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

Interoperable Data Lake (IDL)

LEONARDO: Carolina Berucci
CHERRYDATA: Chiara Francalanci
INAF: Cristina Knapic, Deborah Busonero
INFN: Daniele Spiga

Spoke 3 General Meeting, Elba 5-9 / 05, 2024

Scientific Rationale

"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"

- **WP1:** Test solutions for managing data in a geographically distributed environment by building end-to-end prototype and testbeds to demonstrate - to optimize big data storage and efficiency of data retrieval, exploiting state-of-the-art cloud-based technologies
- **WP2:** Define a data model for organizing, find and access archived data; design and commissioning of a relational database for metadata management identification and provisioning of data
- **WP3, WP4:** New techniques of block-chain and web-based stacks like Object Storage, will be used, tuned and linked together
- **WP5:** Simulation of state of art algorithms for processing of space-based sensors data for SSA and evaluation of the computational load

IG Call3 : IGUC project additional use case, extension of WP2 (D. Busonero)

LEONARDO

ThalesAlenia Space a Thales / Leonardo company

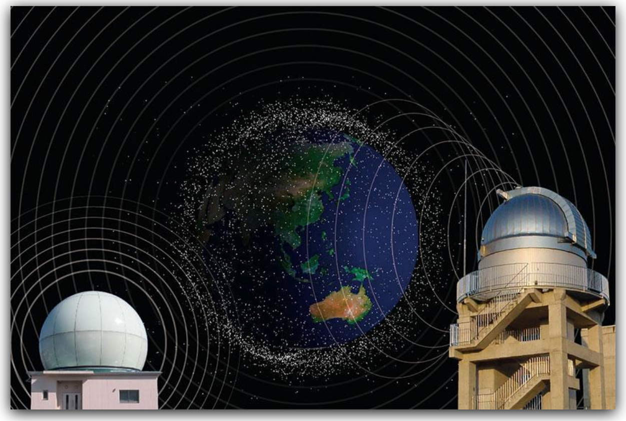
INAF

INFN Istituto Nazionale di Fisica Nucleare

Timescale & Milestones

WP	Del. code	Definition	Del. date
1	**TN1**	**Report the deployment solution used to implement the Data Managementnd results of the functional tests**	**M12**
1	TN2	Report the deployment solution used to implement the Data Management and results of the functional tests	M24
2	**TN1**	**Dataset identified and data model vs1**	**M12**
2	TN2	Technical report of the database	M18
2	TN3	Final report on database and data model vs1.1	M24
2	TN2	Workflow definitions for data object tracking, and implementation in the licensed software	M18
3	**TN1**	**Functional, Analysis, document**	**M6**
3	TN2	Workflow definitions for data object tracking, and implementation in the licensed software	M18
4	**TN1**	**Report the deployment solutions on the Datalake, with initial evaluation of results**	**M12**
4	TN2	Technology tracking report on the blockchain solutions maturity	M24
5	**TN1**	**sensors technologies and data typologies report**	**M8**
5	TN2.1	Algorithms and simulator design and verification report	M18
5	TN2.2	Update - Algorithms and simulator design and verification report v.2	M24

Use case: Space Situational Awareness (SSA)

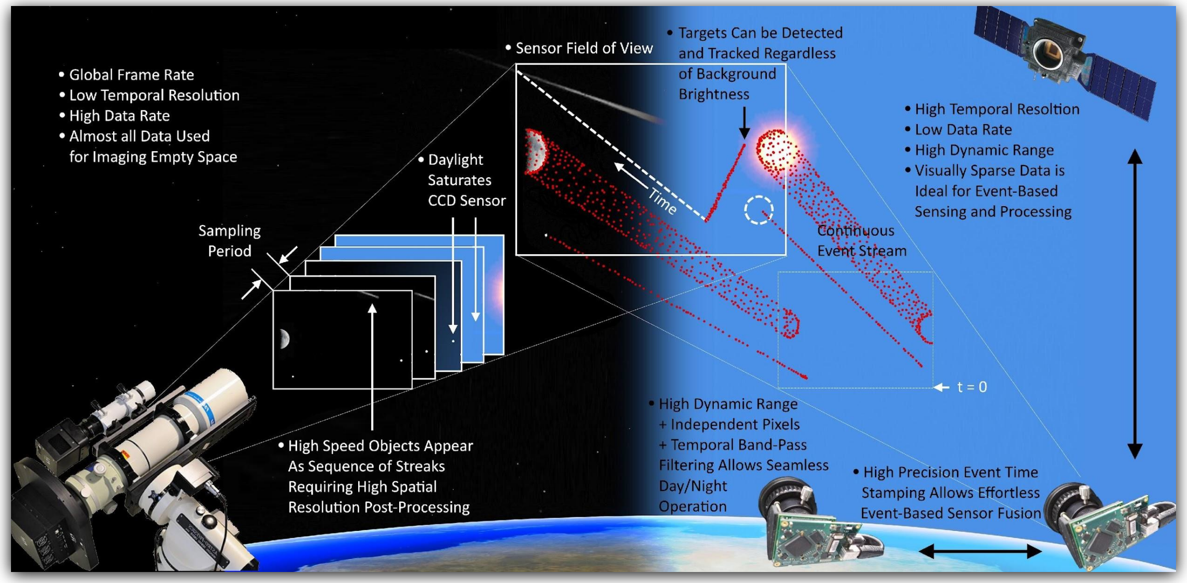


SSA refers to the knowledge of the space environment, including location and function of space objects and space weather phenomena. SSA is generally understood as covering three main areas:

