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

Leonardo - Interoperable Data Lake (IDL)

Carolina Berucci

Spoke 3 General Meeting

12-14 Giugno 2023

Dipartimento di Fisica e Astronomia – Università di Catania

Summary

- Leonardo:
  I. Technological Leadership
  II. Space Sector
- Data Space & High-Level Technical Concepts
- Interoperable Data Lake - General Objectives
- INFN - DataLake & BlockChain system
- INAF - Data Models and metadata definition
- Leonardo - Database technology
- Thales Alenia Space - Architecture and Algorithms for SSA Processing

Technological Leadership

As a **driver of innovation** Leonardo is committed to:

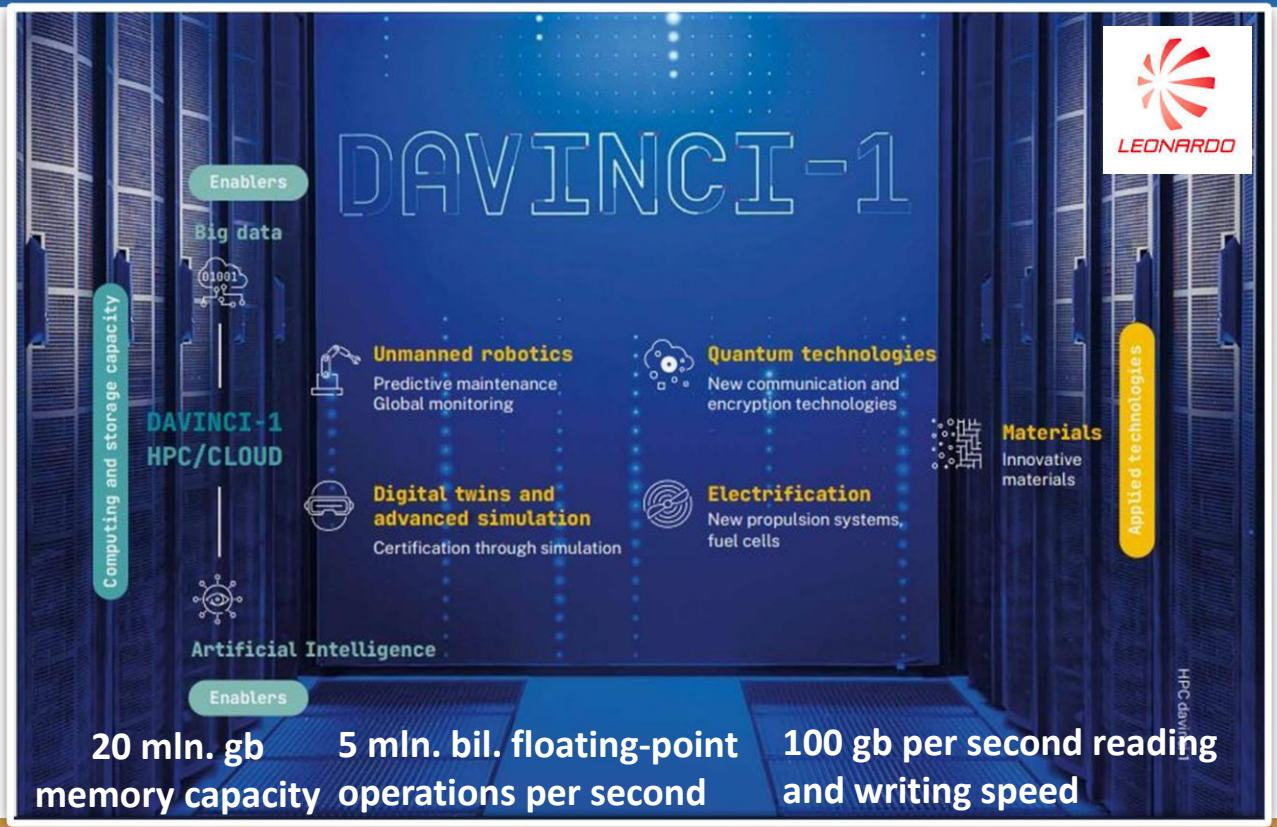
- **digitalisation processes and the enabling technologies**
- technological research

13% OF REVENUES SPENT ON R&D ACTIVITIES

COOPERATION WITH MORE THAN 90 UNIVERSITIES & RESEARCH CENTRES

MORE THAN 400 TECHNOLOGIES IN LEONARDO'S PORTFOLIO

9,600 PEOPLE INVOLVED IN R&D AND ENGINEERING



Space sector

- **Satellite services** for the design, development, launch and in-orbit control of space systems, and services and applications for Earth observation, communications, navigation and satellite localisation.
- **Satellite systems** for telecommunications, navigation, Earth observation, probes and rovers for space exploration, and multi-function orbiting modules.
- **Instruments**: electro-optical and hyperspectral payloads, laser transmitters and robotic systems for space exploration and the study of our planet.
- **Equipment**: photovoltaic panels, atomic clocks, attitude sensors, power distribution systems, orbital micro-propulsion.

