

National Aeronautics
and
Space Administration

EARTHDATA

Evolving Archive guidelines for the cloud

WGISS-61 03/16 - 20/2026

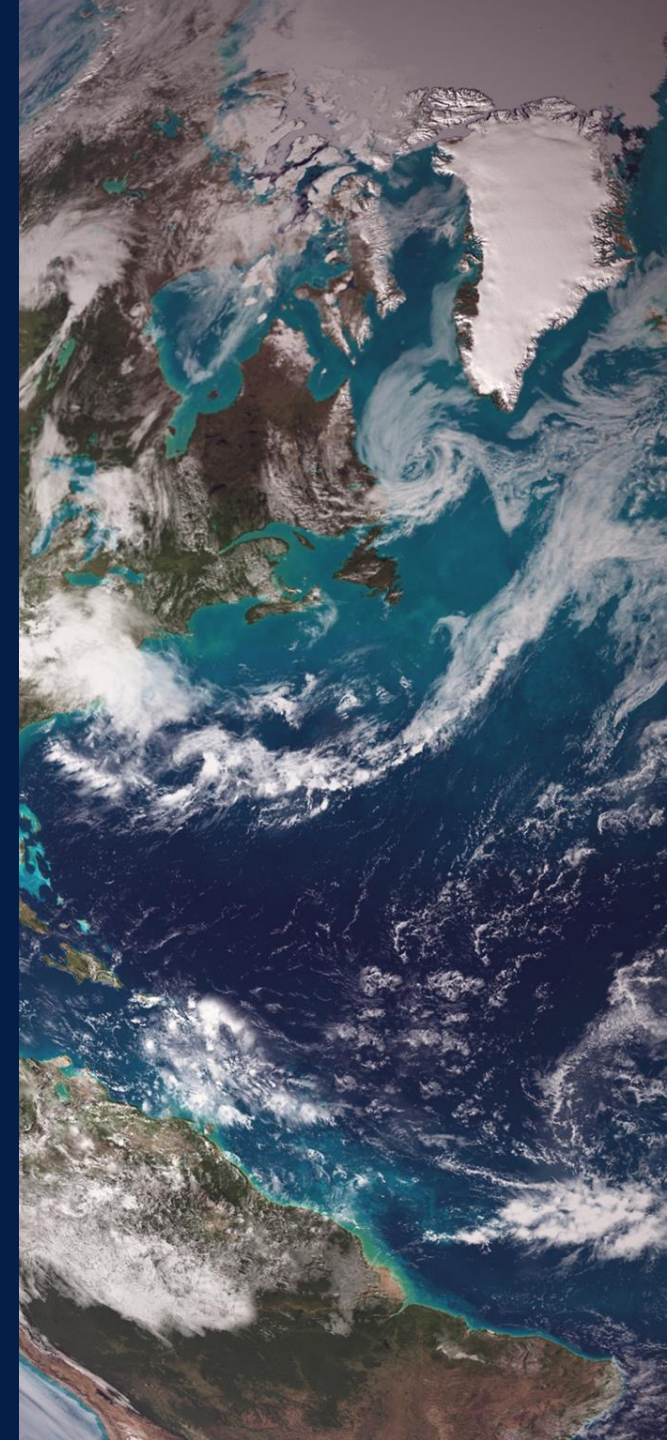
Doug Newman

NASA Goddard Space Flight Center, ESDIS Project



Agenda

- The old risks and mitigations
- The new risks and mitigations
- Our findings
- The future

