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

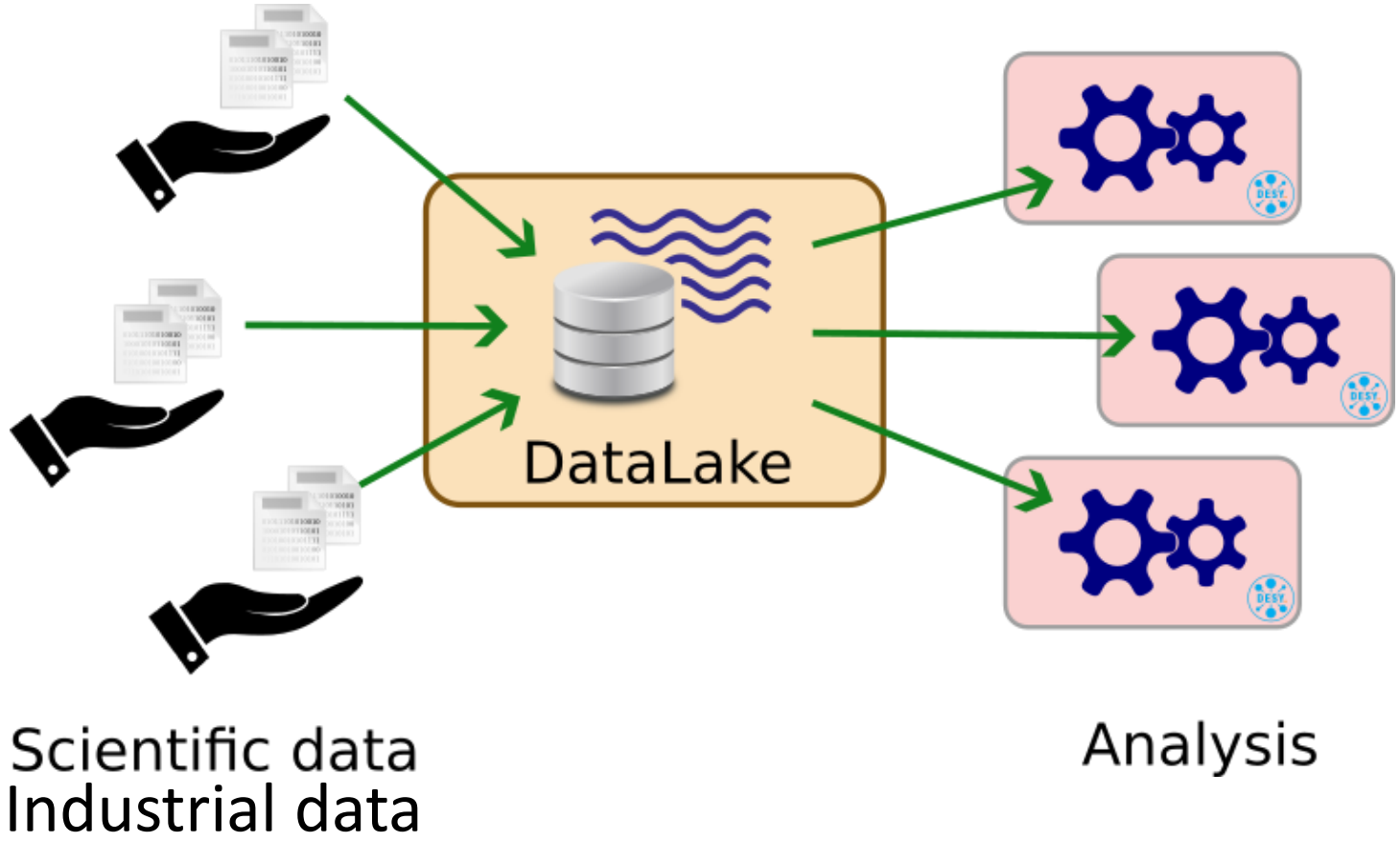
Interoperable Data Lake (IDL)