- ***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)

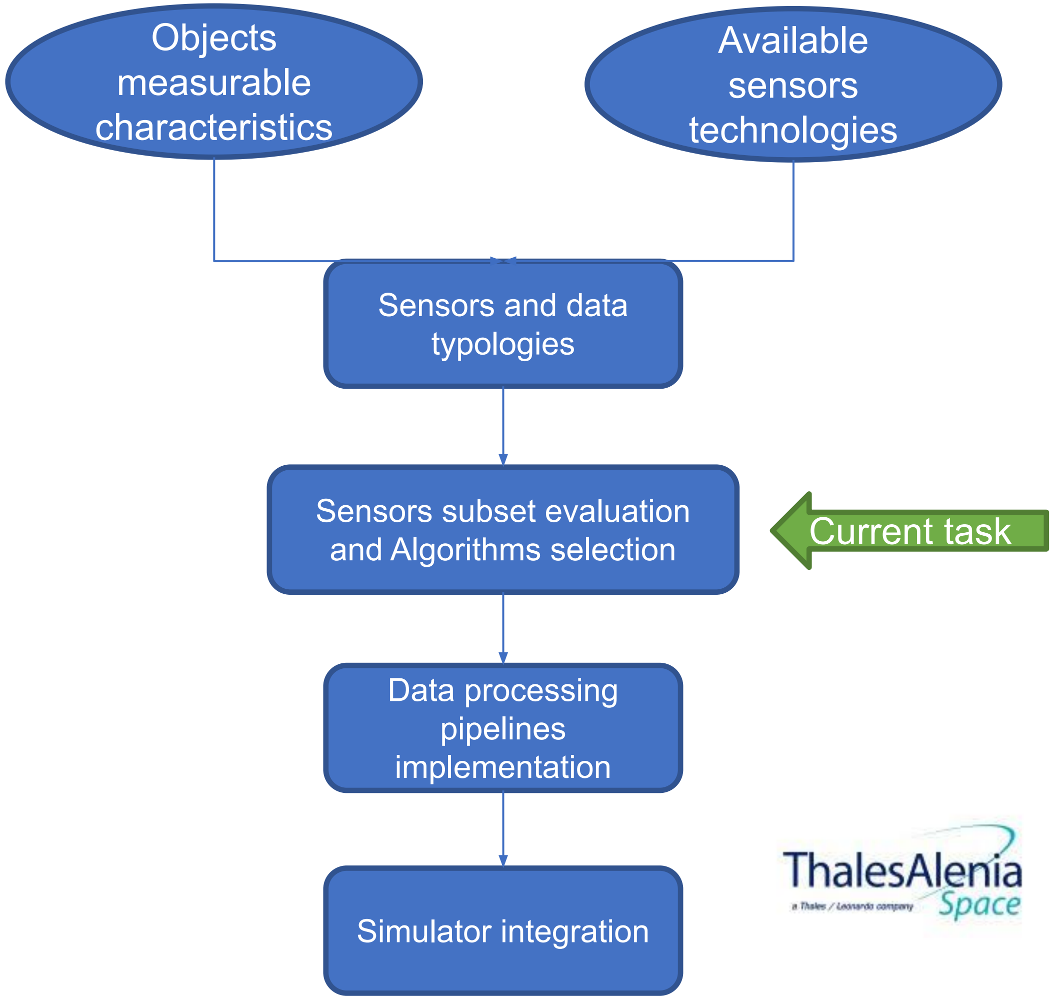
Space sensors in both in Low Earth Orbit (LEO), Medium Earth Orbit (MEO) and Geostationary Earth Orbit (GEO) are suitable to provide:

- **operation and continuity**: space-based surveillance and tracking are not impacted by atmospheric and weather conditions and by day and night cycles
- **Accuracy**: space-based measurements are not affected by the impairments of the atmosphere
- **Global coverage**: space sensors are designed and deployed to complement and augment the coverage of ground-based assets, whose installation/operation is precluded in remote and oceanic areas
- **Responsiveness**: space sensors can improve the revisit of the space volume(s) requiring “constant” monitoring; further, space-based data relay (i.e. inter-satellite links) can provide real time tasking and fast telemetry transmission of/by space sensors.


• Global Frame Rate
• Low Temporal Resolution
• High Data Rate
• Almost all Data Used for Imaging Empty Space
• Sensor Field of View
• Targets Can be Detected and Tracked Regardless of Background Brightness
• High Temporal Resoltion
• Low Data Rate
• High Dynamic Range
• Visually Sparse Data is Ideal for Event-Based Sensing and Processing
• Daylight Saturates CCD Sensor
Time
Sampling Period
Continuous Event Stream
t = 0
• High Speed Objects Appear As Sequence of Streaks Requiring High Spatial Resolution Post-Processing
• High Dynamic Range + Independent Pixels + Temporal Band-Pass Filtering Allows Seamless Day/Night Operation
• High Precision Event Time Stamping Allows Effortless Event-Based Sensor Fusion

WP5 - 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;**
- Starting from the most interesting use cases and the current available technologies, a set of sensors with the data they generate has been identified;
- Data processing pipelines have been constructed and they will serve as an input to start the algorithms investigations;
- Currently passive RF sensor data processing algorithms are under development for evaluation. The objective is to obtain a shortlist of these methods to integrate them in the data processing pipeline.