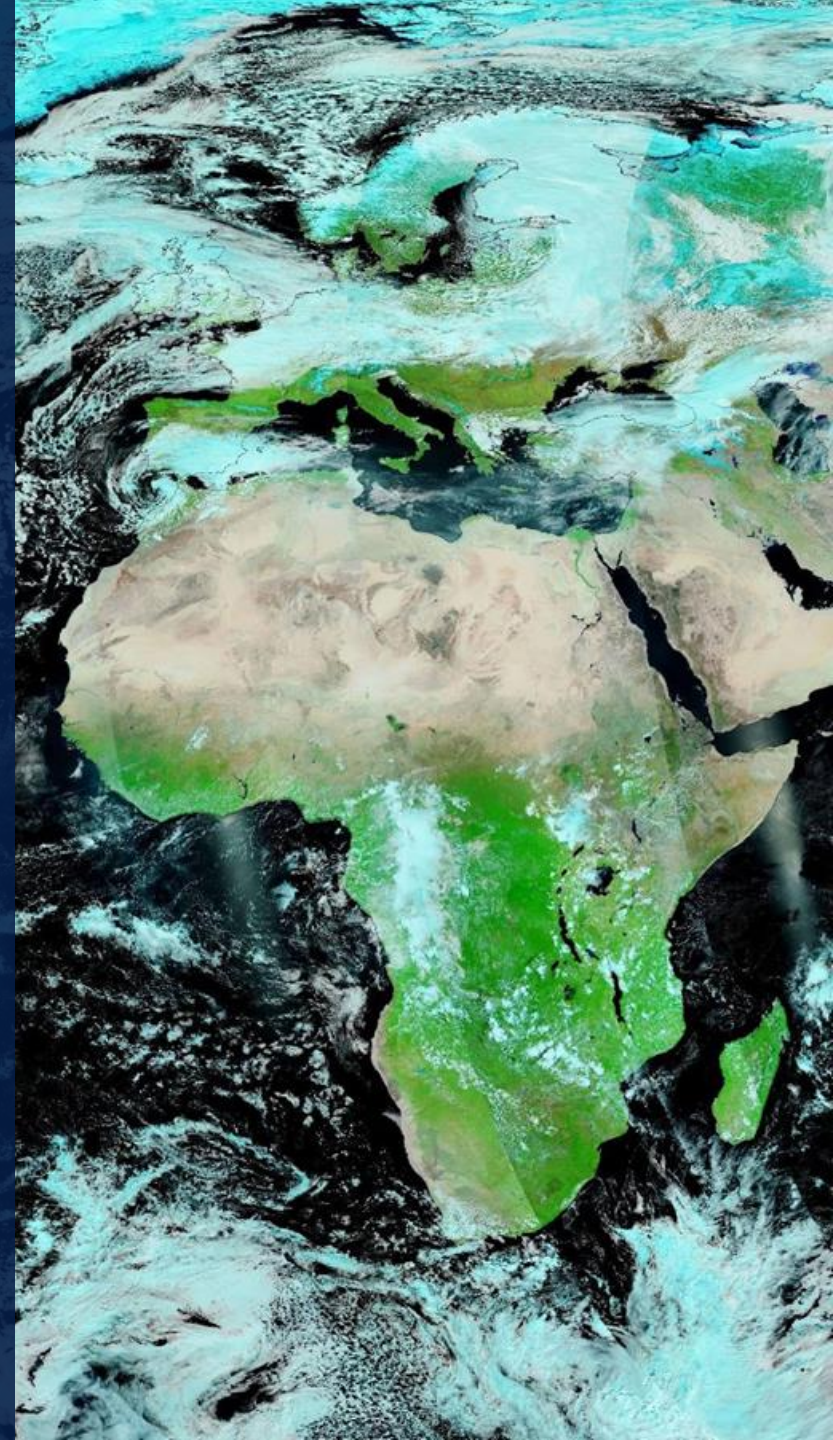


Disclaimer

The following is subject to change as we ~~design~~, implement and test these ideas.



The old risks



3-2-1 strategy

3 total copies of data, stored on 2 different media types, with 1 copy located off-site