Carolina Berucci

"Spoke3 Technical Meeting" ,
Bologna, 17-19 December 2024

Interoperable Data Lake: Overview

Integration with science and industry use cases



LEONARDO

ThalesAlenia Space

INAF

INFN
Istituto Nazionale di Fisica Nucleare

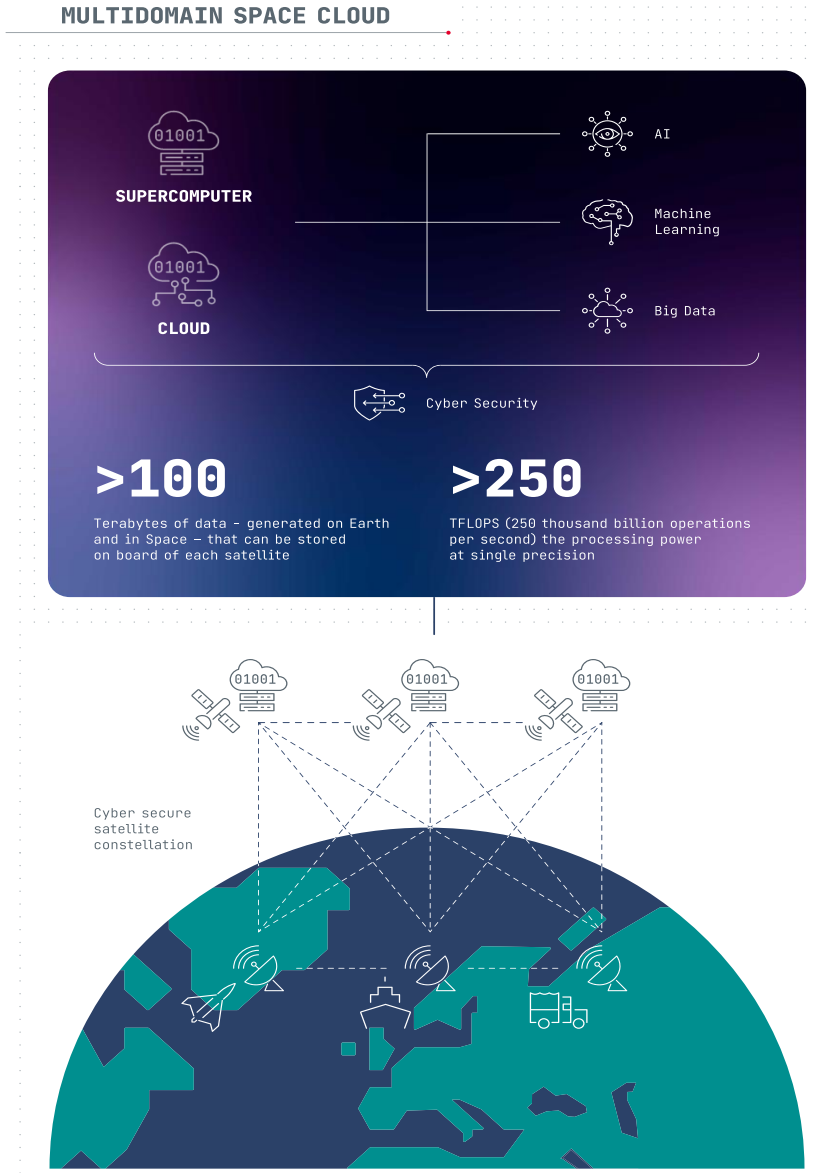
Use case: Space Situational Awareness (SSA)

SSA refers to the knowledge of the space environment, including location and function of space objects and space weather phenomena. SSA is generally understood as covering three main areas:

- ***Space Surveillance and Tracking (SST) of man-made objects -> Space Debris***
- Space WEather (SWE) monitoring and forecast
- Near-Earth Objects (NEO) monitoring (only natural space objects)

