Workshop
Proposed Plan

19 October 2021

Proposed Plan for Collaboration

2022:

- Maintain on-going operations
- Include new data sets
- Complete first cycle of updates for the collaborative hub software / environment

2023:

- Integrate the future data access services interfaces
- On-demand production services / streamlined data access

2024:

- Data relays fully functional as part of the ecosystem
- Supporting services like:
  - Large data bulk transfer
  - Dedicated streamlined access …

New Copernicus Data Access Available

Processors in Python

Support Services

User Management

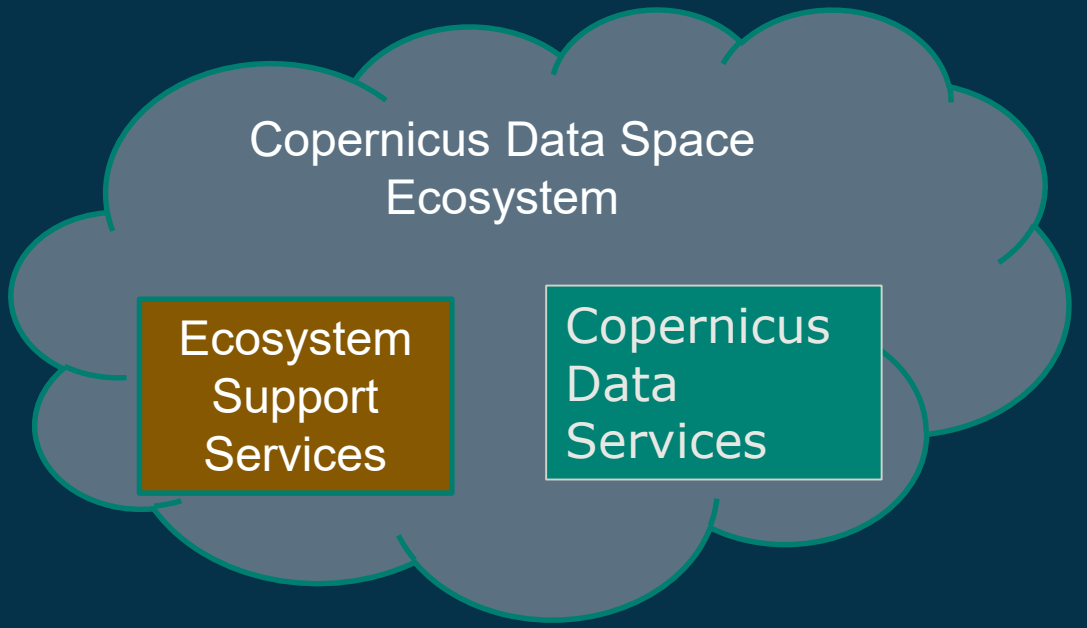
IaaS

Monitoring

Configuration Control

Satellite Operations

Acquisition & data Provision



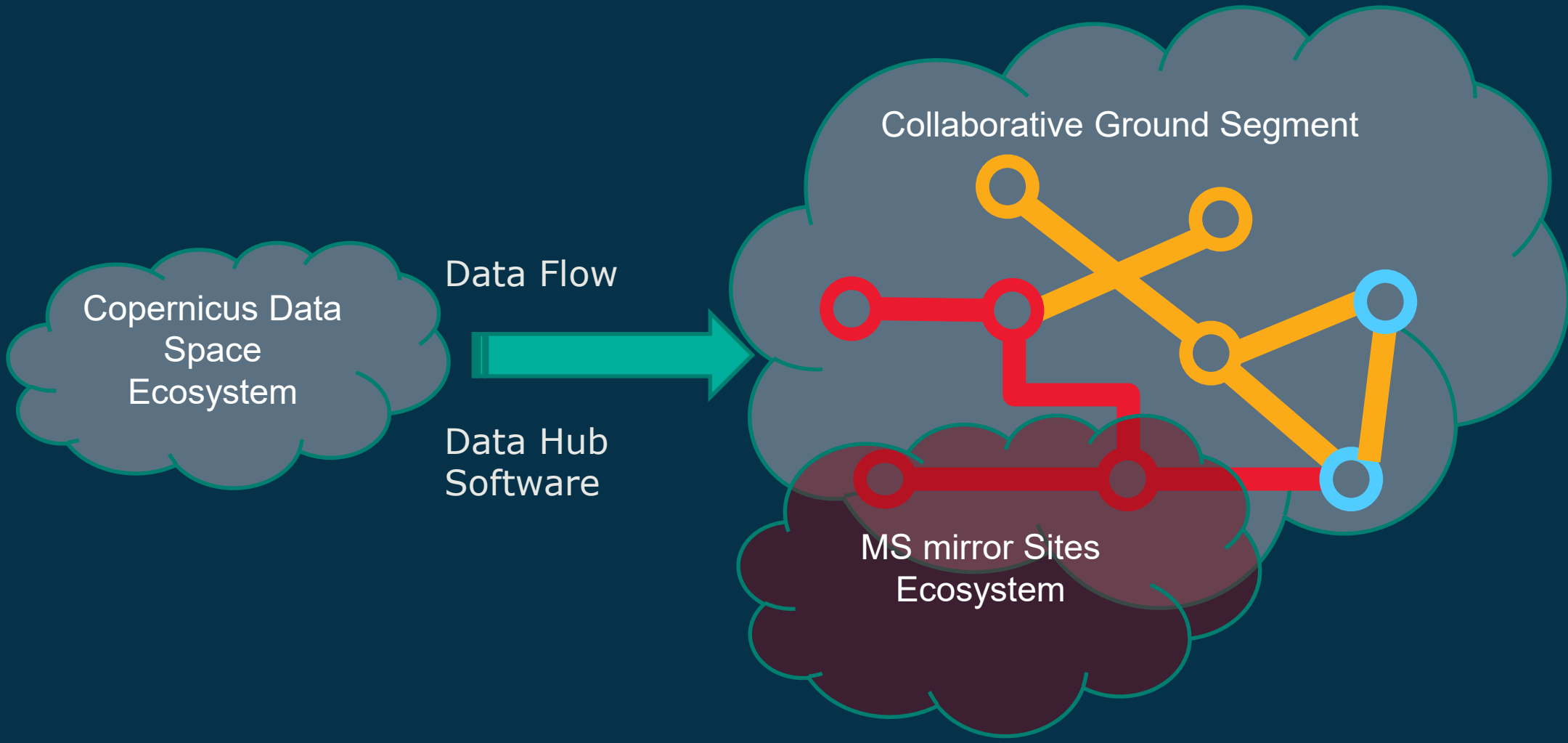
Copernicus Data services

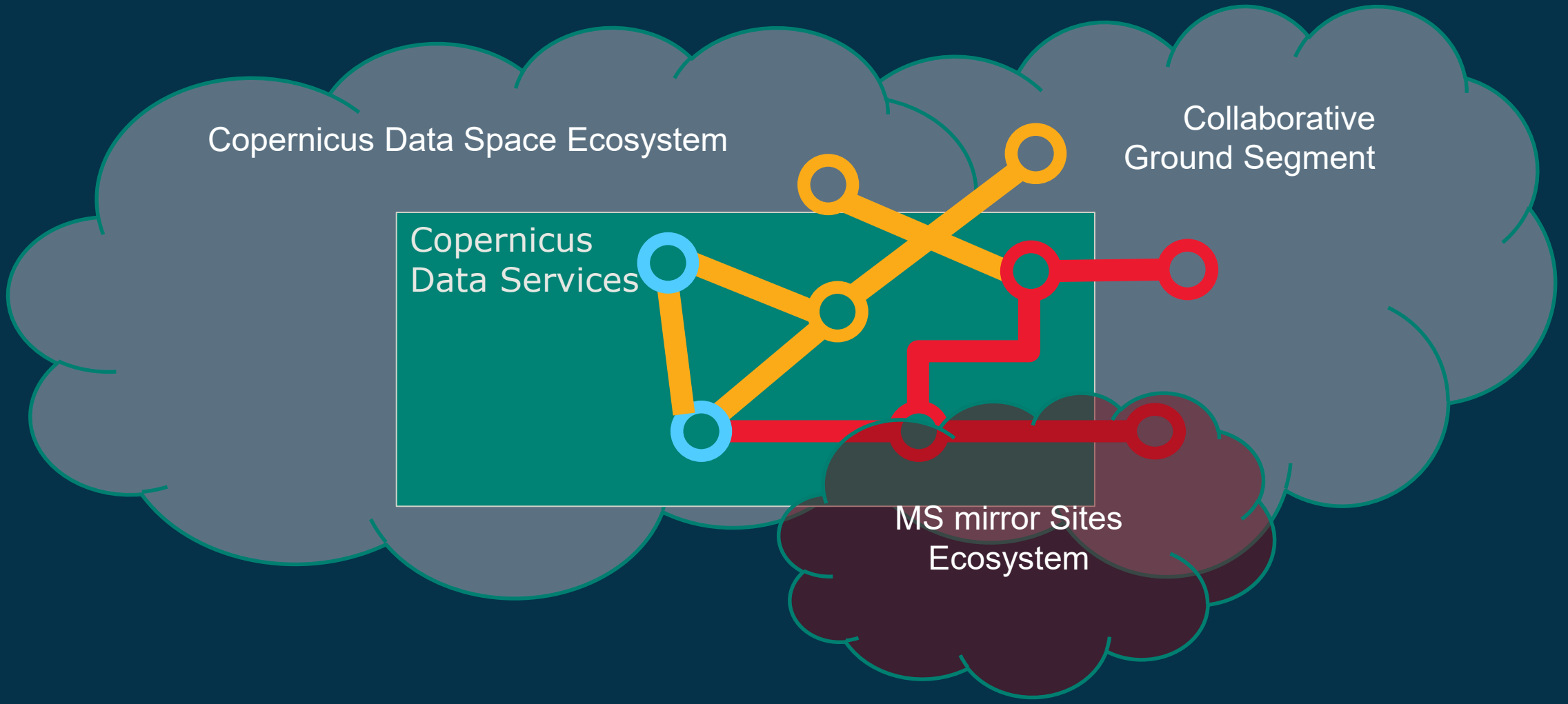
Production

Archiving

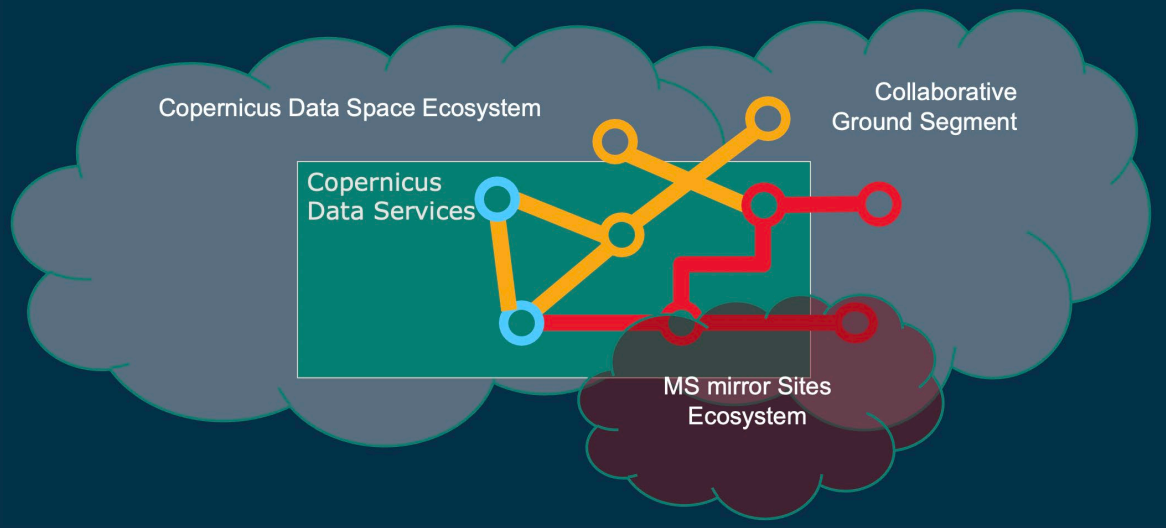
Distribution

...


