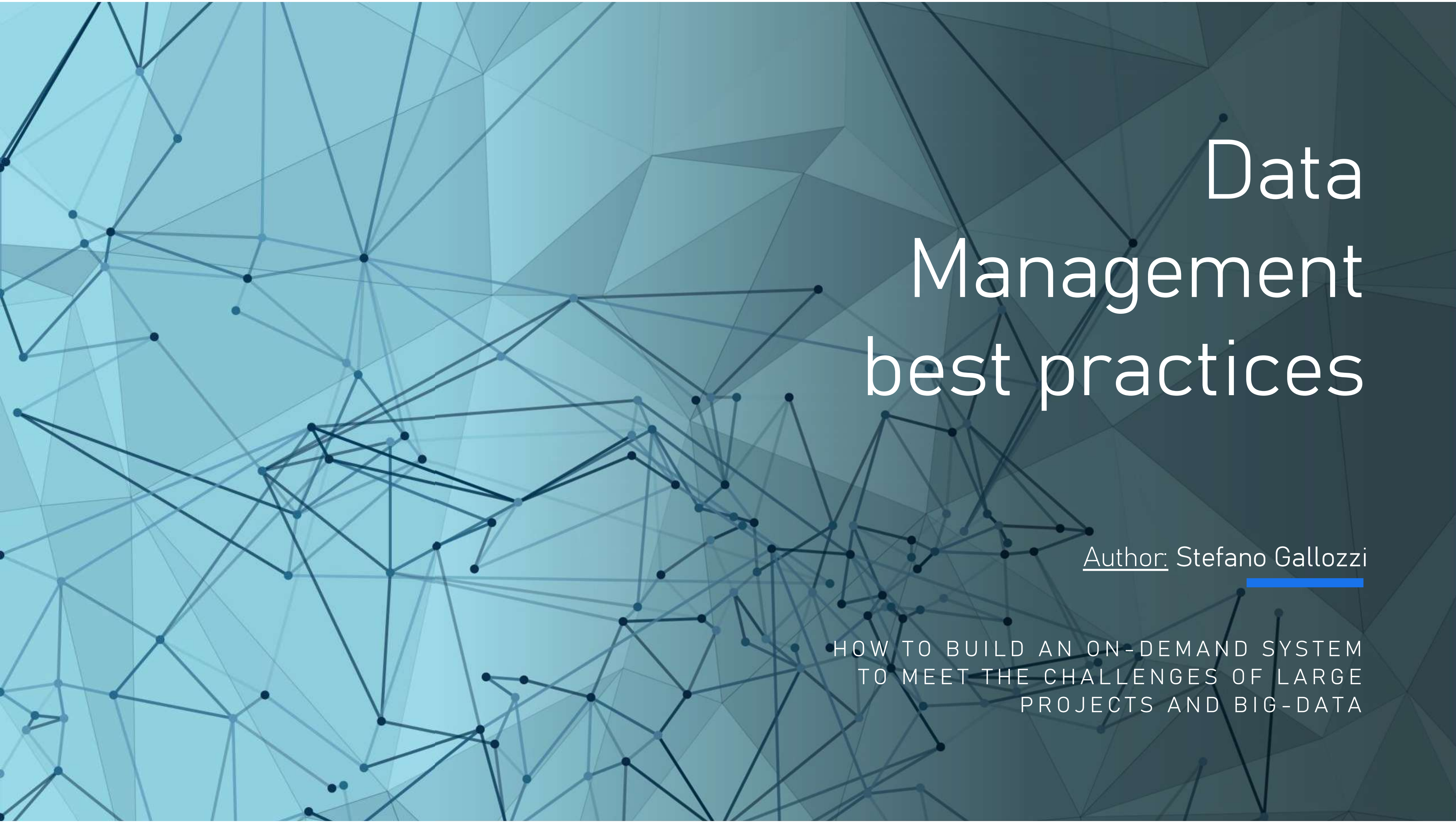
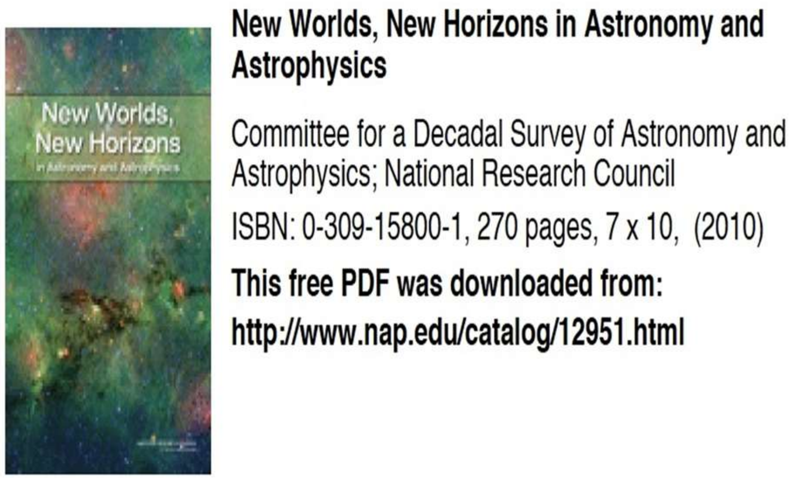


Data Management best practices

Author: Stefano Gallozzi

HOW TO BUILD AN ON-DEMAND SYSTEM TO MEET THE CHALLENGES OF LARGE PROJECTS AND BIG-DATA

Who manage data well can make “good” & “better” science!

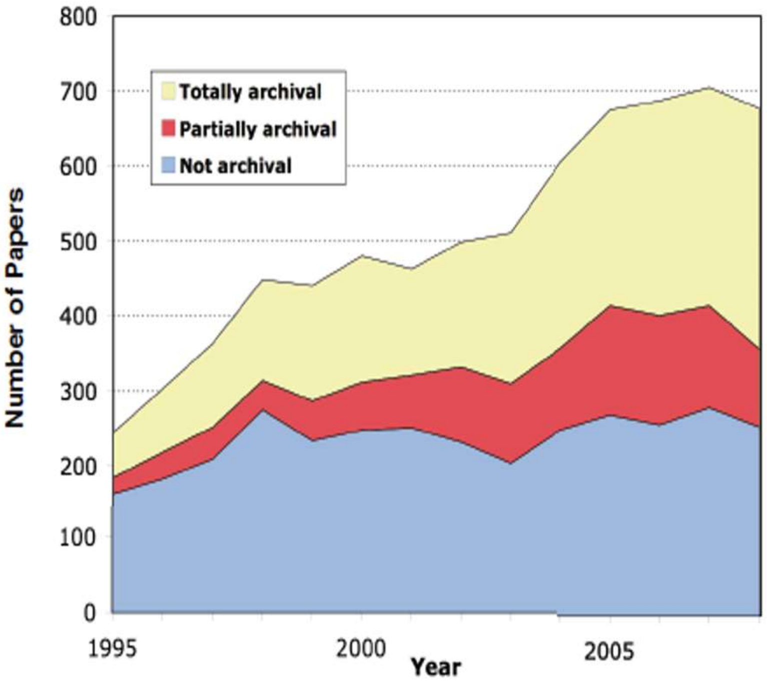


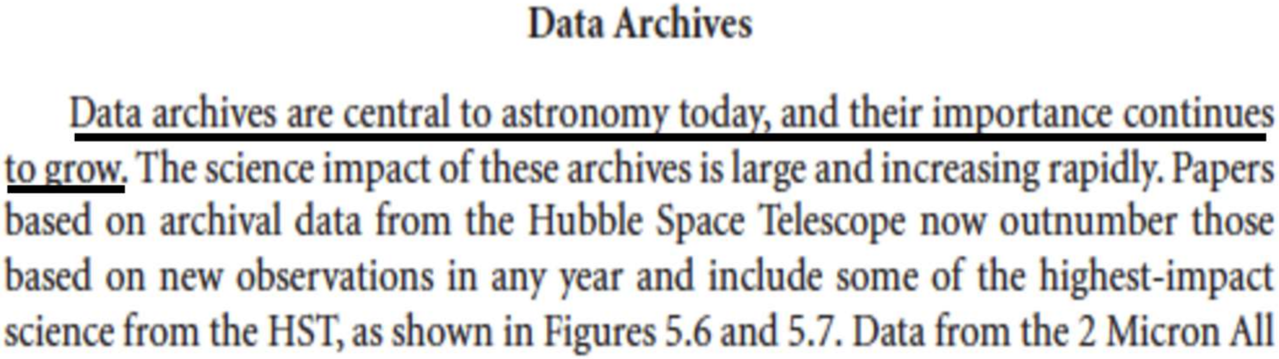
New Worlds, New Horizons in Astronomy and Astrophysics

Committee for a Decadal Survey of Astronomy and Astrophysics; National Research Council

ISBN: 0-309-15800-1, 270 pages, 7 x 10, (2010)

This free PDF was downloaded from:
http://www.nap.edu/catalog/12951.html





Data Archives

Data archives are central to astronomy today, and their importance continues to grow. The science impact of these archives is large and increasing rapidly. Papers based on archival data from the Hubble Space Telescope now outnumber those based on new observations in any year and include some of the highest-impact science from the HST, as shown in Figures 5.6 and 5.7. Data from the 2 Micron All

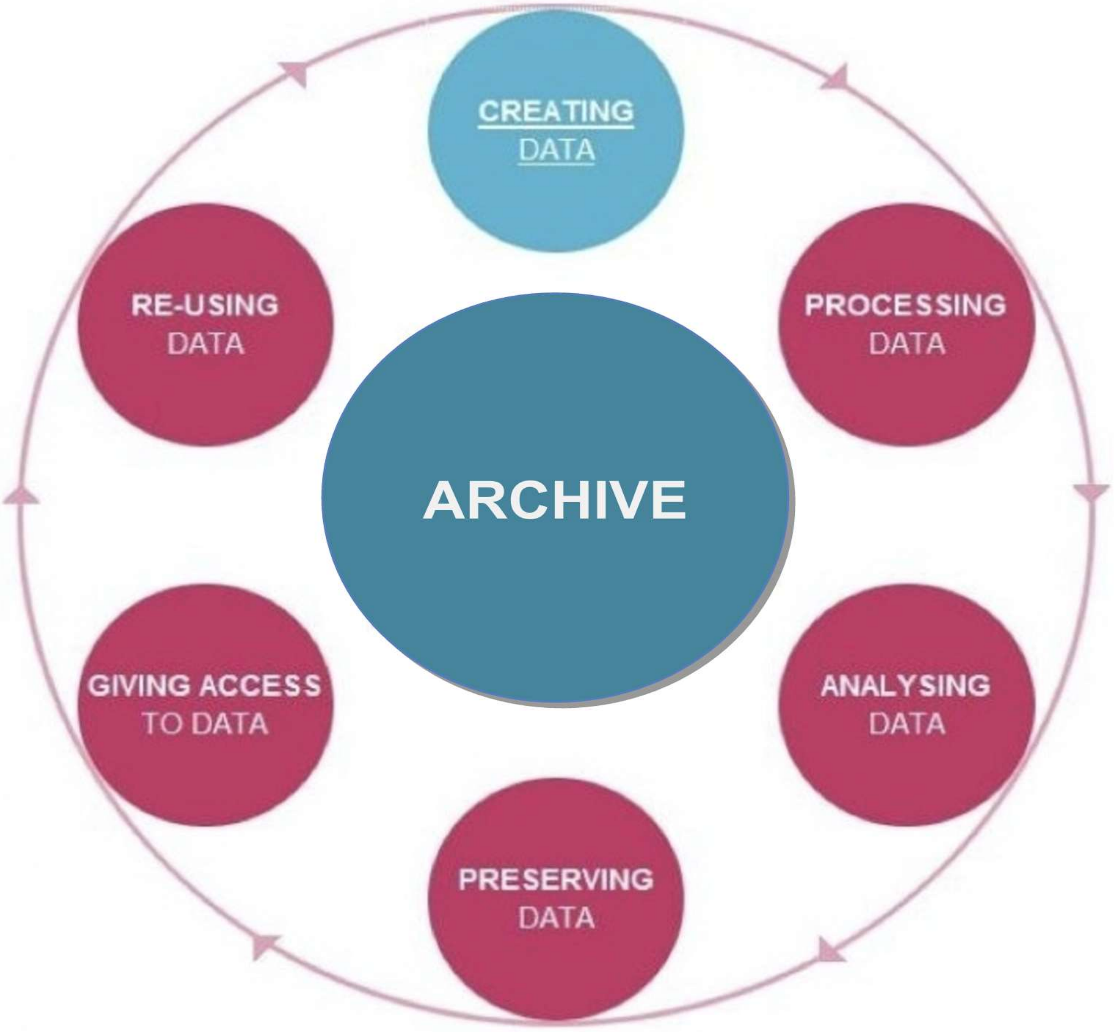
Archive is not a simple Repository!

If you have to handle large amounts of data of different nature, a repository is not enough but you need an archive system with a database catalogue to allow punctual searches on dataset.

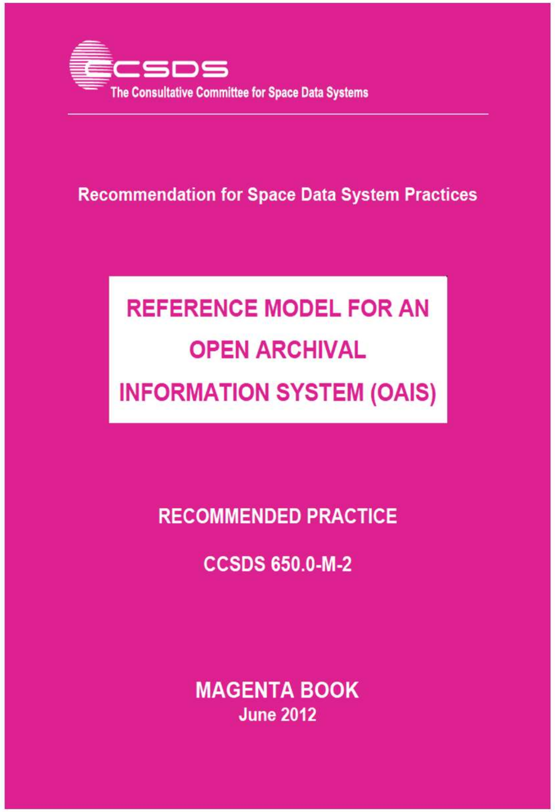
Archive «role» is central in Astrophysics

The major aim of a Scientific Archive is to guarantee data preservation and access information for the Long Term and for all data science products.

The archived information must be also usable by different user categories (data consumers) who are separate in time, space and background from the data producers. Archive must be accessible well beyond the end of the operational life of the observatory.



Open Archival Information System (OAIS)
the Reference Model



CCSDS
The Consultative Committee for Space Data Systems

Recommendation for Space Data System Practices

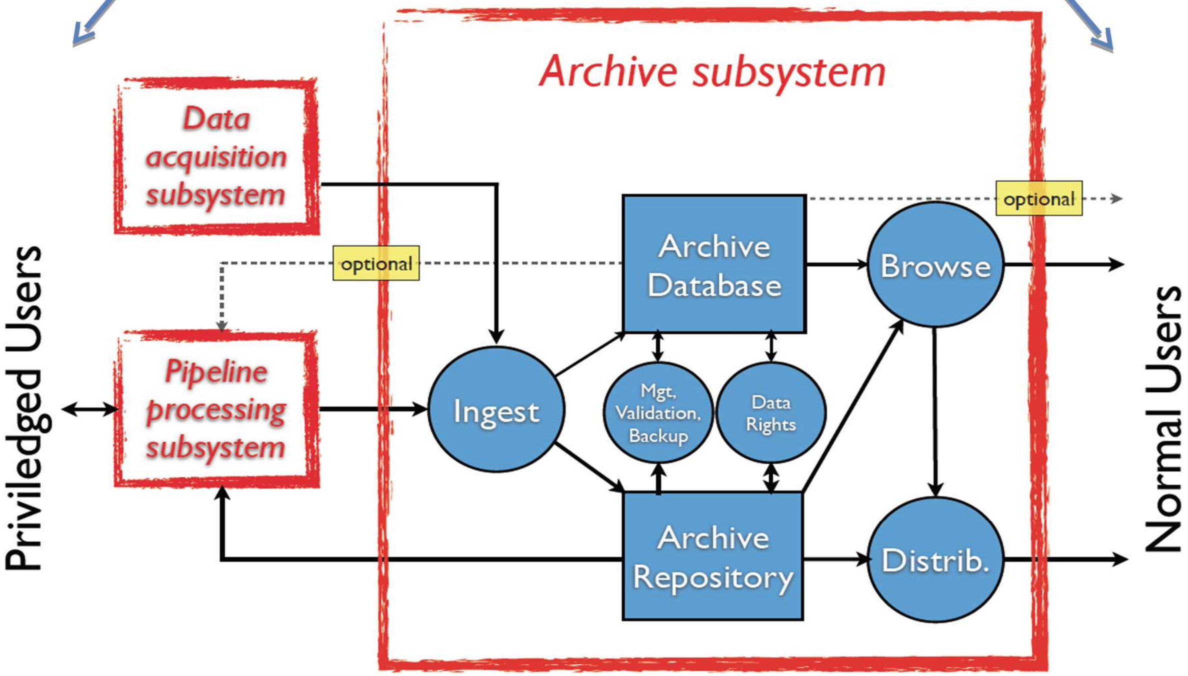
REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

