Data analytics applied to the enhancement and improvement of EO products and services in a big data environment

Alberto Lorenzo-Alonso

13/11/2016

The company

37.000	3.000M€	140	152M€
professionals	revenues	countries	R+D investment

indra

Earth Observations Applications Unit

Actionable information
Security
Digital Data
Coastal monitoring
Disaster Risk Reduction
Land applications
Agriculture
Hazard mapping
Reference mapping
Land use / Land cover
Engines

The platform

An operational platform for the cyclical needs of Earth Observation groups (generation of **downstream products and services**)...

...scalable to massive continent-global-wide production

Land Analytics
Earth Observation Platform

** The Land Analytics EO Platform was developed under partial funding of ESA GSTP 6.2 Programme*

Can be integrated with other engines

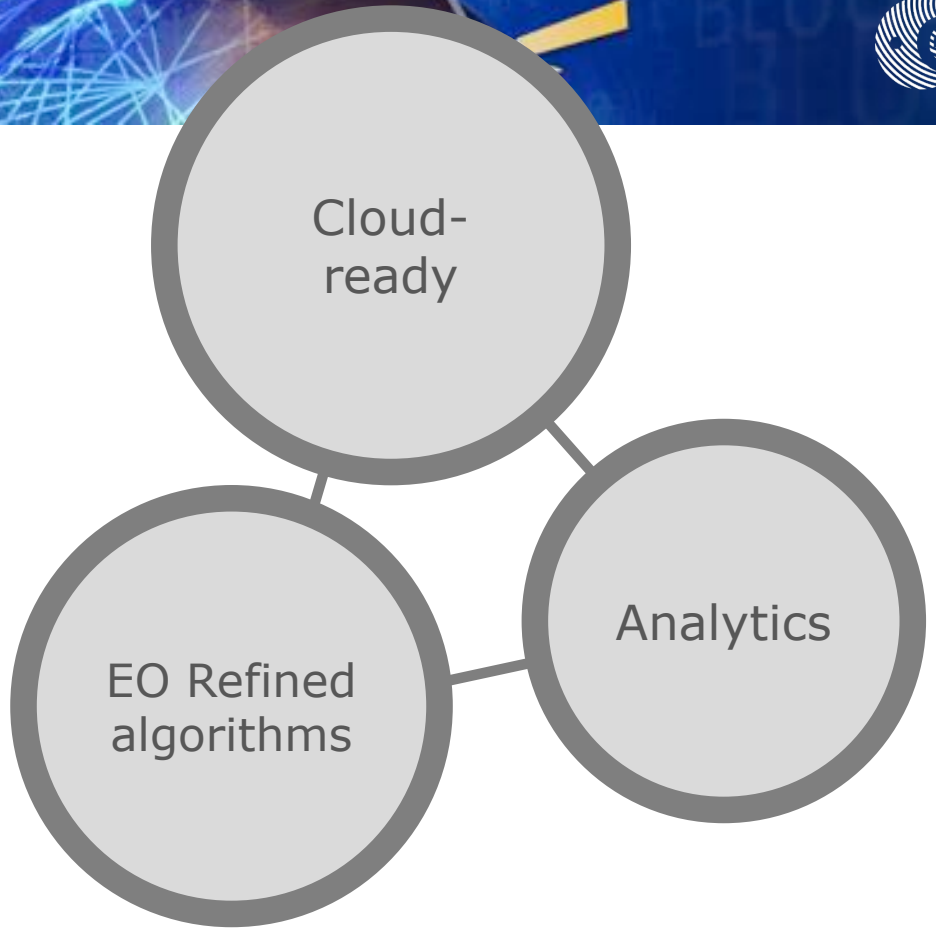
Risk Mapping Engine

AIR INT

Coastal Monitoring Mapping Engine

The platform

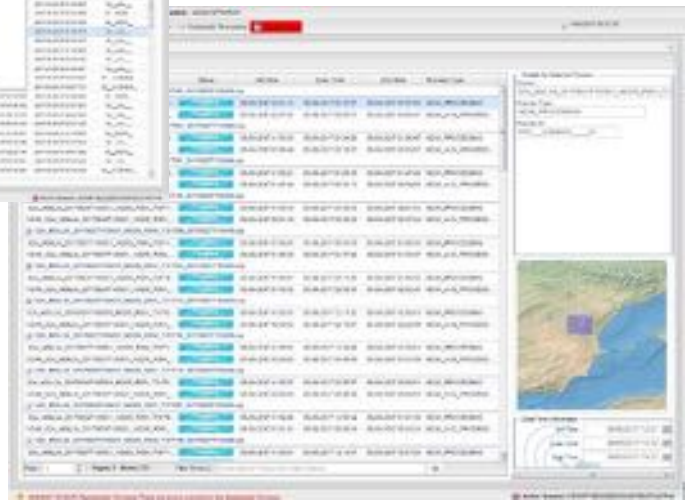
Land Analytics
Earth Observation Platform

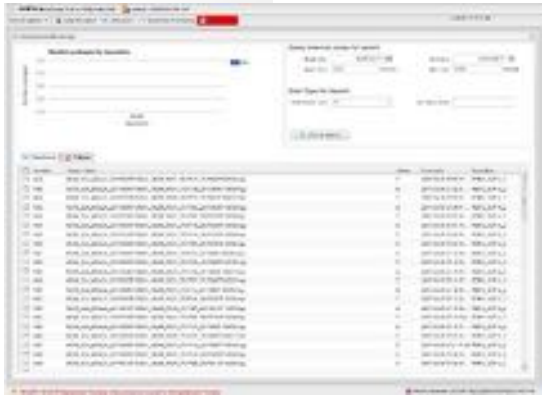


The platform

- Multi-cloud deployment, tested successfully in:
  - Private cloud infrastructure of Indra
  - Amazon Web Services
  - Operationally successful in: Azure
  - DIAS solution as far as Virtual Machines are supported
- Fully scalable:
  - The system allows hundreds of processing hosts running in parallel

Monitoring and control





The platform: scalability

Balancing three types of scalability for each need

Number of CPUs and amount of processors capacity or RAM

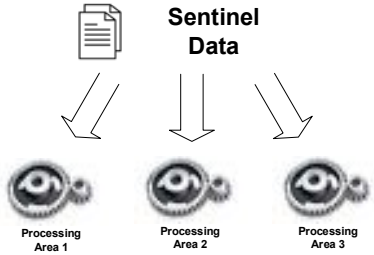
1 CPU controlling each module (data exchange, order handler, scheduler, local archive, monitoring and control)