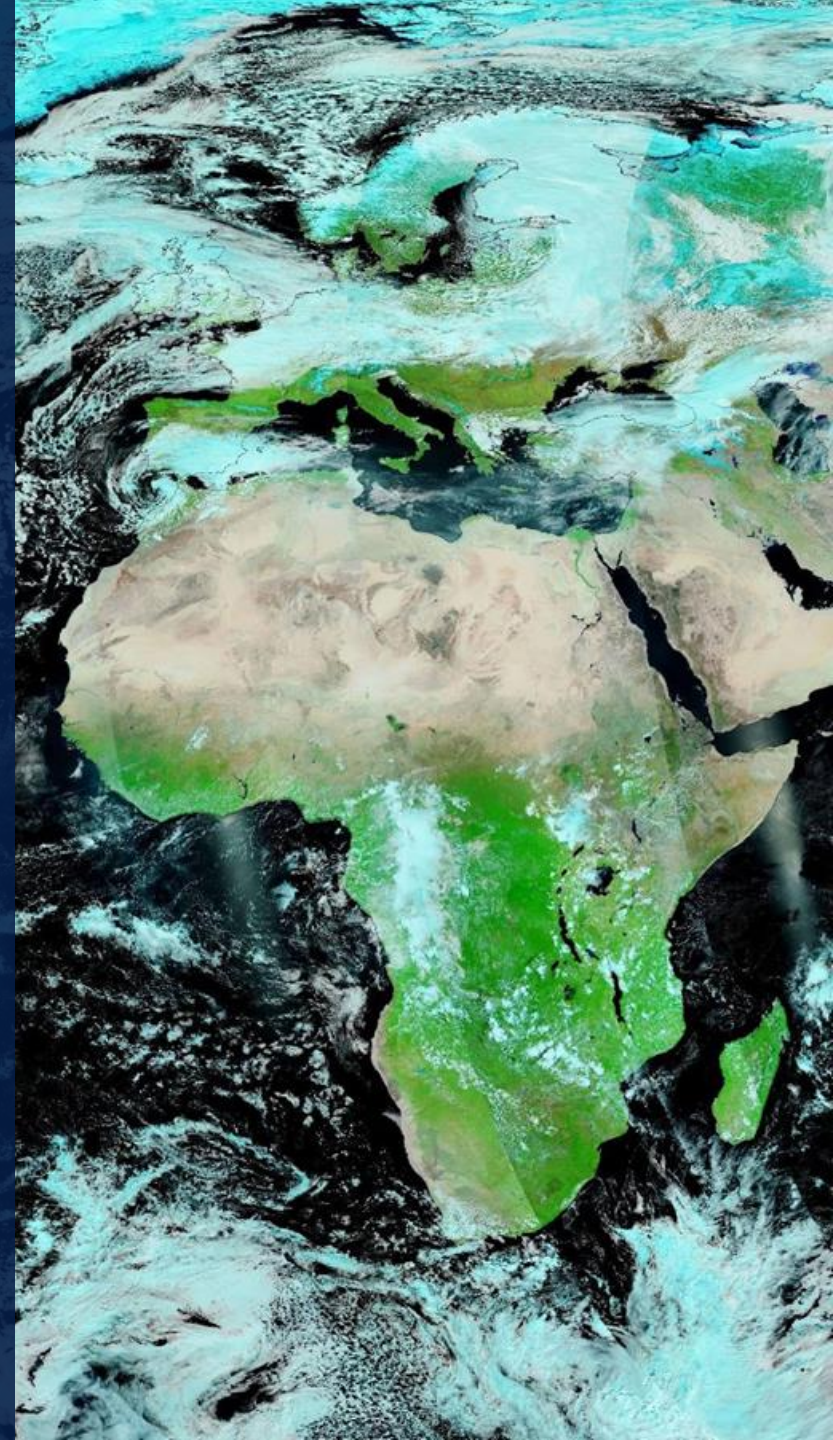
Risk	Mitigation
Disasters	Copy in a location geographically separated from the primary
Media vendor viability	Copy on a different media format
Media vendor cost	Copy on a different media format
Accidental deletion/corruption	Copy in a separate account/system
Malicious deletion/corruption	Copy in a separate account/system



3-2-1 in the cloud

- **‘3 total copies of data’** - Amazon S3 automatically stores at least three copies of each object across multiple physical devices and at least three different Availability Zones
- **‘2 different media types’** - Unknown
- **‘1 copy located off-site’** - ‘AWS Availability Zones in a Region are meaningfully distant from each other, up to 60 miles to prevent correlated failures’

The new risks



The cloud vendor

- The risks of media viability and cost becomes the risk of cloud vendor viability and cost.
- The 'rent rather than own' risk.



