CNES Initiatives on Big Data and Cloud

WGISS

Pierre Lassalle

Thursday, May 2, 2019

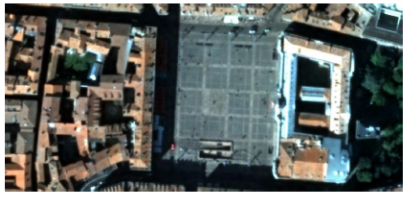
Contents

Context & motivations

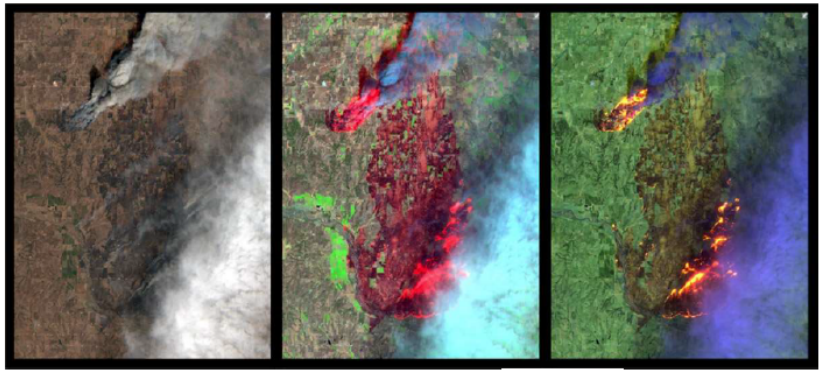
Context

- Massive amount of heterogeneous but complementary data:
  - Spectral, spatial and temporal resolution

Sentinel-2



Pléiades-HR



Sentinel-2

- Institutional and collaborative semantic databases

EU Open Data Portal

OpenStreetMap

géoportail

Context & motivations

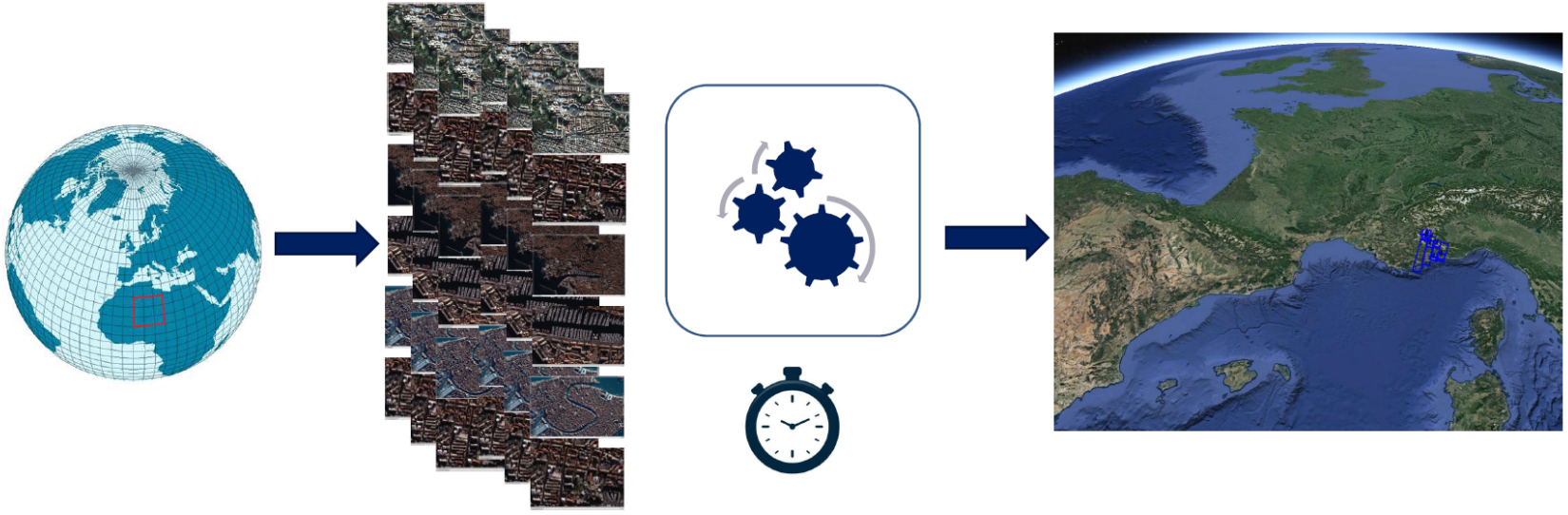
Context

- Increase of data volume illustrated with Sentinel-2/Copernicus:
  - 10 Po of data to process every year
  - Free data with a large spatial coverage and a high revisit frequency