Objects measurable characteristics

Available sensors technologies

Sensors and data typologies

Sensors subset evaluation and Algorithms selection

Current task

Data processing pipelines implementation

Simulator integration

WP2 - Scientific rational

- Define a data model for organizing, find and access archived data; design and commissioning of a relational database for metadata management identification and provisioning of data

Technical Objective of the WP2 is

- data structure and modelization definition;
- relational db optimization;

Methodologies adopted are:

- frequent wp2 internal meeting;
- participation to the general meetings of the project;
- organization of meetings dedicated to specific tasks;
- external expert in the field consultancy (INAF and PoliMi);

Actions:

-Spacecraft constellation sensors data modellization with the goal of objects trajectory computation;
-Definition of data format, structure and architecture of the database and archive;
-Provisioning of sample observative data from Asiago Observatory (optic) and Medicina antennas (radio) in order to organize metadata relevant to evaluate the database performance.

WP2 - Data Models and metadata definition, data archiving

- **Datasets identification**: selection of data coming from radioastronomy, space debris observations & numerical simulations

→ **Tracking Data Message (TDM)**

- **Data Models and metadata definition**: Identification of a suitable data model and metadata definition, Implementation of an ingestion software

→ **Objective:**

- Support the integration and query of data coming from different sources, to achieve maximum effectiveness and efficiency in data provisioning and exploitation.

- **Requirement analysis (for DB)**: choice, implementation, testing and verification of the database technology to be integrated on top of the data-lake.

→ **Requirement analysis:**

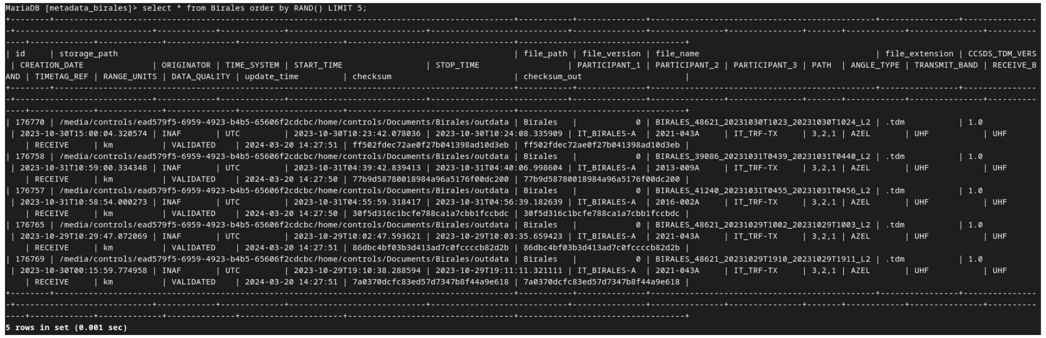
- Gather information necessary for the definition, implementation, testing and verification of the database technology to be integrated on top of the data-lake.
- Analysis and definition of the required data processing activities: preprocess, cleaning, normalize, rescale data, with a focus on data models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

- **Database deployment in ICSC infrastructure**: deployment of the ingestion software and the chosen DB into the ICSC infrastructure
- **Validation and testing**: performance test of the processing pipelines on a hybrid HPC/cloud

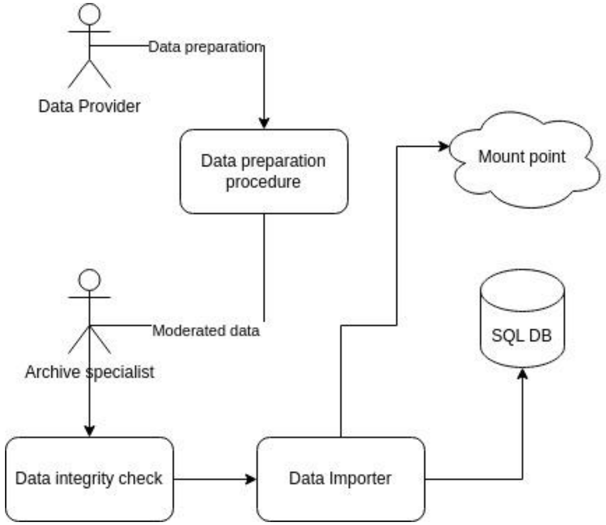
WP2 - Modelization and optimization activities

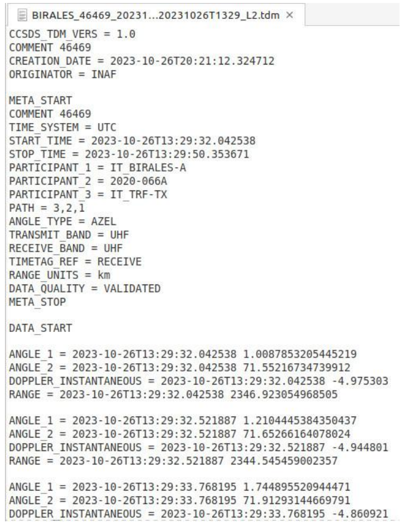
Technical solutions:

- provisioning and organization of teleconf for the explanation of the current state of the art formats of similar detections (ground based);
- analysis of the data structure and discussion about useful features and limitations;
- interaction with sensor specialists to understand if models and standard formats fit the needs;
- support in db set up and performances evaluation;
- support in queries definition;
- sharing of knowledge and expertise in data handling.



```
MariaDB [metadata_birales]> select * from Birales order by RAND() LIMIT 5;
+--------+------------------------------------------------------------------------+-----------+--------------+-----------------------------------------------+----------------+---------------
-+----------------------------+------------+-------------+----------------------------+----------------------------+---------------+---------------+---------------+-------+------------+---------------+----------
----+-------------+-------------+--------------+---------------------+----------------------------------+----------------------------------+
| id     | storage_path                                                           | file_path | file_version | file_name                                     | file_extension | CCSDS_TDM_VERS
 | CREATION_DATE              | ORIGINATOR | TIME_SYSTEM | START_TIME                 | STOP_TIME                  | PARTICIPANT_1 | PARTICIPANT_2 | PARTICIPANT_3 | PATH  | ANGLE_TYPE | TRANSMIT_BAND | RECEIVE_B
AND | TIMETAG_REF | RANGE_UNITS | DATA_QUALITY | update_time         | checksum                         | checksum_out                     |
+--------+------------------------------------------------------------------------+-----------+--------------+-----------------------------------------------+----------------+---------------
-+----------------------------+------------+-------------+----------------------------+----------------------------+---------------+---------------+---------------+-------+------------+---------------+----------
----+-------------+-------------+--------------+---------------------+----------------------------------+----------------------------------+
| 176770 | /media/controls/ead579f5-6959-4923-b4b5-65606f2cdcbc/home/controls/Documents/Birales/outdata | Birales   |            0 | BIRALES_48621_20231030T1023_20231030T1024_L2 | .tdm           | 1.0
 | 2023-10-30T15:00:04.320574 | INAF       | UTC         | 2023-10-30T10:23:42.078036 | 2023-10-30T10:24:08.335909 | IT_BIRALES-A  | 2021-043A     | IT_TRF-TX     | 3,2,1 | AZEL       | UHF           | UHF
   | RECEIVE     | km          | VALIDATED    | 2024-03-20 14:27:51 | ff502fdec72ae0f27b041398ad10d3eb | ff502fdec72ae0f27b041398ad10d3eb |
| 176758 | /media/controls/ead579f5-6959-4923-b4b5-65606f2cdcbc/home/controls/Documents/Birales/outdata | Birales   |            0 | BIRALES_39086_20231031T0439_20231031T0440_L2 | .tdm           | 1.0
 | 2023-10-31T10:59:00.334348 | INAF       | UTC         | 2023-10-31T04:39:42.839413 | 2023-10-31T04:40:06.998604 | IT_BIRALES-A  | 2013-009A     | IT_TRF-TX     | 3,2,1 | AZEL       | UHF           | UHF
   | RECEIVE     | km          | VALIDATED    | 2024-03-20 14:27:50 | 77b9d58780018984a96a5176f00dc200 | 77b9d58780018984a96a5176f00dc200 |
| 176757 | /media/controls/ead579f5-6959-4923-b4b5-65606f2cdcbc/home/controls/Documents/Birales/outdata | Birales   |            0 | BIRALES_41240_20231031T0455_20231031T0456_L2 | .tdm           | 1.0
 | 2023-10-31T10:58:54.000273 | INAF       | UTC         | 2023-10-31T04:55:59.318417 | 2023-10-31T04:56:39.182639 | IT_BIRALES-A  | 2016-002A     | IT_TRF-TX     | 3,2,1 | AZEL       | UHF           | UHF
   | RECEIVE     | km          | VALIDATED    | 2024-03-20 14:27:50 | 30f5d316c1bcfe788ca1a7cbb1fccbdc | 30f5d316c1bcfe788ca1a7cbb1fccbdc |
| 176765 | /media/controls/ead579f5-6959-4923-b4b5-65606f2cdcbc/home/controls/Documents/Birales/outdata | Birales   |            0 | BIRALES_48621_20231029T1002_20231029T1003_L2 | .tdm           | 1.0
 | 2023-10-29T10:29:47.072069 | INAF       | UTC         | 2023-10-29T10:02:47.593621 | 2023-10-29T10:03:35.659423 | IT_BIRALES-A  | 2021-043A     | IT_TRF-TX     | 3,2,1 | AZEL       | UHF           | UHF
   | RECEIVE     | km          | VALIDATED    | 2024-03-20 14:27:51 | 86dbc4bf03b3d413ad7c0fccccb82d2b | 86dbc4bf03b3d413ad7c0fccccb82d2b |
| 176769 | /media/controls/ead579f5-6959-4923-b4b5-65606f2cdcbc/home/controls/Documents/Birales/outdata | Birales   |            0 | BIRALES_48621_20231029T1910_20231029T1911_L2 | .tdm           | 1.0
 | 2023-10-30T00:15:59.774958 | INAF       | UTC         | 2023-10-29T19:10:38.288594 | 2023-10-29T19:11:11.321111 | IT_BIRALES-A  | 2021-043A     | IT_TRF-TX     | 3,2,1 | AZEL       | UHF           | UHF
   | RECEIVE     | km          | VALIDATED    | 2024-03-20 14:27:51 | 7a0370dcfc83ed57d7347b8f44a9e618 | 7a0370dcfc83ed57d7347b8f44a9e618 |
+--------+------------------------------------------------------------------------+-----------+--------------+-----------------------------------------------+----------------+---------------
-+----------------------------+------------+-------------+----------------------------+----------------------------+---------------+---------------+---------------+-------+------------+---------------+----------
----+-------------+-------------+--------------+---------------------+----------------------------------+----------------------------------+
5 rows in set (0.001 sec)
```