>2 million

radar images acquired by **COSMO-SkyMed**, the constellation of ASI and the Italian Ministry of Defence

>50%

of the living space of the International Space Station developed by **Thales Alenia Space**

>50

atomic clocks on board the Galileo constellation

- Telespazio (67%)
- Thales Alenia Space (33%)
- Avio (29,6%)

Telespazio relevant assets & skills

Profound knowledge of Thematic End User needs

In several vertical applications (maritime, agriculture, forestry, environmental monitoring, infrastructure and asset management, cultural heritage, insurance, D&I)

HPC / Cloud / Big Data / Microservices

Management and exploitaiton of scalable computing infrastructure and processing platforms, on prem and in various cloud environments

Remote sensing Opt./SAR data processing

Optical (Multispectral/Hyperspectral) / SAR signal processing to derive model based observations and measurements and monitoring capacity from single image or multitemporal time series

Modelling (Hydraulic/Climate/ Agriculture) and scenario simulation

Design and implementation of modelling and scenario simulation tools to build flood risk and impact scenarios, including climate change effects

3D modelling

Desing and implementation of detailed 3D models from satellite/aerial/drone imagery at multiple scales and LOD.

Artificial Intelligence

ML/DL networks applied to several geosptial challenges (object detection, anomaly detection, trend analysis, ..) and data (satellite, eaerial ,drone imagery, social media, marine traffic, …)

Data Space & High-Level Technical Concepts

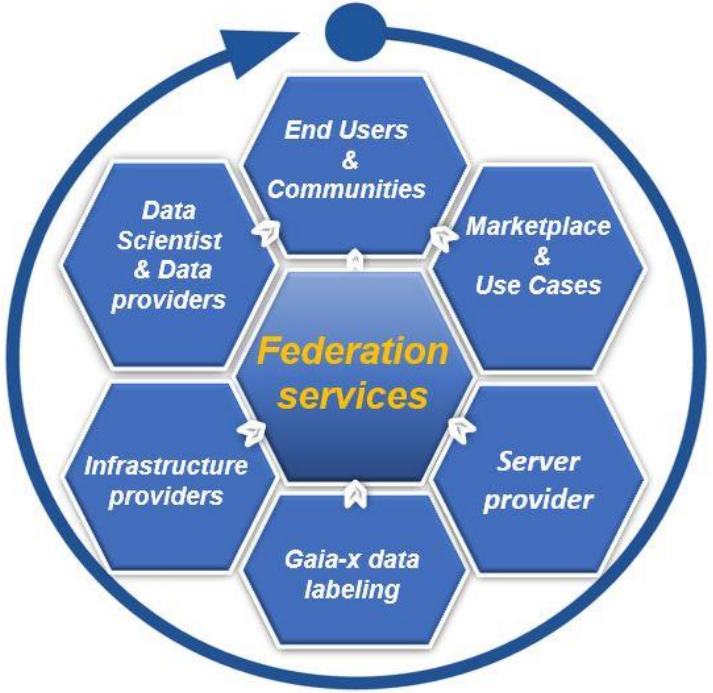
Data Space Concept

Our objective is to facilitates access to space data, to foster innovation and to help attract new potential partnership

The Space industry is characterized by the **sharing economy**, there is a more **open**, wider market participant **outreach** that includes **academic**, **commercial** and **governmental players.**

Our solution is a data exchange platform, builds on **existing infrastructures (CN HPC),** where **data acquirers** and **data providers meet to source**, distribute, **exchange** data **securely**, in compliance with **Gaia-x regulations**

The initiative combines **platforms** and different **ecosystems** that all follow a common set of rules and policies and deploy a common marketplace to enable a **Federated Data Ecosystem**



Space Economy Ecosystem

Building a digital infrastructure to bring user to the data

Academia/ Research
More than 400 scientific papers were published focusing on Future EO Earth Explorers