Collaborative Ground Segment
Copernicus Data Space Ecosystem
Data Flow
Data Hub Software
MS mirror Sites Ecosystem


Copernicus Data Space Ecosystem
Collaborative
Ground Segment
Copernicus
Data Services
MS mirror Sites
Ecosystem

What does it mean?

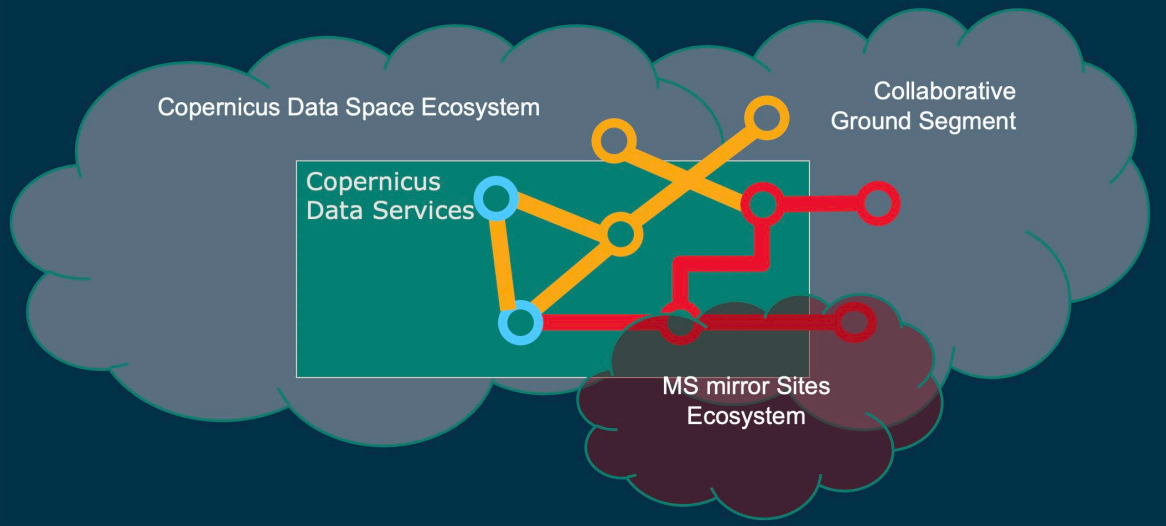


Copernicus Data Space will recognise the collaborative Ground Segment as a potential participant

Copernicus Data Space services will become implicitly available to Collaborative GS

e.g. from news to traceability, from data access to data storage

What does it mean?



Through Collaborative Programme component (end 2024-2025) ESA will support:

- Data relays integration into the ecosystem
- Usage of ecosystem services (e.g. traceability, on-demand production, storage, news, marketplace,…)
- Development of new collaborative services

Collaborate Not dissolve into:

- Authority for respective level of integration always on the MS side

From yesterday Presentation

How to make available Proposed Data from CNES

.

Other points

Traceability Service
for EO Data Products

Update

Collaborative Ground Segment Workshop

ESA-ESRIN, 18-19 October 2021

Christophe Demange

GAEL Systems

Summary of Traceability Service's Objectives

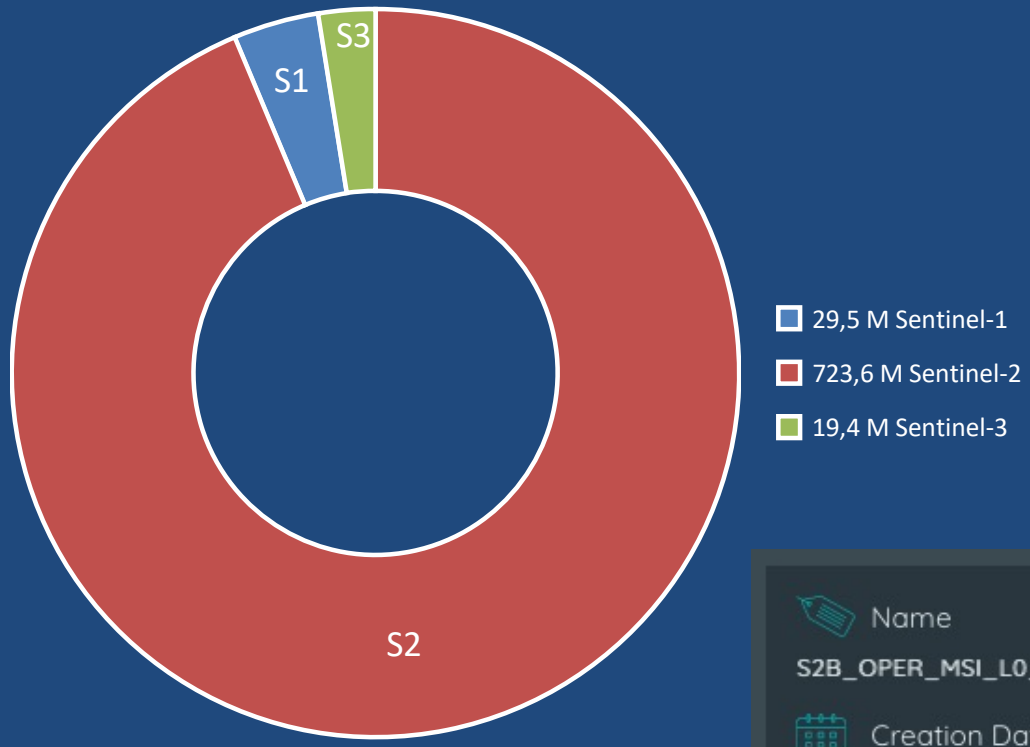
- To record product lifecycle related events
- To allow retrieving product history
- To verify if a product copy is genuine
- To detect potential data tampering
- To enhance trust on official processing chains
- To allow notifying about important changes

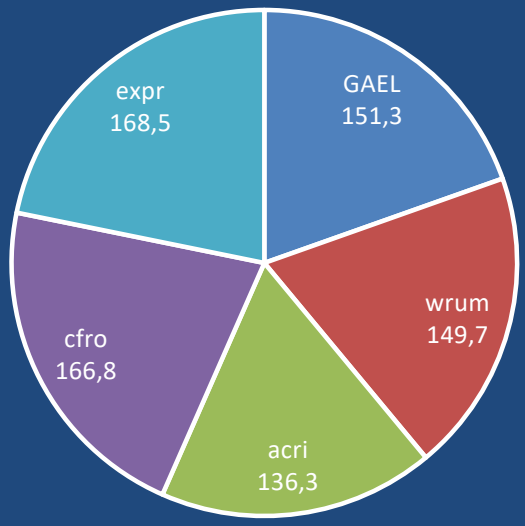
Current usage of the Traceability Service

- From Data Bridge Service:
  - Ingestion of S1, S2 & S3 Level-0 and AUX data
  - Transcription of S2 Logical products to PDI
  - Metadata extraction / Index
  - Sign and add new Traces

- From new LTA Services:
  - Download products from Data Bridge (& Production Services)
  - Verify existing Traces (for products coming from DBS)
  - Sign and add new Traces

772 million Traces so far !





Main features

- Service Owner / Service Providers
- Service Type Schema / Event Type
- Fronted API & UI (add, search, verify, key request)
- Tracetool (client side)
  - Hash & sign to prepare new trace
  - Optional sub-product elements in hashlist
  - Legacy mode to compute former hash algos
- Demo Service for newcomer testing/on-boarding

  `https://demo.trace.gael-systems.com/`

Traceability API examples

base URI : `https://trace.gael-systems.com/trace-api`

Find Traces by hash:

```
<baseURI>/Traces?$filter=hash eq
'ba3dae25d4d7a7032fc13fbf27d072a90df382998ce25a663bc7d81ef61fdc67'
```

Find Trace by product name (using json output):

```
<baseURI>/Traces?$format=json&$filter=name eq
'S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07.tar'
```

Verify a Trace (e.g. signed key is not revoked):

```
<baseURI>/Traces(21fb6171-547d-4ebb-82ae-84a6b21dfcfe)/Datac.Quality()
```

Sample Trace content


```
{
   id: "21fb6171-547d-4ebb-82ae-84a6b21dfcfe",
   date: "2021-09-21T23:22:18.893Z",
   hash: "ba3dae25d4d7a7032fc13fbf27d072a90df382998ce25a663bc7d81ef61fdc67",
   hashAlgo: "GAEL_HASH_SHA3-256_V2",
 - hashList: [
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B01.bin:4eb7b6fc0069556b3a68f5717a9fdd1ae264efd901e3ddb2000d090d61409fae",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B02.bin:ac53510dd9676725002236138188bed8d474f47612365de4c80eee502d10cea67",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B03.bin:162450cda4008301a3e7ab34518b3f4985ff34f30db086a7c58045908a4f27ad",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B04.bin:74d274a0a13b994a51a825fd113d9b9956430914197881c8a6d069298d909801",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B05.bin:a1d8dc8798f83cf2d0623edfef9fa643f88ed447fb462dd2da4e4b305a97fc2c",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B06.bin:82d1bdbfbb33c7653f4884fca45e21a16f010a11e6acac5b0eed5842214924d9",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B07.bin:466a543ebcd84458a758ddbc13fdde7a7e38fa5be7b65f8096c263cae5955114",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B08.bin:889059225199032 48d2b1dd54466d3e8929f1c064f07d623d515cdde4626a9aa",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B09.bin:f1bcc5b78a2fb8c28d7259200a5cc6c373a2bfca0ae082b3ff516dbb65fcf3b9",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B10.bin:0c3c25b19b8a9a509fa7989ae61c8742a55c8f1c7337b9376648a52bb5fb21fd",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B11.bin:7a5c0831be7a59c9c92673f2d8cb8a0a33551069f9dca5750289c55fbbb36d9f",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B12.bin:1d8abf816965b599c905d4d91bfe78234502be5596c8bccd8671fcf880b55904",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/IMG_DATA/S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_B8A.bin:5413b410e2a4bf8a8c41c7c944b730b029f8fb957d3881cc4d6ec0e8ebe4ccc3",
       "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07/manifest.safe:418daa7803f41e2ffd2399cdd467a7358bd931246d69c27913cdf6a8f718e293"
   ],
   signature:
   "LS0tLS1CRUdJTiBQR1AgU01HTkFUVVJFLS0tLS0KCmlRSXpCQUFCQ2dBZEZpRUVnMXpKWlltWlpLeFY1THNkdm01c1RjeWNuSkFGQW1GS2FTb0FDZ2tRdm01c1RjeWMKbkpBUWp4QUFra2twdUZoVERacUdxU1JqZHhuaC9WcjVaTmJkdml5ZGRCa0E2djErT1RSUUtsNjVrcGV3YnJlcgo4YXBI
   signatureAlgo: "GAEL_SIGN_V2",
   size: 18388992,
   name: "S2A_OPER_MSI_L0__GR_SGS__20190324T154736_S20190324T092305_D09_N02.07.tar",
   serviceContext: "DataBridgeService",
   serviceType: "Long Term Archiving",
   serviceProvider: "GAEL",
   publicKey: null,
   publicKeyFingerprint: "835CC965899964AC55E4BB1DBE6E6C4DCC9C9C90",
   eventType: "CREATE",
   referenceId: "ff7c2125-8d64-335c-884c-789f3f9b1be0",
   beginningDateTime: "2019-03-24T09:23:05.000Z",
   cloudCover: 0,
   dataSource: "INDRA",
   datastripId: "S2A_OPER_MSI_L0__DS_SGS__20190324T154736_S20190324T091226_N02.07",
   endingDateTime: "2019-03-24T09:23:05.000Z",
   instrumentShortName: "MSI",
   orbitNumber: 19595,
   platformSerialIdentifier: "A",
   platformShortName: "SENTINEL-2",
   processingCenter: "SGS_",
   processingDate: "2019-03-24T15:47:36.000Z",
   processorVersion: "02.07",
   productGroupId: "GS2A_20190324T085621_019595_N02.07",
   productType: "MSI_L0__GR",
   qualityStatus: "NOMINAL"
}
```

