



Data Processing infrastructures toward a Datalake

JCAD 2019 : JP Gleyzes





Launcher networks



Main Command and Control rooms



Satellite Control Centers



IT and Computing

IT support to Space missions



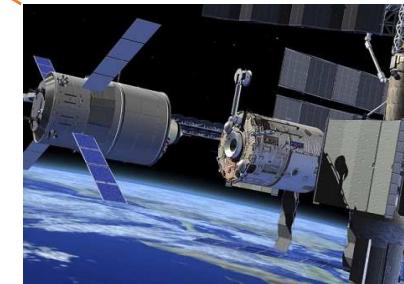
Mission Centres



Stations Networks



Laboratories and equipments



Partners networks

CNES Datacenter overview



HPC (HAL 2019)

- 434 Tflops (**+134**)
- 12 400 cores (**+4000**)
- **8,2 PB (+2) GPFS / 300 TB (+100)** burst buffer/ 200GBs bandwidth (**+100**)
- Infiniband low latency network
- 8 GPGPU Nvidia Volta 100 + 8 GPGPU Nvidia Tesla T4 (**+12**)
- **[NEW]** Deep learning server : DGX-1 (170 teraflops)

Restricted supercomputer (Ktulu)

- Smaller than HAL

GAIA (Hadoop)

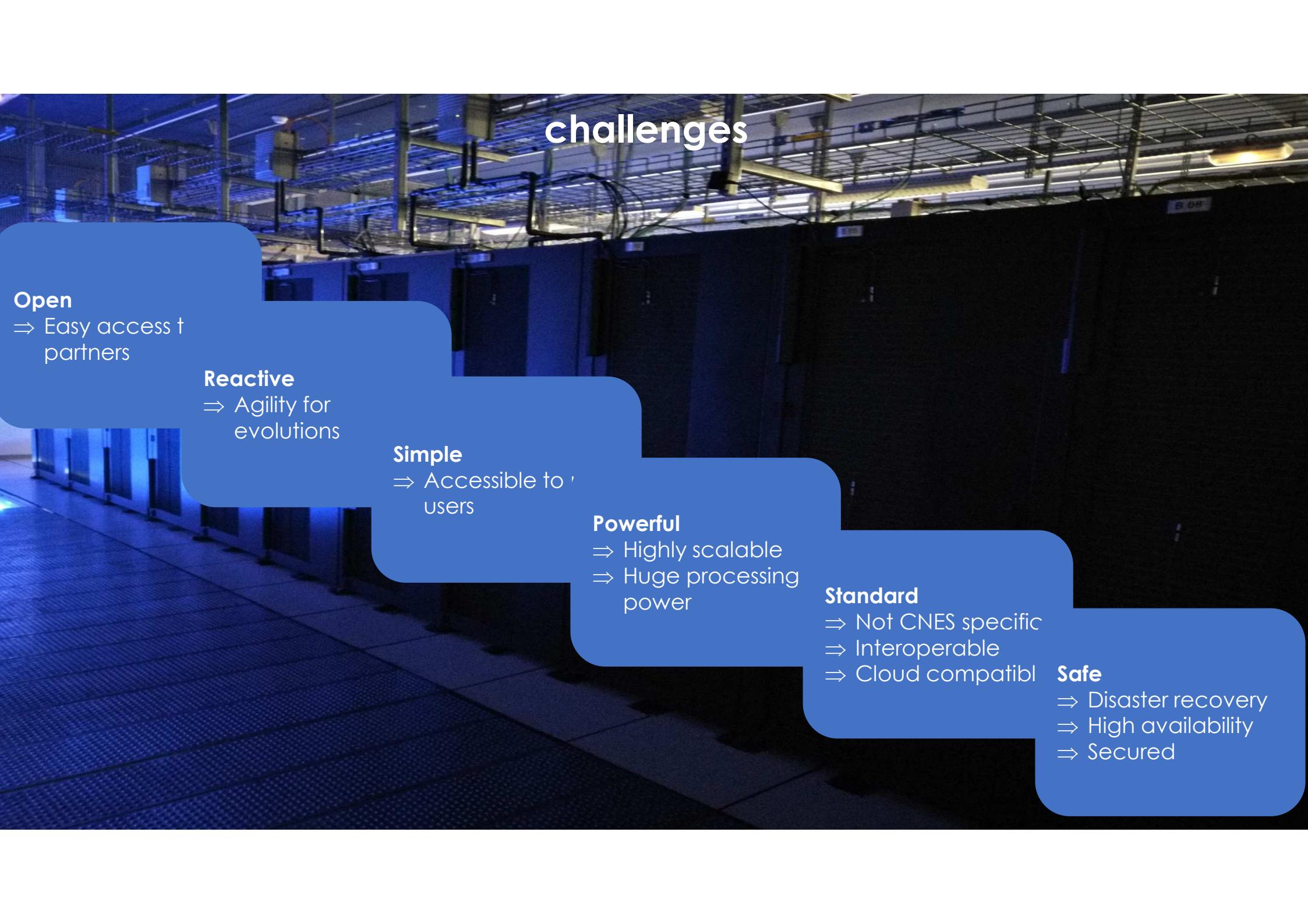
300 servers (2018)

