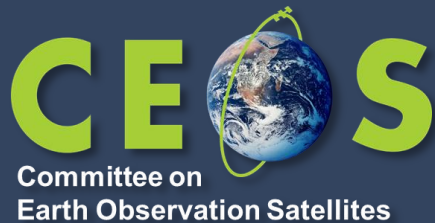


Study for Big Data



Yousuke Ikehata, JAXA
Agenda Item 7.2
WGISS-61
16-20 March, 2026
Dehradun, India

Big Data

What does this mean for WGISS?



Robert Fletcher, UKSA
Agenda Item 7.2
WGISS-60
13-17 October, 2025
Oberpfaffenhofen, Germany



Committee on
Earth Observation Satellites

Relevance to CEOS WGISS



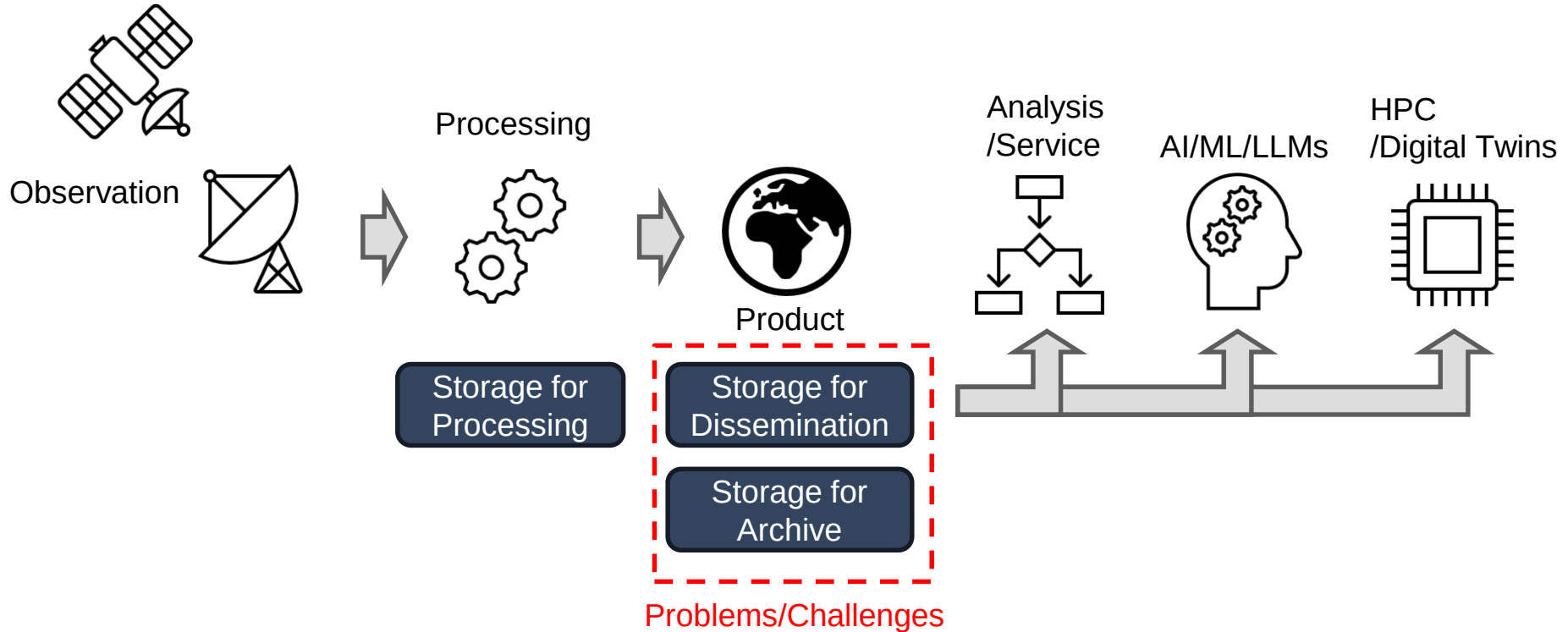
- The five Vs align closely with the CEOS agencies data storage scenarios
- EO data generates some of the worlds largest data sets, 100s of PB and increasing
- We have knowledge of managing Big Data, but how much do 'we' work with it?
- Big Data is closely aligned with AI and ML technologies
- How can WGISS Agencies share our combined expertise across the wider sector?
 - Should we include dedicated Big Data sections within the IHB & AI/ML white papers?
 - Do we need to consider High Performance Computing as a processing engine?
 - Do we include Big Data as a dedicated topic within an existing sub working group?
 - Could we develop a CEOS lessons learnt on working with Big Data?
 - Obvious links into Digital Twins and their utilisation of Data Lakes