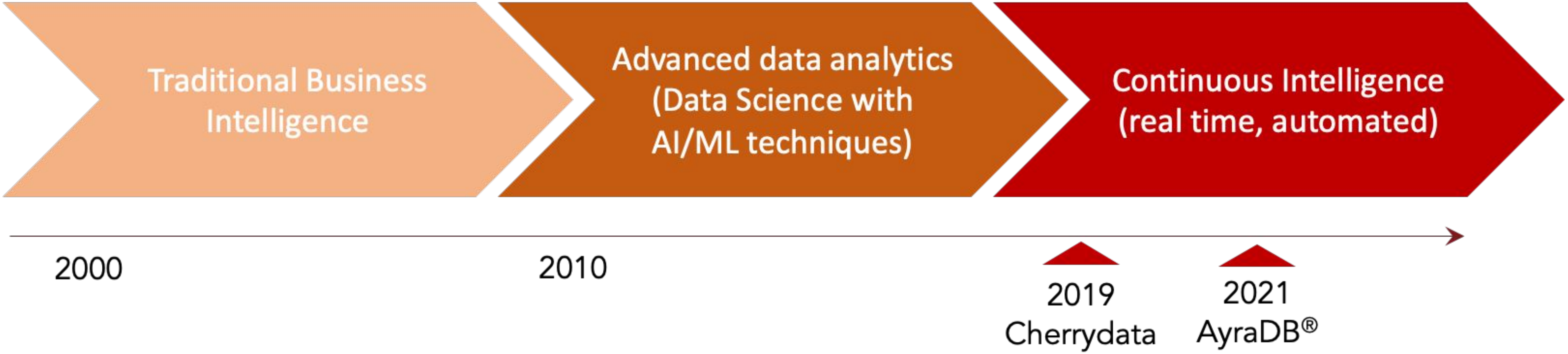


```
CCSDS_TDM_VERS = 1.0
COMMENT 46469
CREATION_DATE = 2023-10-26T20:21:12.324712
ORIGINATOR = INAF

META_START
COMMENT 46469
TIME_SYSTEM = UTC
START_TIME = 2023-10-26T13:29:32.042538
STOP_TIME = 2023-10-26T13:29:50.353671
PARTICIPANT_1 = IT_BIRALES-A
PARTICIPANT_2 = 2020-066A
PARTICIPANT_3 = IT_TRF-TX
PATH = 3,2,1
ANGLE_TYPE = AZEL
TRANSMIT_BAND = UHF
RECEIVE_BAND = UHF
TIMETAG_REF = RECEIVE
RANGE_UNITS = km
DATA_QUALITY = VALIDATED
META_STOP

DATA_START

ANGLE_1 = 2023-10-26T13:29:32.042538 1.0087853205445219
ANGLE_2 = 2023-10-26T13:29:32.042538 71.55216734739912
DOPPLER_INSTANTANEOUS = 2023-10-26T13:29:32.042538 -4.975303
RANGE = 2023-10-26T13:29:32.042538 2346.923054968505

ANGLE_1 = 2023-10-26T13:29:32.521887 1.2104445384350437
ANGLE_2 = 2023-10-26T13:29:32.521887 71.65266164078024
DOPPLER_INSTANTANEOUS = 2023-10-26T13:29:32.521887 -4.944801
RANGE = 2023-10-26T13:29:32.521887 2344.545459002357

ANGLE_1 = 2023-10-26T13:29:33.768195 1.744895520944471
ANGLE_2 = 2023-10-26T13:29:33.768195 71.91293144669791
DOPPLER_INSTANTANEOUS = 2023-10-26T13:29:33.768195 -4.860921
```