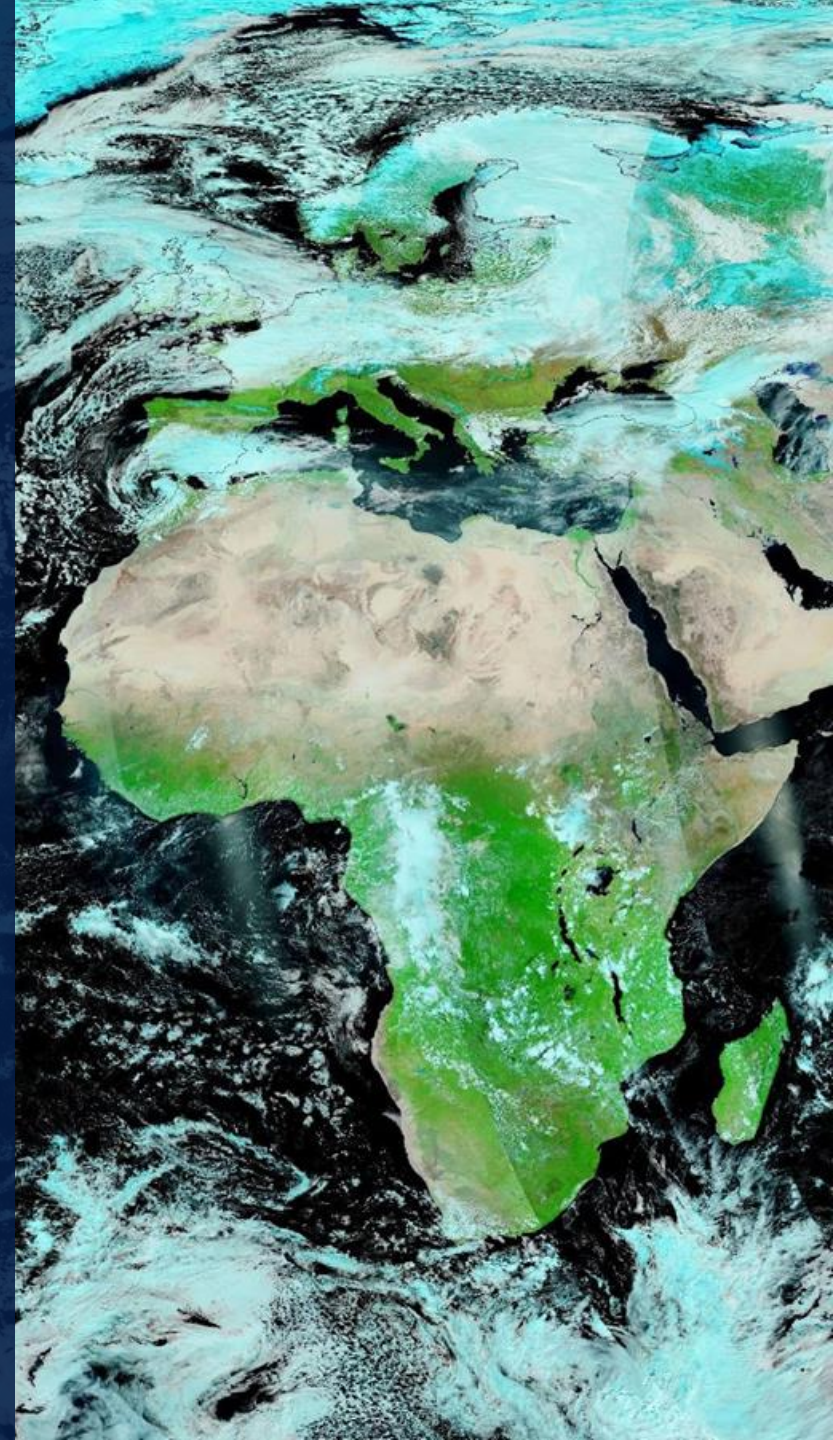
Cloud strategy

Risk	Mitigation
Disasters	multi-AZ cloud storage* 'S3 ... redundantly store objects on multiple devices across a minimum of 3 Availability Zones in an AWS Region'
Cloud Vendor viability	Accept/keep copy with another cloud vendor/retrieve from source
Cloud Vendor cost	Long term contracts
Accidental deletion/corruption	S3 versioning + object lock
Malicious deletion/corruption	S3 versioning + object lock



Where we are today

Some course corrections!



Reducing cost: S3 versioning

The goal is not to store copies of data.

The goal is not to lose data.

Lazy backup:

When data deletion is attempted, don't delete it, mark it as deleted for a period of time, then delete it.

Reducing risk: S3 versioning + Object Lock

- Lifecycle policies will ‘age off’ legitimate deletes
- Object lock reduces the risk of accidentally or maliciously
 - Deleting versions created by versioning
 - Deleting buckets - which wipes out versioning
- The two person solution - this is our preferred solution and we are lobbying AWS to provide it
- We have work to do in detecting accidental/malicious deletions



Reducing risk: Replace on-premises copies

- Cloud vendors offer SLAs we cannot even come close to (nine nines) on-premises
- Why do we have an on-premises copy instead of a cloud backup with a different vendor?
 - Equivalent of '2 different media' from 3-2-1 but necessary?
 - We are wary of renting storage as opposed to owning it

Reducing cost: S3 Intelligent Tiering

We have observed between **40-60%** cost savings on our archive costs since the implementation of S3INT across all earthdata cloud.