WGISS-60, 13-17 October, 2025

Slide 8

- ❖ The EO data is growing exponentially.
 - Already surpassed over 100s PBs
 - Cloud storage is good for keeping scalability
 - **Bring processing to data**
- ❖ Big data analysis including AI/ML is also popular.
 - Especially, Foundation Models get attentions rapidly.
 - **Bring Data to AI** (= Move Data to Compute)
- ❖ Data provider : Egress & Duplication
- ❖ Data consumer : IO Tuning

Types of storage



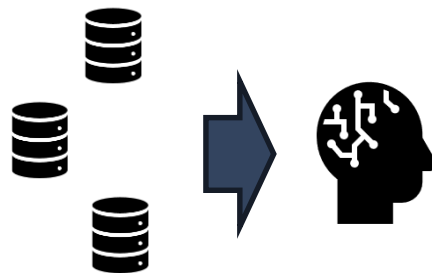
- ❖ Background: On premises -> Cloud
 - Bring users to the data
 - Reduce the need for large data transfers
 - Take advantage of cloud computing
- ❖ However
 - New usage :AI/ML and Digital Twins
 - High speed access
 - Huge data egress
 - Platform
 - Vendor lock-in – hard to migrate
 - Siloed data access
 - Dispersed user base

- ❖ Current situation:
 - Cloud and on premises are used.
 - Scalability: Cloud > On premises
- ❖ Problem
 - (On premises) Scalability
 - Over 100s PB scale
 - (Cloud) Hard to migrate
 - Vendor lock-in

- ❖ Models & Codes has compatibility across the different cloud platforms and AI environments.
 - Bring AI to the Data
 - Federated AI
- ❖ EO data providers should have coordinate common metadata and APIs based on FAIR principles.
 - STAC, WMS/WMTS, EOAPI

❖ Centralized AI

- Bring Data to the AI
- Collect data



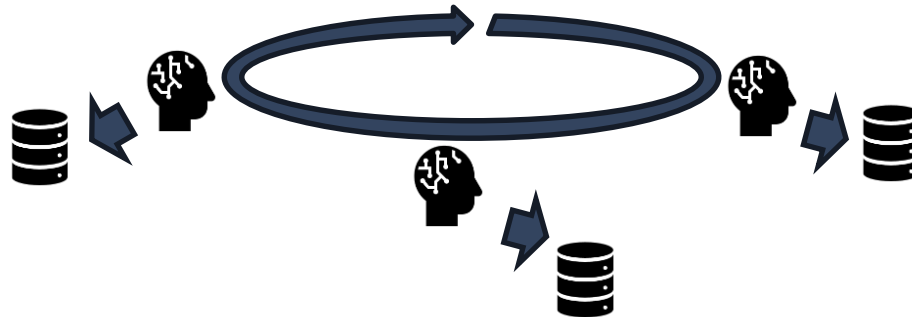
❖ Data centric AI

- Bring AI to the Data.
- Learning is localized
- AI portability (Duplicate AI infra.)