RECOMMENDED PRACTICE

CCSDS 650.0-M-2

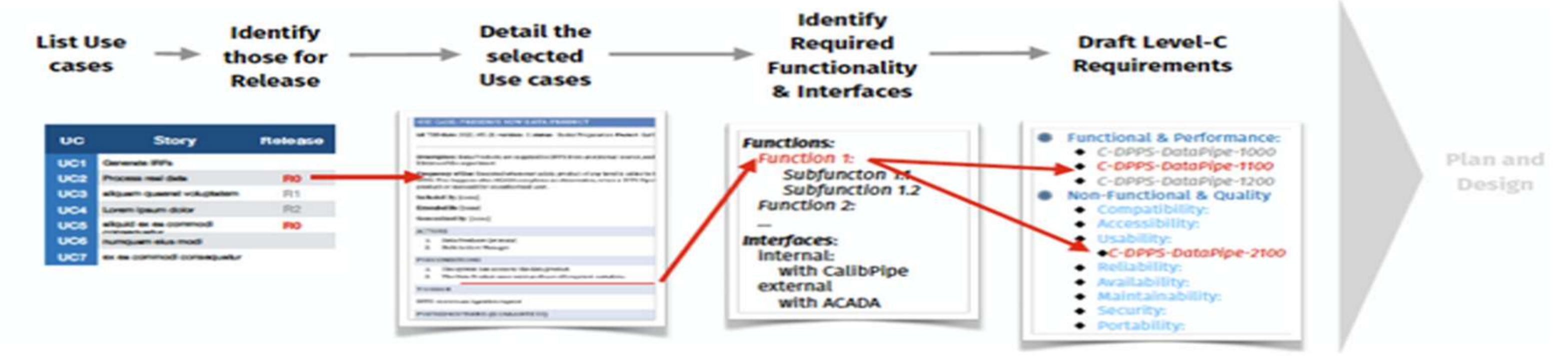
MAGENTA BOOK
June 2012

Data Producers and **Data Consumers** are rigidly separated by the **Archive System**



Good Data Management pass throught good Project Management

1. Write UCs
2. Extract function tree from the UCs
3. Extract functional and non-functional* requirements
4. Extract any interfaces (internal and esternal) implied
5. For each interface write a requirement
6. Translate requirements into an accurate system design



Critical Continuous Integration & Iteration Assessments

- ➔ Find Prototypes
- ➔ Fix Computing Computing Model
- ➔ Evaluate Costs
- ➔ Hardware Procurement
- ➔ Pre-Production
- ➔ Production

Make a simple Workflow for each use-case (i.e. Ingestion)

- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ Avoid SPOF
- ➔ Think on Security
- ➔ Find suitable SFTW products
- ➔ Finad & Purchase HW
- ➔ Buildup the System

Data Producers:
From Cameras
& Workload (sim & Pipes)

Data Products:
From lev0-1-2-3
to science ready

Storage & Computing Distributed
Different Datacenters

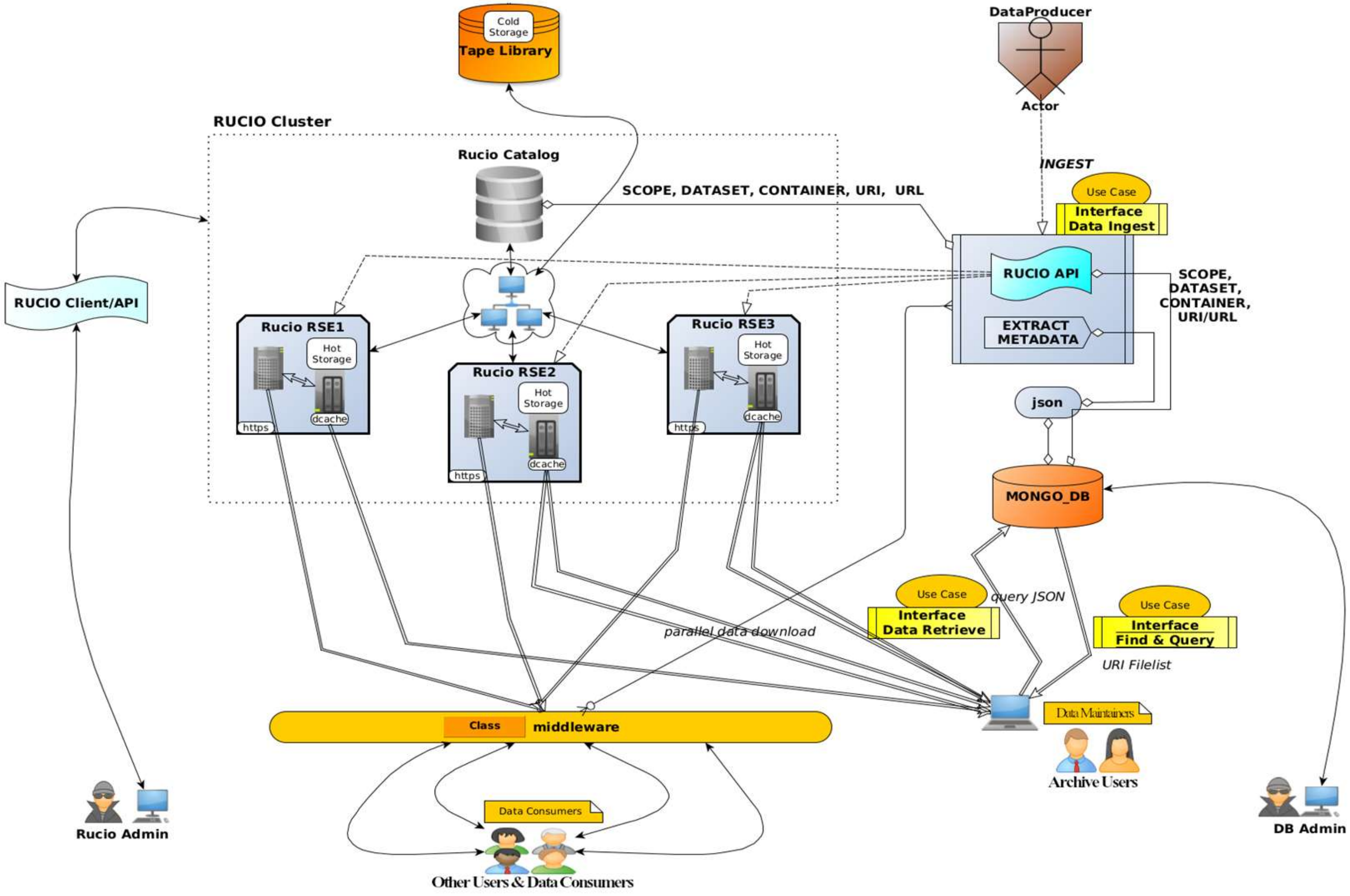
Access Protocols:
Security
Connections
and A&A

Database Catalog:
Should Distributed and
follow Datacenters

Scalability:
Hardware and Resource
Horizontal Scalability

Interfaces:
To Browse Metadata and Access
Data by Sinple Queries

Make a simple Workflow for each use-case



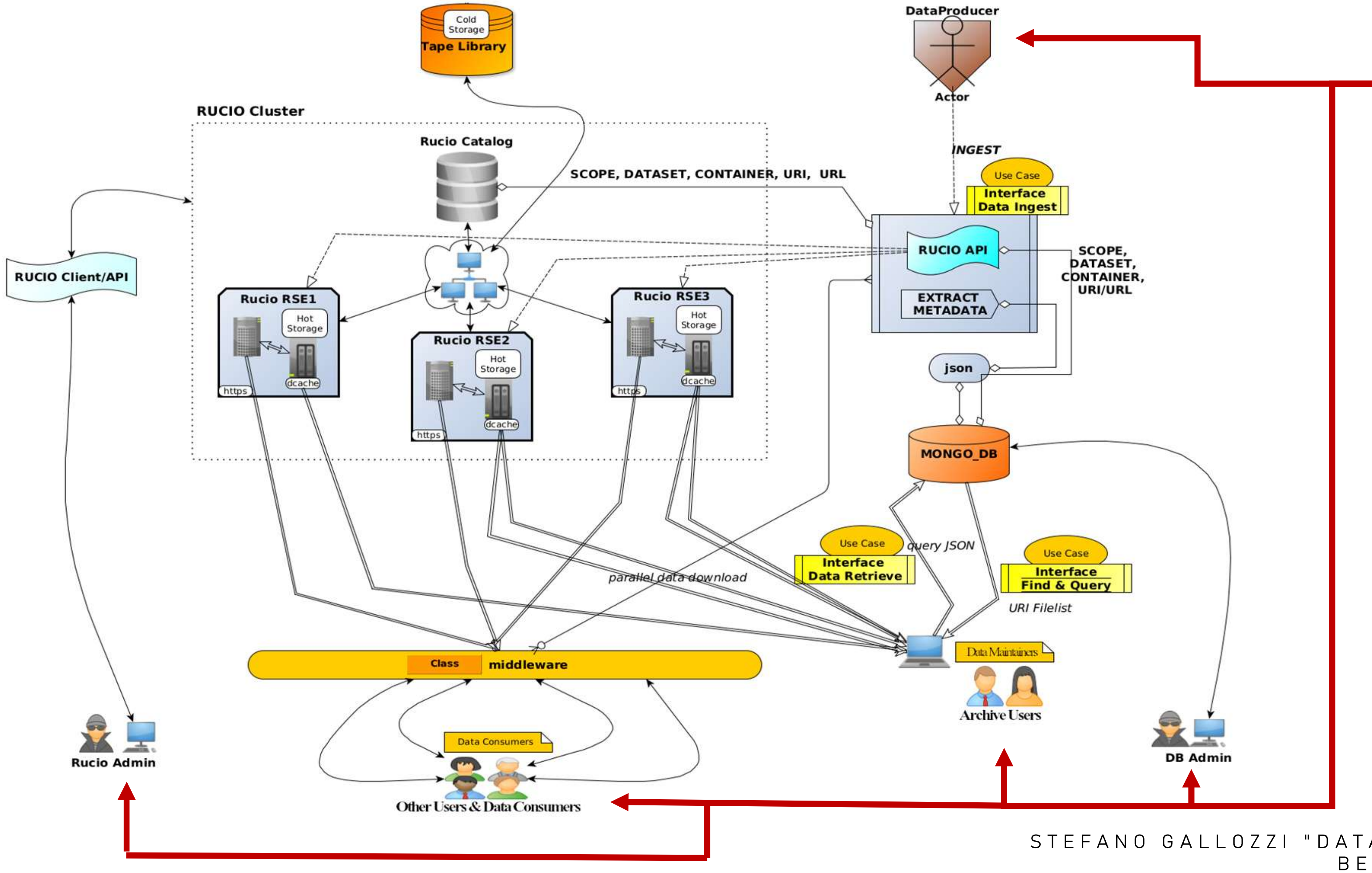
➔ Find Archive Users

➔ Define suitable Storage Sys

➔ Divide into Atomic Functions

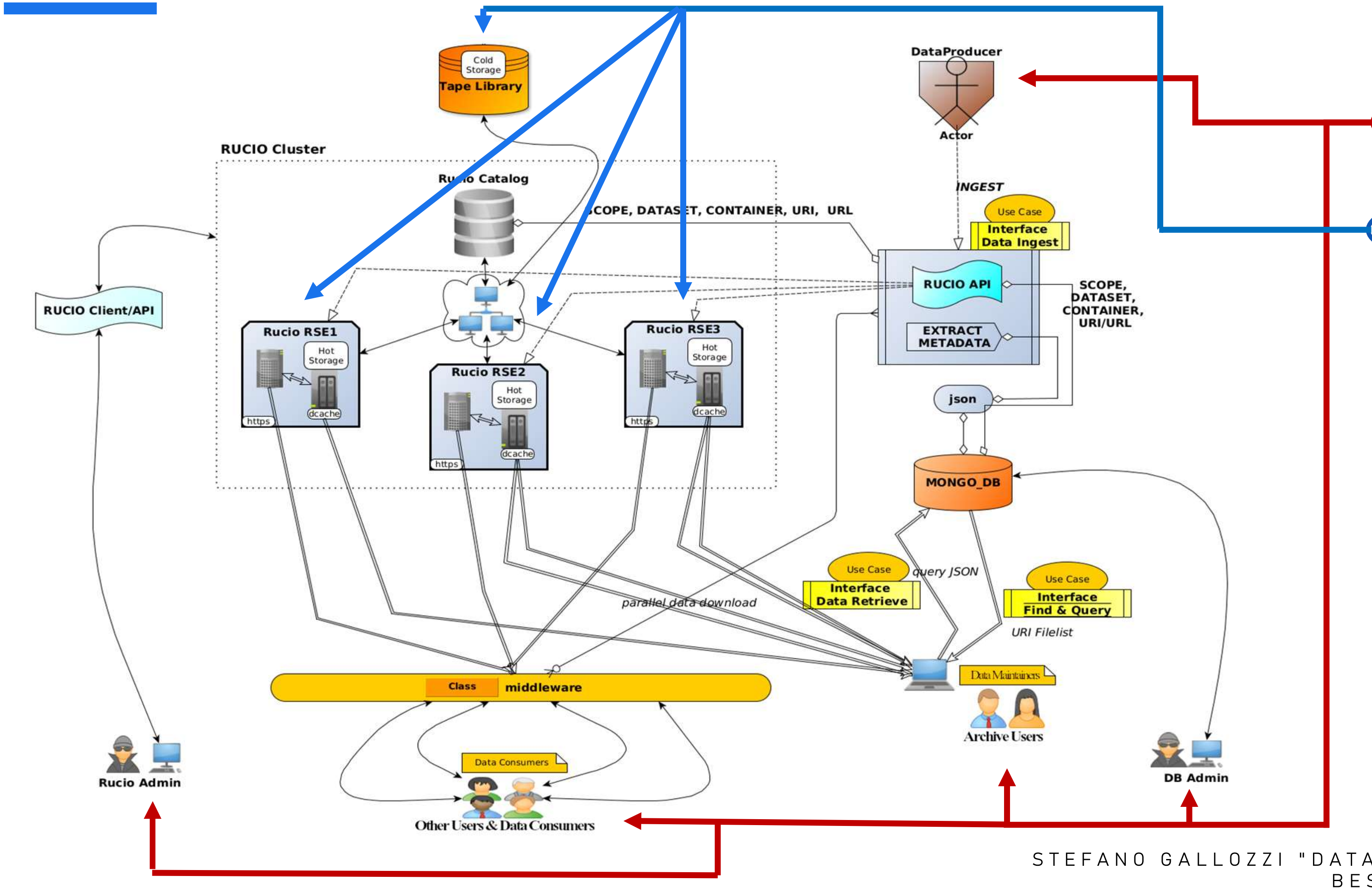
➔ Identify Use Cases

Make a simple Workflow for each use-case



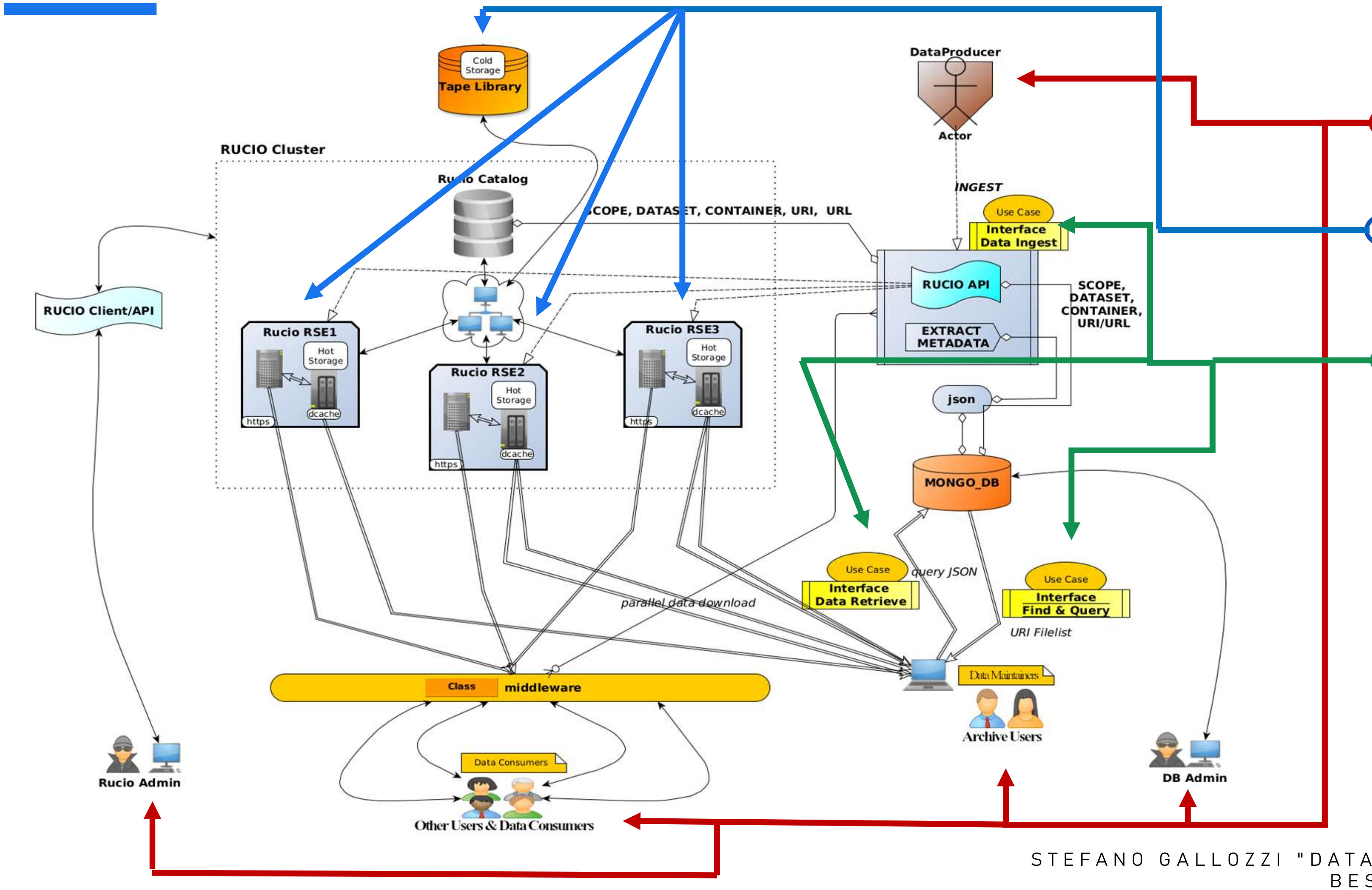
- ➔ Find Archive Users
- ➔ Define suitable Storage Sys
- ➔ Divide into Atomic Functions
- ➔ Identify Use Cases