- 6500 cores
- 10 TB RAM
- 120 Tflops
- 3 PB

PEPS

IBM – HSM Storage

- Disk cache 3PB
- 16PB tapes (HPSS)
- 2x10Gbs/s Ethernet



challenges

Open

⇒ Easy access to partners

Reactive

⇒ Agility for evolutions

Simple

⇒ Accessible to users

Powerful

⇒ Highly scalable
⇒ Huge processing power

Standard

⇒ Not CNES specific
⇒ Interoperable
⇒ Cloud compatible

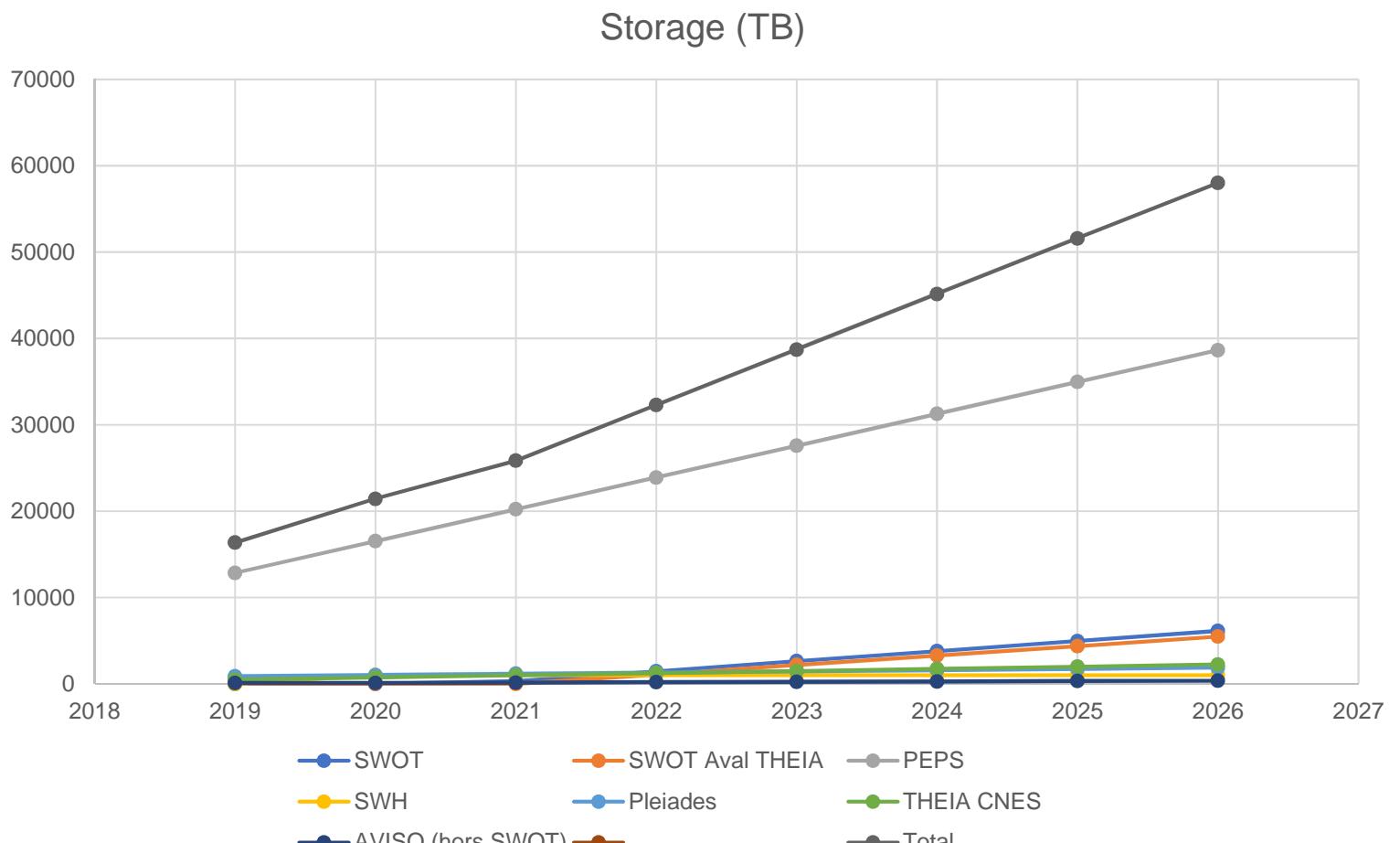
Safe

⇒ Disaster recovery
⇒ High availability
⇒ Secured

Roadmap and questions

Large scale