DB Optimization activity will be given by Cherrydata

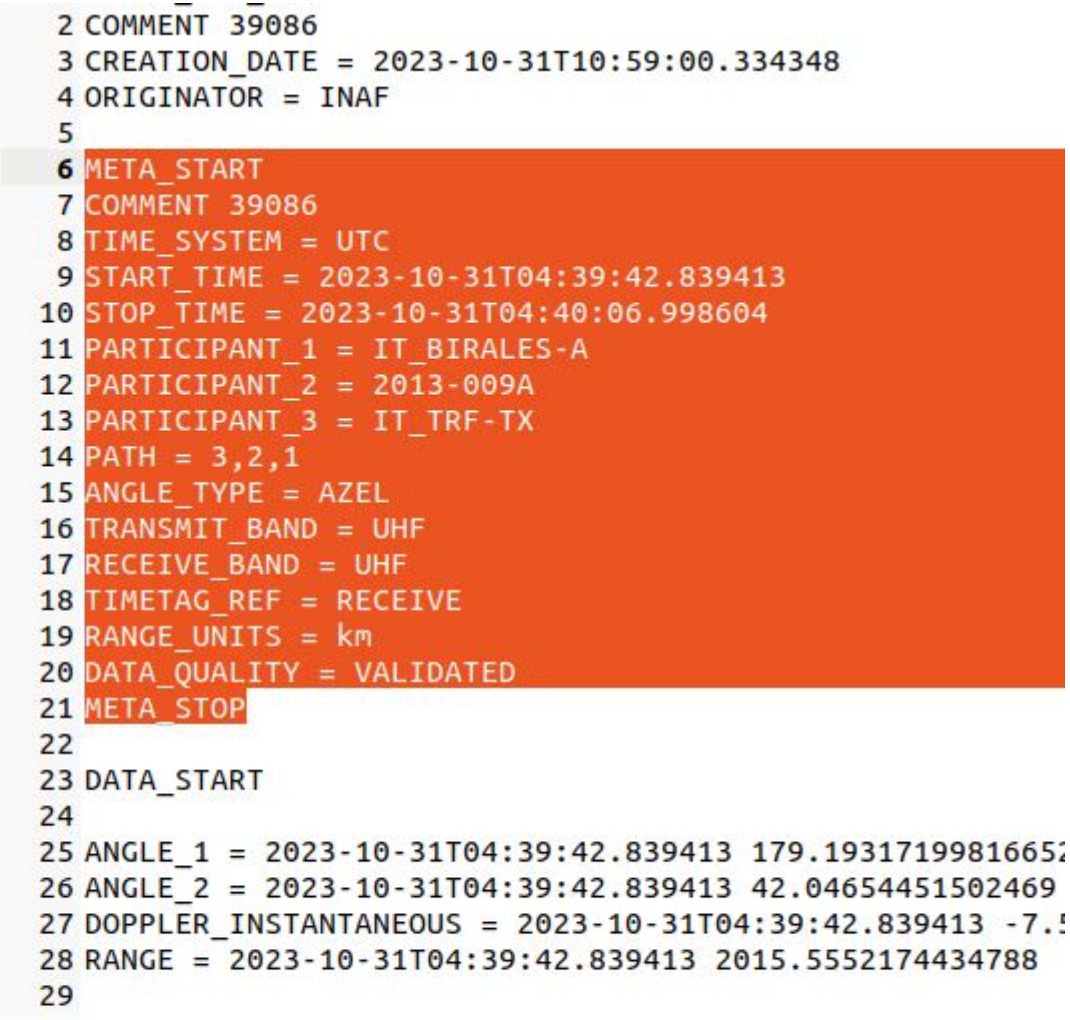
Cherrydata - Introduction

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

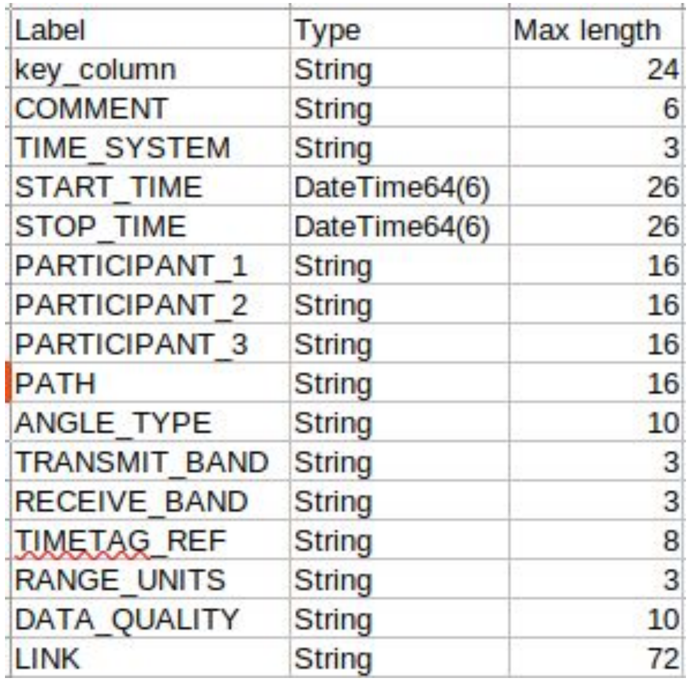

Traditional Business Intelligence
Advanced data analytics (Data Science with AI/ML techniques)
Continuous Intelligence (real time, automated)
2000
2010
2019 Cherrydata
2021 AyraDB®

AyraDB has been tested on Leonardo Davinci-1 supercomputer in 2022, as part of Euro NCC project. Cherrydata is involved in IDL as technology provider, to test AyraDB in the context of storing and querying astrophysical data and satellite measurements.

WP2 – Benchmarking Queries with AyraDB in IDL (1)



```
COMMENT 39086
CREATION_DATE = 2023-10-31T10:59:00.334348
ORIGINATOR = INAF

META_START
COMMENT 39086
TIME_SYSTEM = UTC
START_TIME = 2023-10-31T04:39:42.839413
STOP_TIME = 2023-10-31T04:40:06.998604
PARTICIPANT_1 = IT_BIRALES-A
PARTICIPANT_2 = 2013-009A
PARTICIPANT_3 = IT_TRF-TX
PATH = 3,2,1
ANGLE_TYPE = AZEL
TRANSMIT_BAND = UHF
RECEIVE_BAND = UHF
TIMETAG_REF = RECEIVE
RANGE_UNITS = km
DATA_QUALITY = VALIDATED
META_STOP

DATA_START

ANGLE_1 = 2023-10-31T04:39:42.839413 179.19317199816652
ANGLE_2 = 2023-10-31T04:39:42.839413 42.04654451502469
DOPPLER_INSTANTANEOUS = 2023-10-31T04:39:42.839413 -7.5
RANGE = 2023-10-31T04:39:42.839413 2015.5552174434788