Make a simple Workflow for each use-case



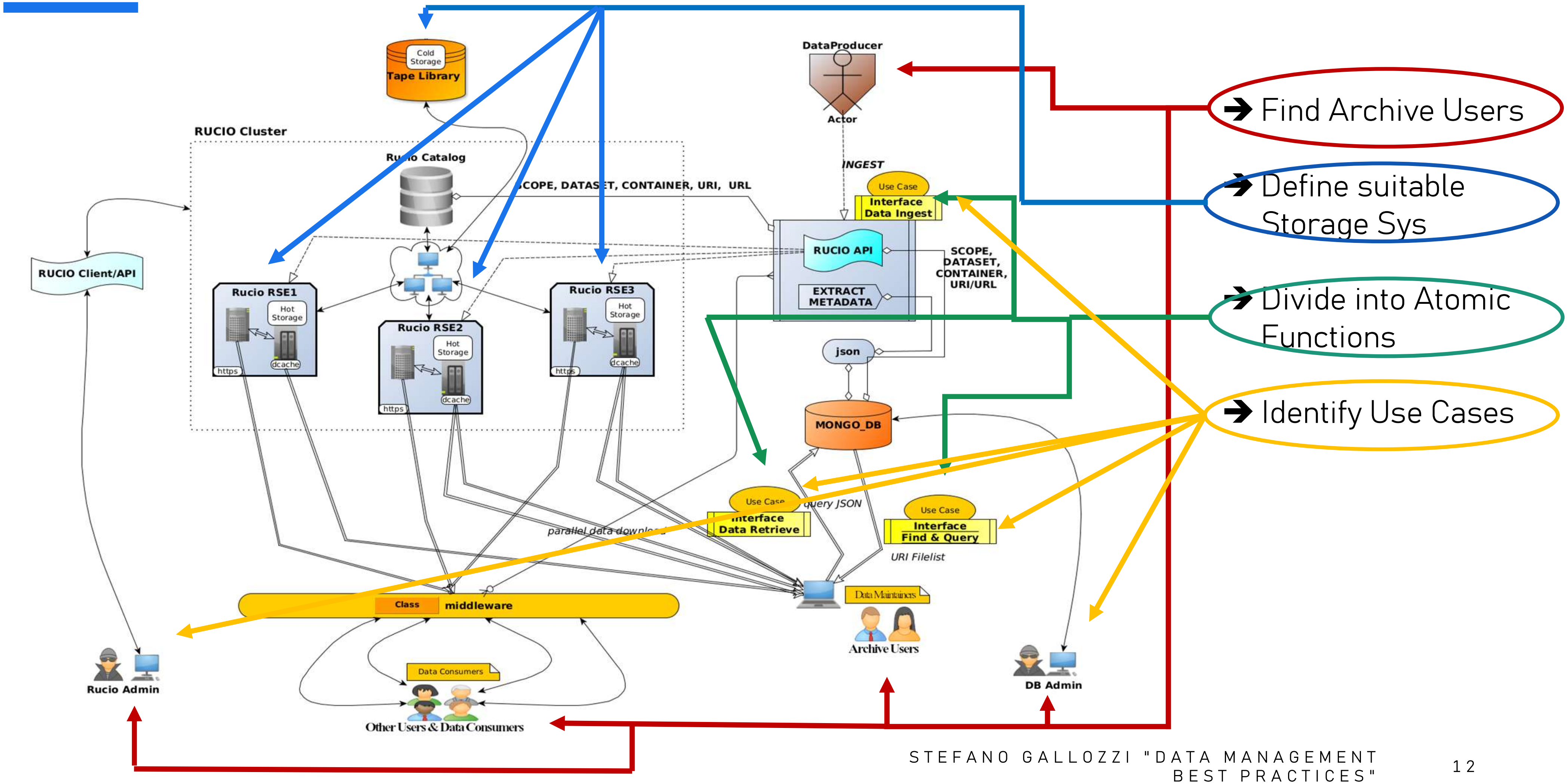
- ➔ Find Archive Users
- ➔ Define suitable Storage Sys
- ➔ Divide into Atomic Functions
- ➔ Identify Use Cases

Make a simple Workflow for each use-case



- ➔ Find Archive Users
- ➔ Define suitable Storage Sys
- ➔ Divide into Atomic Functions
- ➔ Identify Use Cases

Make a simple Workflow for each use-case



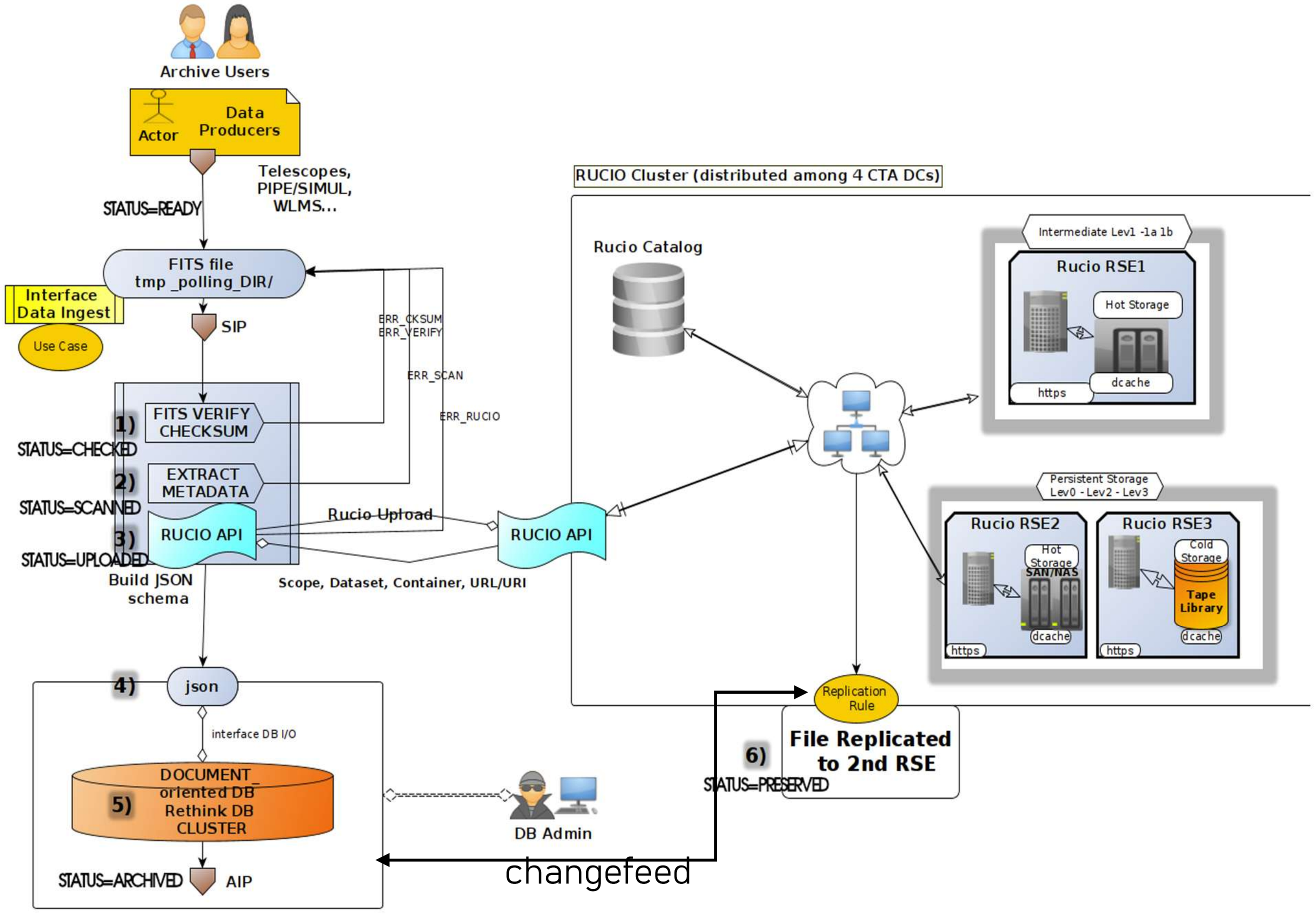
➔ Find Archive Users

➔ Define suitable Storage Sys

➔ Divide into Atomic Functions

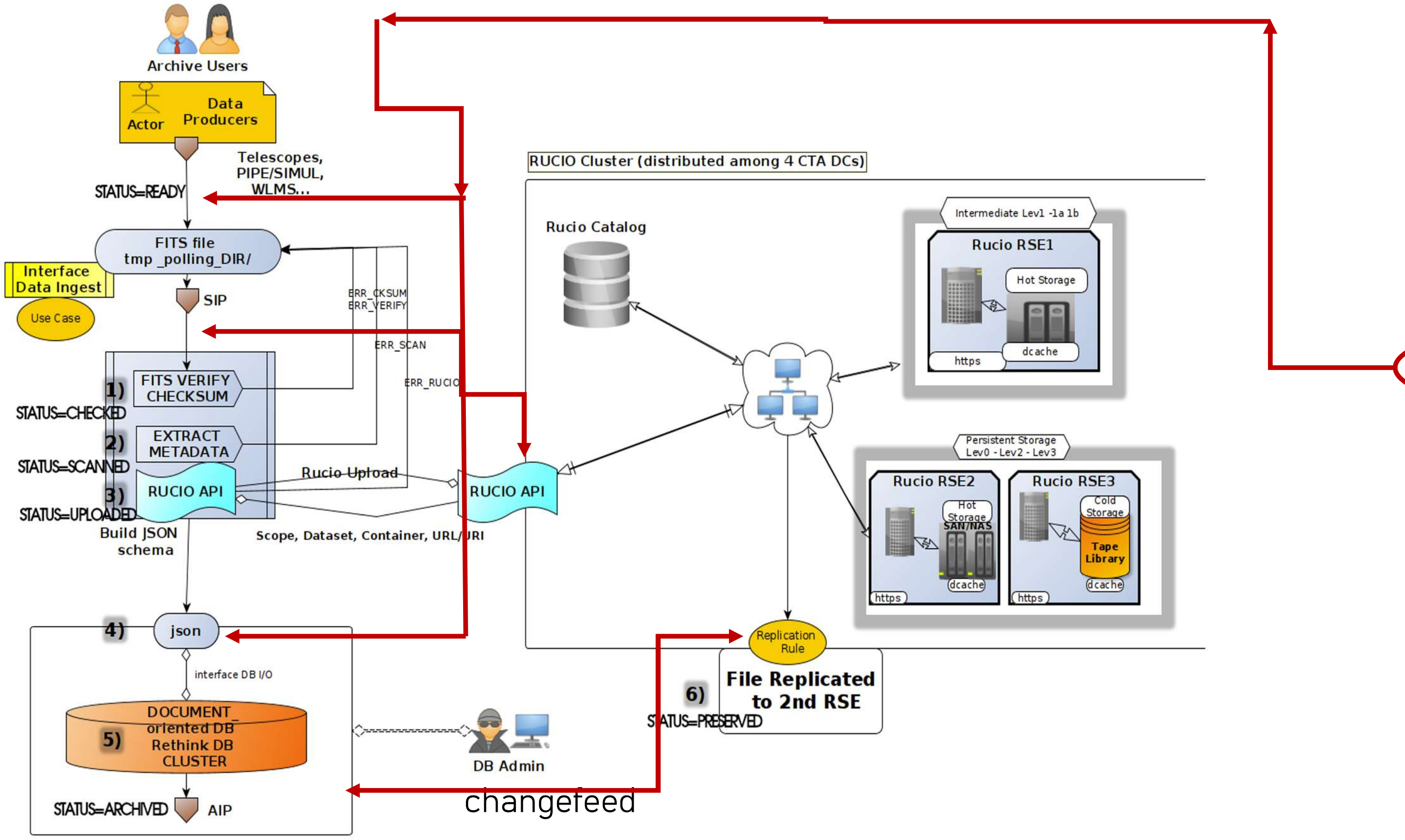
➔ Identify Use Cases

Analyse Prototypes: Use Case INGEST



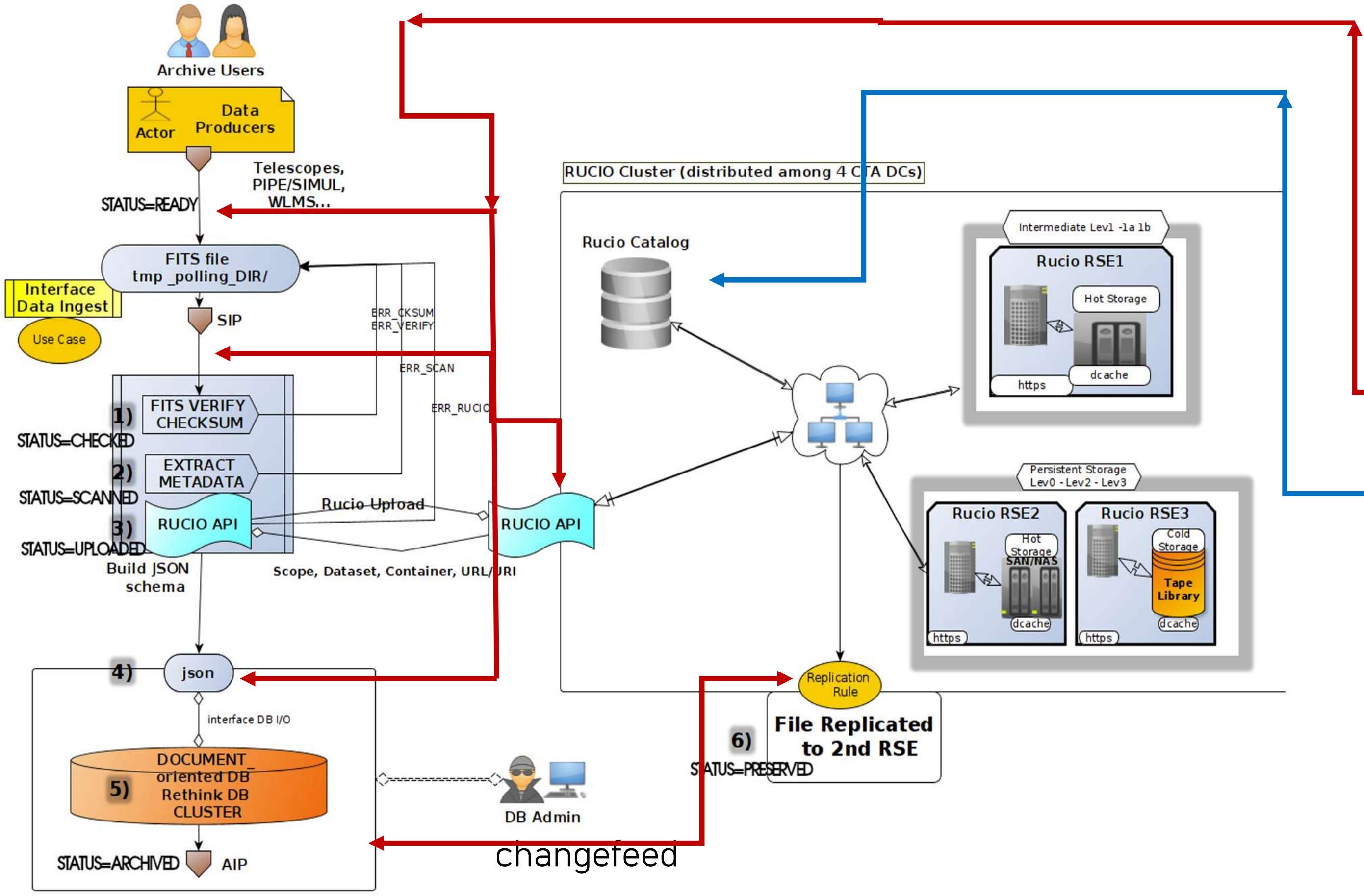
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case INGEST



