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

WP4 Status Report

Cristina Knapic - INAF OATs

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

Scientific Rationale

- Design and implementation of a Data Lake VO and astrophysic standard compliant to store and distribute scientific data (mainly outcome of key science projects) of:
  ○ synthetic and simulated data;
  ○ observation data;
  ○ reduced data;
- Design and implementation of a performant and open source based infrastructure for Big Data databases

The archive of the different key science projects data will be modelled to be Virtual Observatory compliant to improve the data interoperability. The ICSC infrastructure will be the final deployment infrastructure in order to benefit of the computation and high storage capacity and the new architecture will be tested to verify the performances will be in line with the expectations.

Technical Objectives, Methodologies and Solutions

The goal of the WP4 is the implementation of a data lake for the Big Data. Expectations in production are several Terabyte of online interoperable data.
To reach the goal we subdivided the efforts in 3 tasks:

i) Modellization (data model and data management plan)
   (1) DMs for simulated data;
   (2) DMs for observation data;
   (3) DMs for catalogues;

ii) data lake infrastructure:
   (1) services instead of storage mount points, redundancy and transfer efficiency;
   (2) data workflow implementation;

iii) data portals and data retrievals methods:
   (1) general purpose front end for the customized archives;
   (2) interoperable access to all the archives both for programmatic and human interaction processes;
   (3) Data access, sharing and retrieval policy

Technical Objectives, Methodologies and Solutions

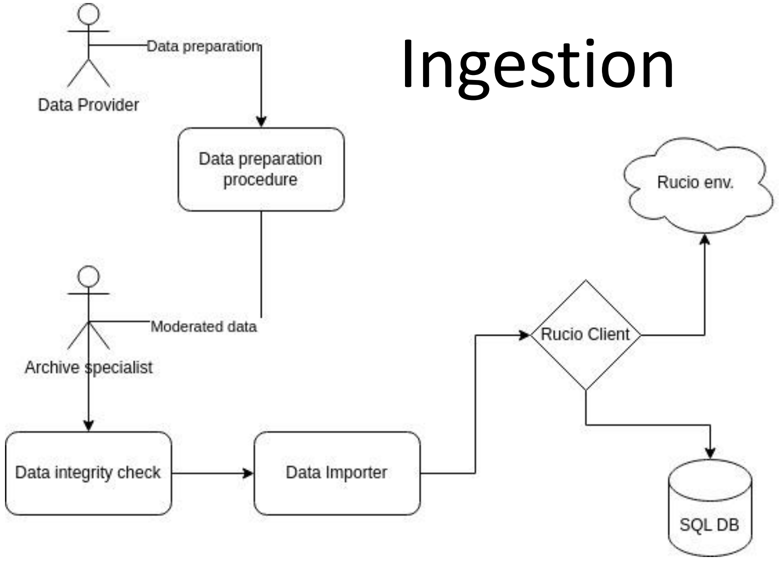
Methodology adopted is:

- regular monthly meetings;
- use of mailing list;
- regular weekly telecon to harmonize technical work;
- strong interaction with data providers in order to move in the direction of standardization of the archives custom structure:
  - procurement of complet sampling dataset;
  - dataset analysis and splitting of content in metadata and data (application of the OAIS principle);
  - customization of data handling software for each dataset (or collection) to handle different data formats;
  - customization of the data lake management software to include the descriptive astrophysical metadata;
  - interaction with the data provider to verify the VO compliancy;
  - analysis of the most frequently used queries to optimize the database for data search;
  - implementation of the archive portal respecting the privacy policy and data structure prescriptions;
  - first data archive release;
  - consequent interaction to adjust and refine goals;

Technical Objectives, Methodologies and Solutions

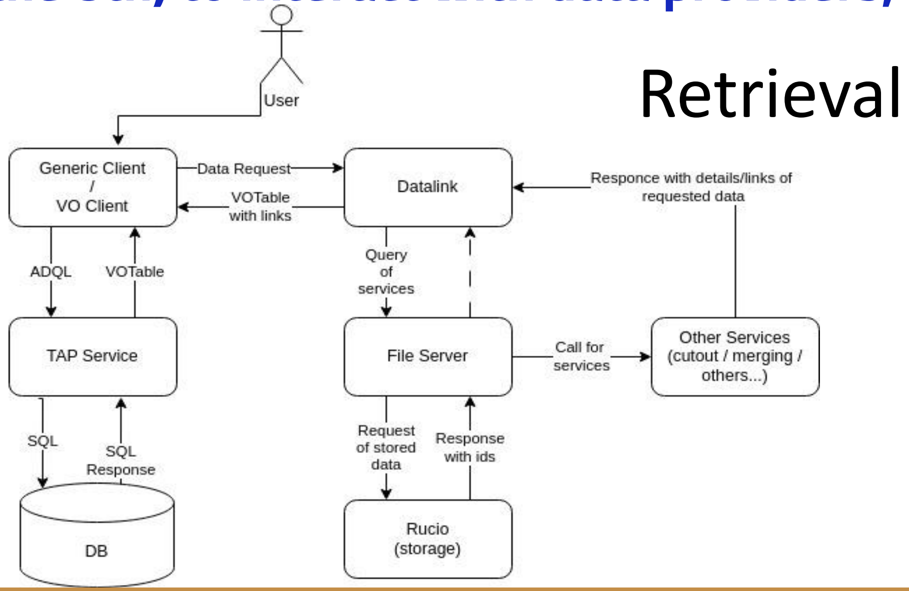
Technical solutions:

- Data Models using OAIS principles, VO suggestions and best practices (see next talks by G. Coran, M. Costantini and S. Gelsumini and A. Adelfio);
- Python based software for data ingestion (closer to scientific software development approach);
- architecture of services discussion and set up;
- Dimensioning of the infrastructure resources needed on VMs;
- Call for PICA;
- Meetings and round table (14:30 on Wednesday the 8th) to interact with data providers;



Ingestion

Data Provider
Data preparation
Data preparation procedure
Archive specialist
Moderated data
Data integrity check
Data Importer
Rucio Client
Rucio env.
SQL DB



Retrieval

User
Generic Client / VO Client
Data Request
VOTable with links
Datalink
Responce with details/links of requested data
ADQL
VOTable
Query of services
TAP Service
File Server
Call for services
Other Services (cutout / merging / others...)
SQL
SQL Response
DB
Request of stored data
Response with ids
Rucio (storage)

Personnel recruited

-INFN recruited for WP4
- ○ 0.5 FTE for Task 4.2 (Rucio Specialist)
- ○ 0.5 FTE for Task 4.1 (DM definition)

- INAF recruited for WP 4
- ○ 1 FTE for Task 4.3 (web development)
- ○ 0.7 FTE for Task 4.1 (DM definition)
- ○ (1 FTE involved as INAF additional contribution and IG IDL from 06/24)

Timescale, Milestones and KPIs

WP 4		Milestone 7: September 2023 - February 2024			
Milestone WP		**Proponent (Name - Institute)**	**Description (short description of the activity the target/KPI addresses)**	**Target**	**KPI**
M7		Andrea Adelfio	First importing in the Rucio Ancillary DB of a collection metadata.	Integration of the metadata descriptors into a distributed infrastructure for data sharing among data centers.	Internal report-
M7		Cristina Knapic	Data Ingestion Software for importing metadata in ancillary Rucio DB and pushing data in Rucio	Inheriting and expanding existing software to import data in Rucio and metadata in DB. Some debugging on Phyton multithreading libraries.	Technical report
M7		Massimo Costantini	Smart web application on top of Rucio for data management.	Deployment and first tests on a Postgress high availability installation, with the PGSpheres, ready for Rucio ancillary DB installation.	Technical Report
M7		Deborah Busonero, Sara Gelsumini	Implementation of a smart approach to the in-memory DB for GAIA	Design and implementation of a smart approach for in-memory GAIA db using open-source db like Postgress.	Technical Report

Timescale, Milestones and KPIs

WP 4		Milestone 8: Jan 2024 - Jun 2024			
Milestone WP		**Proponent (Name - Institute)**	**Description (short description of the activity the target/KPI addresses)**	**Target**	**KPI**
M8		Diego Ciangottini	Implementation of Rucio data lake for several collections, depending on the Key Science Products.	Task 4.2 - Implementation of the data lake	Publication on a conference journa
M8		Massimo Costantini	Implementation of an evolutive prototype web app to export the different data collections.	Task 4.3 - Archives web interface	Technical report (INAF repository) and user manuals
M8		Deborah Busonero Sara Gelsumini	Implementation of a evolutive prototype of the in-memory DB for GAIA	Task 4.3 - Design and implementation of a smart solution to cover the in-memory db for GAIA version2 Evolutive prototype	Publication in ADASS or SPIE conference

Accomplished Work, Results

-Task 4.1

- Definition of use cases DMs for cases FERMI, GAIA, PLUTO and RAMSES;
- Definition of data flow and queries (work in progress);

-Task 4.2

- Implementation of the data lake (Rucio);
- Deployment of archive software (Importer);
- Performance tests (Importer);

-Task 4.3

- Implementation of archive interface (web portal);
- Inclusion of scientific tools for data handling (cutout - work in progress);
- Definition of data policies (work in progress);
- Implementation of data access methods (use of AuthN & AuthZ OIDC/VO compliant);

Accomplished Work, Results

-Task 4.1

- ○ Analysis of the structure of the proposed dataset and definition of the possible DMs;

-Task 4.2

- ○ Rucion installation and set up;
- ○ Analysis of the scripts to upload of the relevant metadata in the ancillary Rucio DB customized per dataset;
- ○ Writing and set up of the archival ingestion software:
  - ■ investigation on multithreading and parallel programming;
  - ■ Performance tests;
- ○ study of the queries in order to optimize db structure and performances;

-Task 4.3

- ○ Analysis of the services architecture and internal relations;
- ○ Data Lake portal definition and implementation;
- ○ Analysis of the access policies features: users, groups and groups of groups;
- ○ Additional services study (merging, cutout, FERMI Tools (?));
- ○ Implementation of VO compliant services;

Next Steps and Expected Results

-Task 4.1

- ○ Inclusion of eventual other KSP in the modelization process;
- ○ Data management Plan for FERMI, GAIA, PLUTO, RAMSES and other cases;
- ○ Definition of data flow and queries;

-Task 4.2

- ○ Verification of a simple case from data ingestion to retrieval;
- ○ Performance tests on the whole system;

-Task 4.3

- ○ Inclusion of scientific tools for data handling (datalink and other scientific tools, connection with WP3 for data analysis and visualization);
- ○ Implementation of data access policies (use of AuthN & AuthZ OIDC/VO compliant);

The most challenging result in WP4 is to be able to handle so different data and purposes on a performant, interoperable and transparent platform.