Storage, archive,
network, infra

Huge (multi so
multi sensors)
(re)processing



Roadmap and questions

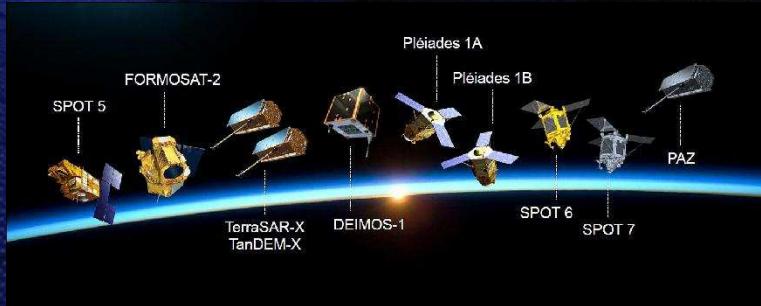
Large scale

Storage, archive
network, infra

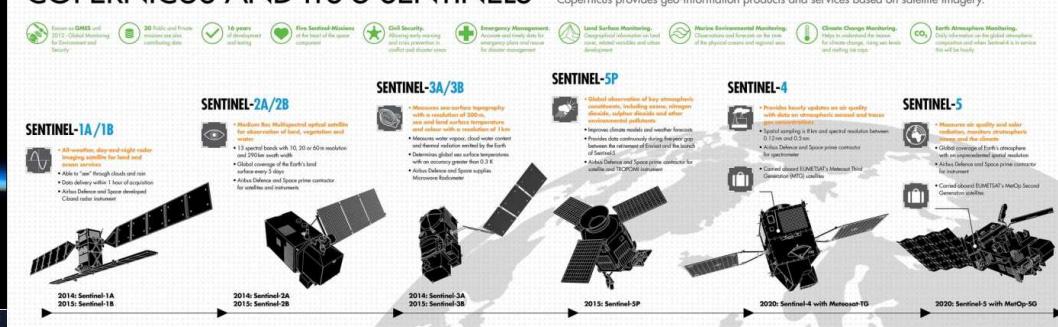
Huge (multi sources,
multi sensors)
(re)processing



Long time series



COPERNICUS AND ITS 5 SENTINELS



Exemple recalage de séries Sentinel-2 (avec ou sans GRI)

Avant recalage



Après recalage



30 ans de données à retraiter → [un timelapse de la terre !](#)