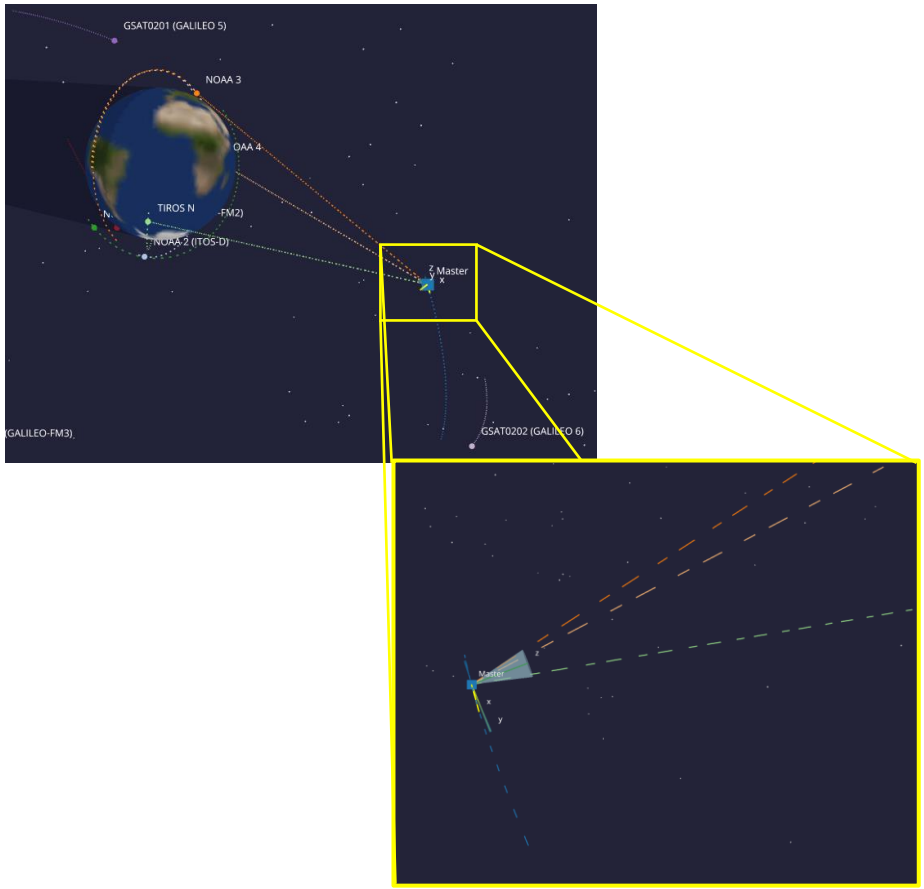
Space sensors in both in Low Earth Orbit (LEO), Medium Earth Orbit (MEO) and Geostationary Earth Orbit (GEO) are suitable to provide:

- **Operation and continuity**
- **Accuracy**
- **Global coverage**
- **Responsiveness**



Architecture and algorithms for data processing

Objective: To build a simulation software able to generate synthetic data coming from space-based SSA sensors whilst evaluating the computational load of the data processing chain



✓ Sensors and algorithms have been identified, the research conducted has been delivered inside the first deliverable of the WP and a report.

✓ The simulator has been designed to be composed by independent modules:
- Objects state module: tasked with objects orbit and attitude simulation and event handling
- RF module: tasked with the generation of signals and baseband digital signal processing analysis for feature extraction
- Optical module: tasked with satellites and Resident Space Objects image generation using a GAN algorithm and image feature extraction using a CNN

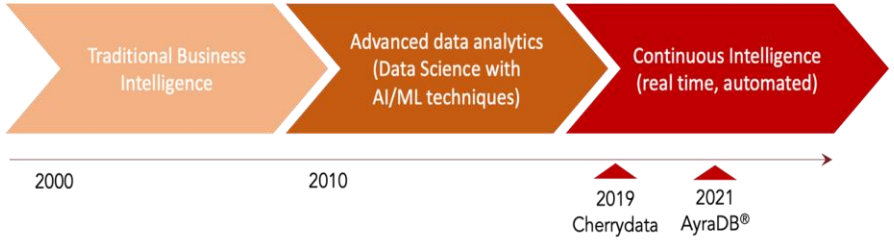
✓ The simulator is currently in its first integration phase (end foreseen in Q1 2025):
- The Objects state module has already been developed and validated
- The RF module is currently under validation using Monte Carlo simulations
- The optical module development is foreseen to be started in Q1 2025