```



Label	Type	Max length
key_column	String	24
COMMENT	String	6
TIME_SYSTEM	String	3
START_TIME	DateTime64(6)	26
STOP_TIME	DateTime64(6)	26
PARTICIPANT_1	String	16
PARTICIPANT_2	String	16
PARTICIPANT_3	String	16
PATH	String	16
ANGLE_TYPE	String	10
TRANSMIT_BAND	String	3
RECEIVE_BAND	String	3
TIMETAG_REF	String	8
RANGE_UNITS	String	3
DATA_QUALITY	String	10
LINK	String	72

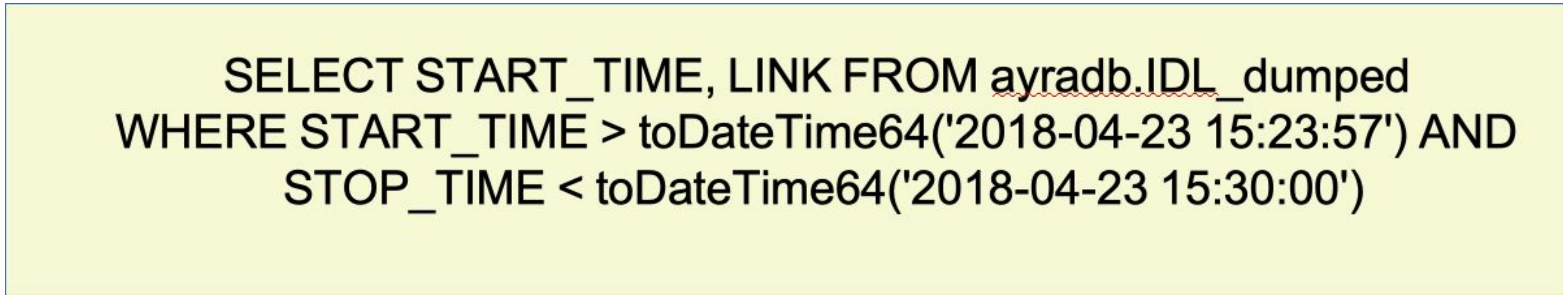
- **Level 2 TDM data**
- **metadata are fields between META_START and META_STOP (see figure on the left)**
- **We have added a field named LINK representing the reference to actual data**
- **We have defined the colums of a table as shown in the figure above**

WP2 – Benchmarking Queries with AyraDB in IDL (2)

- A synthetic dataset of approximately 1 billion records was created, for a total size of approximately 230 GB
- An AyraDB cluster of two machines in the cloud was configured, with 32 GB of RAM and 360 GB of SSD each
- The dataset was loaded onto a distributed table of the cluster

WP2 – Benchmarking Queries with AyraDB in IDL (3)