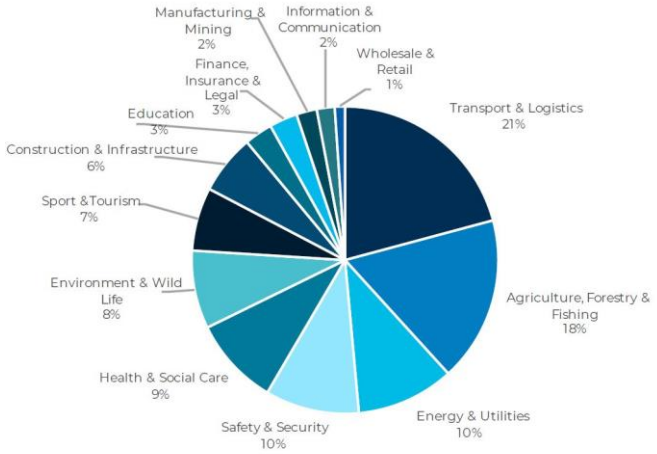
Entreprise/ SMI/Startup
There are 144 companies (ITA) in the downstream segment (included IT providers and system integrators) operating in the sector

Data provider
(National/Local PA) Commercial Data Providers (imagery, analytics, road traffic)

User (end & intermediate) **Communities**

Cloud-to-edge federations ⟷ ***Middleware***

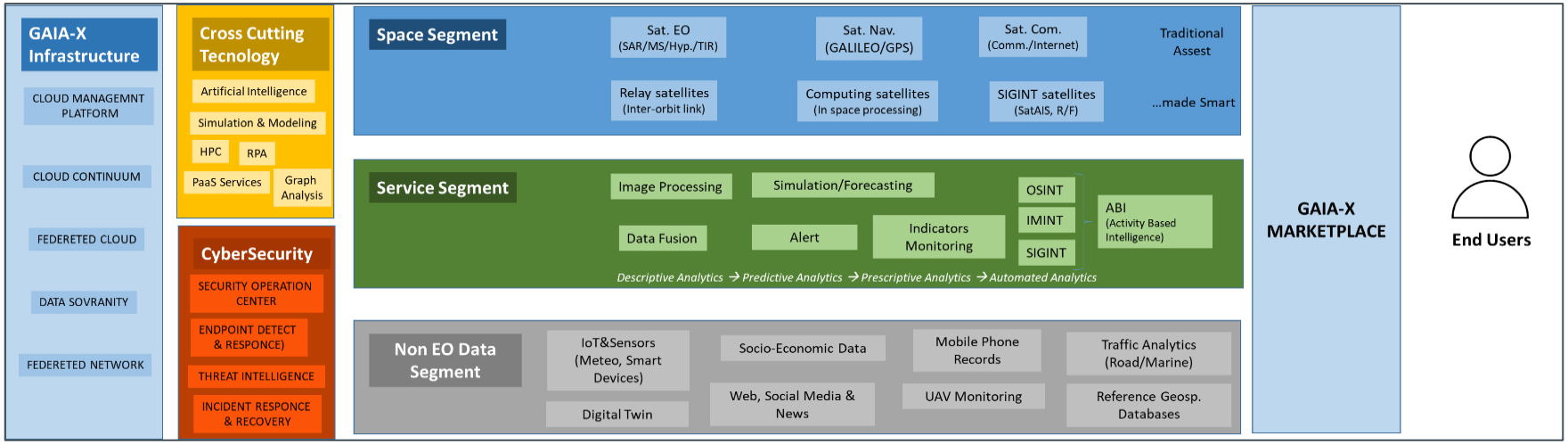
ICSC infrastructure



1008 applications satellite-based that traditionally would have no interest in space

Osservatorio Space Economy 2023

High Level Technical Concepts



- **Cloud & Edge Management Platform** (CMP) capable of effectively implementing a **Multicloud** and **Hybrid approach** integrating also **edge level** as independent computational node
- **HPC at the edge Infrastructure & Services** for computing intensive workloads to carry out HPC directly at the EDGE/DEEP EDGE
- **Federated trust** and **identity management**, access to **AI** and **ML on demand services** and **PaaS** capabilities for **data analytics**
- High **interoperability and portability** of **services** and **data** among all **cloud-edge users** and **providers**

Embedding space ecosystem into data value chain

Data sources → **Standardized data** → **Optimized data** → **Sharing models** → **Data use**

Input sensor and data layer

Space satellite sensor

Observatory

data repository

ESA - DIAS
ASI - SSDC

Aggregation and normalization of data within local, interoperable repositories

Metadata

FAIR data

Findability
Accessibility
Interoperability
Reusability

Sharing prerequisites through harmonized access management

- **Security & access management**
- **Data catalogue**
- **Harmonized Privacy management**