Recent improvements

- Infrastructure scalability enables adding several million traces per day (250/s tested, can do more)
  - New frontend servers, multi-regions
  - Queue management (absorb bursts, bulk operations)
  - DB & Index optimisation (collections, shards)
- Docker version of Tracetool
  - All-in-the-box deployment
  - Isolation to cope with GPG multi-threading limitations and to help large parallelisation

Coming next & under study

- Use X509 certificates to replace GPG keys
  - Benefits of more standard libraries in various languages
  - CA Chain of Trust is more adapted to Service Owner / Providers
- Use faster hash algorithm (e.g. BLAKE3 instead of SHA3-256)
- Improve signature process
- Advanced Search / Full text search
- Notification trigger on specific events (revoke, delete…)
- Python version of Tracetool and/or use of WebAssembly (WASM)?
- Link from our product Catalogs (view traces from product details)
- Mutualized or dedicated platforms (for various Service Owners)
- Traceability-as-a-Service Business Model

Thank you !

Contact us:

info@gael-systems.com

Prototyping an Earth Observation ecosystem
Collaborative GS workshop

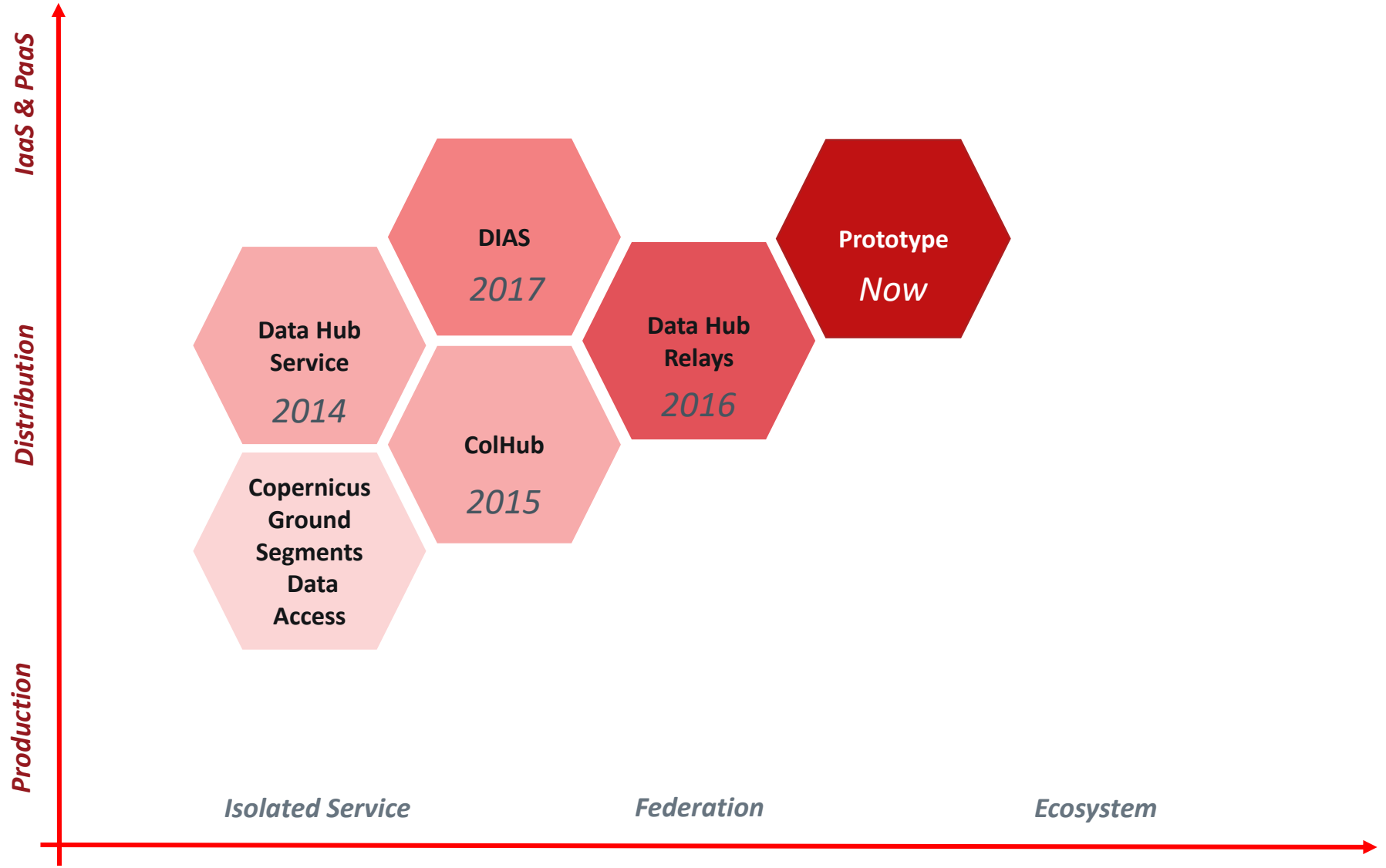
19th October 2021

Gaia Cipolletta, System Engineer
gaia.cipolletta@serco.com

Overview
Meeting Agenda

- Context and objectives
- Main concepts
- Participants Model
- Federations as a use case
- Services baseline
- Proof-of-concept & hands-on session

Services evolution chart



serco

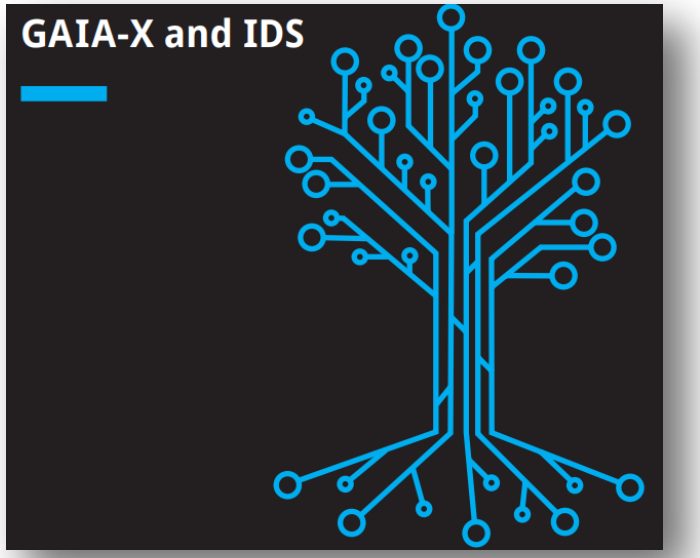
An Ecosystem vision

Context	Initiatives defining an open, secure and transparent environment where to exchange data and resources (e.g. ***Gaia-X*** and ***International Data Spaces***)
Motivation: define an EO ecosystem	Collaborate with other partners
	Federate users, infrastructure and data with open source reference implementations
	Facilitate the identification of user requirements and conceptualise relevant use cases
	Operate in autonomy across the value chain

Gaia-X & the International Data Spaces

Gaia-X and IDS initiatives provide main concepts of an **open, transparent and secure digital ecosystem**.

To this aim, both provide a reference architecture model, governance aspects and usage control policies, in line with **data sovereignty** and **decentralization** principles.