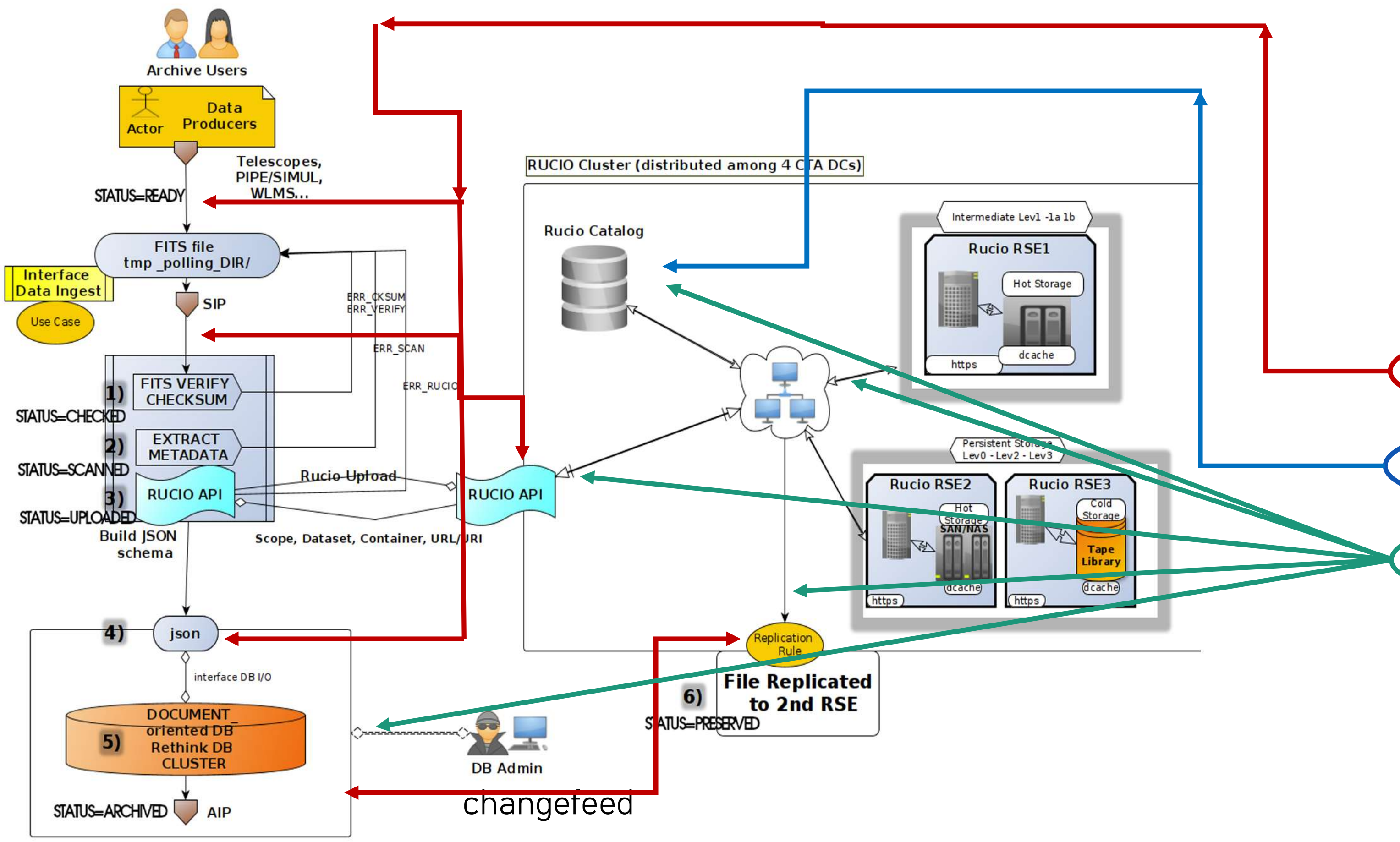
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case INGEST



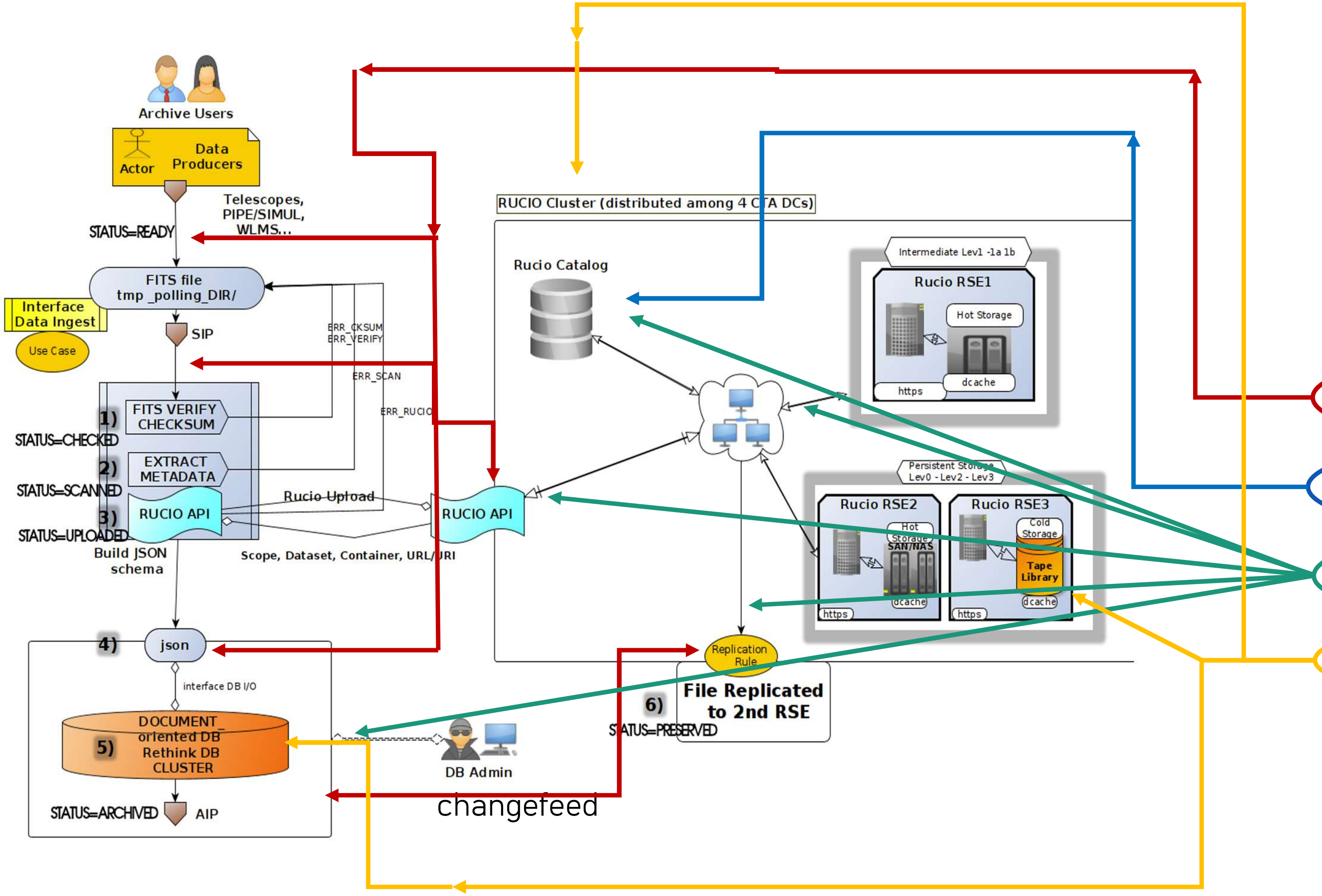
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case INGEST



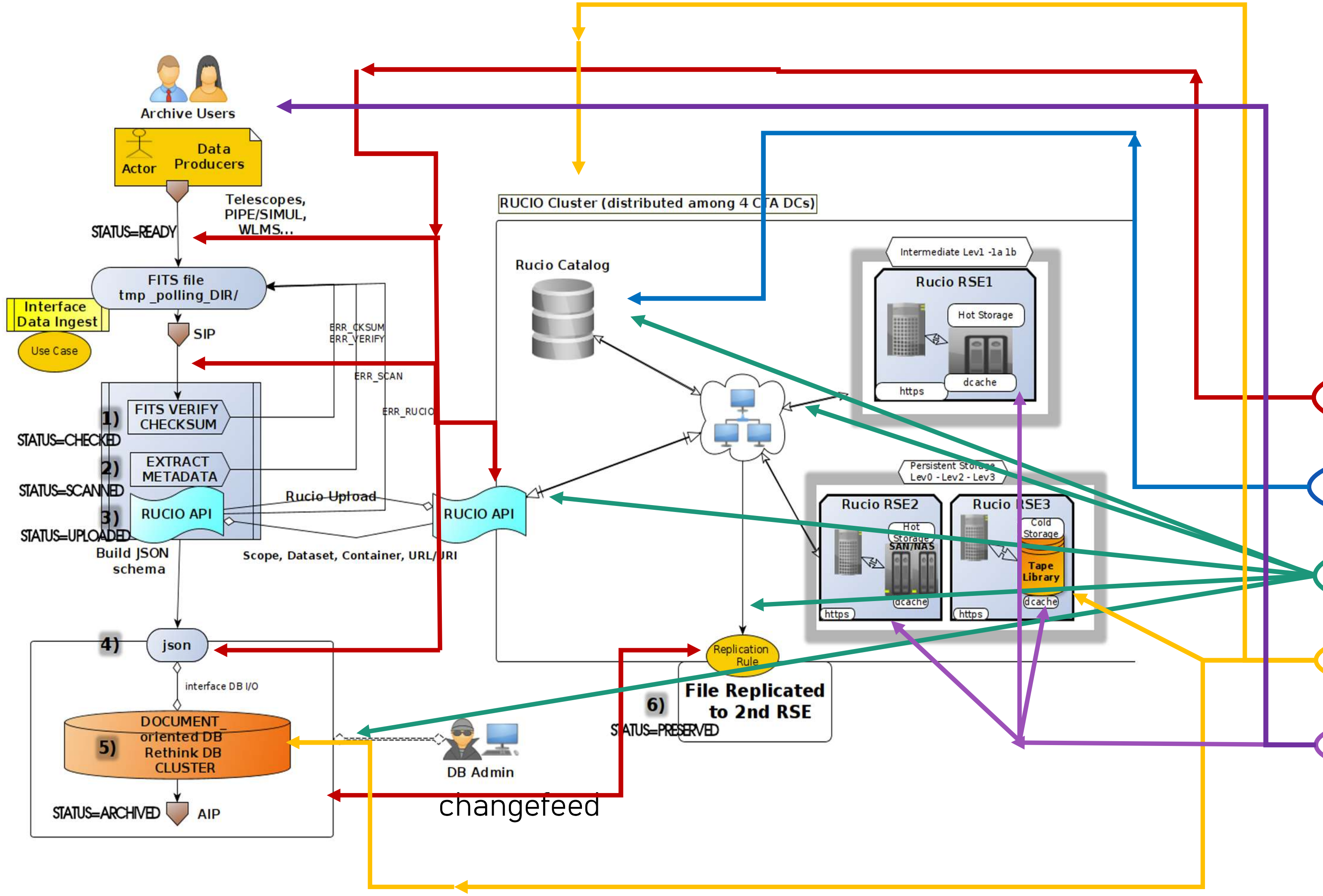
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case INGEST



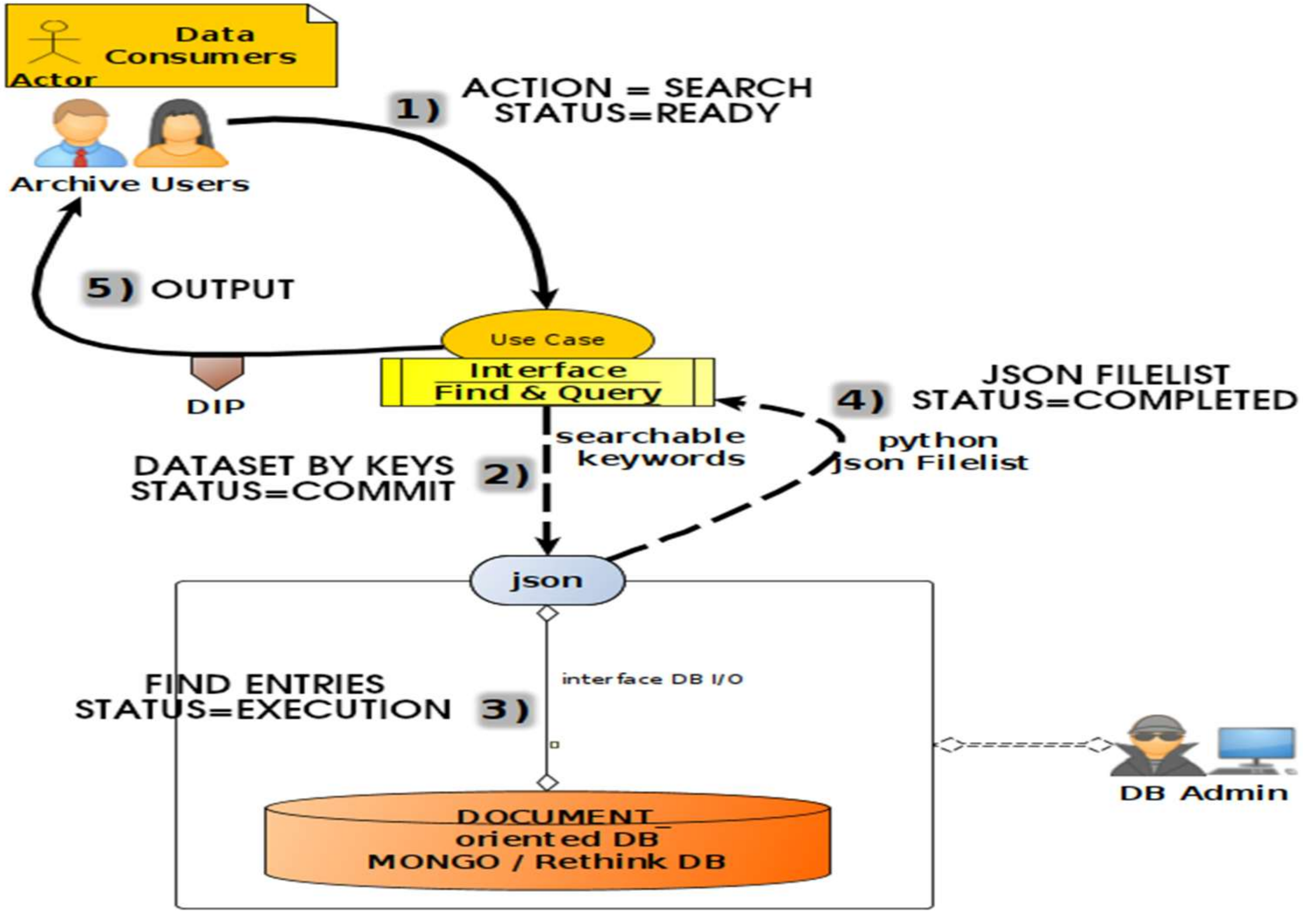
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case INGEST



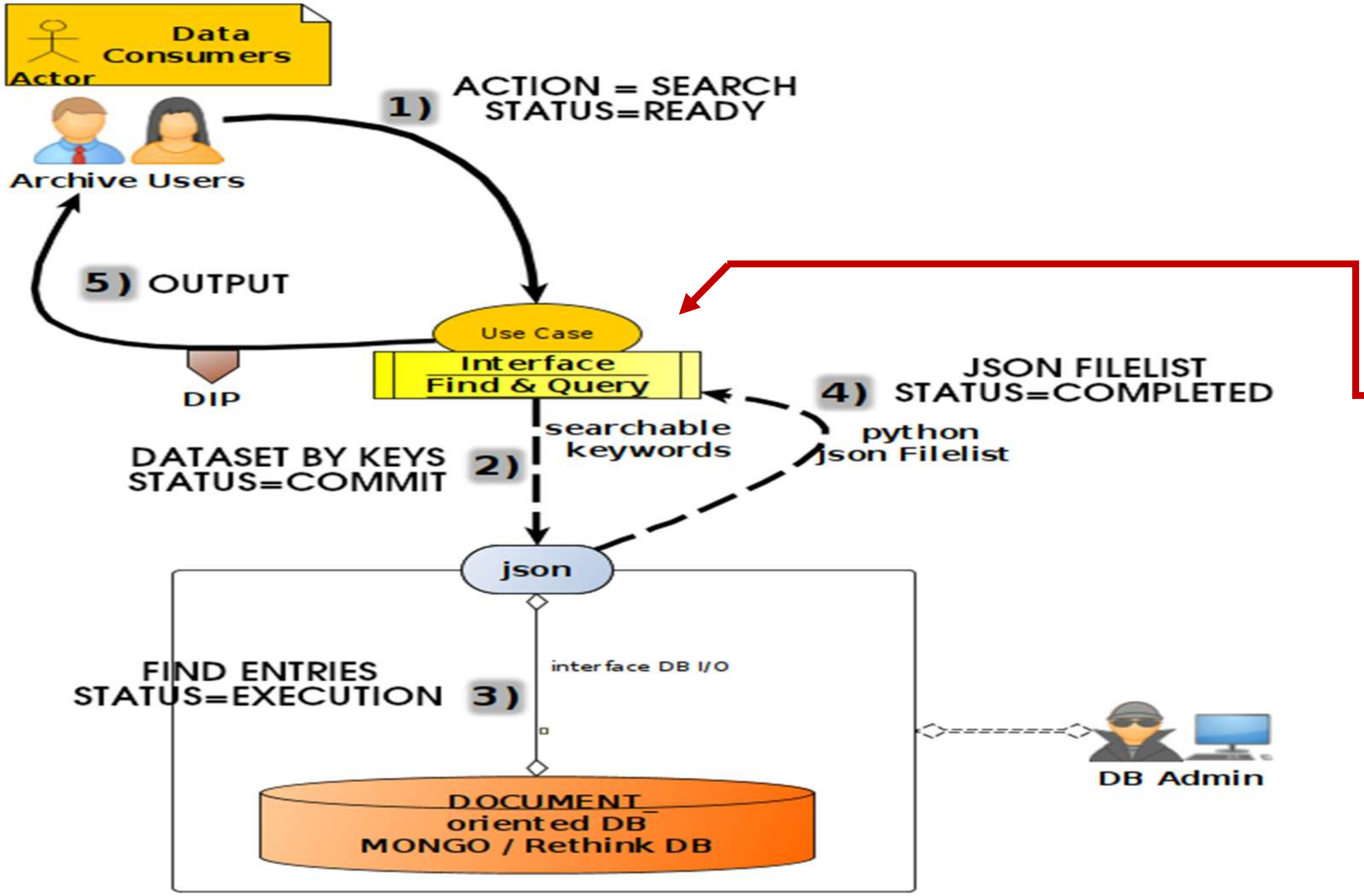
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case SEARCH