Roadmap and questions

Large scale

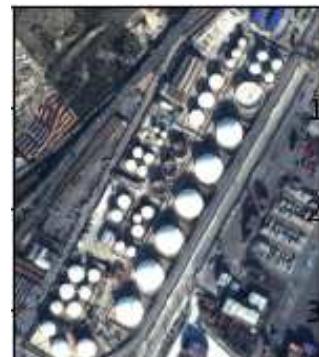
Storage, archive
network, infra

Huge (multi sources -
multi sensors)
(re)processing

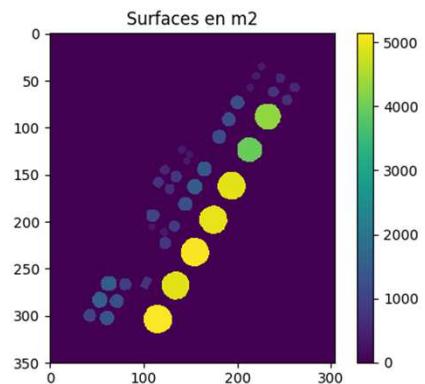
Information = new
“black gold”

Information mining
AI

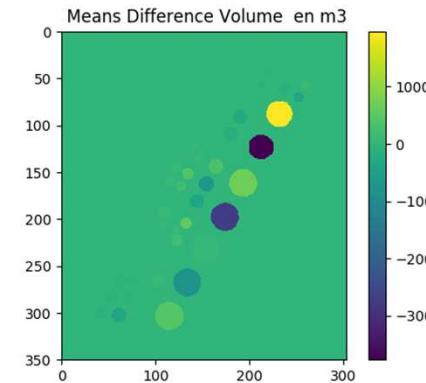
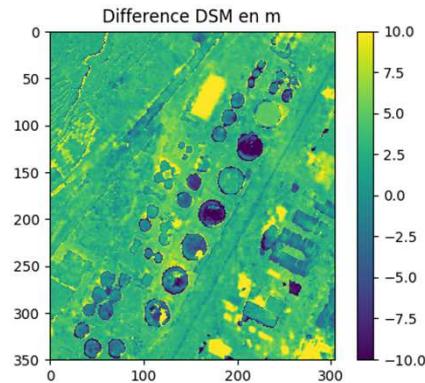
Zones industrielles : détection de changements 3D



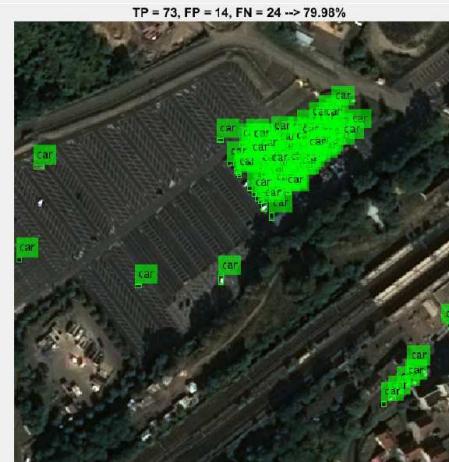
t1



t2



Détection d'objets



Travaux CNES, IRISA, Quantcube, 2019

**Travail en
cours sur le
suivi des
décharges
illégales par IA**



**Travail en
cours sur le
suivi des
décharges
illégales par IA**



Roadmap and questions

Large scale

Storage, archive
network, infra

Huge (multi sources -
multi sensors)
(re)processing

Information = new
“black gold”

Information mining
AI

Full Web model

App shop
Sand boxes
Innovation boosters

Service chaining
Interoperability
Cloud
Streaming of info

IRMA et post IRMA



Sandy Ground, Saint-Martin
Image Pléiades Pleine résolution
AVANT Irma : 12/02/2017

IRMA et post IRMA



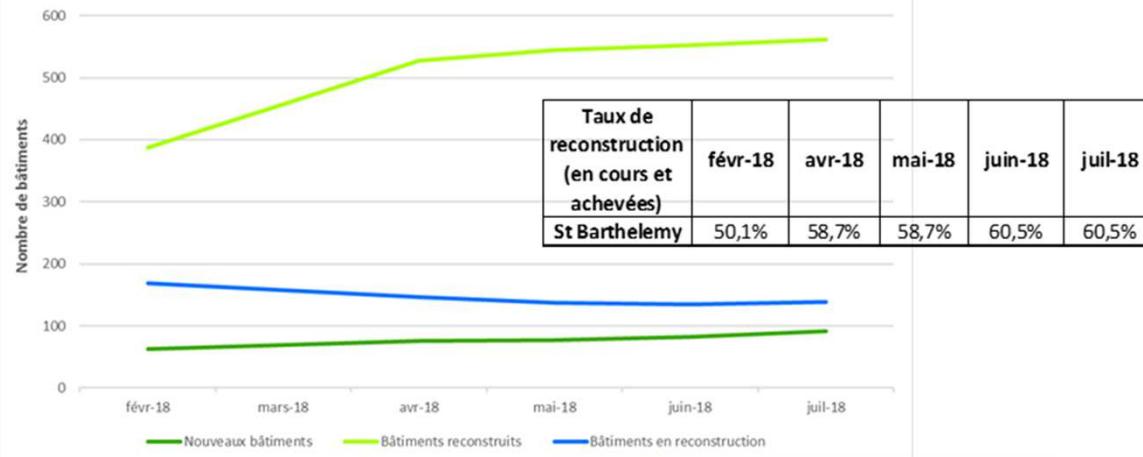
Sandy Ground, Saint-Martin
Image Pléiades Pleine résolution
APRES Irma : 10/09/2017