Reducing cost: Cold storage

We could realize further storage cost savings by putting some of our data in 'cold storage'. I.e. storage that does not allow instant access.

This is not cost effective (for us) right now but may become increasingly attractive as our archive doubles in size with NISAR

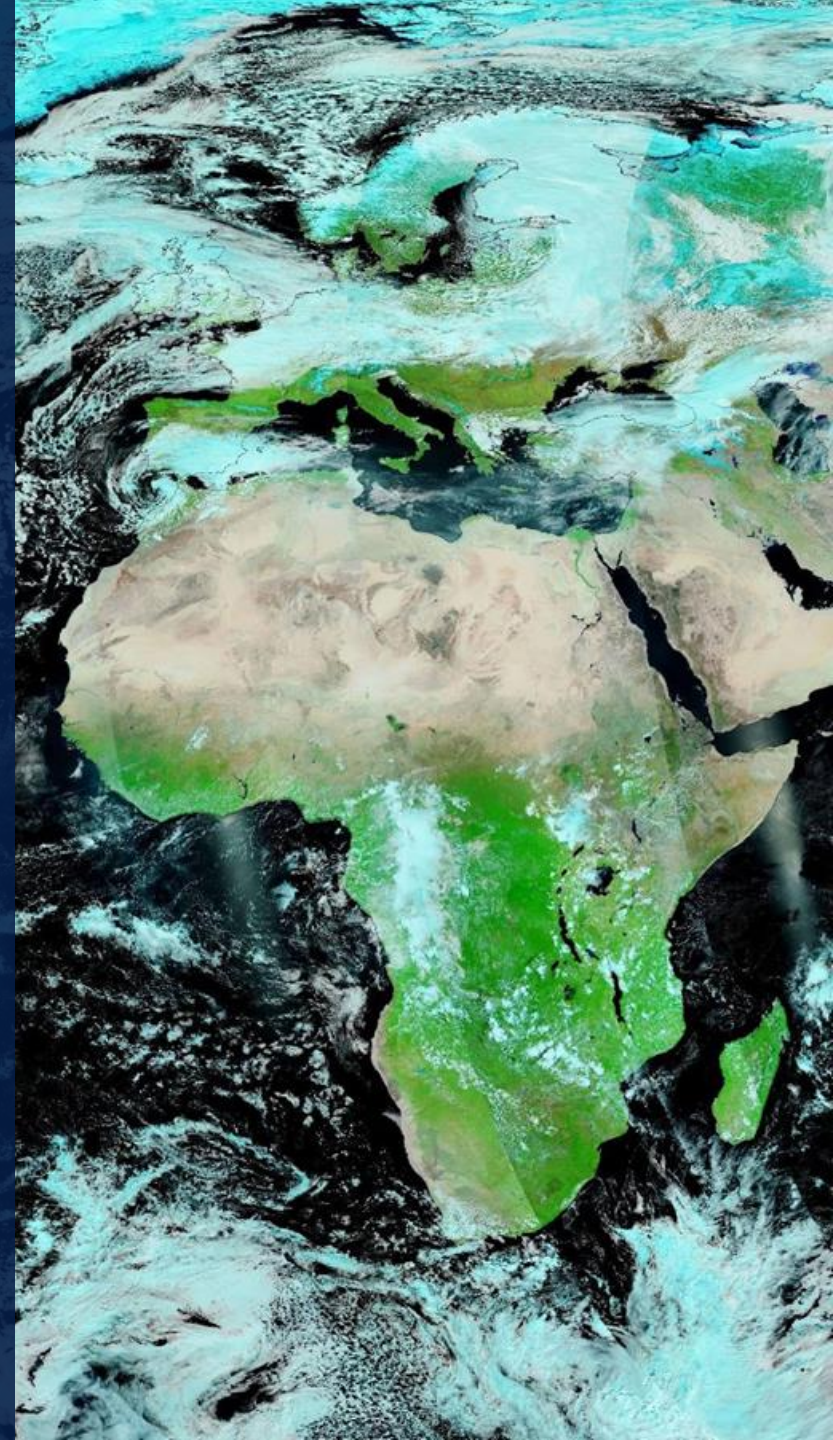
Reducing cost: deduplication & reduction

Why are we keeping this data?