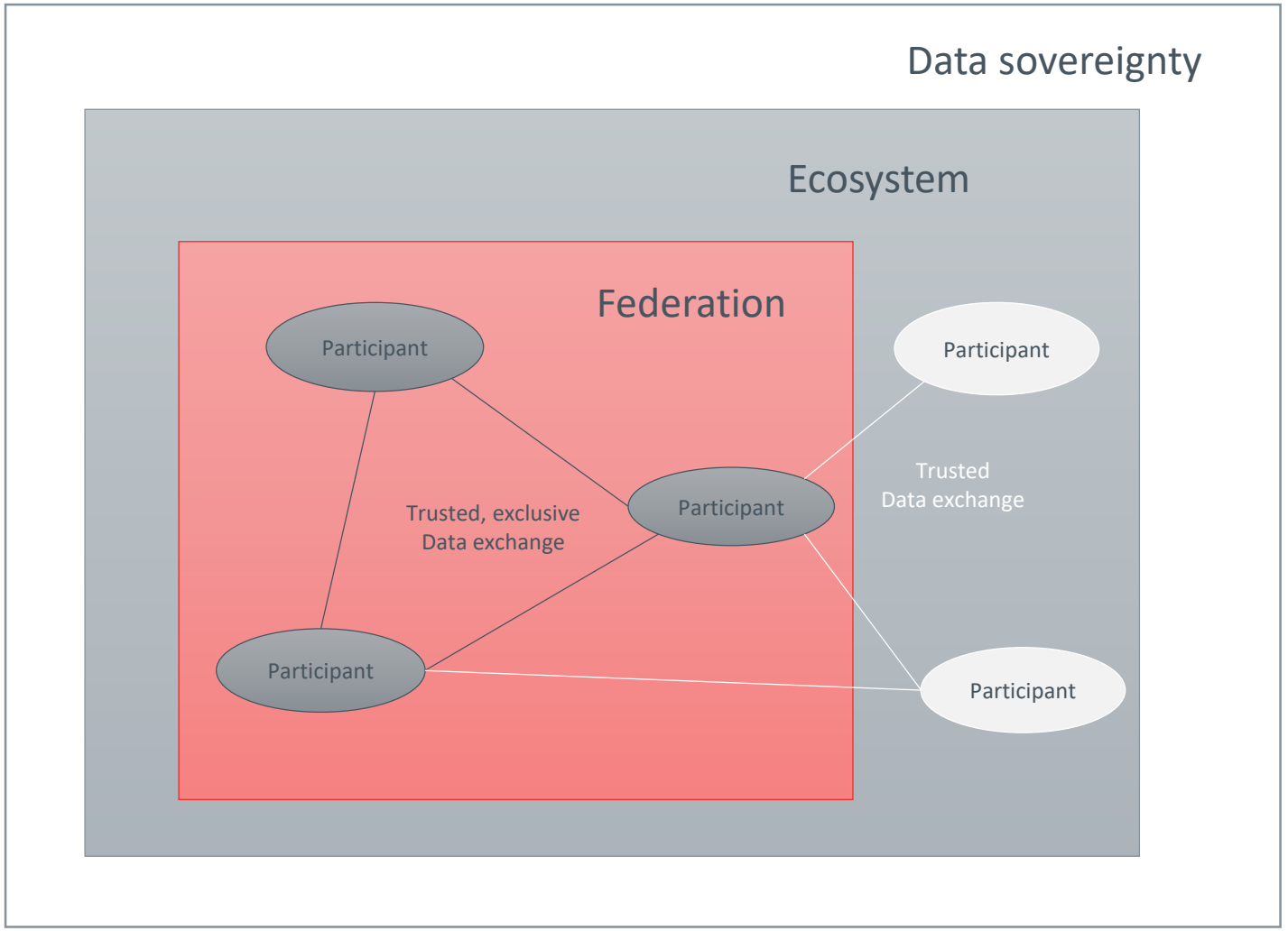


References

- Gaia-X architecture document v21.06
- IDS Reference Architecture Model v3.0

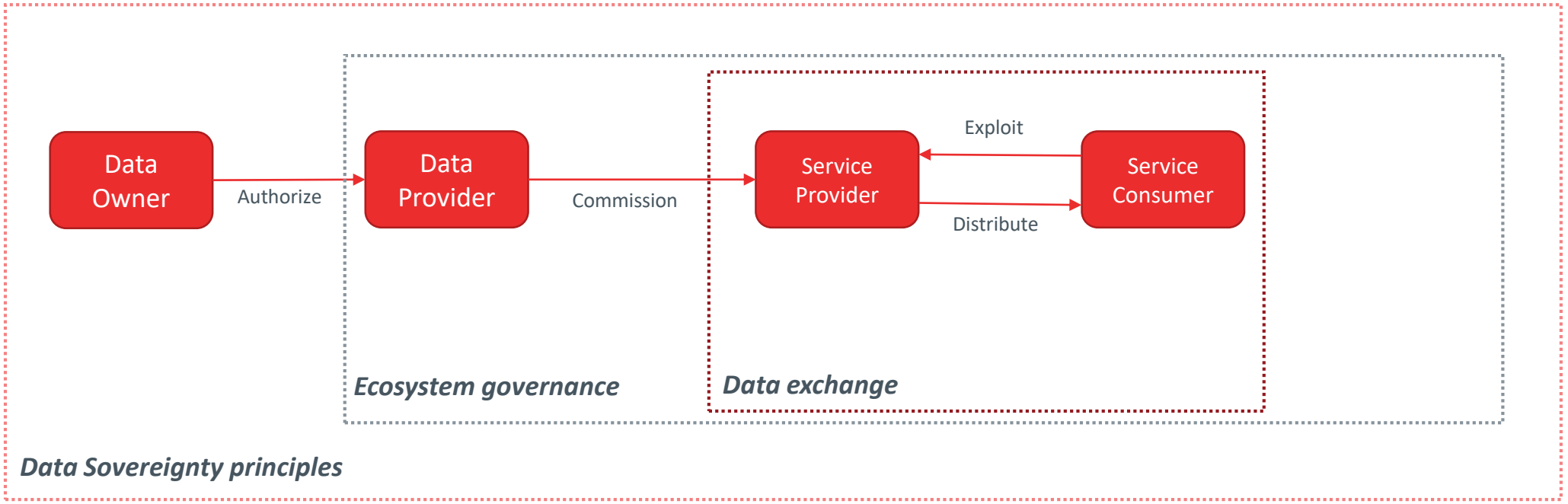
Concepts

- ***Data Sovereignty*** principle for autonomy and self-determination in terms of data exchange
- ***Ecosystem*** defines the governance boundary and the technical context to enable the data exchange
- ***Federations*** follow the idea of a partnership where members can share data and services while applying policies and maintaining sovereignty over data
- ***Participants*** are actors of the ecosystem, showing roles moderating over the interactions of data exchange

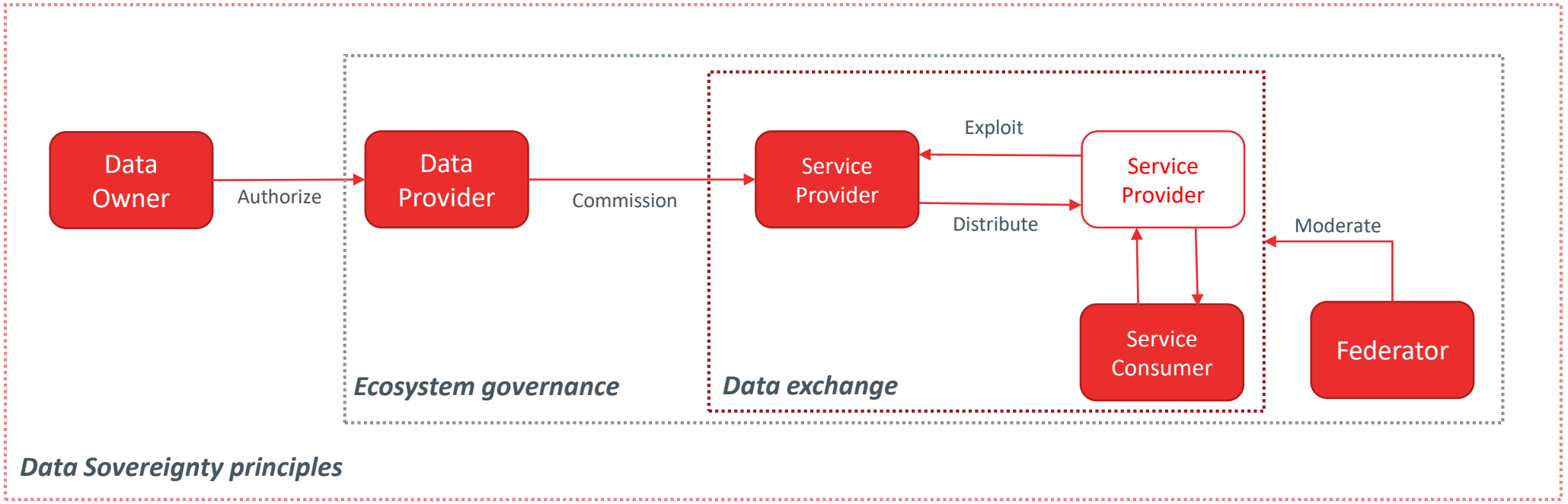


serco

Participants Model



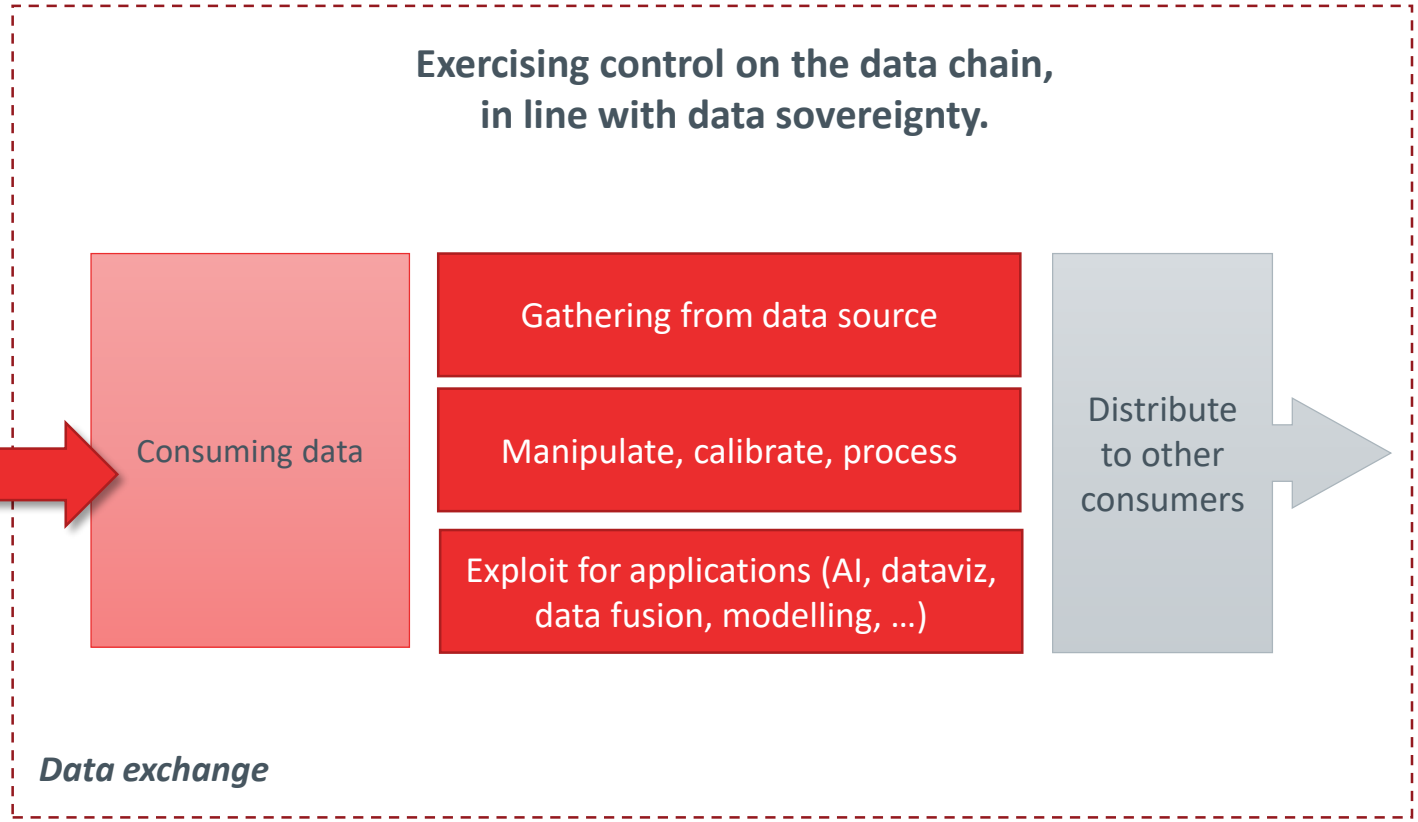
Participants Model



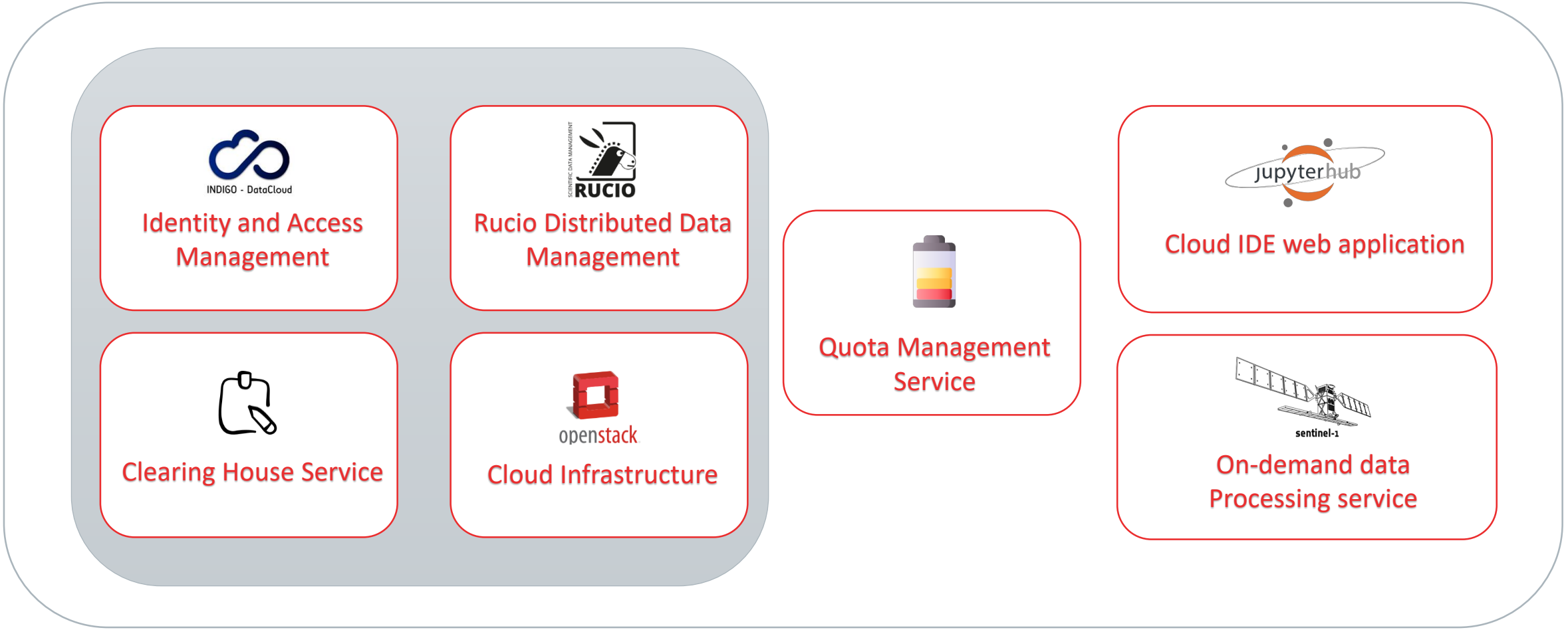
Federations Use Case

Term	Level
Scope	Governance
Group	**Exchange policies/Federation**
Data Space	Data Sovereignty, Operational