Date : 09/12/2024
Ref : non referenziato
Rif. Modello : 87201590-QCI-TAS-IT-007

AyraDB as a metadata database for the "space debris" use case, objectives

- In the IDL system, data are stored on a data lake, while metadata are stored on a dedicated database
- AyraDB (high-performance database designed by Cherrydata, www.ayradb.com) has been chosen as metadata DB
- The objective is to maximise query performance by executing SQL queries on the metadata database (AyraDB) and retrieving from the data lake only the requested data
- The implementation of AyraDB has been designed to minimise response time to queries operating on large tables
- Preliminary tests have been performed on synthetic metadata (1 billion records)

Cherrydata is a startup (and a spinoff of PoliMi), offering consulting, innovation, and research services on big data and analytics.



- *AyraDB has been tested on Leonardo Davinci-1 supercomputer in 2022, as part of EuroCC project.*
- *Cherrydata is involved in IDL as technology provider, to test AyraDB in the context of storing and querying space-based or ground-based measurements.*

LEONARDO

Metadata: Preliminary Results

- Various queries selecting records in a specific time interval has been executed on the 1-billon rows metadata table
- An example of query is the following:

```
SELECT START_TIME, LINK FROM ayradb.IDL_dumped
WHERE START_TIME > toDateTime64('2018-04-23 15:23:57') AND
STOP_TIME < toDateTime64('2018-04-23 15:30:00')
```

- This query has scanned the metadata table and returned 1700 records (out of 1 billion records in the table), in a time of 600 milliseconds

Metadata table

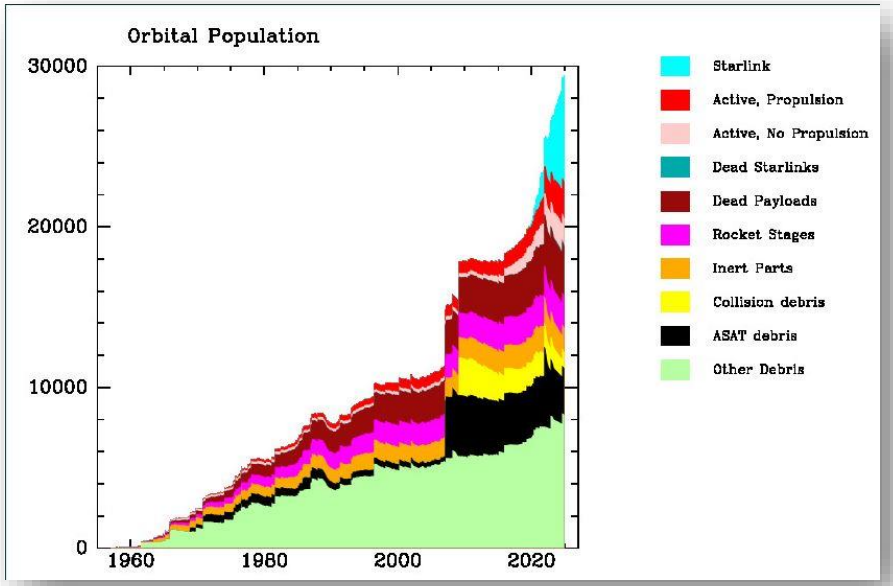
Block	Value type	Posible Values
META_START	string	META_START
COMMENT	string	-
TIME_SYSTEM	String	UTC, TAI, GPS, TT
TIMETAG_REF	String	TX, RX
EPOCH	time	YYYY-MM-DDThh:mm:ss
START_TIME	time	YYYY-MM-DDThh:mm:ss
STOP_TIME	time	YYYY-MM-DDThh:mm:ss
PARTICIPANT_1	String	Participant name or catalogue ID
PARTICIPANT_2	String	Participant name or catalogue ID
PARTICIPANT_3	String	Participant name or catalogue ID
PARTICIPANT_n	String	Participant name or catalogue ID,
PATH	String	-
REFERENCE FRAME	String	ICRF, ITRF2000, EME2000, TEME
SENSOR_TYPE	String	-
MEAS_TYPE	String	ANGLE, ORBIT, RF, PHOTO
MEAS_FORMAT	String	ref to table
MEAS_UNIT	String	
MEAS_RANGE_MIN	list of numbers	-
MEAS_RANGE_MAX	list of numbers	-
MEAS_RANGE_DESC	String	-
MEAS_RANGE_UNIT	String	-
DATA_QUALITY	string	L_n
LINK	string	-
META_STOP	string	META_STOP