Depending on the legal basis and use case

- **National Hubs**
- **Use case**
- **Data Space**
- **Marketplace**

Data, insights, software and algorithms as a service

Providers and research institutions

Insights **Data**

user communities

PAAS **SAAS**

IG Project: Interoperable Data Lake (IDL)

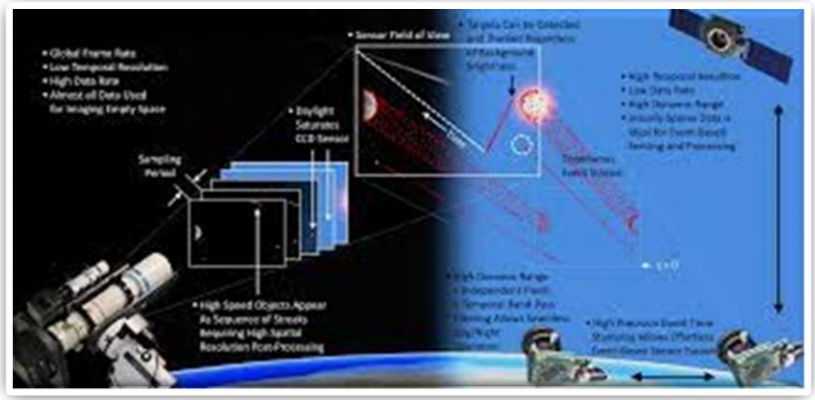
Abstract

The Project aims at creating a Data Lake service, supporting a seamless access to space and ground-based observations and simulated data.

The project addresses the design and commissioning of an interoperable, distributed data archive, relying on state-of-the-art open technologies, supporting both science and industry.

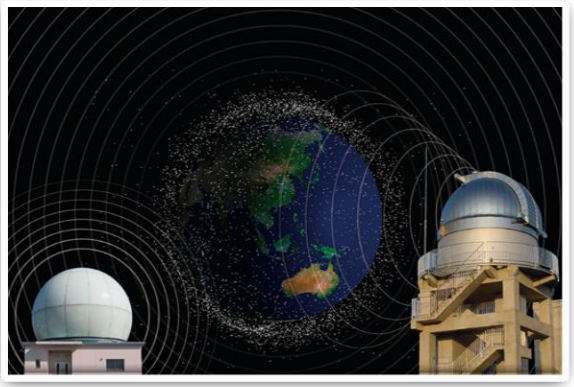
The service will specifically address the challenges related to the big data scenario, in terms of both data management, storage, access, identification and of access to computing resources necessary to process the data.

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 WEather (SWE)** monitoring and forecast.
- **Near-Earth Objects (NEO)** monitoring (only natural space objects)



General Objectives

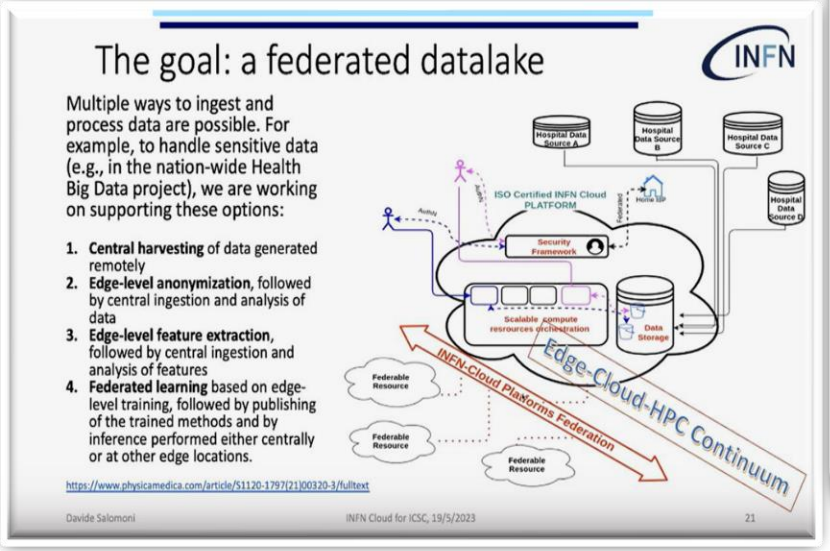
- Organization of radio observation for space situational awareness (SSA) application;
- Improvement of existing catalogs of near-earth object or space debris integrating ground-based observations with space-based simulation;
- Create a federated data infrastructure that follows the FAIR data principles;
- Design and implement a prototype application performing the end-to-end processing chains to demonstrate the Data Management (DM) capability;
- Exploit a cloud-native distributed database to support the integration and query of data coming from different sources;
- Create a mock-up to generate a synthetic data set, define the algorithms of the processing chain and evaluate the computational load.