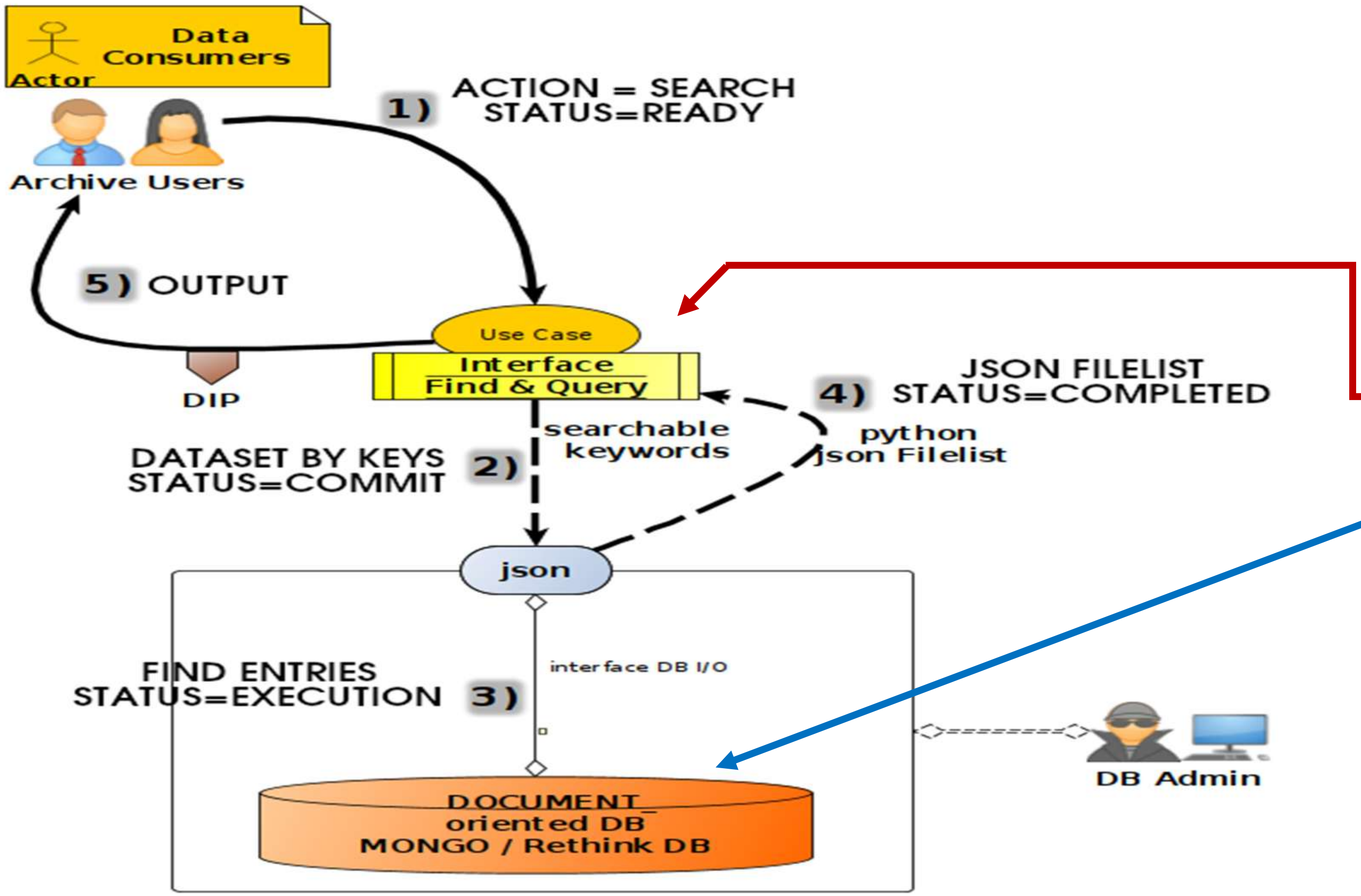
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case SEARCH



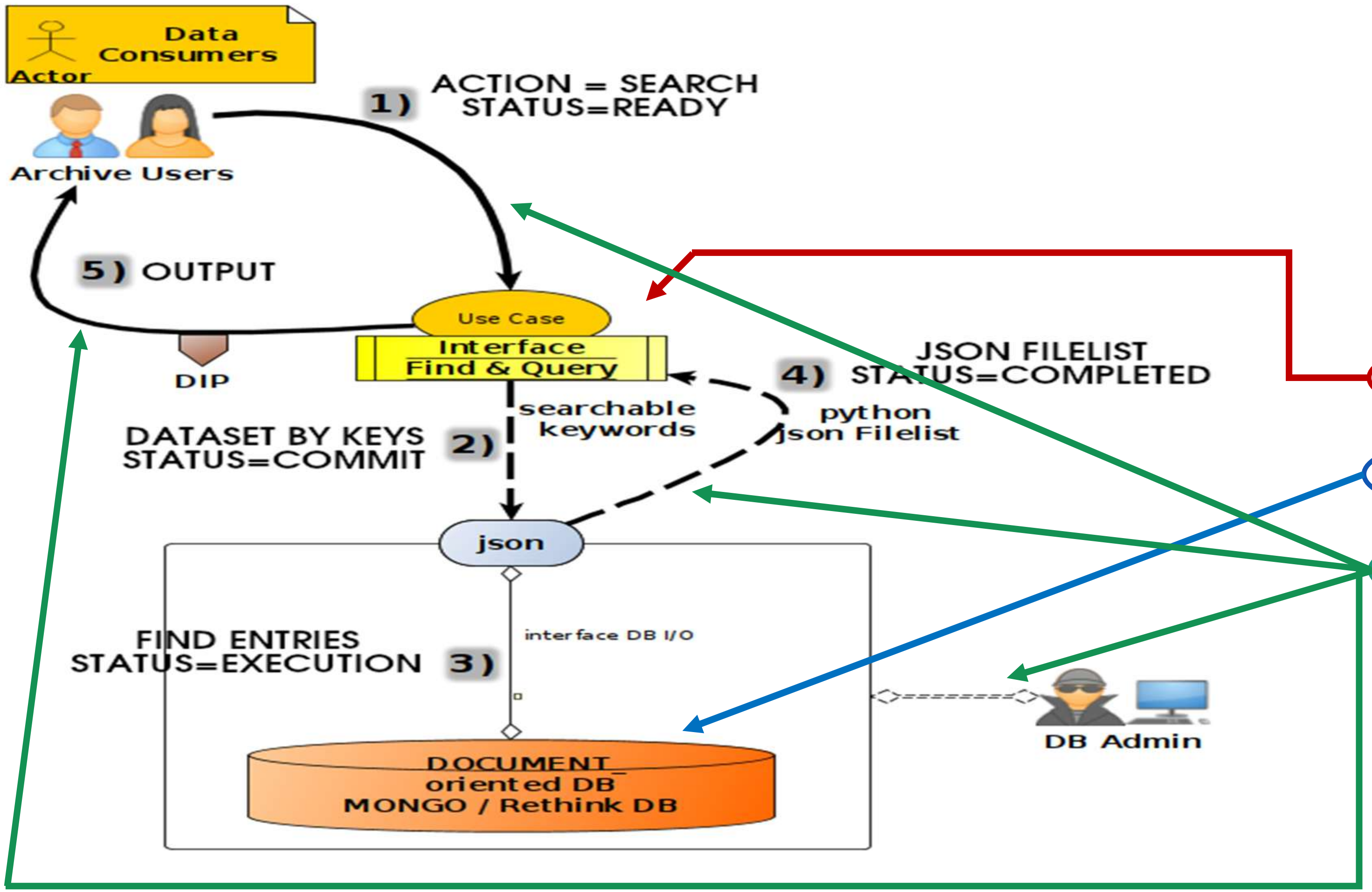
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case SEARCH



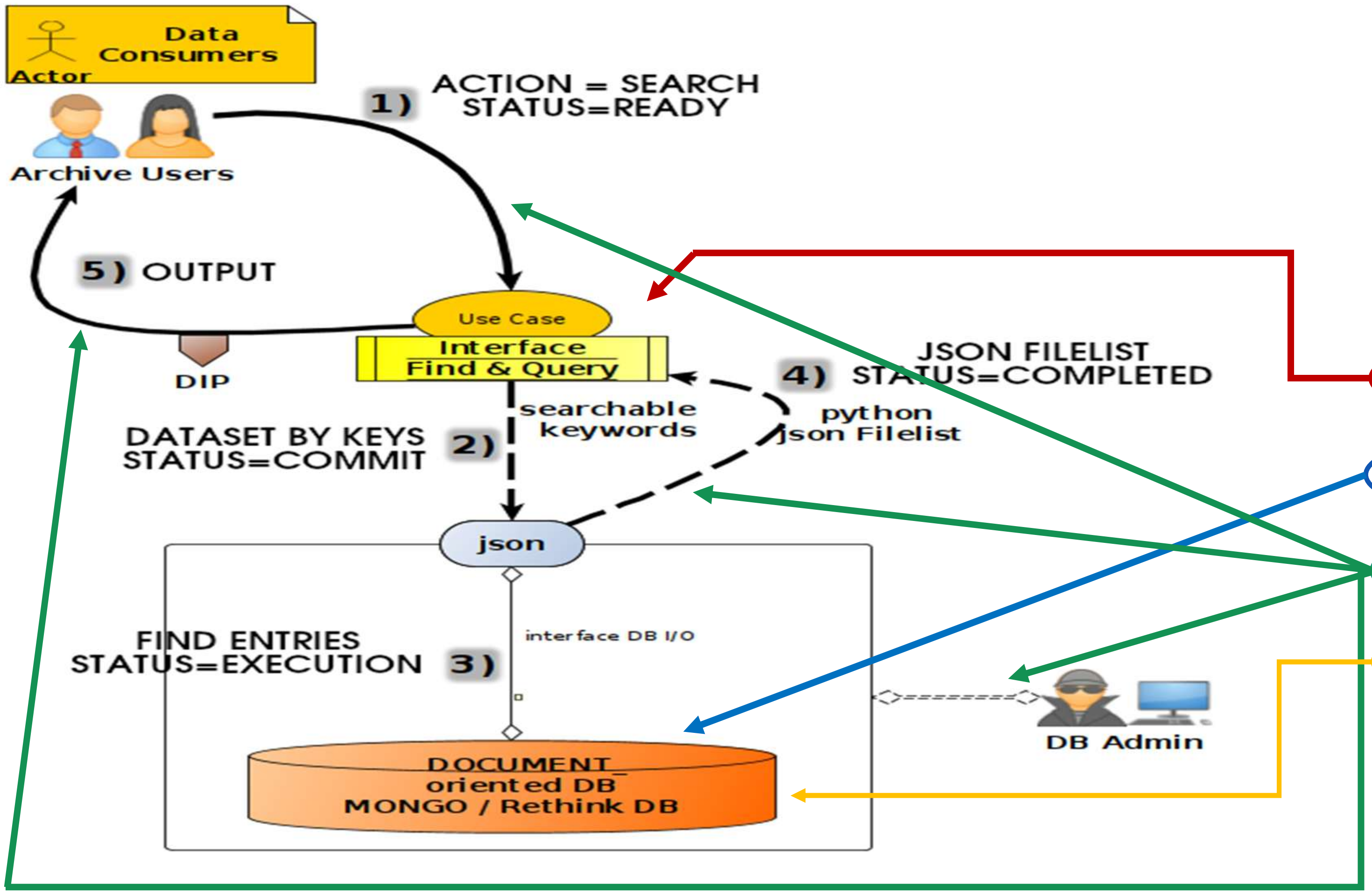
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case SEARCH



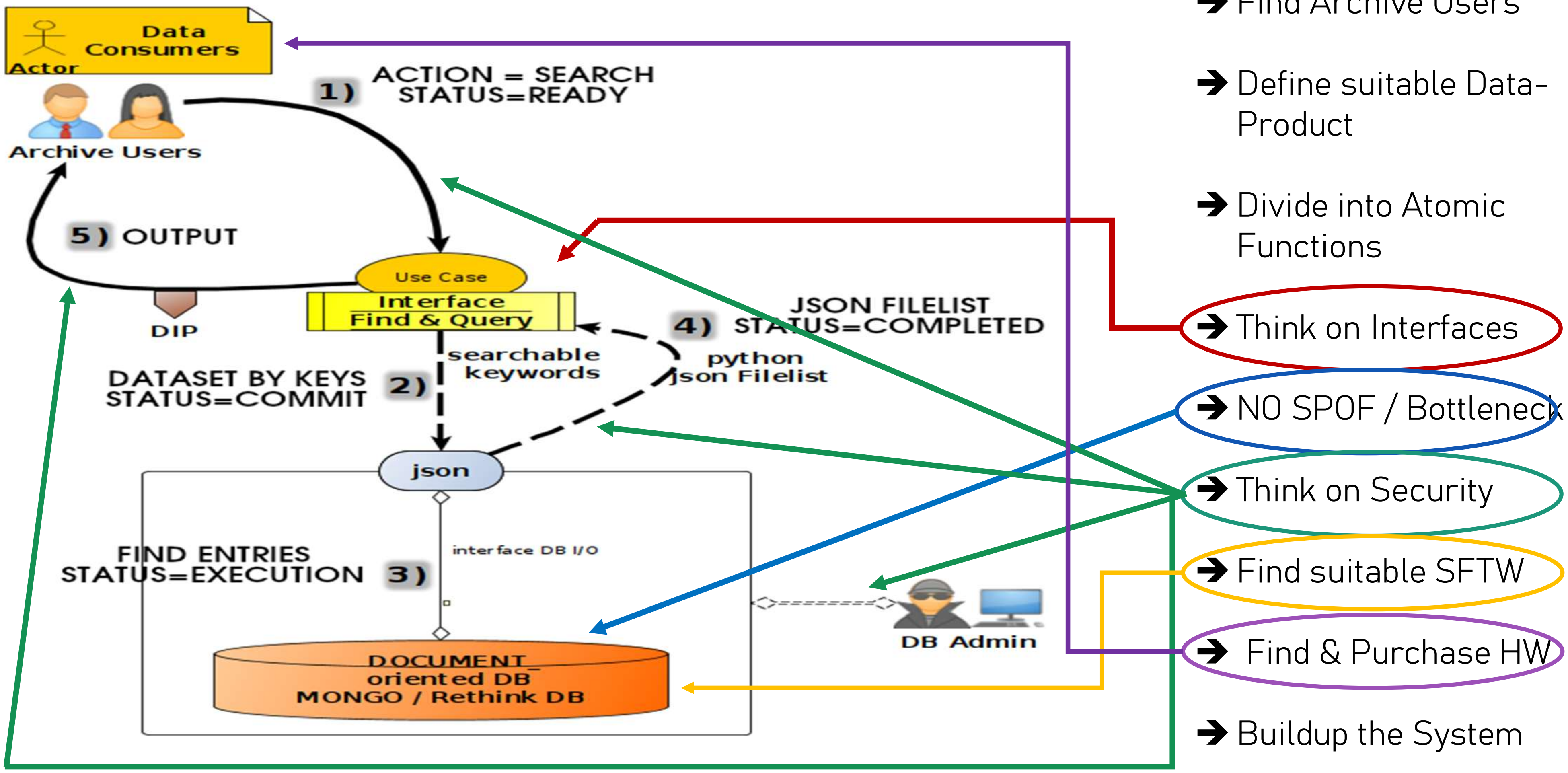
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case SEARCH