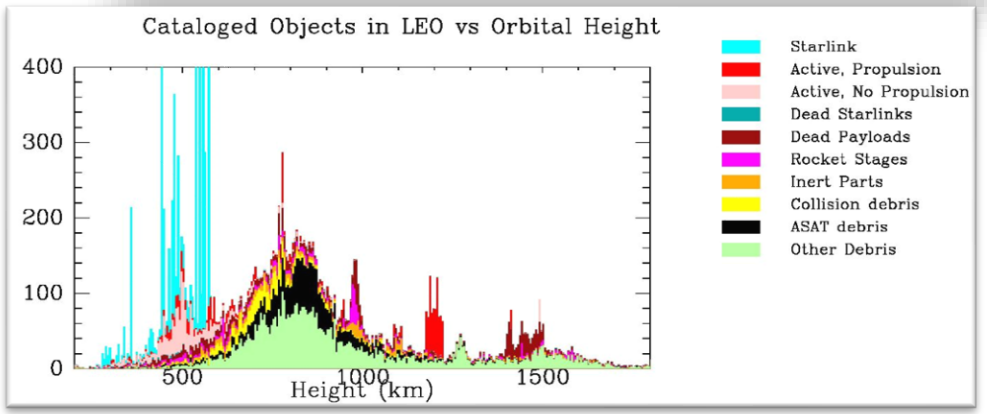
Metadata: Work in Progress

CELESTRAK SINCE 1985

Current work is focused on the following activities:

- Integrating the metadata database into the overall IDL system.
- Testing and benchmarking the overall end-to-end query process (including data retrieval).
- Populating the metadata database with real data (https://celestrak.org/NORAD/elements).
- Planning and executing a broader set of queries on the real dataset:
    - *Participant 1/2/3 - START_TIME - STOP_TIME*
    - *Participant 1/2/3 – EPOCH*
    - *Participant 1/2/3 – SENSOR_TYPE*
    - *Participant 1/2/3 – MEAS_TYPE*
    - *Participant 1/2/3 – MEAS_FORMAT*
    - *Participant 1/2/3 – MEAS_TYPE – MEAS_FORMAT*