❖ Federated AI

- Federated learning.
- Federated analytics/inference

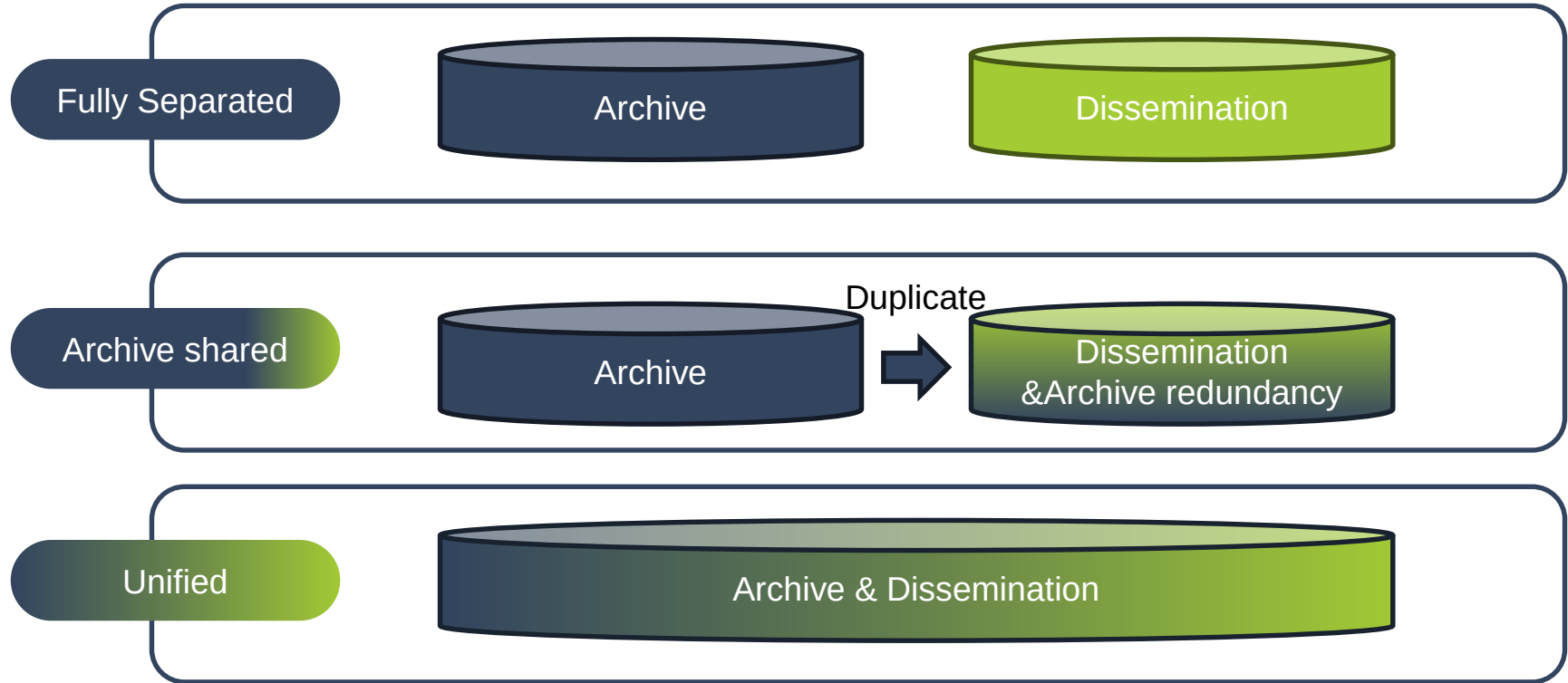


❖ Hybrid infrastructure

- Harmonize on premises & clouds

	Cloud	On premises
Storage	S3, Blob, GCS	Local
Data transfer	Egress fee	Free
Scalability	High	Fixed
Data access	Unfit for random access	No worries
Cost	Good for temporary use	Good for long term use
Operation	Managed	Self manage

Combination



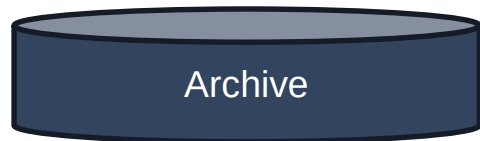
- ❖ “Generally”, data providers might archive 2 or 3 copies.
 - Redundancy
 - Site diversity
 - Media compatibility

Only cloud can provide “Unified”



- ❖ To unify the storage for “Archive & dissemination”, it must have high redundancy.
- ❖ To realize this requirement, only cloud storage can be used.
(99.999999999999%)