Data Space (*ref.* IDS): digital environment comprising data assets object of a secure and standardised data exchange.



Proof-of-concept: a proto-Ecosystem



serco

Introduction to services
Indigo-IAM and Rucio as PoC technical solutions

Indigo-IAM (INFN)

Indigo-IAM is an open source solution for the **Identity and Access Management** provided by the INFN.

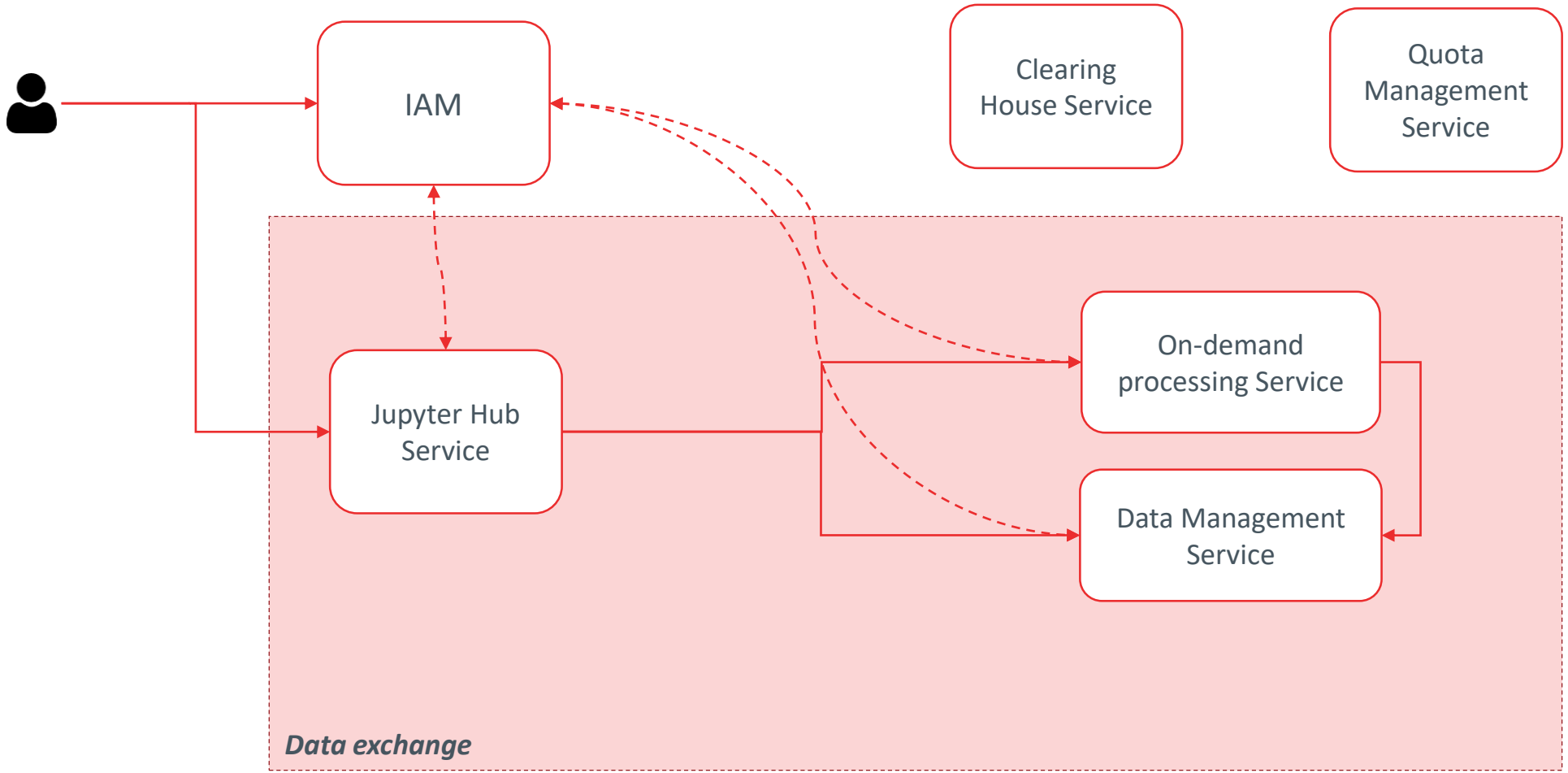
It supports multiple authentication mechanisms (credentials, OIDC, SAML, X.509 certificates) and groups of users.

Rucio (CERN)

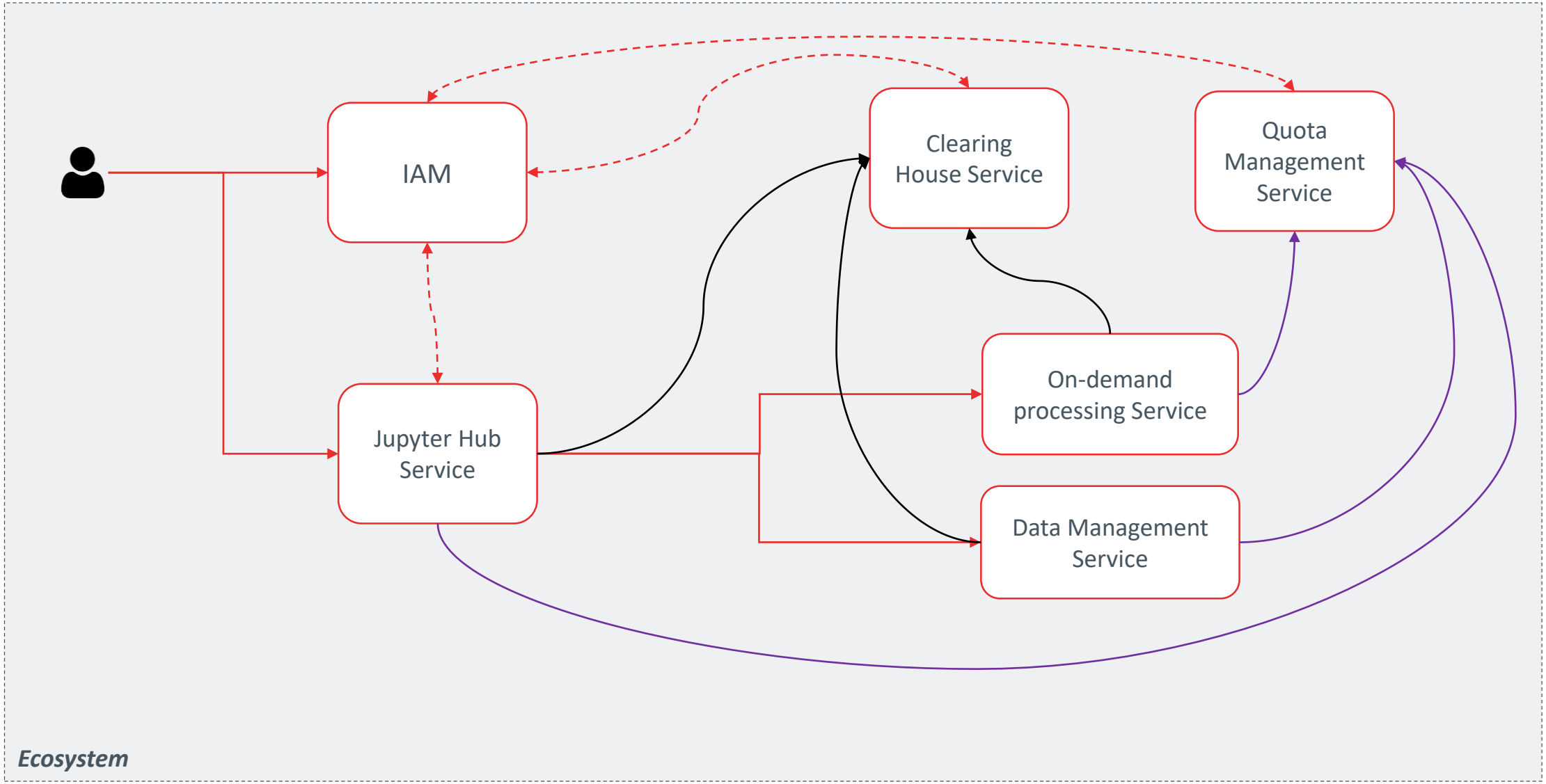
Rucio is an open-source framework to organize, manage, and access data at scale within a distributed storage system.

Rucio is based on a few interesting operational concepts of data management as **Namespaces** (Data Structures), **Storage Elements**, **Accounts** and **Replicas**.