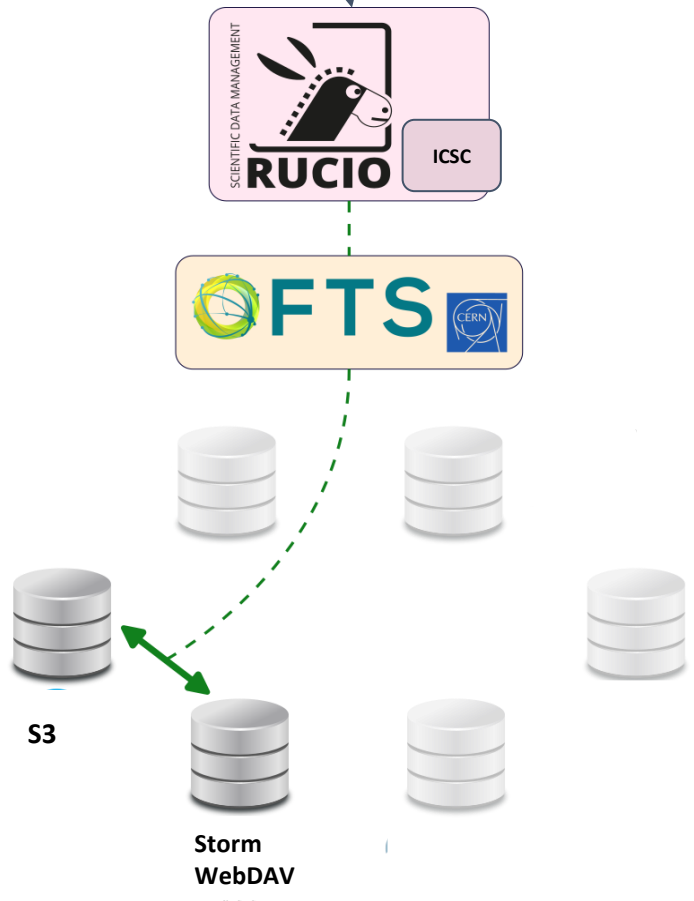


The pillars of the IDL Data Lake

INFN Istituto Nazionale di Fisica Nucleare

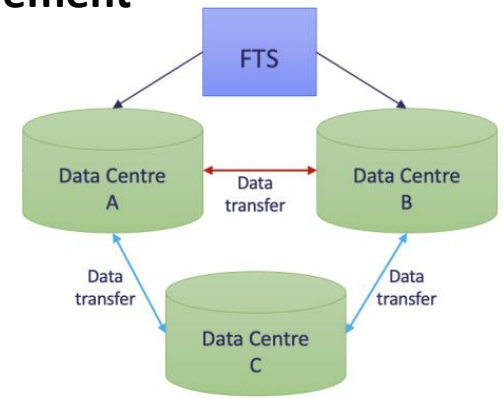
Rucio provides a mature and modular scientific data management federation

- 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 **(ExaBytes)** data management
  - Rucio is free and open-source software licenced under Apache v2.0
  - Open community-driven development process



FTS : File Transfer Service responsible for Bulk data movement

- Efficiently **schedules data transfers**
- **Maximizes** use of available **network** & **storage** resources whilst respecting any limits



Enhancing the existing services

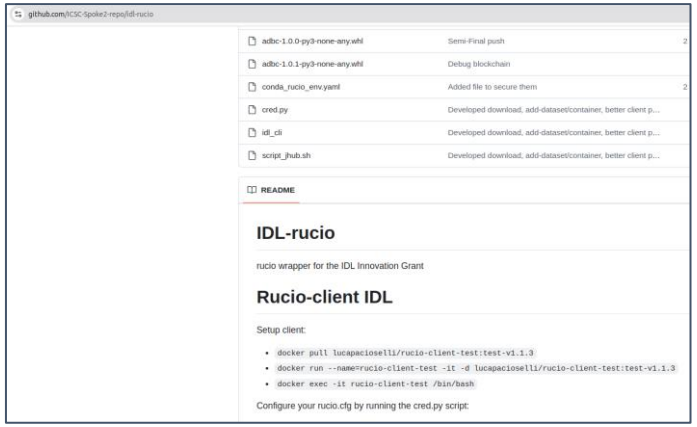
- to provide a seamless integration with external metadata
- to integrate datalake and compute environment

IDL prototype current status

- The first prototype is ready to support an end-to-end test
- Exploiting INFN Cloud resources to develop the prototype (both central services and cloud storage endpoint)
- Documentation being prepared. Together with developed code it will be available on Spoke2 git repo

✓ Deployed an automated (custom Docker image) Rucio server instance on a k8s cluster (nginx, HTTPS, etc…)
✓ Functional prototype of a **DID-metadata plugin** to communicate with an external database (**AyraDB**) provided by CherryData
✓ Support for the BlockChain integration
✓ First prototype of the **IDL Rucio client**