Pléiades © 2017 CNES, distribution Airbus Defense and Space

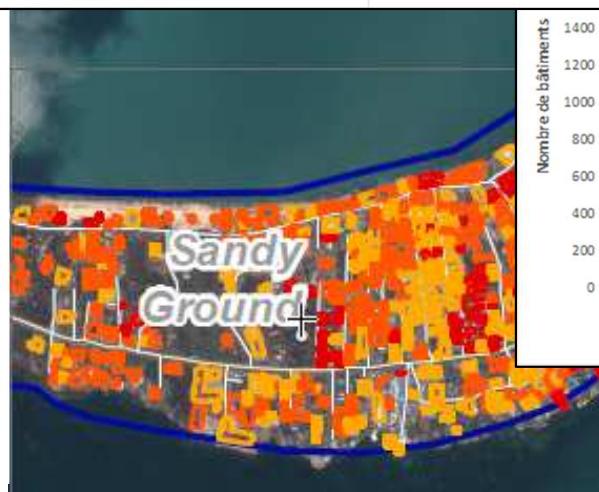
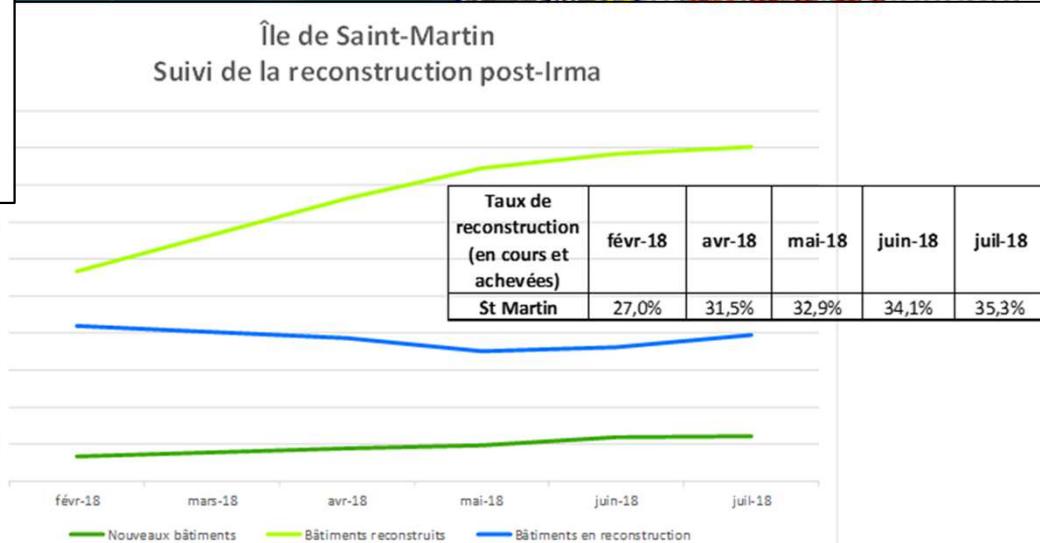
Réponse à

Analyses des images Pléiades des 8 et 10 sept 2017

Île de Saint-Barthélemy Suivi de la reconstruction post-Irma



Île de Saint-Martin Suivi de la reconstruction post-Irma



Opernicus



Monitoring de la température urbaine (îlots de chaleur)



Ecostress + PHR
Février 2019
Température entre 0° et 5°
Négatif à l'extérieur de la ville

Cnes/JPL

Roadmap and questions

Large scale

Storage, archive
network, infra

Huge (multi sources -
multi sensors)
(re)processing

Information = new
“black gold”

Information mining
AI

Full Web model

App shop
Sand boxes
Innovation boosters

Service chaining
Interoperability
Cloud
Streaming of info

New skills, new model

Accounting/billing
Ressource
management
Cloud brokering
Vendor lock in ?