- Data are more and more available:
  - Big Data / Cloud technologies
  - Data access services
  - Data fusion

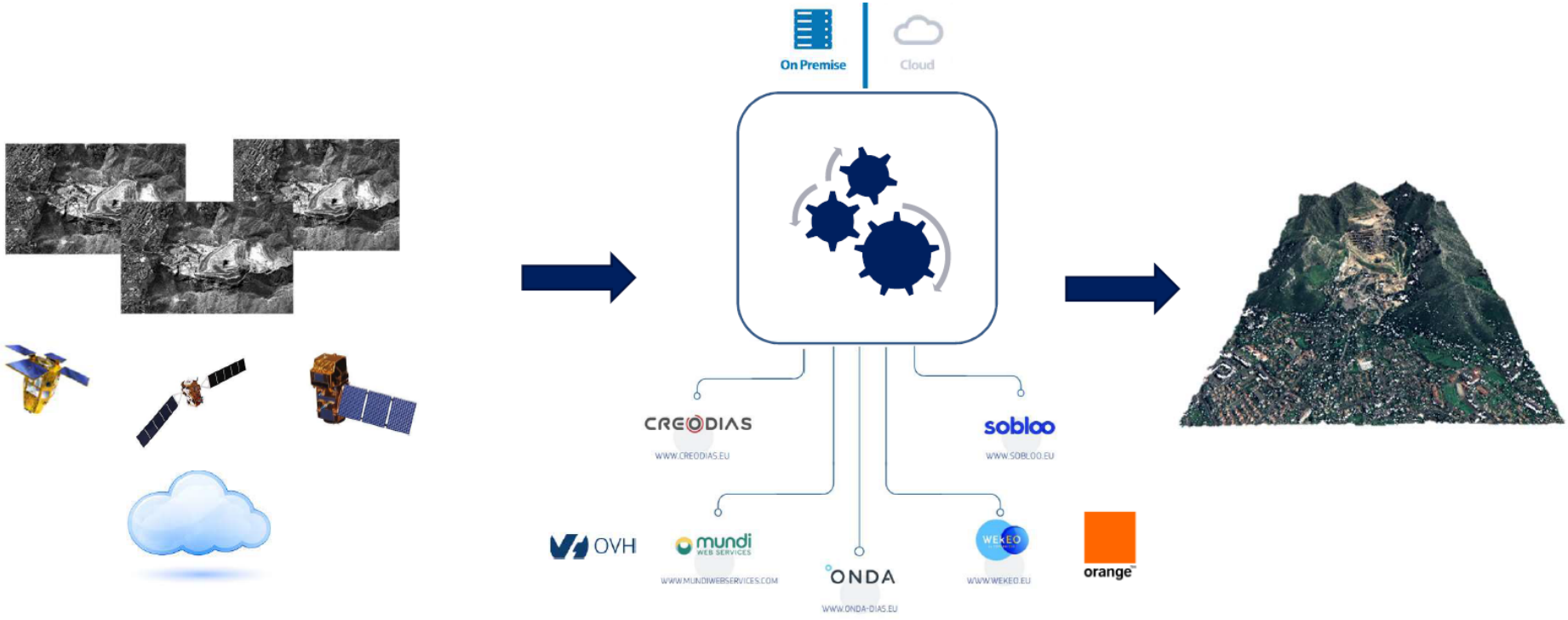
Motivations

- To do large scale complex image processings in near real time at a reasonable cost