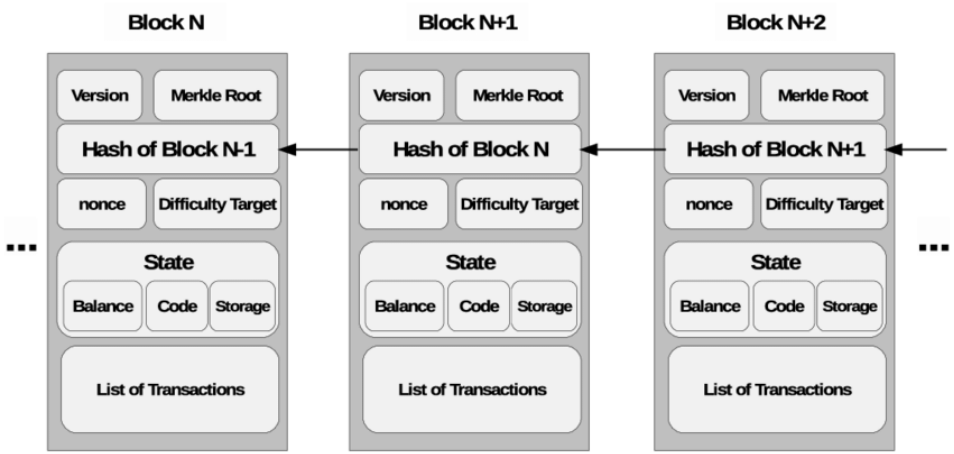
DESCRIPTION	DEPLOYMENT IDENTIFIER	STATUS	CREATION TIME	DEPLOYED AT	ACTIONS
Client Container test	11ef3dcb-3827-4e05-a163-76b2587994cf	CREATE_COMPLETE	2024-07-09 08:14:00	CLOUD-INFN-CATANIA	Details
vm-minio-test	11ef3940-97d9-fb88-a163-76b2587994cf	CREATE_COMPLETE	2024-07-03 13:31:00	CLOUD-INFN-CATANIA	Details
Rucio-VM-8GB	11ef27d1-9bff-427c-ad50-22533e954eeb	CREATE_COMPLETE	2024-06-11 09:04:00	CLOUD-INFN-CATANIA	Details

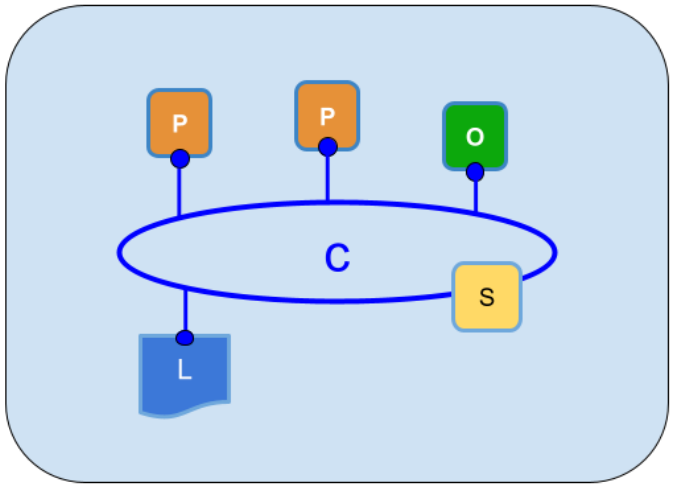


Blockchain Technology

INFN Istituto Nazionale di Fisica Nucleare

Blockchain ensures data integrity and protection against unauthorized tampering leveraging cryptographic techniques





P	Peer
O	Orderer
S	Smart contracts
L	Ledger (Blockchain and World state DB)

- Deployed a Blockchain network
- Integrated with the IDL infrastructure, storing blockchain information during file upload and using it for validation during file retrieval

Thank you for your attention