Sequence of Events



Sequence of Events



Scientific data management with Rucio

Martin Barisits (CERN)

Collaborative Ground Segment WS
19 October 2021

CERN

Rucio in a nutshell

Rucio provides a mature and modular scientific **data management federation**

- **Seamless integration** of **scientific and commercial** storage and their network systems
- Data is stored in **global single namespace** and can contain **any potential payload**
- Facilities can be **distributed at multiple locations** belonging to **different administrative domains**
- Designed with **more than a decade of operational experience** in very large-scale data management

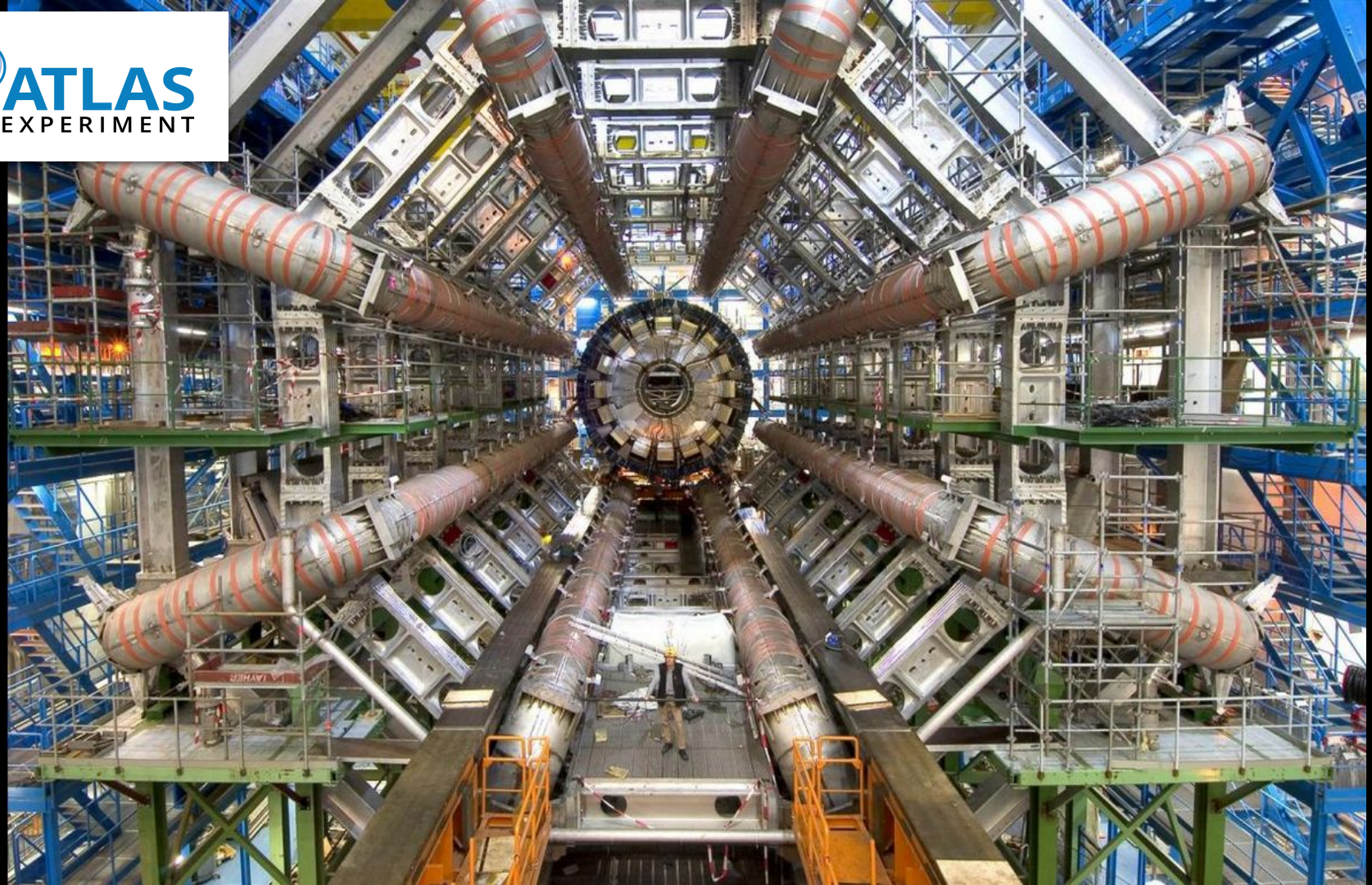
Rucio is location-aware and manages data in a heterogeneous distributed environment

- Creation, location, transfer, deletion, annotation, and access
- **Orchestration of dataflows** with both low-level and high-level policies

Principally developed by and for the ATLAS Experiment, now with many more communities

Rucio is free and open-source software licenced under *Apache v2.0*

Open community-driven development process

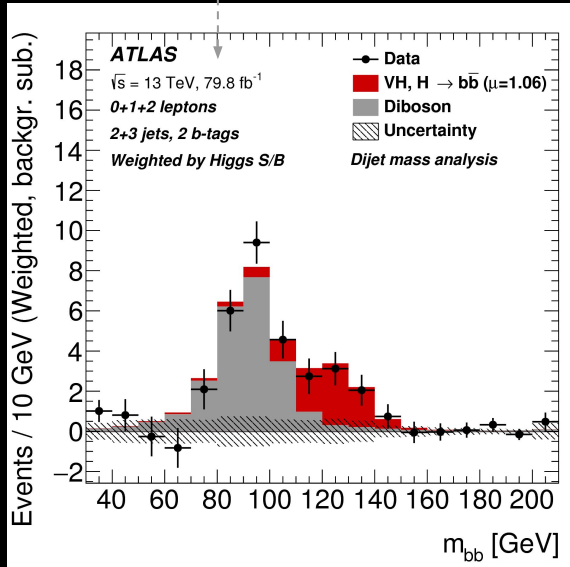

ATLAS
EXPERIMENT

ATLAS EXPERIMENT

Candidate Event:
$pp \rightarrow H(\rightarrow b\bar{b}) + W(\rightarrow \mu\nu)$
Run: 338712 Event: 335908183
2017-10-19 23:31:18 CEST

13 TeV detector data
8 quadrillion collision candidates
92 petabytes
130 million files

13 TeV simulation data
166 petabytes
544 million files



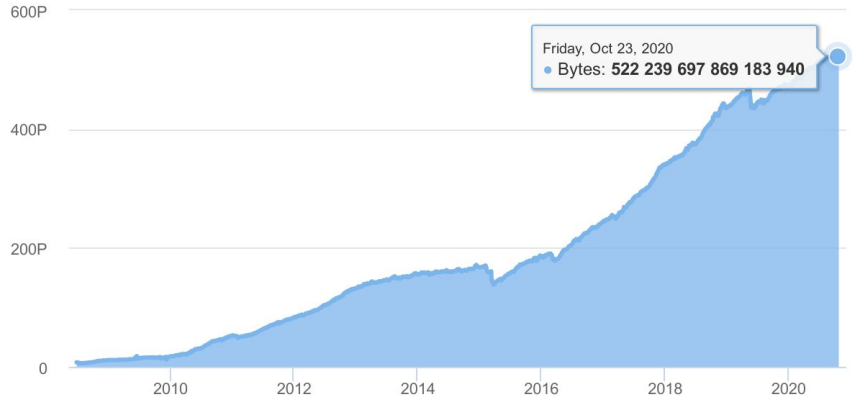
A candidate event display for the production of a Higgs boson decaying to two b-quarks (blue cones), in association with a W boson decaying to a muon (red) and a neutrino. The neutrino leaves the detector unseen, and is reconstructed through the missing transverse energy (dashed line). (Image: ATLAS Collaboration/CERN)

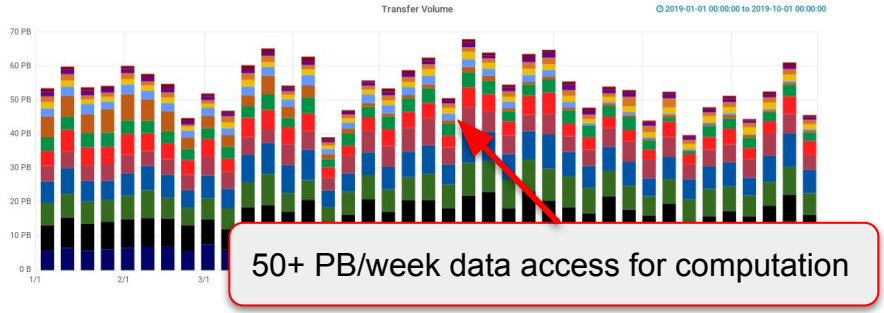
ATLAS

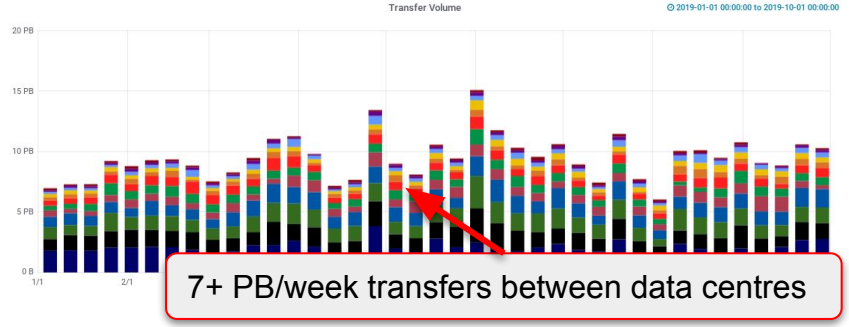
A few numbers to set the scale

- 1B+ files, 500+ PB of data, 400+ Hz interaction
- 120 data centres, 5 HPCs, 2 clouds, 1000+ users
- 500 Petabytes/year transferred & deleted
- 2.5 Exabytes/year uploaded & downloaded

Increase 1+ order of magnitude for HL-LHC







Community


Advanced European Network of E-infrastructures
for Astronomy with the SKA


LCLS


Belle II

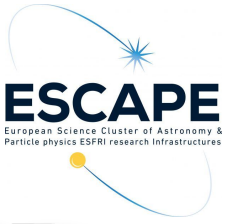

Science & Technology
Facilities Council


COMPASS


CERN

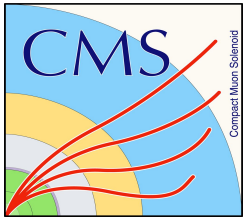

ARC


ATLAS
EXPERIMENT


ESCAPE
European Science Cluster of Astronomy &
Particle physics ESFRI research Infrastructures


Alpha Magnetic Spectrometer
AMS-02
Europe • Asia
North America


Xe
XENON
Dark Matter Project


CMS
Compact Muon Solenoid

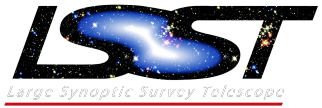

DUNE
DEEP UNDERGROUND
NEUTRINO EXPERIMENT


egi


3D
EISCAT


VIRGO


cta
cherenkov telescope array


LSST
Large Synoptic Survey Telescope


IceCube
SOUTH POLE NEUTRINO OBSERVATORY


FTS


eXtreme DataCloud


Fermilab


BROOKHAVEN
NATIONAL LABORATORY


PAWSEY
supercomputing centre


SKA
SQUARE KILOMETRE ARRAY


LIGO


FOLDING
@HOME


Open Science Grid

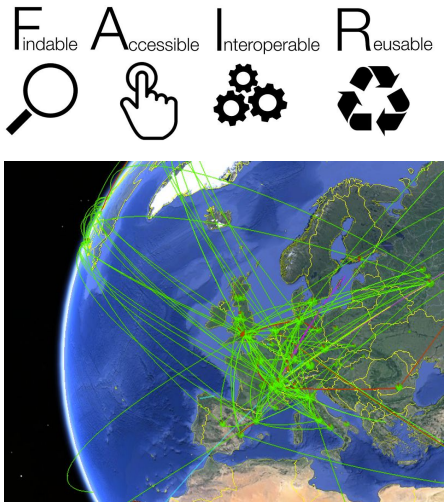

EUXDAT

Rucio main functionalities

Provides many features that can be enabled selectively

More advanced features

- **Horizontally scalable catalog** for files, collections, and metadata
- Transfers between facilities including **disk, tapes, clouds, HPCs**
- **Authentication and authorisation** for users and groups
- **Many interfaces** available, including CLI, web, FUSE, and REST API
- **Extensive monitoring** for all dataflows
- Expressive **policy engine** with rules, subscriptions, and quotas
- Automated **corruption identification and recovery**
- Transparent support for **multihop, caches, and CDN dataflows**
- **Data-analytics based flow control**