1...n CPU for processing module

Number of processing hosts running in parallel

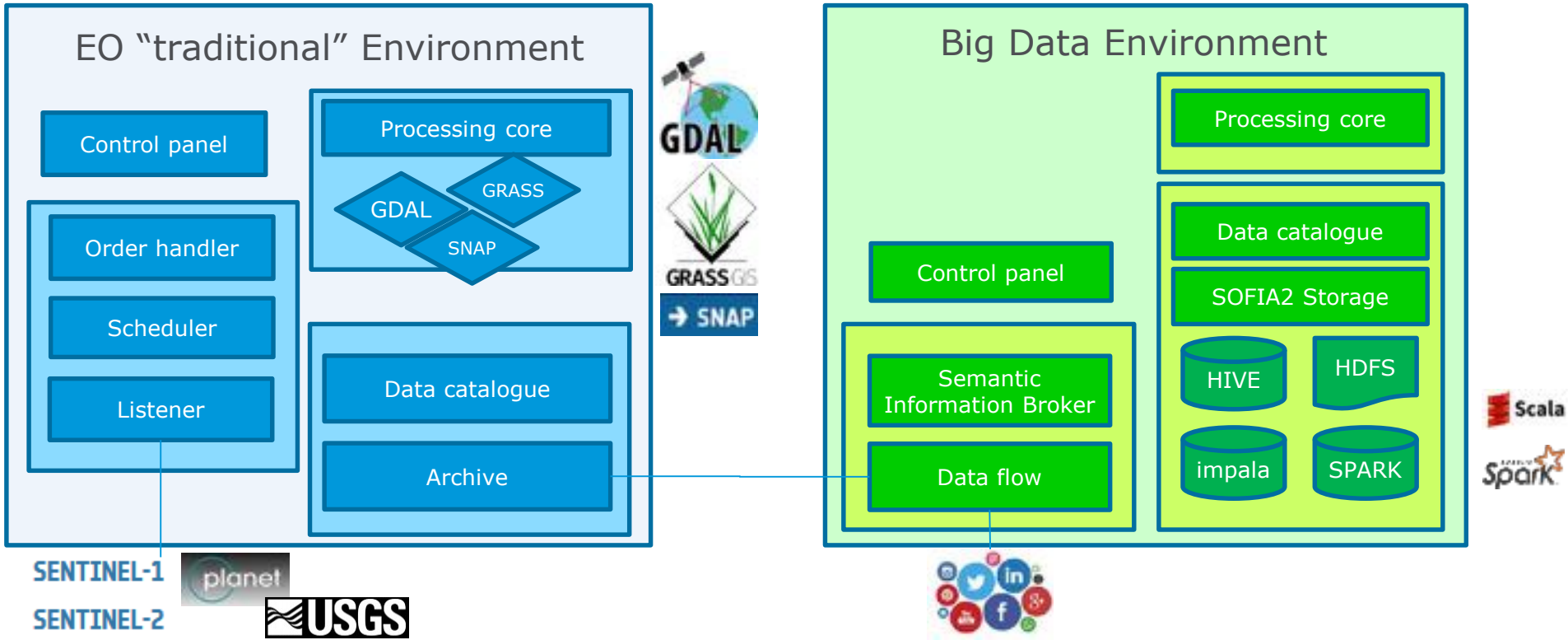
Operationally tested: 33 processors processing SMOS data in parallel.

Several "platforms" installed in parallel processing 1.zones or 2. periods



Areas and periods must be wide enough

The platform: SW architecture, the best of two worlds



EO “traditional environment”: multi-functional module

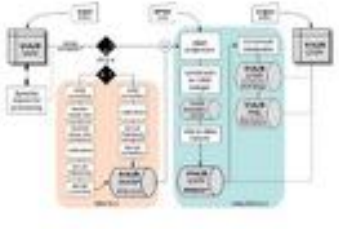
Cross-sensor

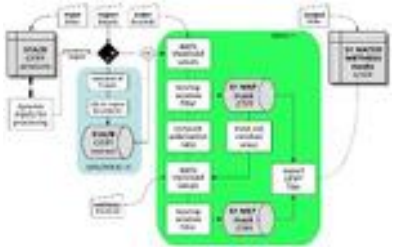
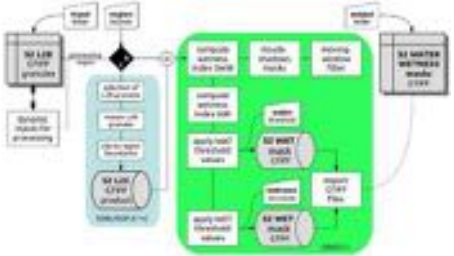
- Sentinel-1, Sentinel-2, SMOS, Landsat, Planet Scope (automatic accessing and ingestion of a selected AOI and time-range)

Multi-EO-processor

- Plug-in system for inserting processors based on open-source software (GRASS, GDAL, SNAP)