LEONARDO
ThalesAlenia Space
INAF
INFN Istituto Nazionale di Fisica Nucleare

Activities & Work Packages structures of the project

INFN – WP1: DataLake & WP3,4: BlockChain system

Integration with Spoke 2 Research Program



Activity #1 (DataLake)

INFN proposes to test solutions for managing data in a geographically distributed environment to exploit the Data lake. All the services will be deployed in the form of containers managed via container orchestrators.

The prototype will show the feasibility to deploy PaaS services for the actual processing of the data ingested into the Data lake.

Activity #3 (BlockChain system for the data lake)

Deployment of a blockchain network to execute fundamental CRUD (**C**reate, **R**ead, **U**pdate, **D**elete) operations, catering to the requirements of the project. This blockchain network will be accessible through a REST API with an HTTPS backend, facilitating seamless communication between the blockchain and databases.

INAF – WP2: Data Models and metadata definition

Integration with Spoke 3 Research Program

Spoke 3 - WP4 - Big Data Management, Storage and Archiving

Objectives and Methodologies

- **Objective 2.** Big data processing and visualization, via adopting innovative approaches for the analysis of large and complex data volumes and for their exploration.
- **Objective 3.** High Performance storage, Big Data management, and archiving applying the Open Science principles and implementing them in the Big Data Archives.
- ACO-S will promote the FAIRness of the research outputs and services across research communities involved in the project

Activity #2 (Data Models and metadata definition)

Study and implementation of an IVOA/FAIR compliant data model for the correct description and handling of the data sets.

The data model will include all relevant information about data, software for data reduction, analysis, data policy, and all the information for data filtering for retrieval.

Leonardo - Cloud-native distributed database
WP2: Data Models and metadata definition, data archiving

LEONARDO

Objective. *Support the integration and query of data coming from different sources, complemented with simulated data, with a performance that enables novel application with real-time requirements in SSA use case, to achieve maximum effectiveness and efficiency in data provisioning and exploitation*

Topics

- **Requirement analysis:** Gather information necessary for the definition, implementation, testing and verification of the database technology to be integrated on top of the data-lake. Analysis and definition of the required data processing activities: preprocess, cleaning, normalize, rescale data, with a focus on data models and metadata definition.
- **Database deployment in ICSC infrastructure, validation and testing**

  Implementation of the data ingestion service, to collect data efficiently from the instrument/sensor network. Integrate the components of the prototype system with their applications and data platform and deploy them on the Consortium infrastructure.

Thales Alenia Space

WP5: Architecture and Algorithms for SSA Processing

Simulation of state of art algorithms for processing of space based sensors data for SSA and evaluation of the computational load. The simulator will provide:

- Information to support architectural trade offs in terms of optimal split between on board and on ground processing
- A reference to evaluate complex scenarios with multiple sensors processing

Topics

1. Investigate existing and upcoming sensors typologies and technologies for space based ssa applications, identify measurable space objects characteristics they can detect and the data typologies they generate
2. Selection of a representative subset of sensors and identification of the state of art algorithms for data processing
3. Design and implementation of a mock up (simulator) which is capable to generate a synthetic data set for the selected sensors, implement the algorithms of the processing chain and evaluate the computational load

Expected Results

- Design and commissioning of an interoperable, distributed data archive in the perspective on the ICSC HPC, Cloud and Data infrastructure;
- Availability (to both scientific and industrial users) of effective interoperable services for the storage and the processing of the data;
- Experimentation and Demonstration of the use of blockchain sw stacks for the certification / tracking of valuable datasets;
- Updating existing debris databases with new space–based observations and observations from space;
- Creation of a public data archive of state-of-the-art astrophysical simulations;
- Experimentation of technological solutions capable to support science in large astronomy observatories like LOFAR and the SKA.

Spoke Leader INAF
Co-Leader INFN

Soggetti Pubblici Affiliati	
Università Roma Tor Vergata	
Università di Trieste	
Università di Torino	
Università di Catania	
Scuola Normale Superiore - Pisa	
Sissa - Trieste	

Imprese Aderenti allo Spoke	
Intesa SanPaolo	
UnipolSai	
Sogei	
IFAB	
Leonardo	

Thank you for your attention!