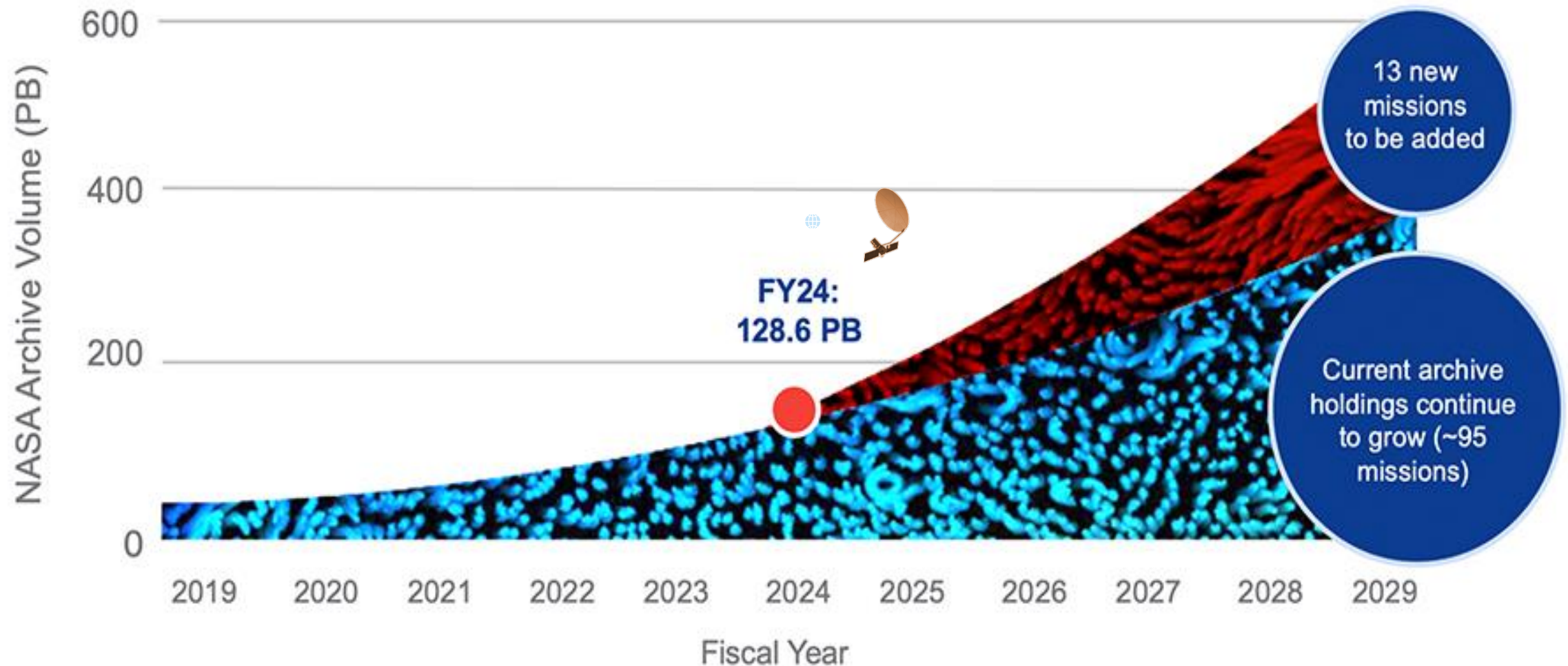
- Do we have better versions? Get rid of it
- Is it *archived* somewhere else that is trusted? Get rid of it



The future



Earth Science Data Archive Growth Projection



What can we afford?

NISAR alone, over its lifespan, will generate **140 PB** of data.

Storage of 140 PB of data in S3 standard access will cost you: **\$3 million per month**

This figure does not include egress of that data :)

(30% of our volume is generally distributed per month: **\$5.4 million per month**)

Can we afford 3 or even 2 copies? Not for NISAR.