- Lack of unique and complete data referential
- Data access is complex (« by project » infrastructure, multiple API)
- Multiple data format
- Difficulty to download external data, access to external catalogs
- Multiple data distribution channels

User point
of view

- Data distribution across too many platforms
- Data duplication
- Same data downloaded multiple times
- Data silos

IT point of
view

Datalake : objectives

Datalake CNES :

- Virtually 50 -100PB : Dynamic buffer from multiple external Datasources (host most frequently accessed data)
- Break Data silos
- Multi temporal / multi sources Data repository
- Datacentric architecture

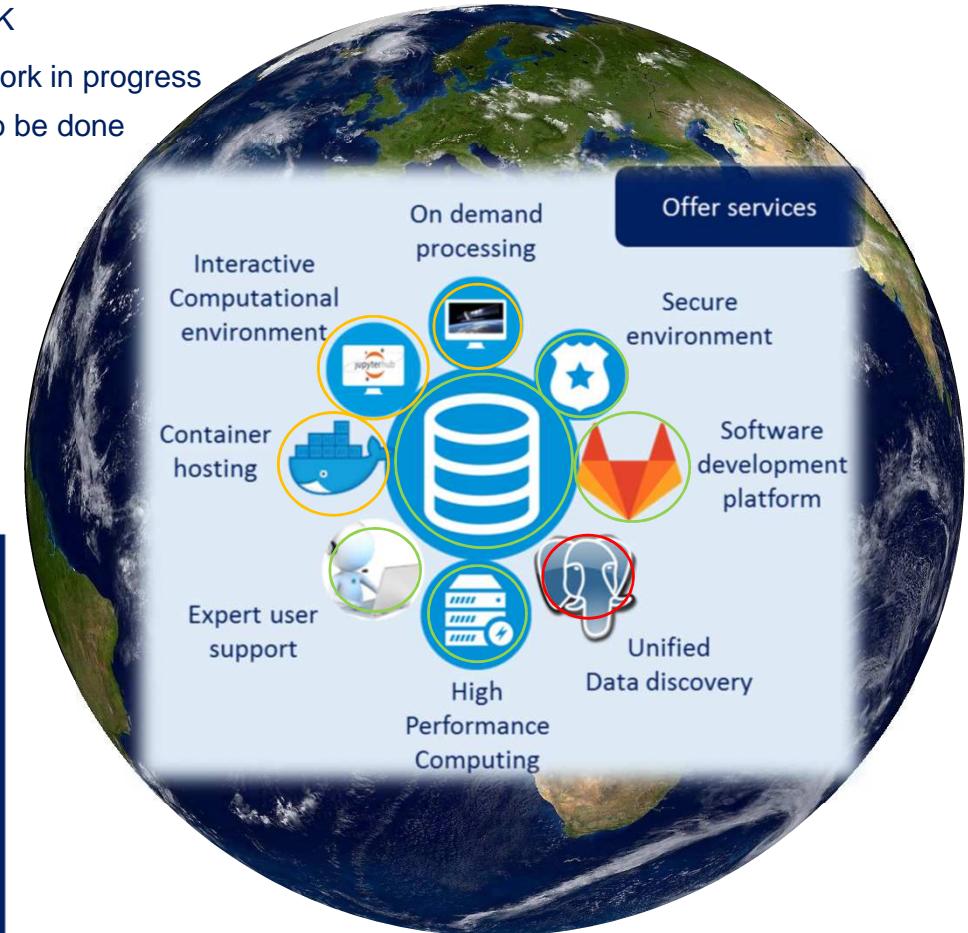