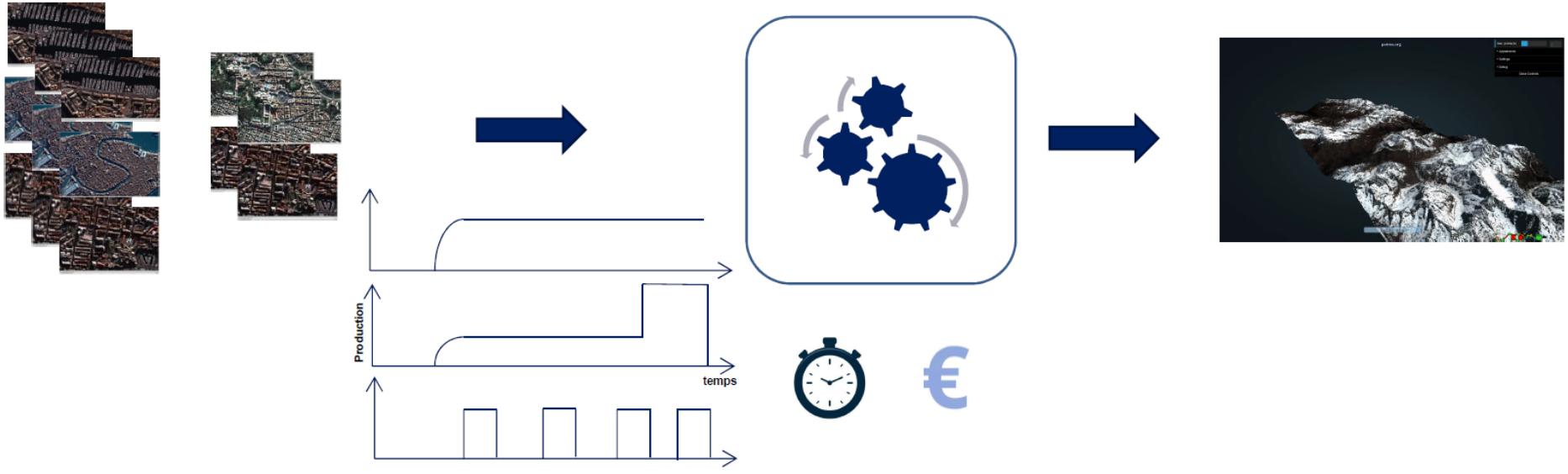
Motivations

- To host data and algorithms on Cloud platforms:
  - Development of applications close to data location
  - Enhance fusion of multi-source data
  - Open for collaborations



Motivations

- To optimise the use of resources:
  - Access to necessary data only
  - Reduction of the processing cost:
    - Elaborate production strategies
    - Find a compromise cost / execution time



Contents

Recent R&D studies on Big Data processing

- Image processing using Cloud & Big Data technologies
  - Analysis of the benefits and impacts of using big data technologies in cloud environment for image processing
  - Redefine software architecture and flow logic of Image Algorithmic Software to adapt to Big Data technologies
  - Push image processing chains on Cloud
    - Identification of good practices for generalization
- 3D Image processing using Big Data technologies on the cloud
  - Validate the concepts studied on an operational platform
  - Modelize and develop the future image processing chains