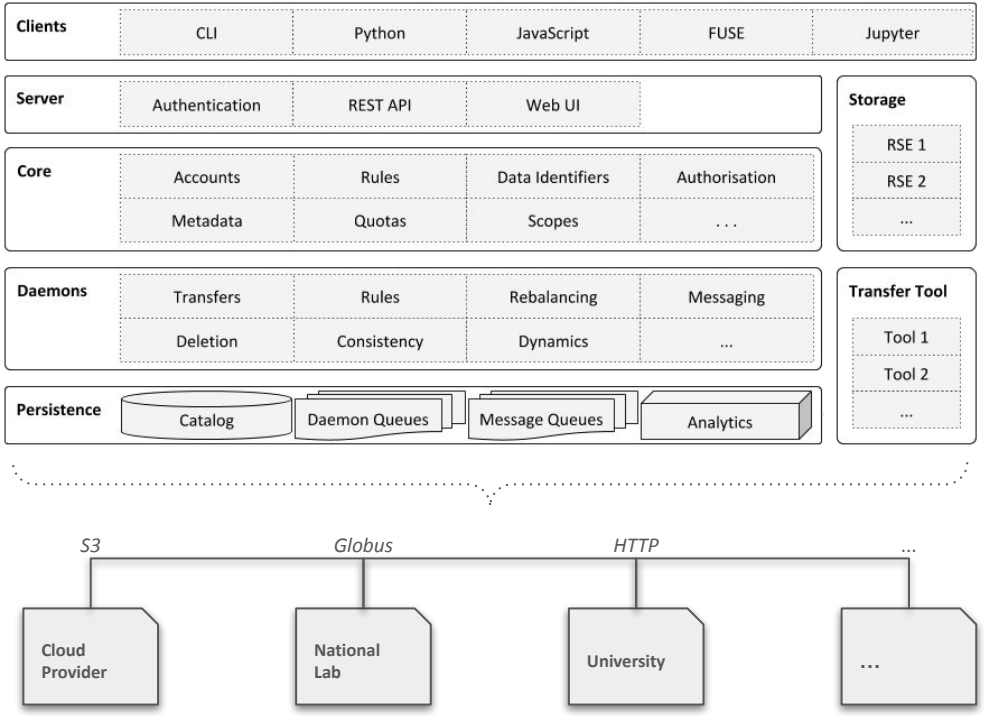


Rucio is not a distributed file system, it connects existing storage infrastructure over the network

No Rucio software needs to run at the data centres

Data centres are free to choose which storage system suits them best

High-Level Architecture



Horizontally scalable component-based architecture

Servers interact with users
- HTTP API using REST/JSON
- Strong security (X.509, SSH, GSS, OAuth2, …)
- Many client interfaces available

Daemons orchestrate the collaborative work
- Transfers, deletion, recovery, policy, …
- Self-adapting based on workload

Messaging support for easy integration
- STOMP / ActiveMQ-compatible protocol

Persistence layer
- Oracle, PostgreSQL, MySQL/MariaDB, SQLite
- Analytics with Hadoop and Spark

Middleware
- Connects to well-established products, e.g., FTS3, XRootD, dCache, EOS, Globus, …
- Connects commercial clouds (S3, GCS, AWS)

Summary

Rucio is an open, reliable, and efficient data management system

- Supporting the world's largest scientific experiments, but also a good match for smaller sciences
- Extended continuously for the growing needs and requirements of the sciences

Strong cooperation between physics and multiple other fields

- Diverse communities have joined, incl. astronomy, atmospheric, environmental, …
- Community-driven innovations to enlarge functionality and address common needs

Benefit from advances in both scientific computing and industry

- Lower the barriers-to-entry by keeping control of data in scientist hands
- Seamless integrations with scientific infrastructures and commercial entities
- Detailed monitoring capabilities and easy deployment have proven crucial

Thank you!

Website	http://rucio.cern.ch
Documentation	https://rucio.cern.ch/documentation
Repository	https://github.com/rucio/
Images	https://hub.docker.com/r/rucio/
Online support	https://rucio.slack.com/messages/#support/
Developer contact	rucio-dev@cern.ch
Journal article	https://doi.org/10.1007/s41781-019-0026-3
Twitter	https://twitter.com/RucioData


CERN

The INDIGO Identity and Access Management Service

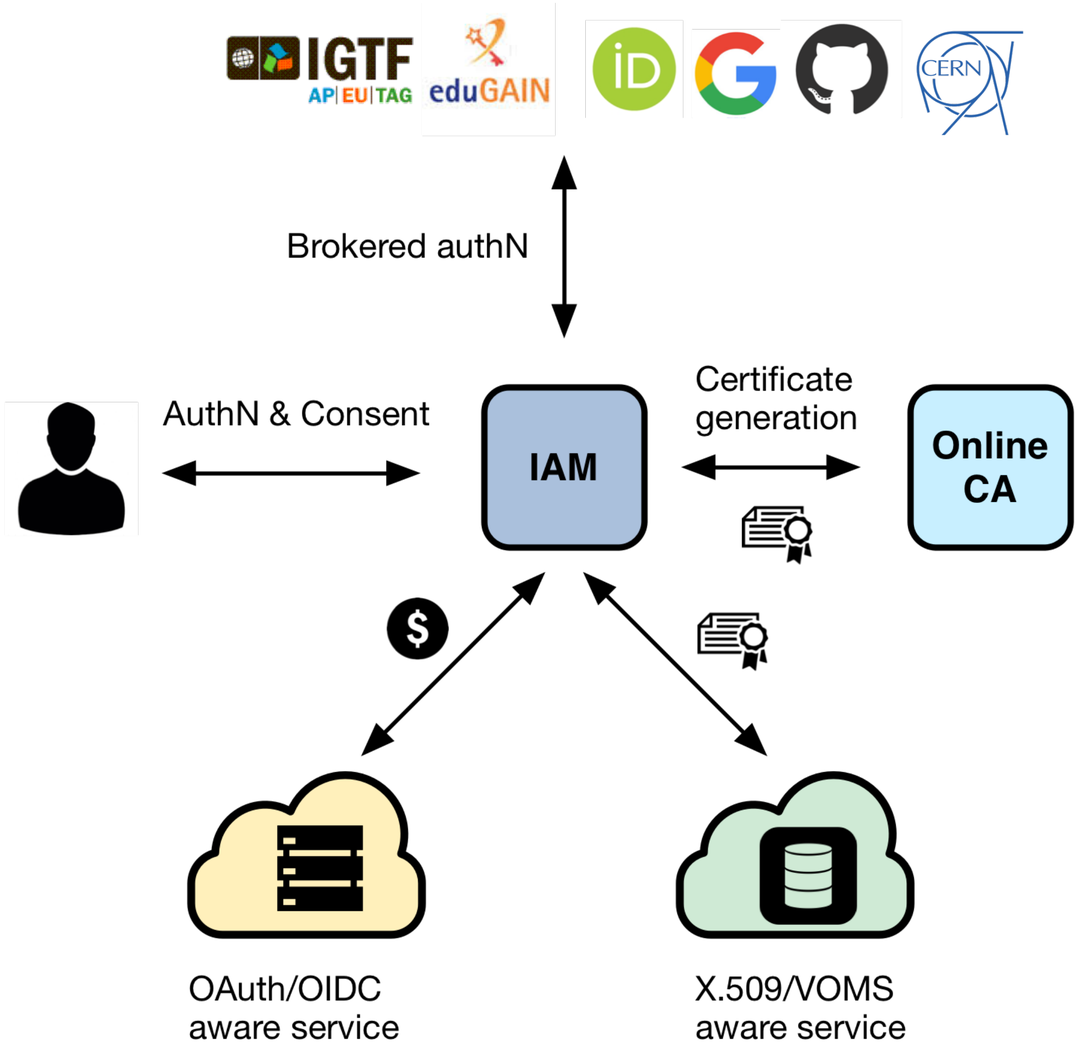
Andrea Ceccanti
INFN-CNAF

October 19th, 2021
Copernicus Collaborative Ground Segment Users Workshop

INDIGO Identity and Access Management Service

An authentication and authorization service that

- supports **multiple authentication mechanisms**
- provides users with a **persistent, organization scoped** identifier
- exposes **identity information**, **attributes** and **capabilities** to services via **JWT** tokens and standard **OAuth & OpenID Connect** protocols
- can integrate existing **VOMS**-aware services
- supports **Web** and **non-Web access**, **delegation** and **token renewal**

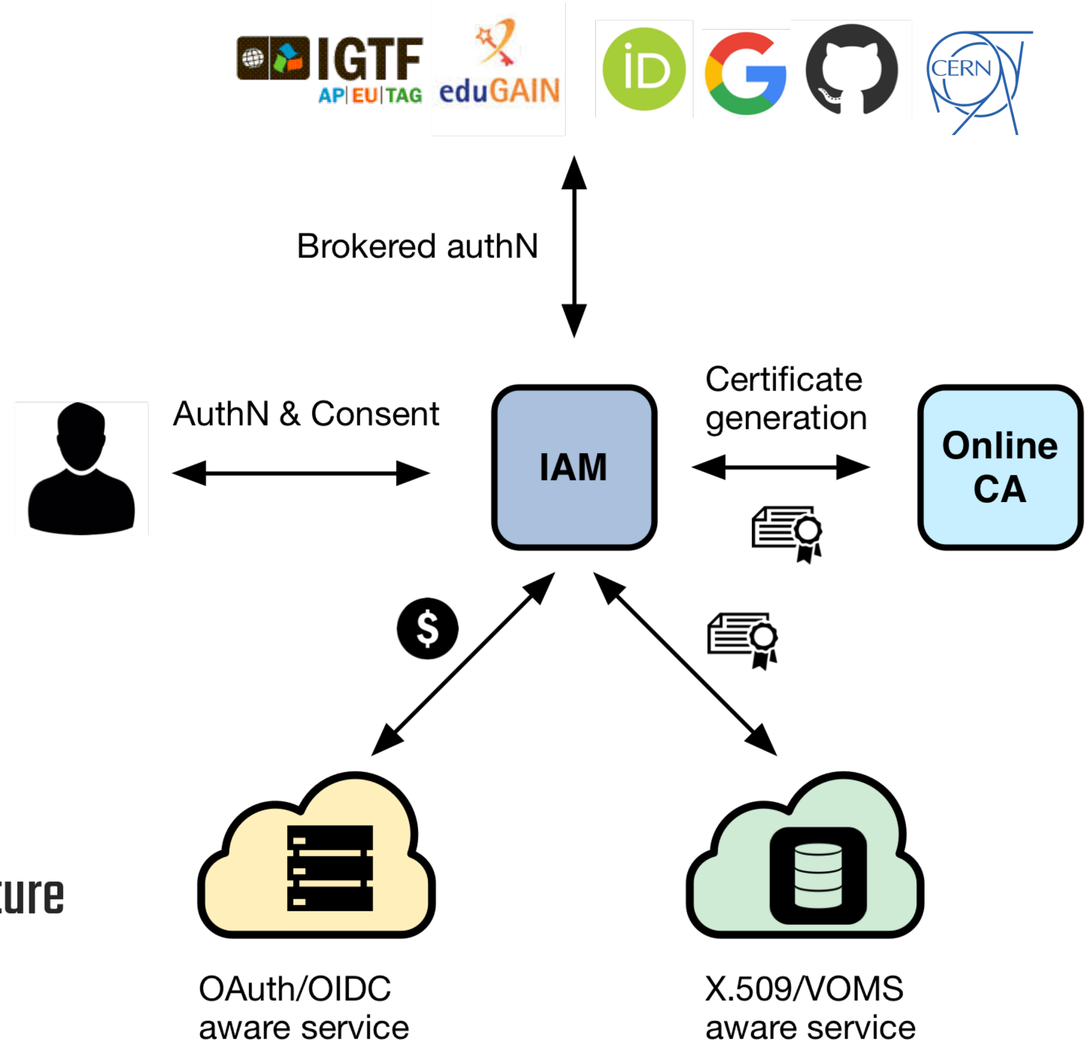


INDIGO Identity and Access Management Service

First developed in the context of the **H2020 INDIGO DataCloud** project

Selected by the Worldwide LHC Computing Grid (WLCG) management board to be the core of the future, **token-based** WLCG AAI

Sustained by INFN for the foreseeable future, with current support from:



Easy integration with relying services

Standard OAuth/OpenID Connect enables **easy integration** with off-the-shelf services and libraries.