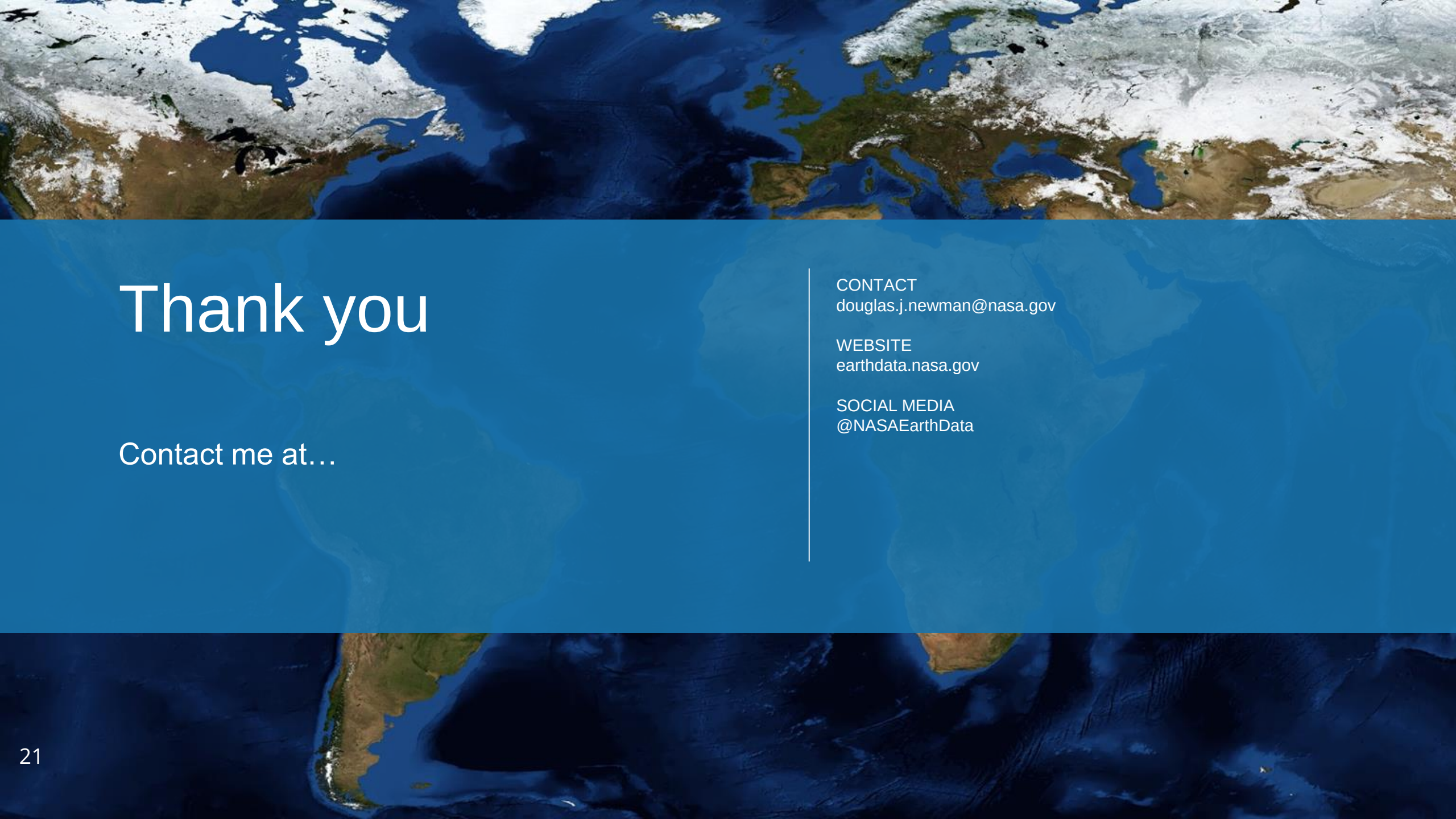
The conventional wisdom is that you should still backup your data on-premises when using the cloud. We are not conventional. Our budgets are not in lock-step with our data growth.



And another thing... cloud-optimized data

- How do we provide an optimal experience for cloud computing with their 'adjacent' compute?
- How do we maintain the experience of traditional users at the same time?
- How do we do the above without replicating data in different formats?





Thank you

Contact me at...

CONTACT
douglas.j.newman@nasa.gov

WEBSITE
earthdata.nasa.gov

SOCIAL MEDIA
[@NASAEarthData](https://twitter.com/NASAEarthData)