On premises & Cloud



On premises

X

On premises

➡ On premises *Case1*

On premises

X

Cloud

➡ Hybrid *Case2*

Cloud

X

On premises

➡ ☹️ ? Unlikely

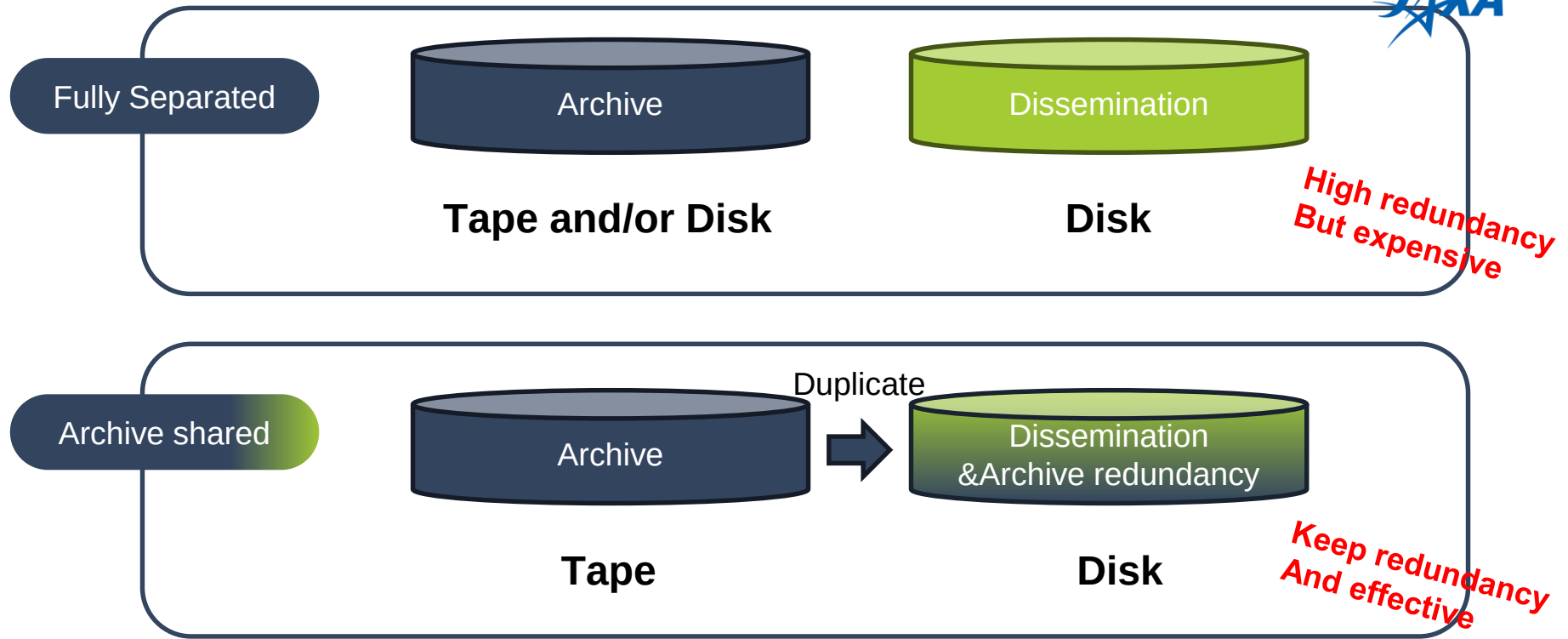
Cloud

X

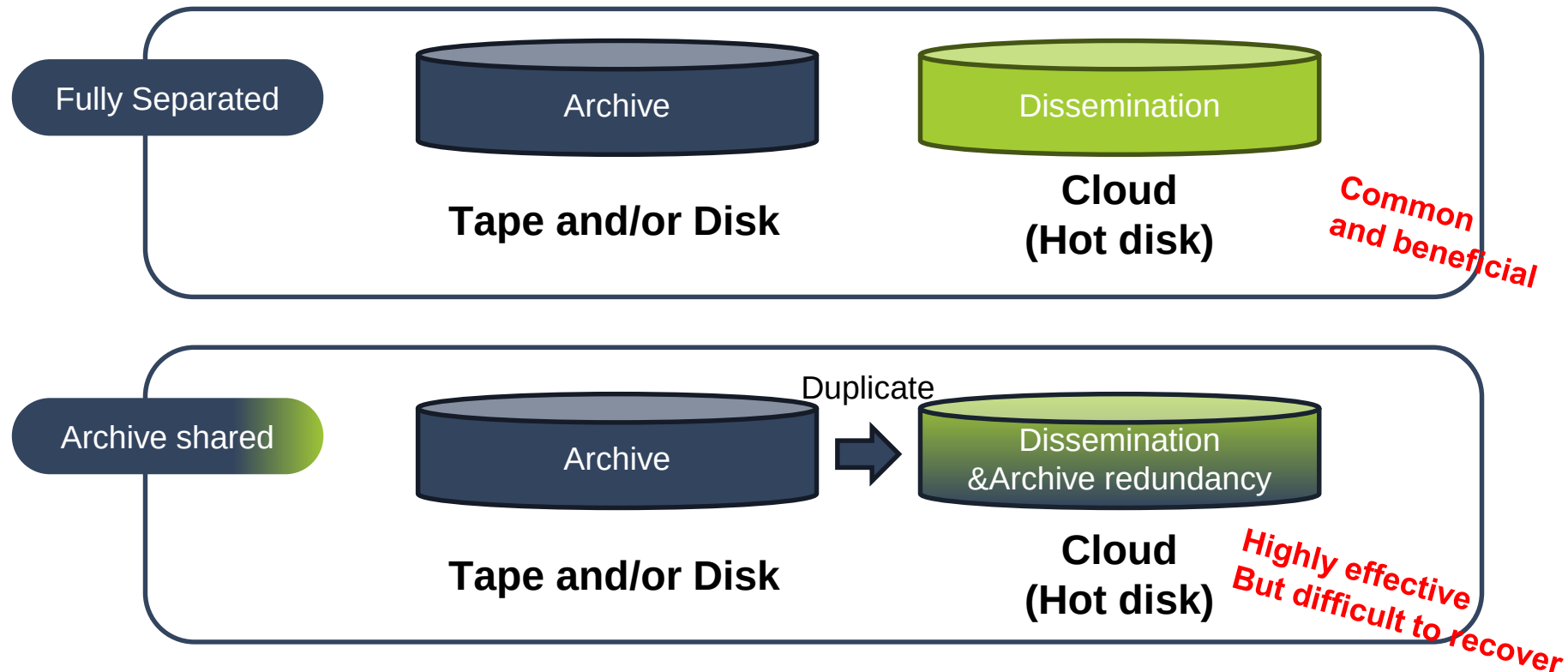
Cloud

➡ Cloud *Case3*

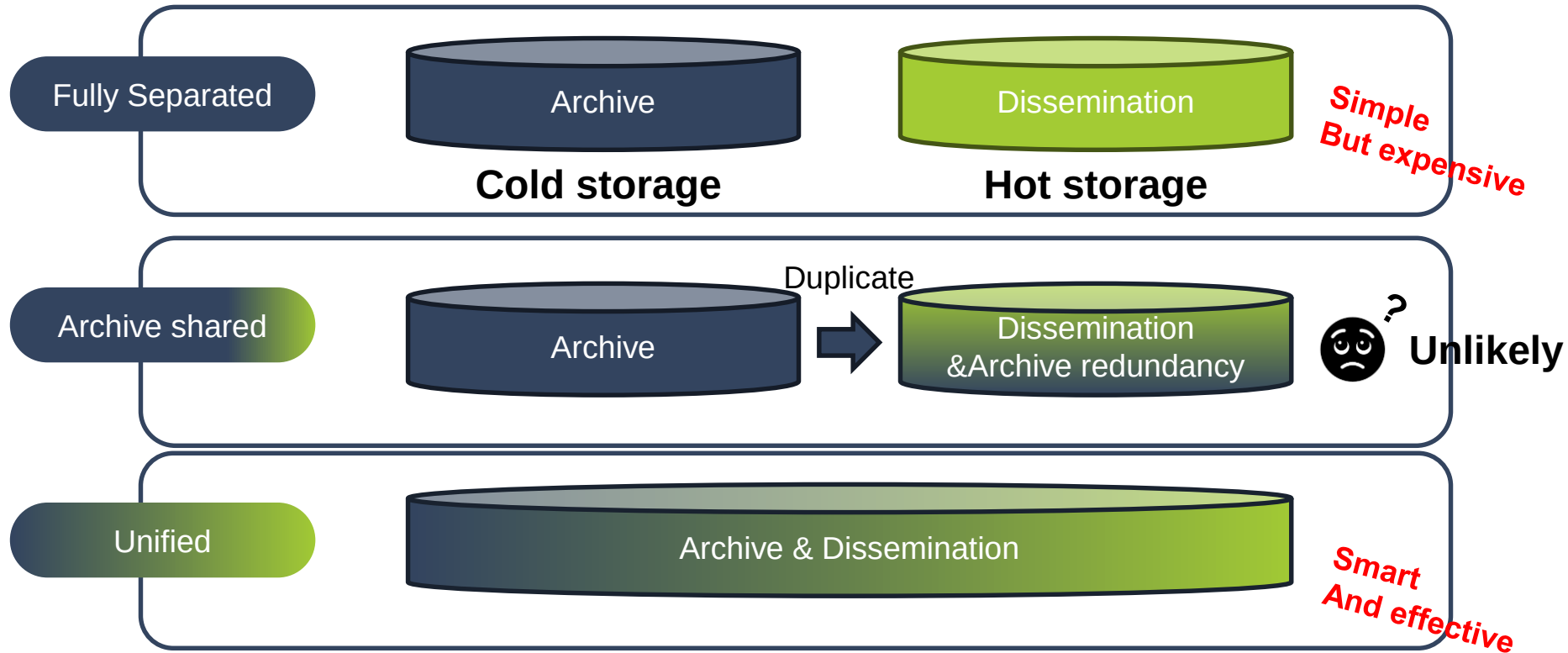
Case 1 : On premises



Case 2: Hybrid



Case 3: Cloud



Case	Combination	Pros.	Cons.
Case 1	Fully separated	High redundancy	Expensive
	Archive shared	Effective cost and reliability	
Case 2	Fully separated	Effective cost and reliability	
	Archive shared	Highly effective	Difficult to recover(read back)
Case 3	Fully separated	Simple	Expensive
	Archive shared	N/A	
	Unified	Effective	Vendor lock

- ❖ Infrastructure and architecture can be imagined.
- ❖ Cost evaluation is most important, but it is difficult to estimation.
 - How long should it be preserved before reprocessing?
 - How large is the egress?