- Based on the characteristics of the dataset, it was decided to execute SQL SELECT queries, filtering the records based on START_TIME (and/or STOP_TIME)
- An example of query is:



```
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 scans the table and returns approximately 1700 records, in a time of approximately 600 milliseconds

Data Lake (WP1,WP3,WP4) - Scientific rational

Test solutions for managing data in a geographically distributed environment (aka the DataLake) by building end-to-end prototype and testbeds to demonstrate the capability to analyze the astrophysical observations and simulations available data in a cloud environment.

- **The Data Management capability**
  - Store/inject data (meant as files or data objects) in the DataLake
- **The data and compute integration for a effective processing and analysis**
  - deploy Platform as a Service (PaaS) services for the actual processing of the data ingested into the Datalake.

Define a strategies and solutions to certify that objects in the Datalake (i.e. datasets, single files archives) have not been corrupted or modified without permission

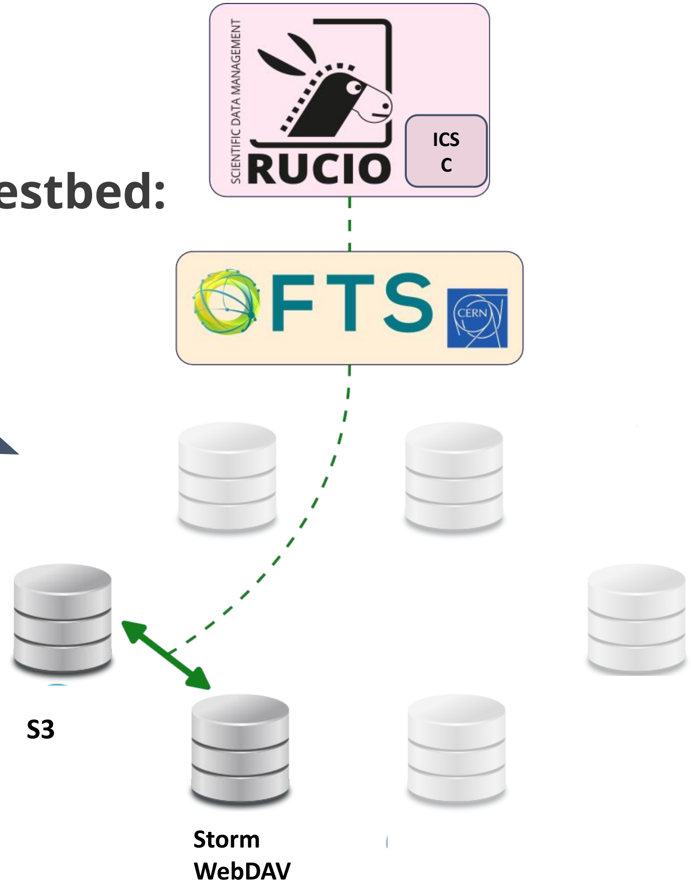
- as well as to guarantee that an object is physically in a certain location on a certain filesystem.
- all this must include the capability to trace the entire history of the data, including who made the changes.

The DataLake

Data Management Services: data orchestration and data transfers

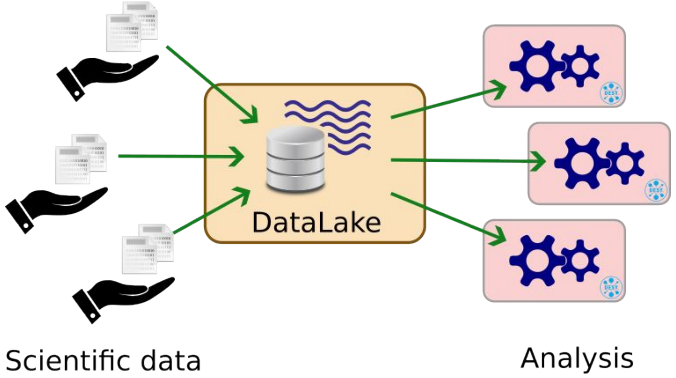
Fully aligned and sinergic to the Spoke0 model

Establishing the DataLake testbed:



Integration with science use-cases

- Collect exemplar datasets
- The datasets will be injected in the Datalake building science driven playground



Technical Objectives, Methodologies and Solutions

1. To deploy the Data Management system, storage backends and basic compute services in order to build the DataLake
2. To deploy a blockchain private service in order to deal with object certification
3. to implement the linking logic between the two system by using metadata enrichment
   a. i.e. using HASH/UUID
4. Open the DataLake to the real tests. Astrophysical observations and simulations are used for the system evaluation

Where we are

1. Resources to deploy the building blocks of the overall system are coming (through RAC)

2. All the technologies have been identified
    a. deploying procedure and related recipes have been “locally tested”

3. Structure of the test dataset successfully defined
    a. Starting with “dummy” dataset to validate the Datalake system

4. Integration logic between Datalake and certification services based on metadata is in early definition phase
    a. there are technical open topics to be further evaluated as soon as the Datalake PoC will be ready

The first PoC of the Datalake is expected by the end of May

About data certification

A virtuous example of a research topic of interest not only in the context of the **industrial environment** (i.e to prevent fraud and corruption) but also in **several scientific domains:**

- data provenance
- data analysis reproducibility
- scientific output certification

IGUC - Interoperability Data Lake for Gaia Use Case

- ❑ IGUC is an additional WP (WP2_G) of the IDL project.
- ❑ The project expands the IDL project by bringing a new typology of astronomical big data offering a new challenge in big data management and recovery: the Gaia use case.
- ❑ Gaia data present similar data field relations and characteristics with the data typology the company is interested in.
- ❑ Excellent case for testing new solutions of data management and data retrieving initially established in the contest of IDL project, stressing the performance of the technological solution.

LEONARDO

INAF

IGUC - Interoperability Data Lake for Gaia Use Case

❑ The specific technological goal of IGUC project is to identify and implement additional database and data management solution to complement the traditional ones. The purpose of this activity is to support the integration and query of data coming from different sources, with a performance that enables novel application with real-time requirements, to achieve maximum effectiveness and efficiency in data provisioning and exploitation.

LEONARDO

❑ The Gaia INAF team scientific goal is to do a further step in the implementation of the innovative platform dedicated to Gaia's legacy located at DPCT, showing the best solutions to retrieve billion of data for analyzing portions of the sky (tenth of square degrees) to identify significant variations of sources, to support science as discovery and characterization of cosmological gravitational waves or new earth-like planets.

INAF

IGUC - Interoperability Data Lake for **G**aia **U**se **C**ase

Partners involved: ONLY Leonardo/Cherry Data and INAF/Gaia team

LEONARDO

□ The Data are covered by an **NDA - NO PUBLIC DATA**

↓

An ad-hoc undisclosure agreement is foreseen

Database deployment in ICSC infrastructure **BUT IGUC experiment will carry on a dedicated INAF/Leonardo infrastructure due to the data policy.**

INAF

The agreement provides that once the project is concluded the data will be deleted.

IGUC - Chronological description of the project activities:

The main project activities are supposed to start at **M4** after gathering the Gaia use case requirements and the definition of the HW infrastructure features to realize the experiment.

WP Deliverables and milestones:
M1-M2
Project Kick-off
Gaia Use case requirements gathering, data model and metadata definition (INAF)
M4
Benchmark requirements definition: definition of the metric to obtain an estimate of the performances invariant with respect to the execution platform (INAF-Leonardo)
M4 deliverable: Technical report (Leonardo-INAF)
M13
Implementation of the Gaia PoC on the Gaia Legacy prototype infrastructure at DPCT (INAF)
Implementation of the Gaia PoC on the INAF HW infrastructure/Da Vinci (Leonardo-INAF)
M13 deliverable: Report including hw and sw architecture description and verification tests (Leonardo-INAF)
M15
Final benchmark results (Leonardo-INAF)
M15 deliverable: Report on final results (Leonardo-INAF)

LEONARDO

INAF