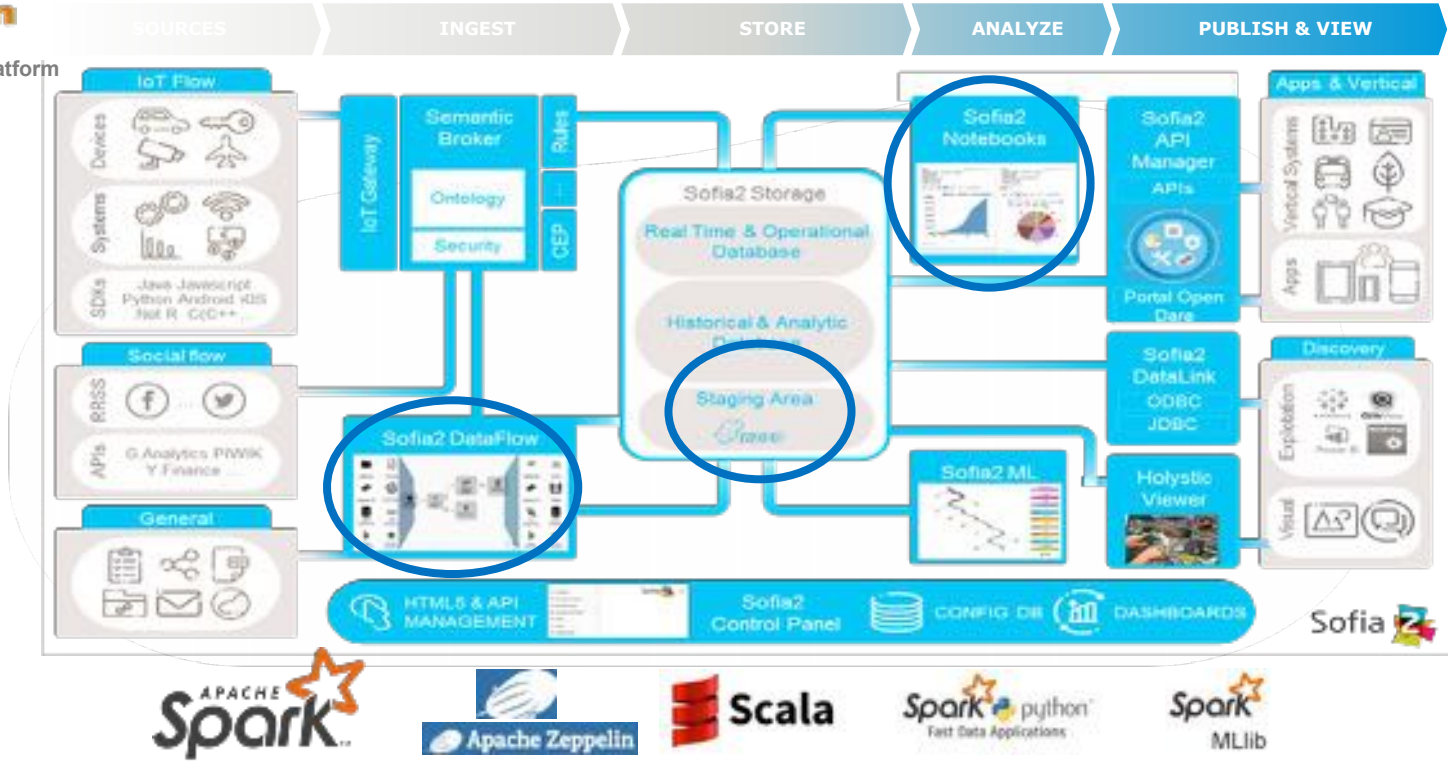
“Typical” pre-processing of SAFE packages, application of geo-biophysical algorithms into Analysis Ready Data



Big data environment

tmforum
Open Digital Ecosystem Platform 2016 Award



Sofia 2

http://sofia2-analytic.cloudapp.net/demowebenergiav2/login.html

Conversion of EO intermediate products into sparse matrix

Images are converted into a sparse matrix. This procedure is optimized thanks to Spark RDD parallelized structure taking few milliseconds to be carried out.

```
select * from total_area where Orbit = "R051" and Tile = "T31TBG"
```