OK

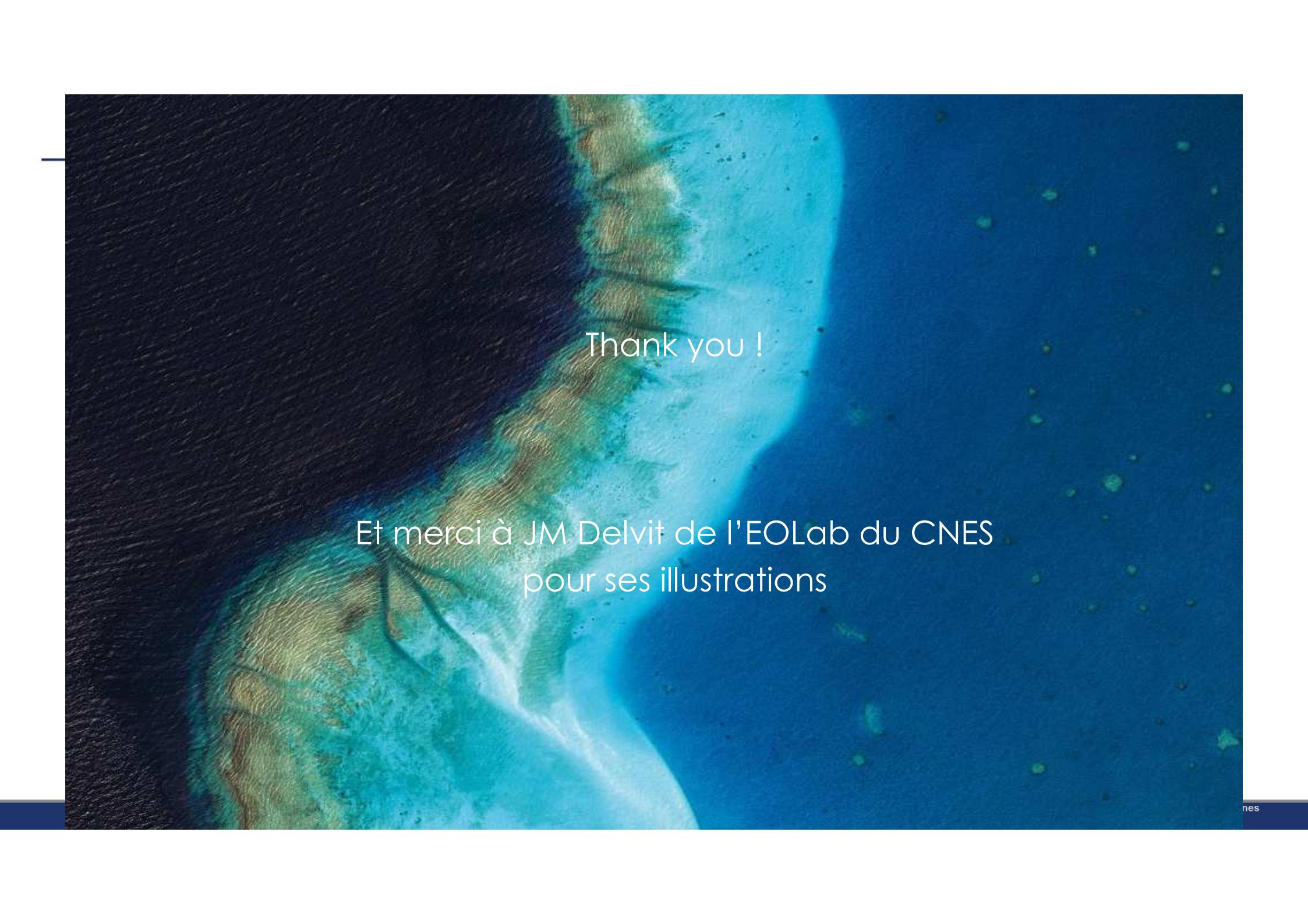
Work in progress

To be done

Technical challenges

- Provide users with *complete/ standard / flexible* tools to innovate from spatial data (Notebook &co)
- New parallel framework : Dask, Spark
- 1B+ files filesystem
- Cloud interoperability



The background image shows an aerial view of a coastal region. The land is covered in green vegetation, and the surrounding water is a deep blue. There are some small, isolated green dots scattered across the dark blue water, possibly representing small islands or artificial structures.

Thank you !

Et merci à JM Delvit de l'EOLab du CNES
pour ses illustrations