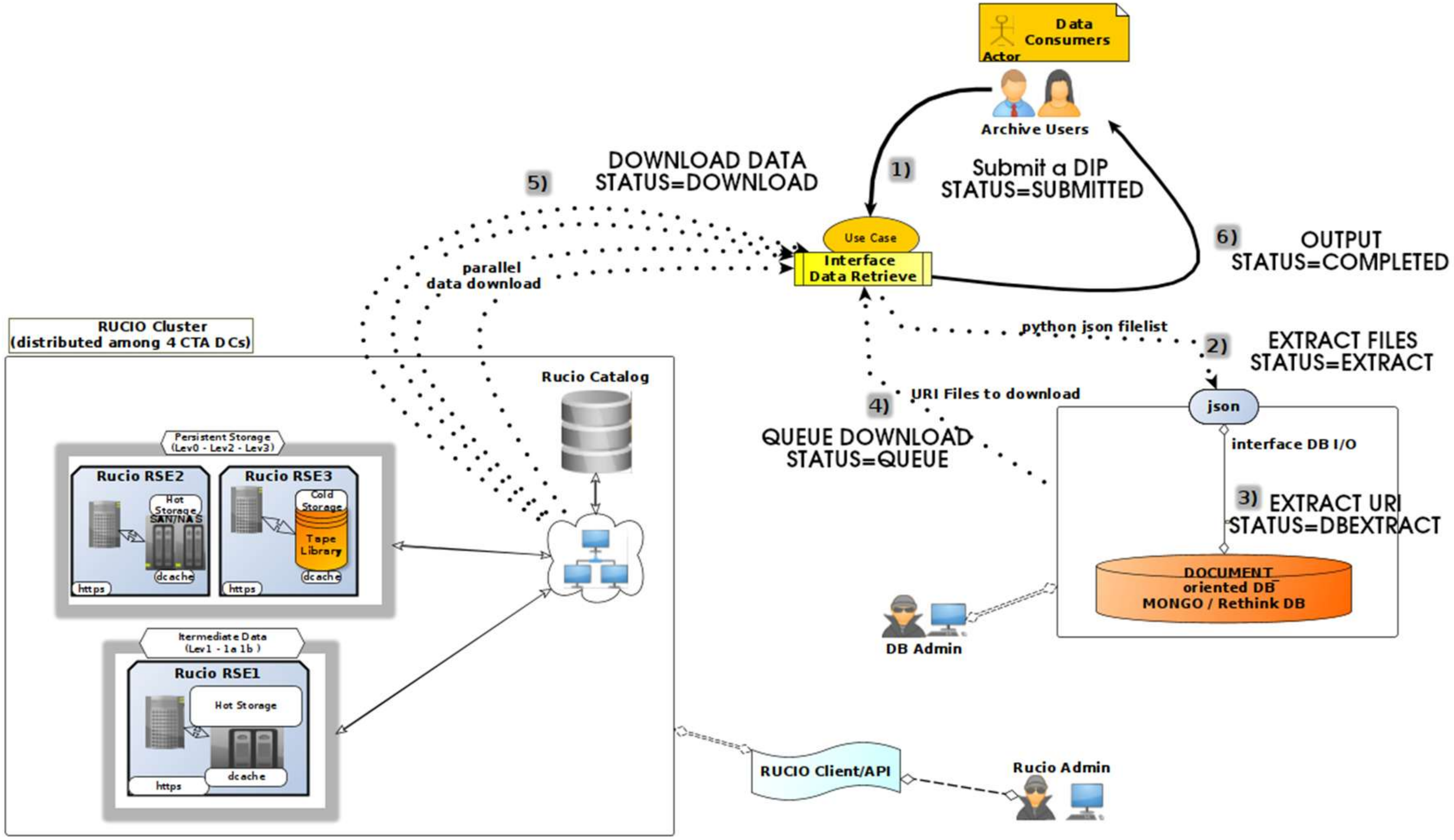
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case SEARCH



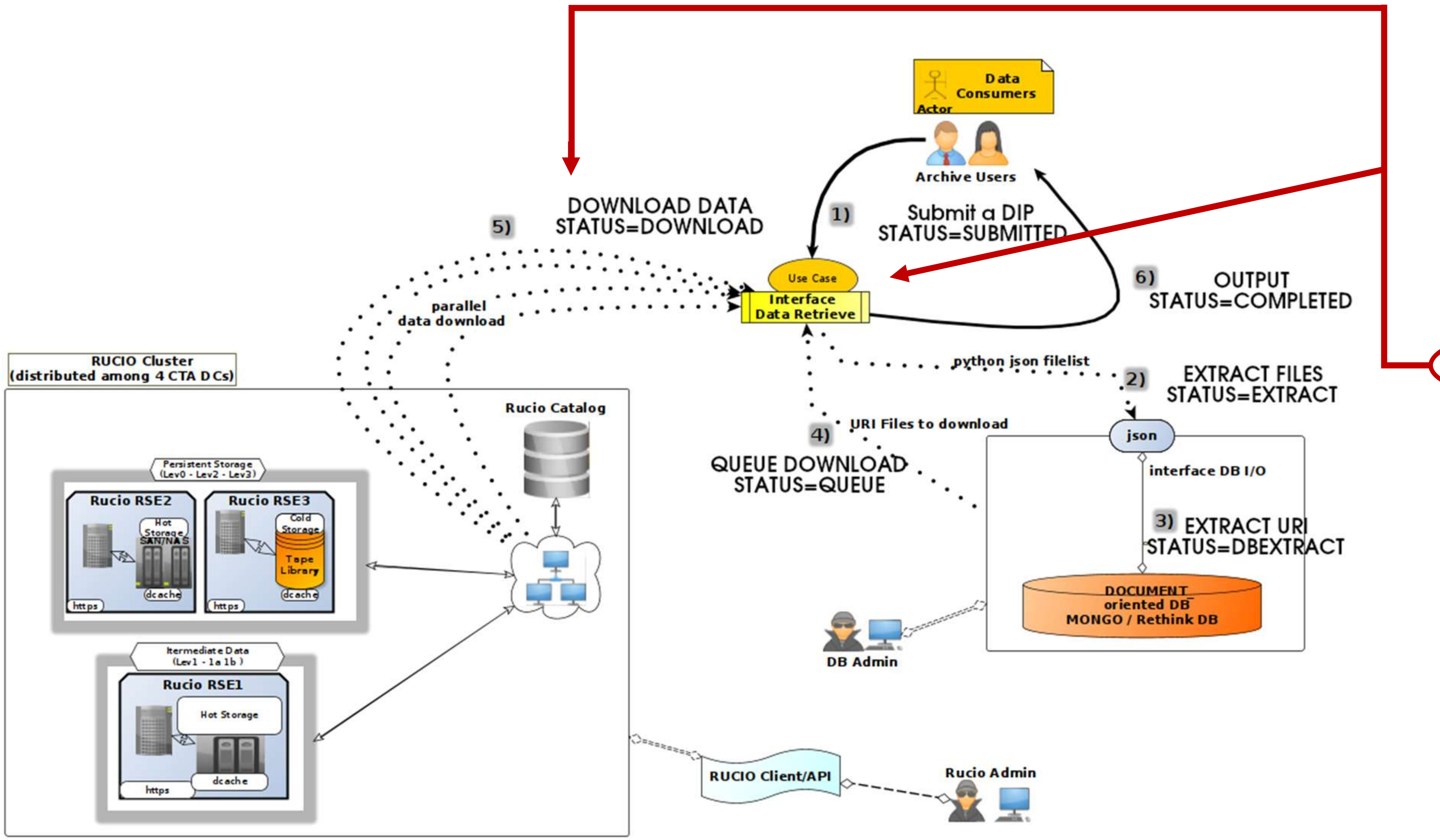
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case RETRIEVE



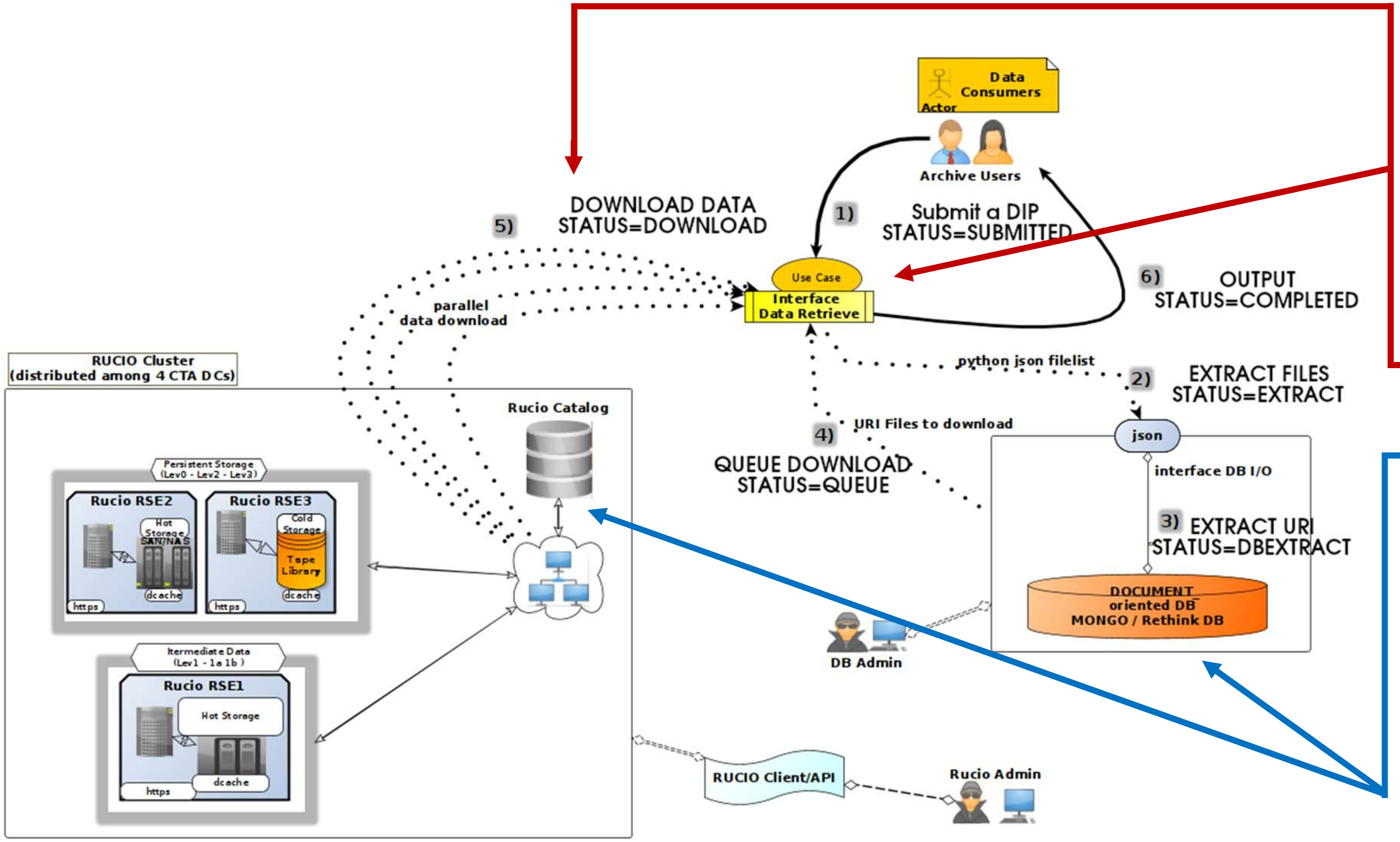
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case RETRIEVE



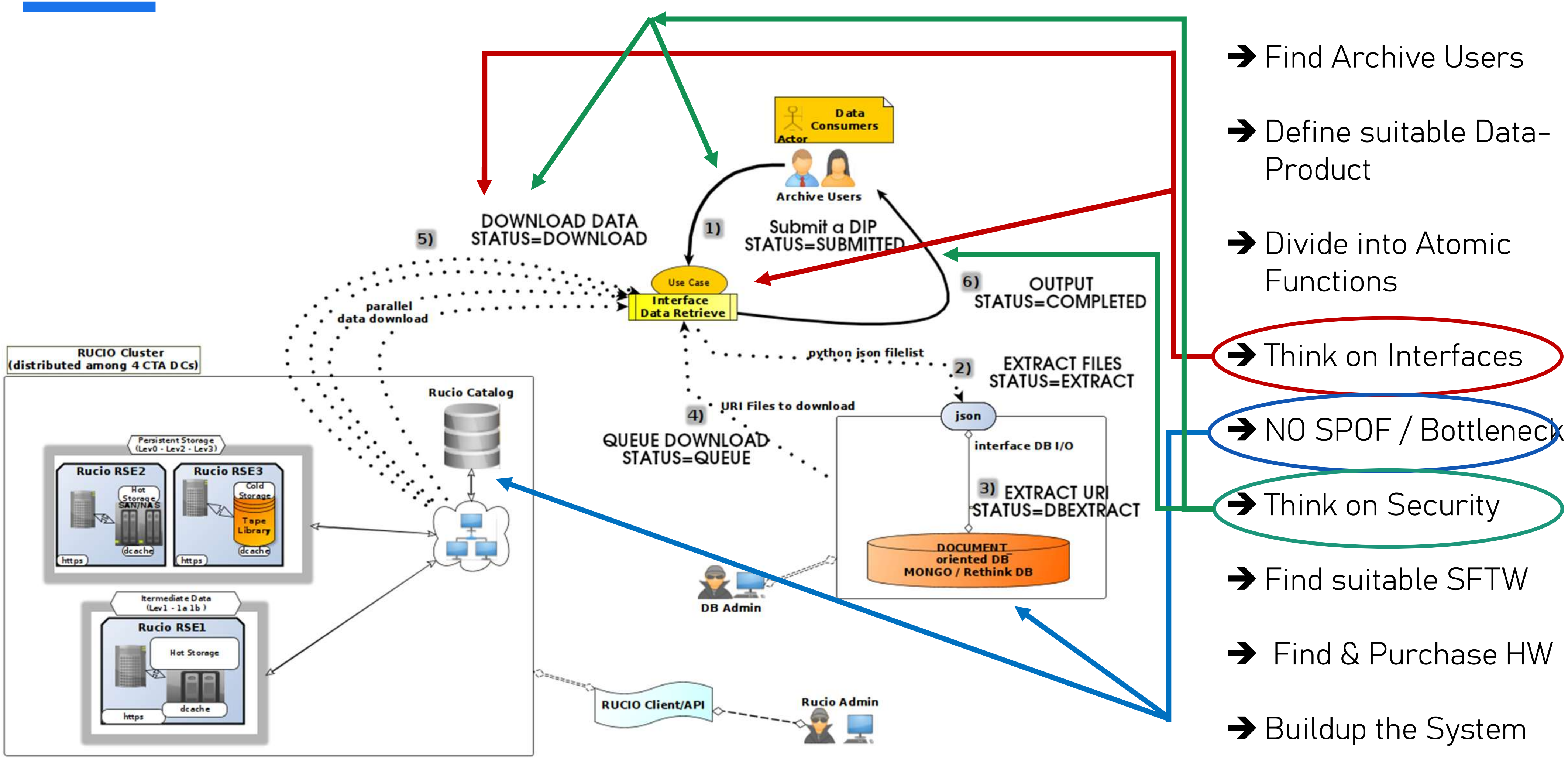
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case RETRIEVE



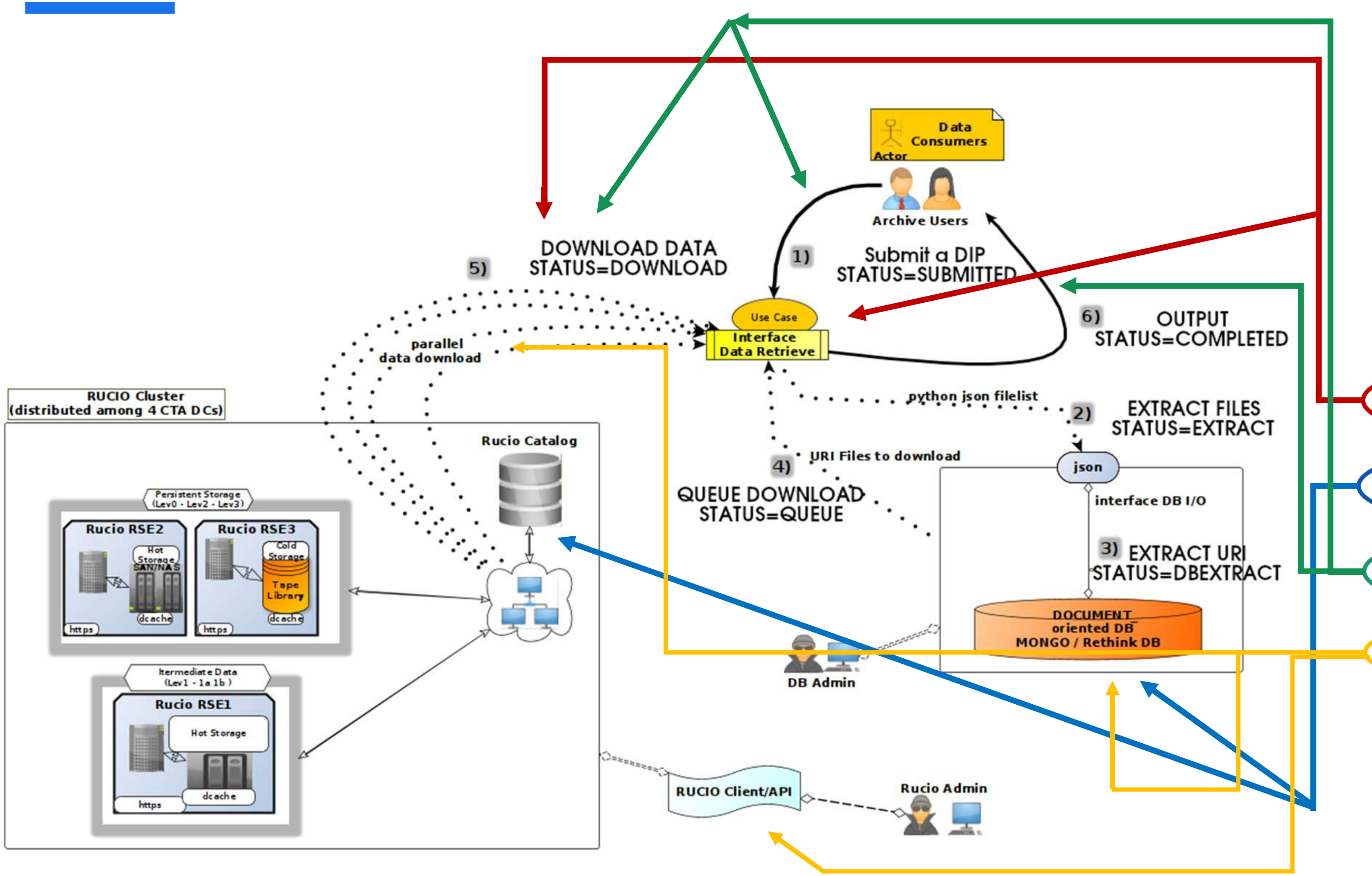
- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case RETRIEVE



- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Analyse Prototypes: Use Case RETRIEVE



➔ Find Archive Users

➔ Define suitable Data-Product

➔ Divide into Atomic Functions

➔ Think on Interfaces

➔ NO SPOF / Bottleneck

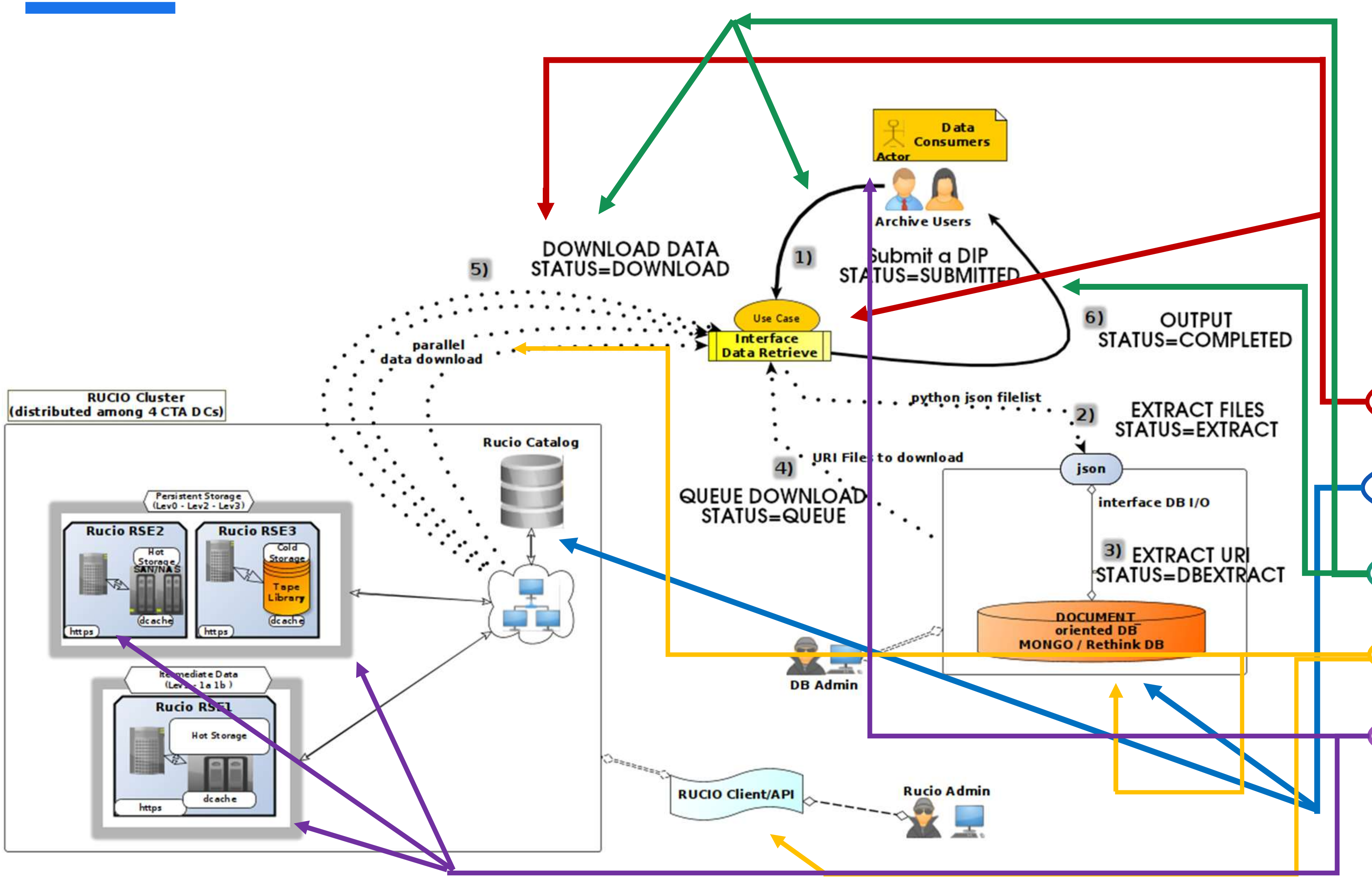
➔ Think on Security

➔ Find suitable SFTW

➔ Find & Purchase HW

➔ Buildup the System

Analyse Prototypes: Use Case RETRIEVE

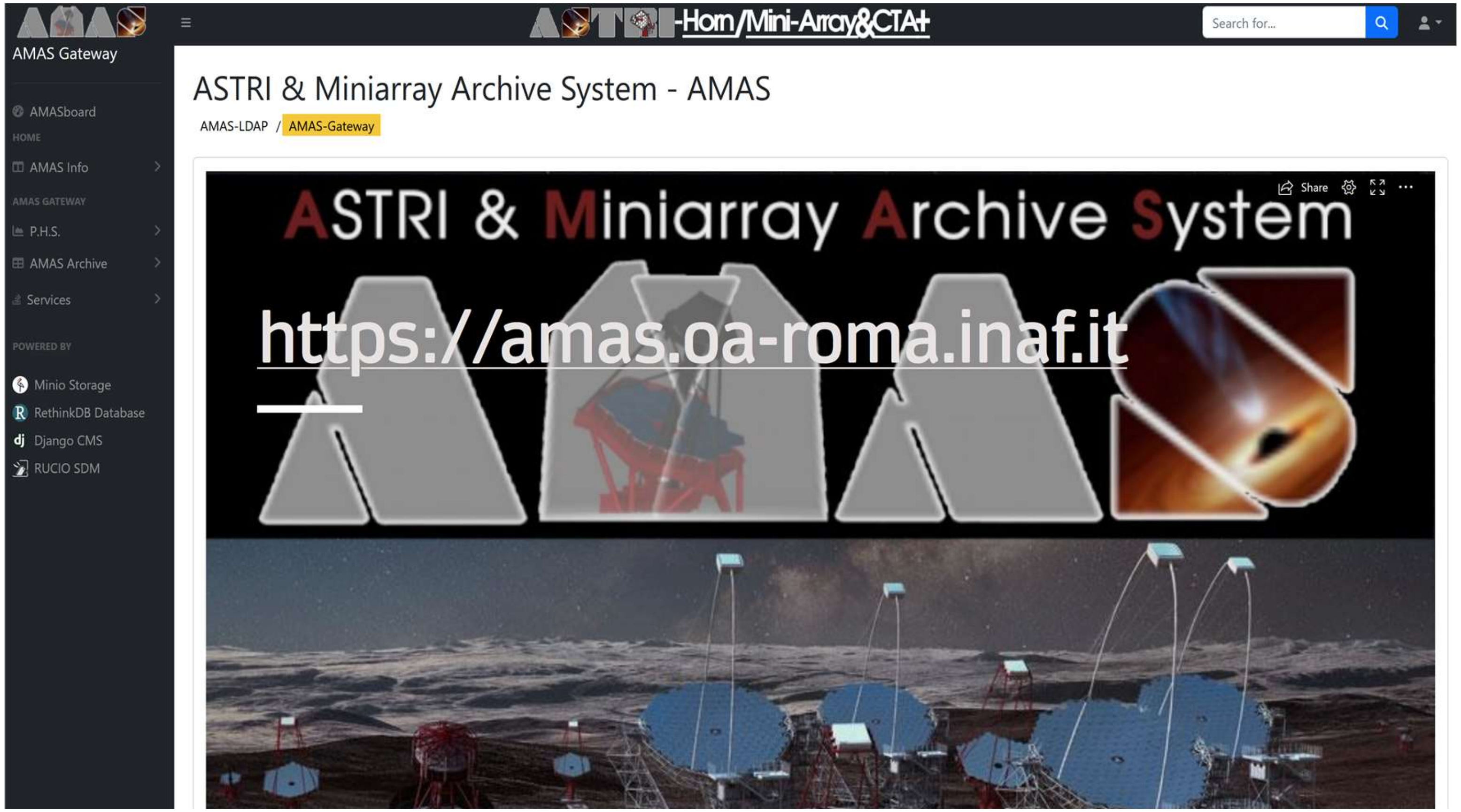


- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ NO SPOF / Bottleneck
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Purchase Hardware to match Requirements

- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ Avoid SPOF
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Put Everything Together and Build-up a Protoptype System



ASTRI & Miniarray Archive System - AMAS

ASTRI & Miniarray Archive System

https://amas.oa-roma.inaf.it

- ➔ Find Archive Users
- ➔ Define suitable Data-Product
- ➔ Divide into Atomic Functions
- ➔ Think on Interfaces
- ➔ Avoid SPOF
- ➔ Think on Security
- ➔ Find suitable SFTW
- ➔ Find & Purchase HW
- ➔ Buildup the System

Best Practices
_vs_ Real Life

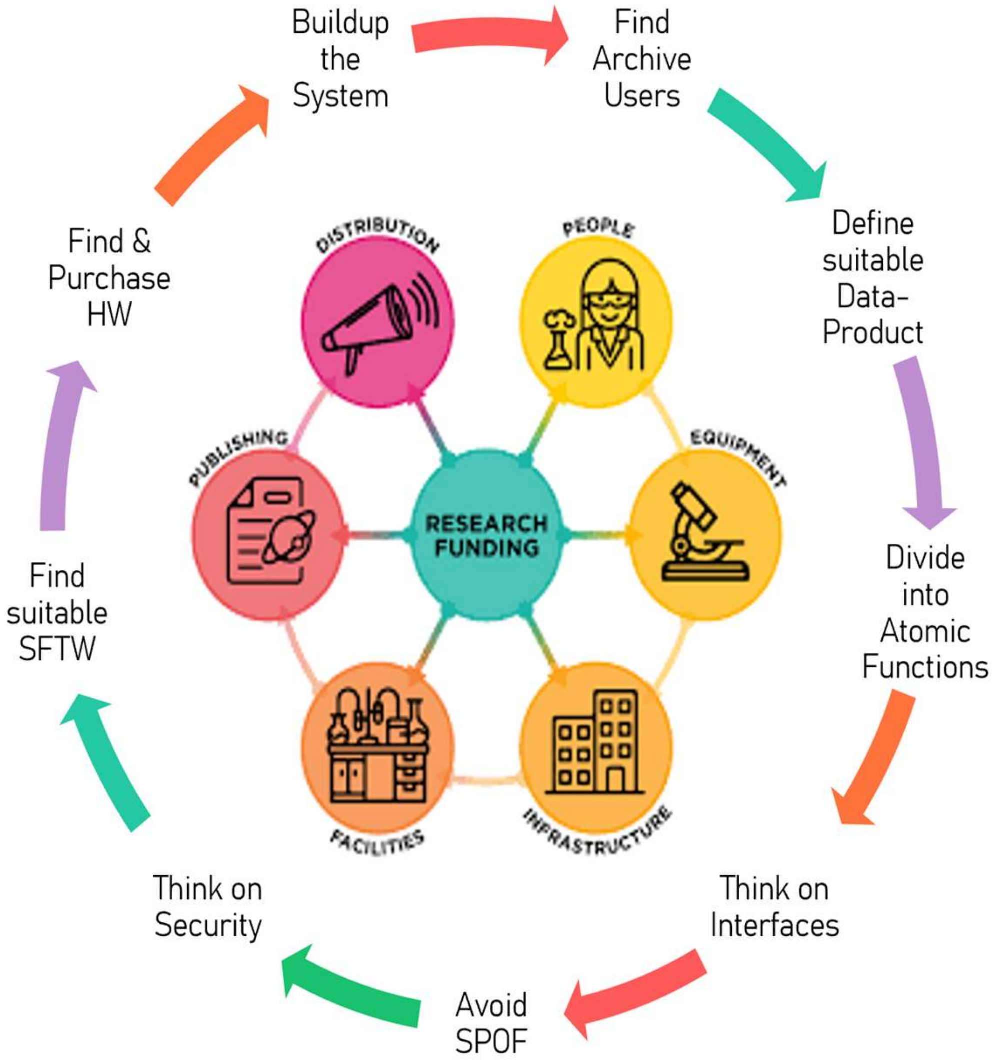
Scientific ideas are usually limited by the cost of a Scientific Archive

The Hardware Procurement is driven by the Scientific Project Funds

It is common to build a system starting from an hardware scratch.

Use Cases and Requirement satisfaction is demanded by Virtual configurations: “as a Service”

New Common Paradigm -> IaaS, Paas, SaaS, AaaS.