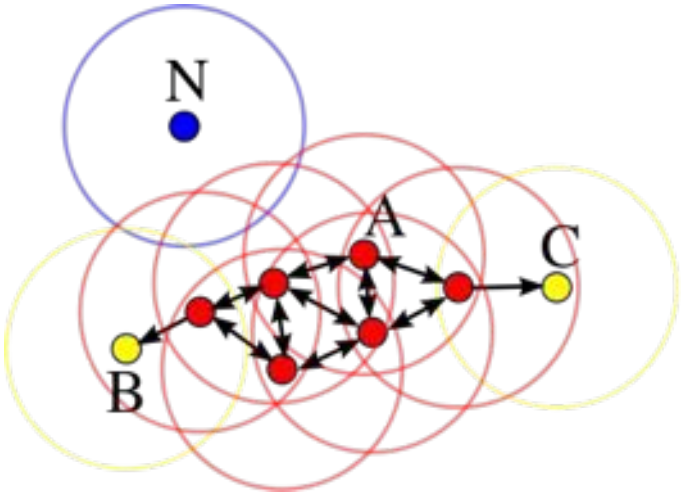
File	Date	Orbit	Tile	Area	TotalArea	Perc
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170426T105031_N0204_R051_T31TBG_20170426T105321.parquet	20170426	R051	T31TBG	3524.0	1.205604E8	2.92301618110
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170506T105031_N0205_R051_T31TBG_20170506T105029.parquet	20170506	R051	T31TBG	845240.0	1.205604E8	0.00701092564
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170516T105031_N0205_R051_T31TBG_20170516T105322.parquet	20170516	R051	T31TBG	574912.0	1.205604E8	0.00476866367
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170526T105031_N0205_R051_T31TBG_20170526T105518.parquet	20170526	R051	T31TBG	845665.0	1.205604E8	0.00701445084
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170605T105031_N0205_R051_T31TBG_20170605T105303.parquet	20170605	R051	T31TBG	542470.0	1.205604E8	0.00449957033
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170625T105031_N0205_R051_T31TBG_20170625T105322.parquet	20170625	R051	T31TBG	449783.0	1.205604E8	0.00373076897
/GSTPLAND/PROCESS/TRATAMIENTO/NDWI_S2A_MSIL2A_20170705T105031_N0205_R051_T31TBG_20170705T105605.parquet	20170705	R051	T31TBG	830296.0	1.205604E8	0.00688697117

Analytics: catalogue of algorithms

Algorithms: extended catalogue to choose from

- **Classification:** Logit, CART, SVC, NNC, …
- **Clustering:** K-Means, Hierarchical Clustering, Optimized Mixed Clustering
- **Regression:** Gaussian processes, Relevance Vector Machine, XGBoost...
- **Dimension reduction:** PCA, MCA, Rotation Varimax, …
- **Anomaly detection**: 1-class SVM, Attribute bagging, Fuzzy Logic, …
- **Attribute importance** Random Forest FR, Logit AI, Lasso Tuning, …
- **Association rules**: Apriori, Eclat, FP Growth, …
- **Natural language:** NER, Dirichlet Topic Modeling, Neural Network Classification…
- **Graphs mining:** Frequent Subgraph Mining, Lynk Analysis, Path-Based Algorithm
- **Deep Learning:** Convolutional Networks, Recurrent Networks

Case of use: DBSCAN CLUSTER algorithm for detection of features

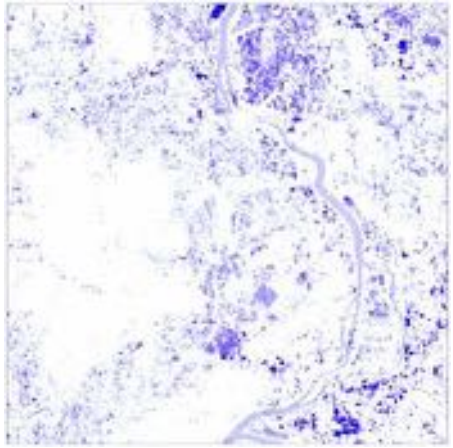


- A = **core point** (the main feature);
- B = **border point** (points located at the border of the main feature, having less than "x" neighboring points)
- N = **noise point** (Any point that is not within the cluster core nor in the border