IAM has been successfully integrated with

- Openstack, Atlassian JIRA & Confluence, Moodle, Rocketchat, Grafana, Kubernetes, JupyterHub, dCache, StoRM, XRootD (HTTP), FTS, RUCIO, HTCondor






Use case:
The Worldwide LHC Computing Grid

WLCG

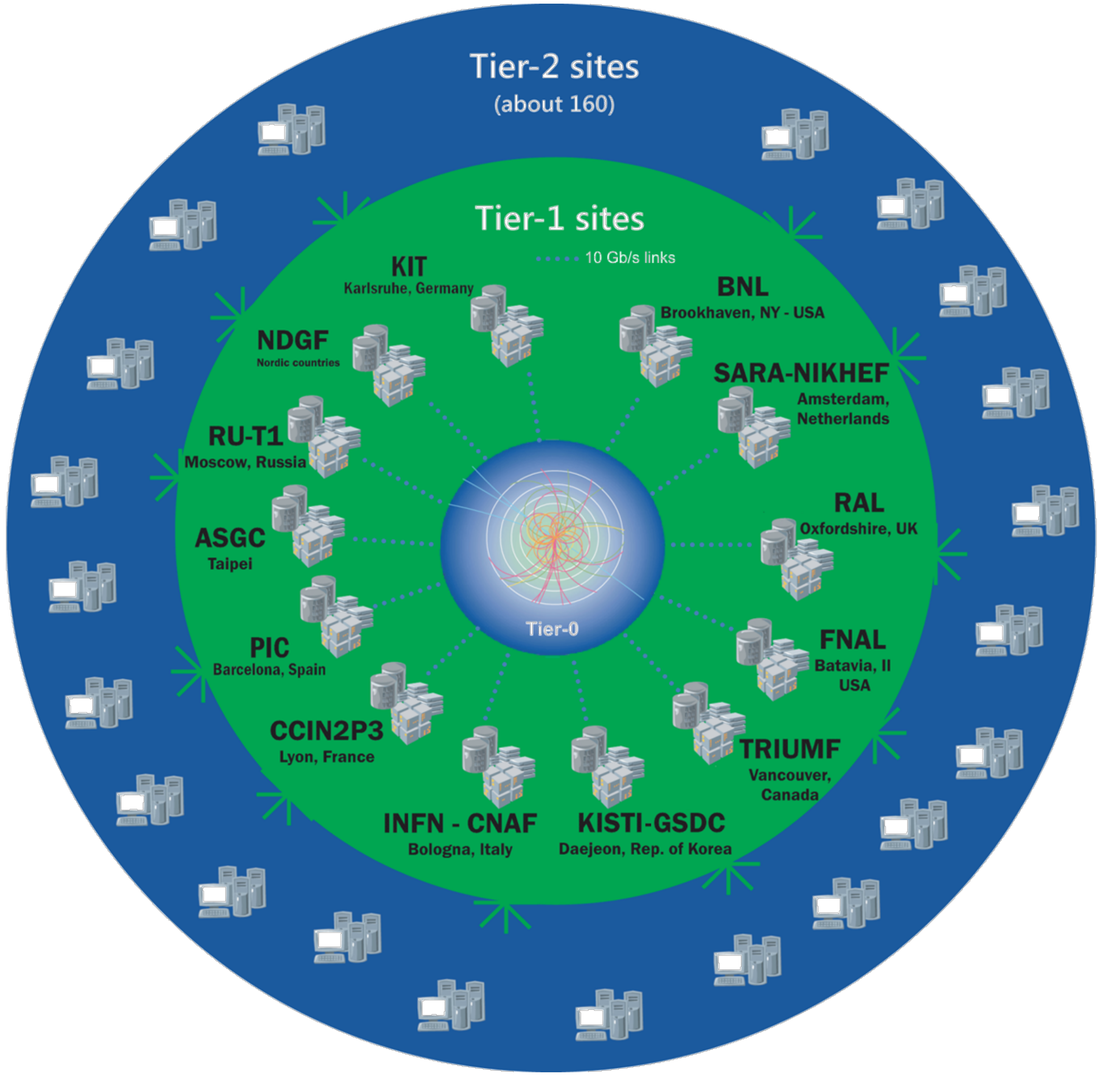
The Worldwide LHC Computing Grid (WLCG), is a distributed computing infrastructure arranged in tiers – giving a community of over 12,000 physicists near real-time access to LHC data.

167 sites, 42 countries

~1M CPU cores

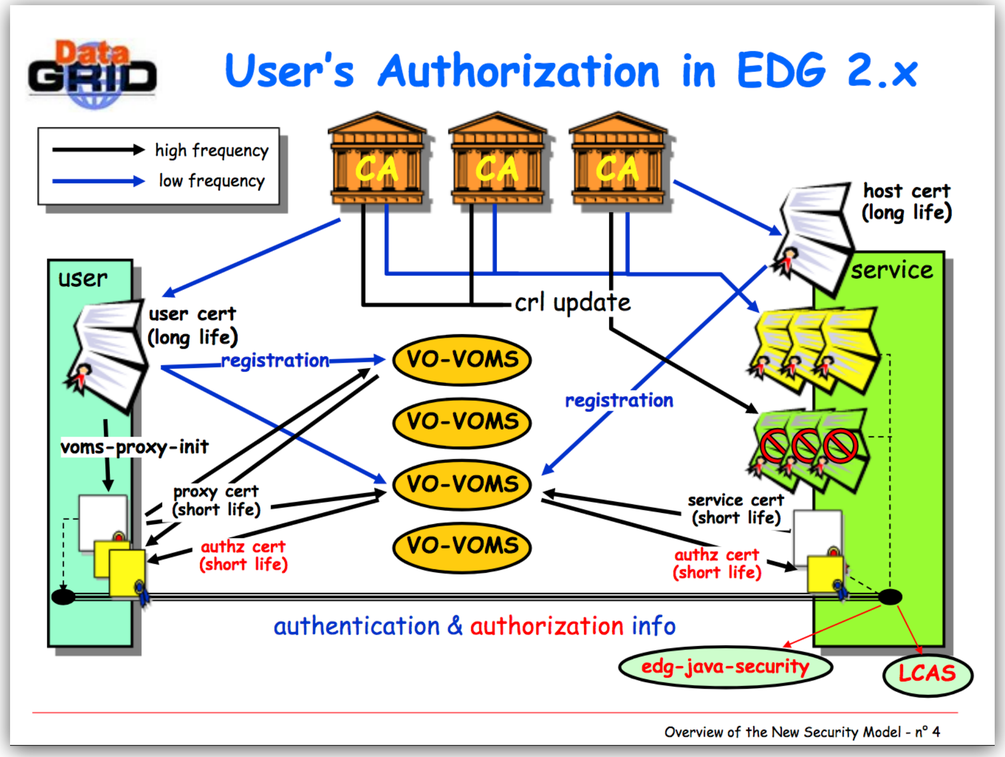
~1EB of storage

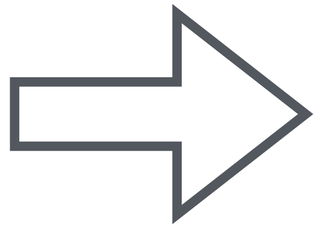
> 2 million jobs/day

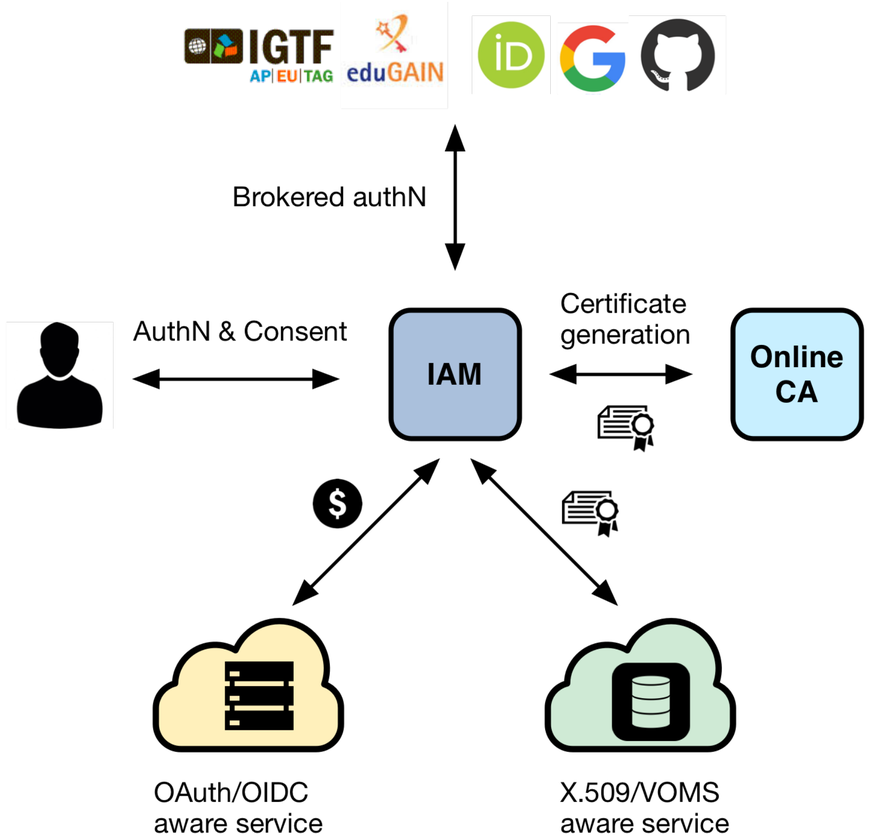


See https://wlcg-public.web.cern.ch/ for a nice presentation of the WLCG

Objective: evolution of the WLCG AAI beyond X.509







One slide summary

To access computing and storage resources in the WLCG today you use a **VOMS proxy certificate** (a short-lived X.509 credential), which provides information about **who you are**, **for which Virtual Organization (VO) you're acting** and **what you can do on the infrastructure** (i.e., VOMS groups and roles)

In the near future we will use **tokens,** which will provide more or less the same information

Tokens are obtained from a **VO token issuer** (e.g., IAM) using **OpenID Connect**

Tokens are **sent to services/resources following OAuth recommendations** (e.g., embedded in the header or an HTTP request)

Tokens are **self-contained**, i.e. their **integrity and validity can be verified locally** with no callback to the token issuer

Thanks for your attention.
Questions?

References

IAM Website: https://indigo-iam.github.io

IAM @ GitHub: https://github.com/indigo-iam/iam

IAM Users workshop slides: https://indico.cern.ch/event/970568/

IAM in action video: https://www.youtube.com/watch?v=1rZlvJADOnY

Contacts:

- andrea.ceccanti@cnaf.infn.it
- iam-support@lists.infn.it
- indigo-iam.slack.com