Buildup the System

Find Archive Users

Define suitable Data-Product

Divide into Atomic Functions

Think on Interfaces

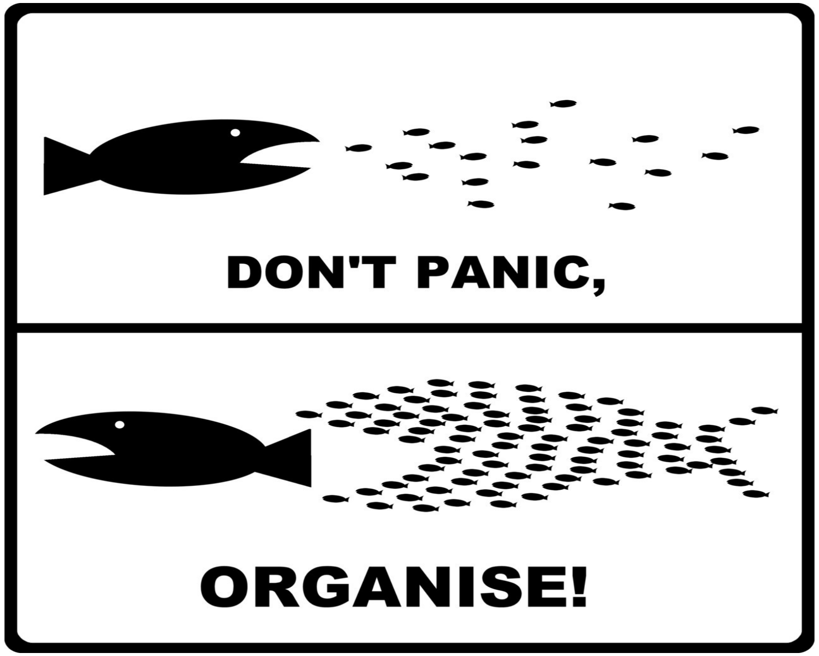
Avoid SPOF

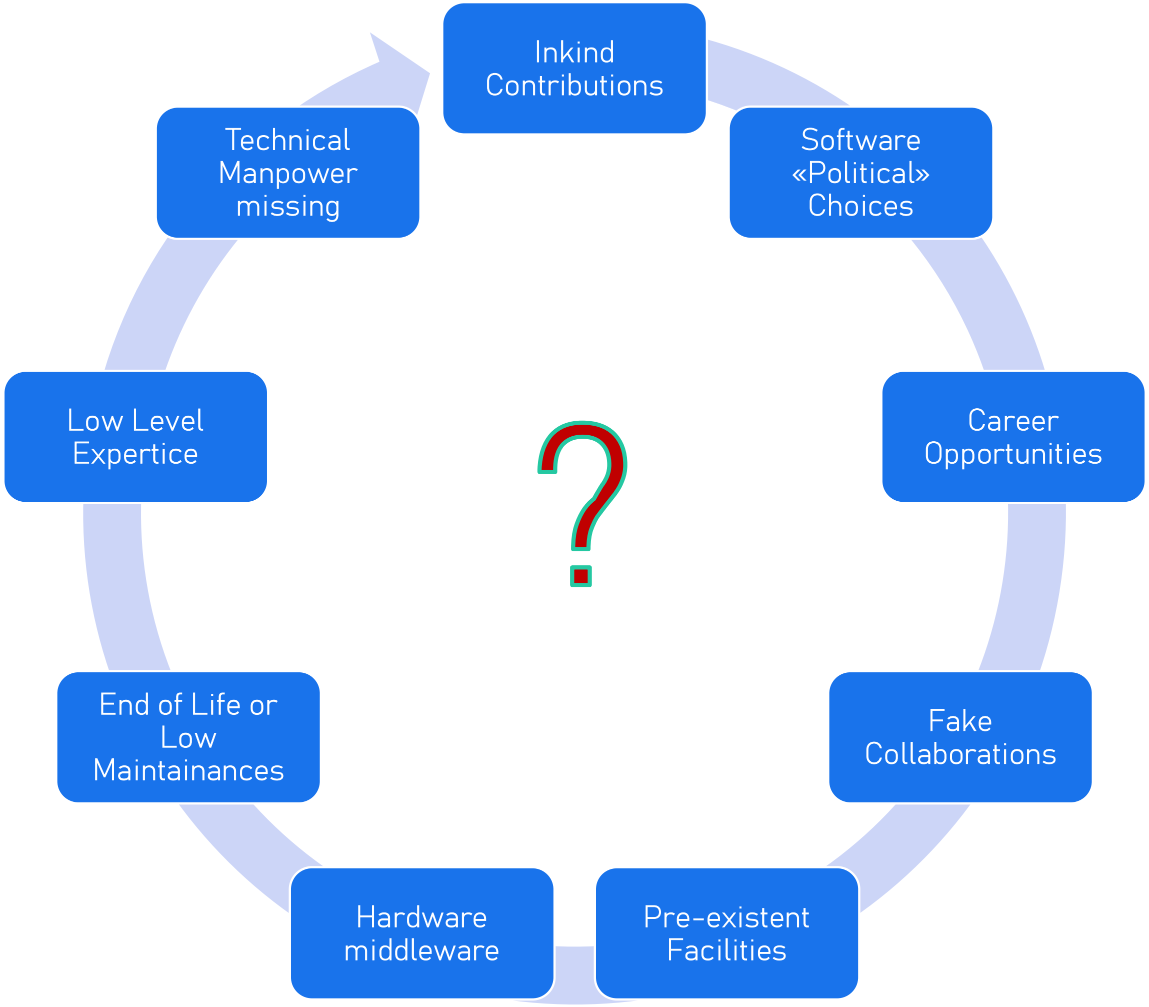
Think on Security

Find suitable SFTW

Find & Purchase HW

Problem Making and Problem Solving in Big Projects





Dedicated Mailing List

for

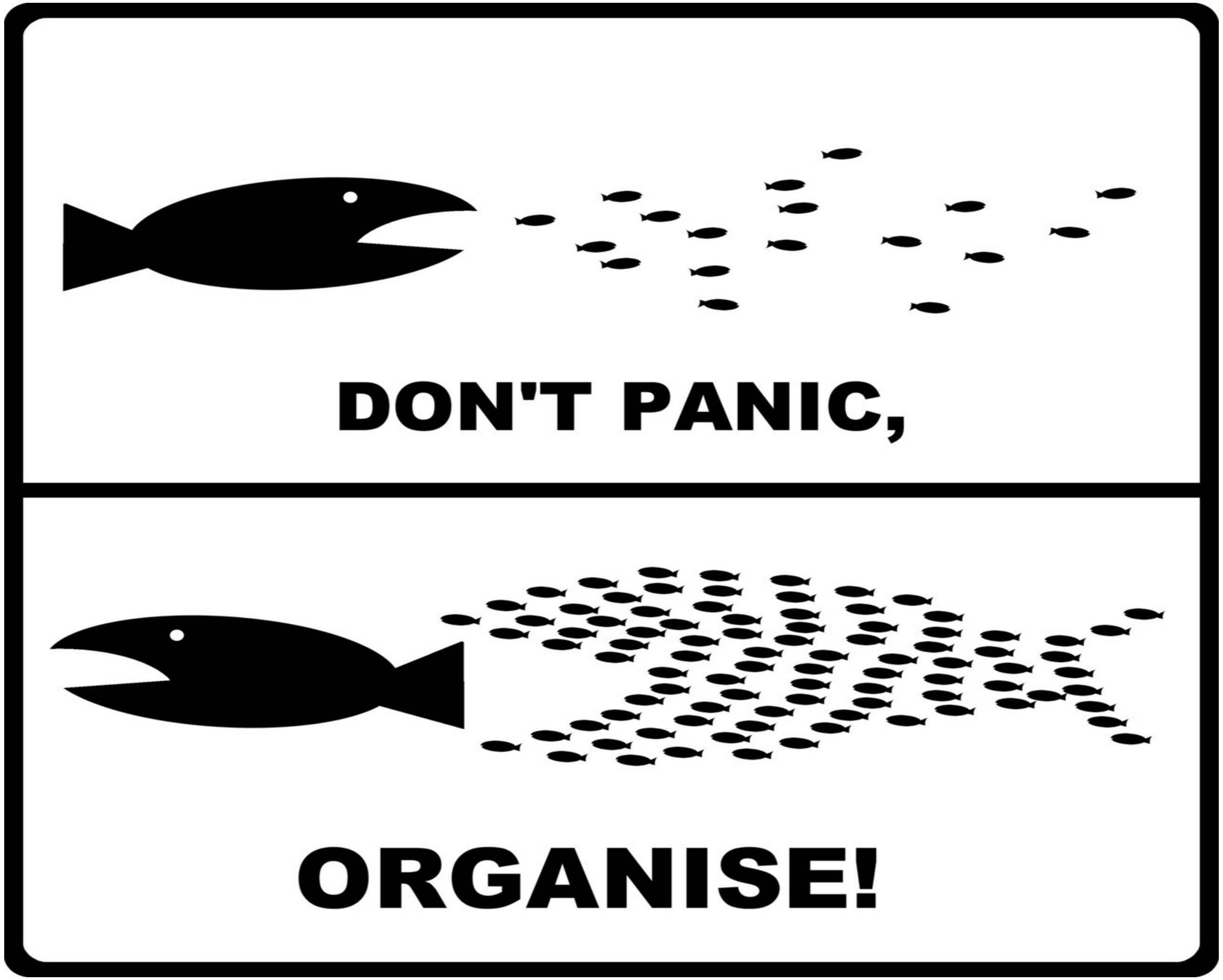
DataManagement Systems

of USC-C

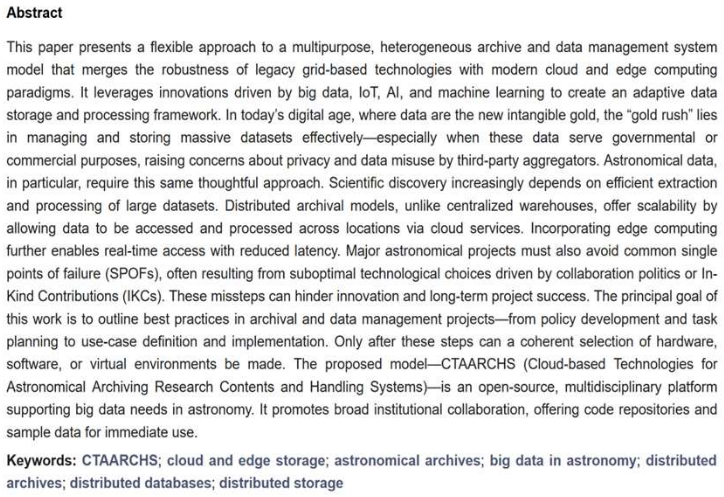
dms_uscc@inaf.it

(previous “dms_usc8”)

Join to share Expertise



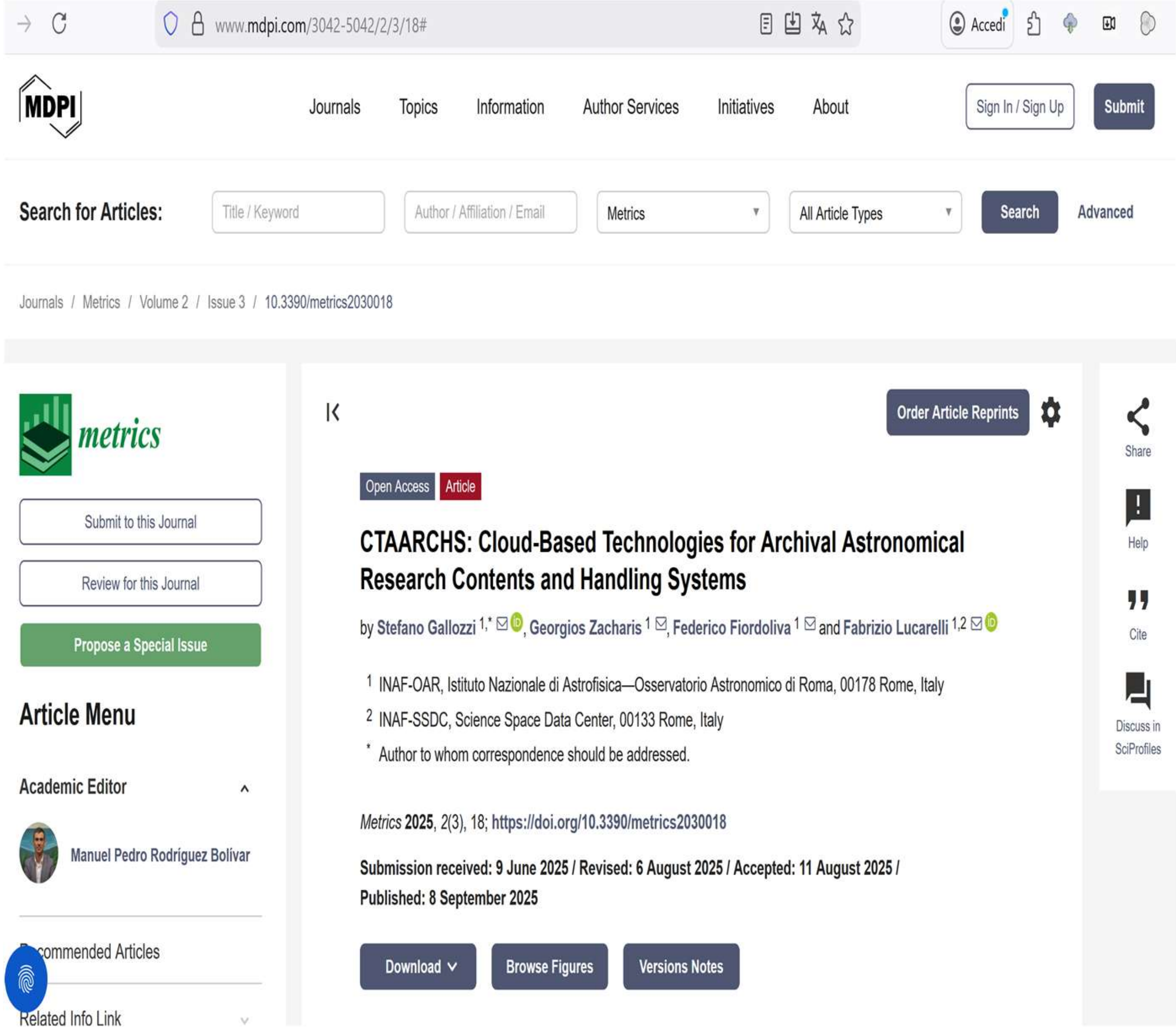
A Summary Compendium



Abstract

This paper presents a flexible approach to a multipurpose, heterogeneous archive and data management system model that merges the robustness of legacy grid-based technologies with modern cloud and edge computing paradigms. It leverages innovations driven by big data, IoT, AI, and machine learning to create an adaptive data storage and processing framework. In today's digital age, where data are the new intangible gold, the "gold rush" lies in managing and storing massive datasets effectively—especially when these data serve governmental or commercial purposes, raising concerns about privacy and data misuse by third-party aggregators. Astronomical data, in particular, require this same thoughtful approach. Scientific discovery increasingly depends on efficient extraction and processing of large datasets. Distributed archival models, unlike centralized warehouses, offer scalability by allowing data to be accessed and processed across locations via cloud services. Incorporating edge computing further enables real-time access with reduced latency. Major astronomical projects must also avoid common single points of failure (SPOFs), often resulting from suboptimal technological choices driven by collaboration politics or In-Kind Contributions (IKCs). These missteps can hinder innovation and long-term project success. The principal goal of this work is to outline best practices in archival and data management projects—from policy development and task planning to use-case definition and implementation. Only after these steps can a coherent selection of hardware, software, or virtual environments be made. The proposed model—CTAARCHS (Cloud-based Technologies for Astronomical Archiving Research Contents and Handling Systems)—is an open-source, multidisciplinary platform supporting big data needs in astronomy. It promotes broad institutional collaboration, offering code repositories and sample data for immediate use.

Keywords: CTAARCHS; cloud and edge storage; astronomical archives; big data in astronomy; distributed archives; distributed databases; distributed storage



CTAARCHS: Cloud-Based Technologies for Archival Astronomical Research Contents and Handling Systems

by Stefano Gallozzi [1,*], Georgios Zacharis [1], Federico Fiordoliva [1] and Fabrizio Lucarelli [1,2]

[1] INAF-OAR, Istituto Nazionale di Astrofisica—Osservatorio Astronomico di Roma, 00178 Rome, Italy

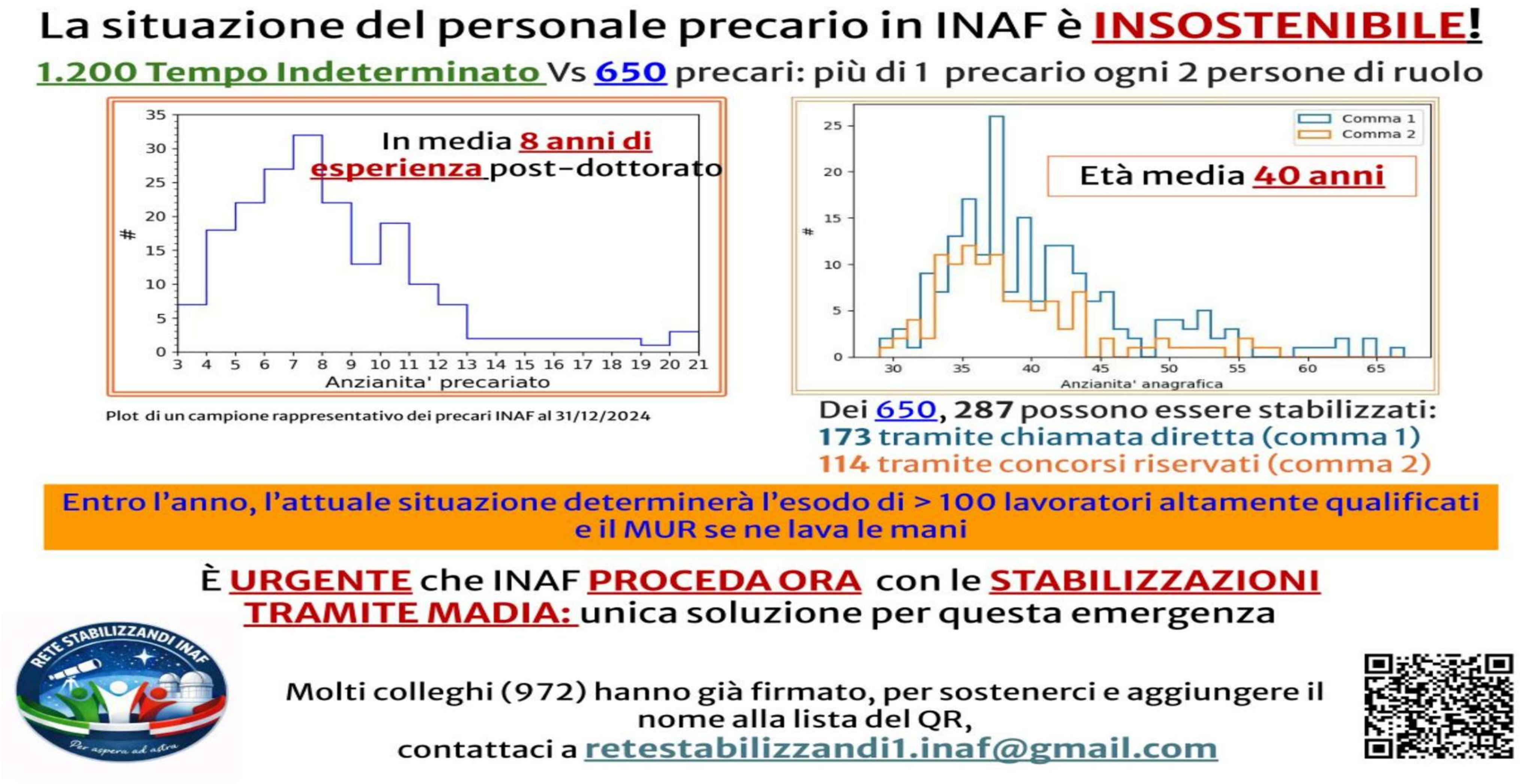
[2] INAF-SSDC, Science Space Data Center, 00133 Rome, Italy

[*] Author to whom correspondence should be addressed.

Metrics 2025, 2(3), 18; https://doi.org/10.3390/metrics2030018

Submission received: 9 June 2025 / Revised: 6 August 2025 / Accepted: 11 August 2025 / Published: 8 September 2025

https://www.mdpi.com/3042-5042/2/3/18#



La situazione del personale precario in INAF è **INSOSTENIBILE!**

1.200 Tempo Indeterminato Vs **650** precari: più di 1 precario ogni 2 persone di ruolo

In media **8 anni di esperienza** post-dottorato

35
30
25
20
15
10
5
0
#
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Anzianita' precariato

Plot di un campione rappresentativo dei precari INAF al 31/12/2024

Età media **40 anni**

Comma 1
Comma 2
25
20
15
10
5
0
#
30 35 40 45 50 55 60 65
Anzianita' anagrafica

Dei **650**, **287** possono essere stabilizzati:
173 tramite chiamata diretta (comma 1)
114 tramite concorsi riservati (comma 2)

Entro l'anno, l'attuale situazione determinerà l'esodo di > 100 lavoratori altamente qualificati e il MUR se ne lava le mani

È **URGENTE** che INAF **PROCEDA ORA** con le **STABILIZZAZIONI TRAMITE MADIA:** unica soluzione per questa emergenza

Molti colleghi (972) hanno già firmato, per sostenerci e aggiungere il nome alla lista del QR,
contattaci a retestabilizzandi1.inaf@gmail.com

Thanks for your attention!