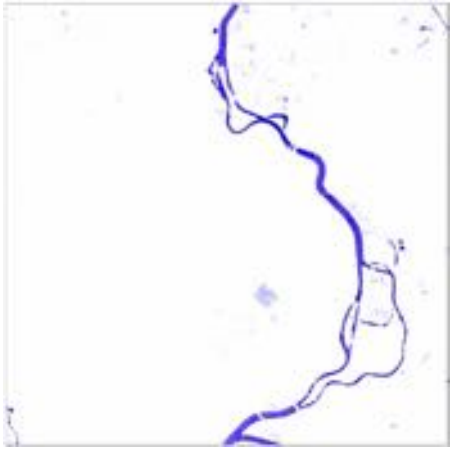
- **Core points** are given a water body code
- **Noise points** can be compared with ancillary data for a better discrimination
- A threshold can be applied to detect **border points** that eventually belong to a certain water body

Demonstration on water

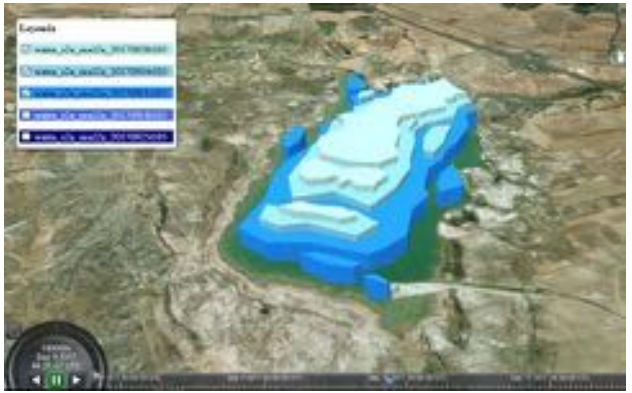
- **Basic statistics per pixel**: Water occurrence-persistence, Water occurrence change intensity, flood frequency, Water seasonality, etc.
- **Detection and naming of water bodies** (a feature will be given the same code according to time series analysis)



Wetness presence index



Water presence index



Dynamic mapping of water bodies

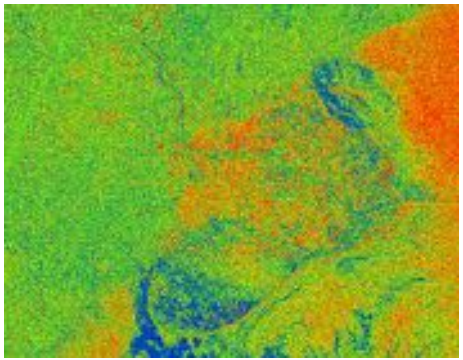
https://development.onesaitplatform.com/web/demo_webmap_big_data_spain_2018_rc2_mod__B/index.html

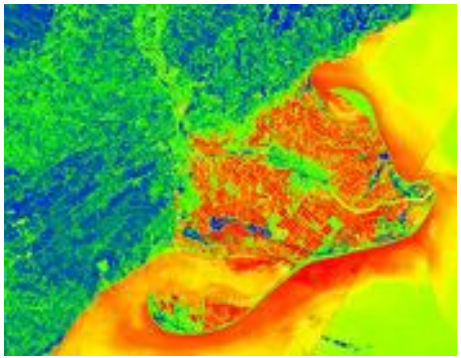
Predictive models example

Training with ancillary data + application of the prediction model

To use the predictive model with dynamic or static variables:

- Example: tide calendar + weather events + inundated area = prediction of inundated area when tide and weather events coincide





Industrial view

- Three models:
  - **Static model** (final result to be downloaded or acceded)
  - **On-demand requests** acceding remotely to hive tables by simple querying (area or time)
  - **Dynamic Water Body Monitroing** (A new image is downloaded, processed and the product published in a few minutes -> wetlands layer is updated continuously)
- Two types of clients:
  - Institutional
  - Private

Client: SQL querying