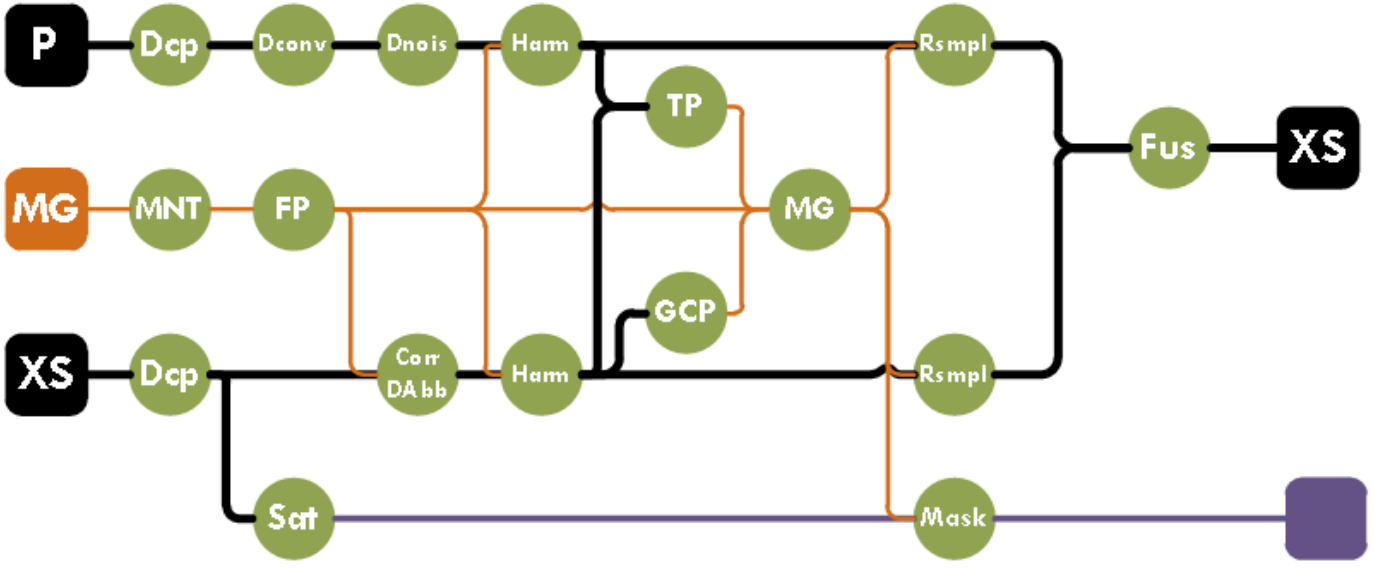


Illustration of a processing chain for Pleiades Ground Segment

Contents

New generation of image processing chains

Main goal

- Adapt image processing algorithms to new flow logic paradigm of recent Big Data frameworks
  - Directed Acyclic Graph (generalization of Map/Reduce) in Apache Sparks
  - Dask task graph proposed by Dask framework

Main constraints

- Selection of a Big Data framework
  - Need for maturity
  - Separation of data management and core algorithms
  - Handle node failure and preserve data integrity
  - Need for simple work orchestration

Current decision

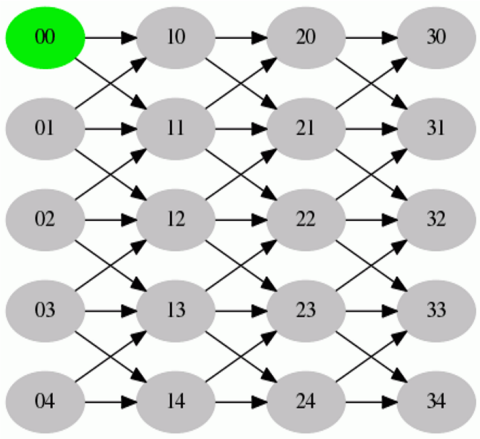
- Selection of Apache Spark framework but Dask is getting more and more our attention
  - Jobs are brought to node where data is located
  - Data tiling and aggregation are executed by Spark function
  - Data propagation is optimized
  - Scalability is ensured
  - Flexibility to architecture variablity

Contents

DAG adaptation of image processing algorithms

Adaptation of "legacy" code

- Tile management
- A process must follow the requirements:
  - Describes the input tiles as a list of coordinated tiles and their size
  - Describes the outputs as a list of coordinated tiles and their sizes
- Able to build the graph of the tasks (DAG)



Execution strategy

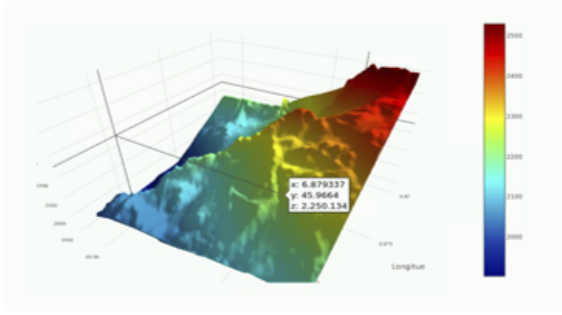
- Instantiate a pool of worker nodes to feed the graph
- Handle data lifecycle:
  - To remove temporary data when not needed
- Depth-first exploration of the graph
  - Release memory as soon as possible and reuse newly free node workers for other tasks

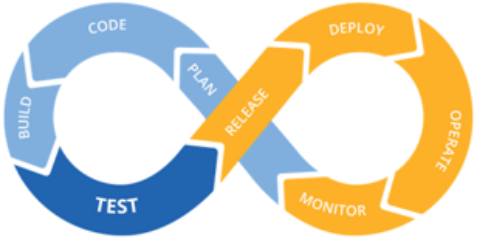
Functional approach

- $driver.getFlow() -> ortho() -> write() -> run()$
- Modularity thanks to lazy evaluation

Contents

Ongoing activity





Main goal

- Validate the ability of massive image production using Spark in a Cloud production-like environment
- Validate the flexibility and scalability of such technologies
- Develop image processing chain for Big Data and Cloud environment
- Introduction of DevOps tools and methods in a production-like environment (define operational concepts)

Establish a new reference image processing framework for new Earth Observation missions

Thank you for your attention

pierre.lassalle@cnes.fr