

# BIG DATA AND CLOUD PLATFORMS

---

Cluster migration - Based on a true story

# Migration

## Goals

- Evaluating the costs for a cloud/on-premises data platform
- Real-world case study
- Fill in this table

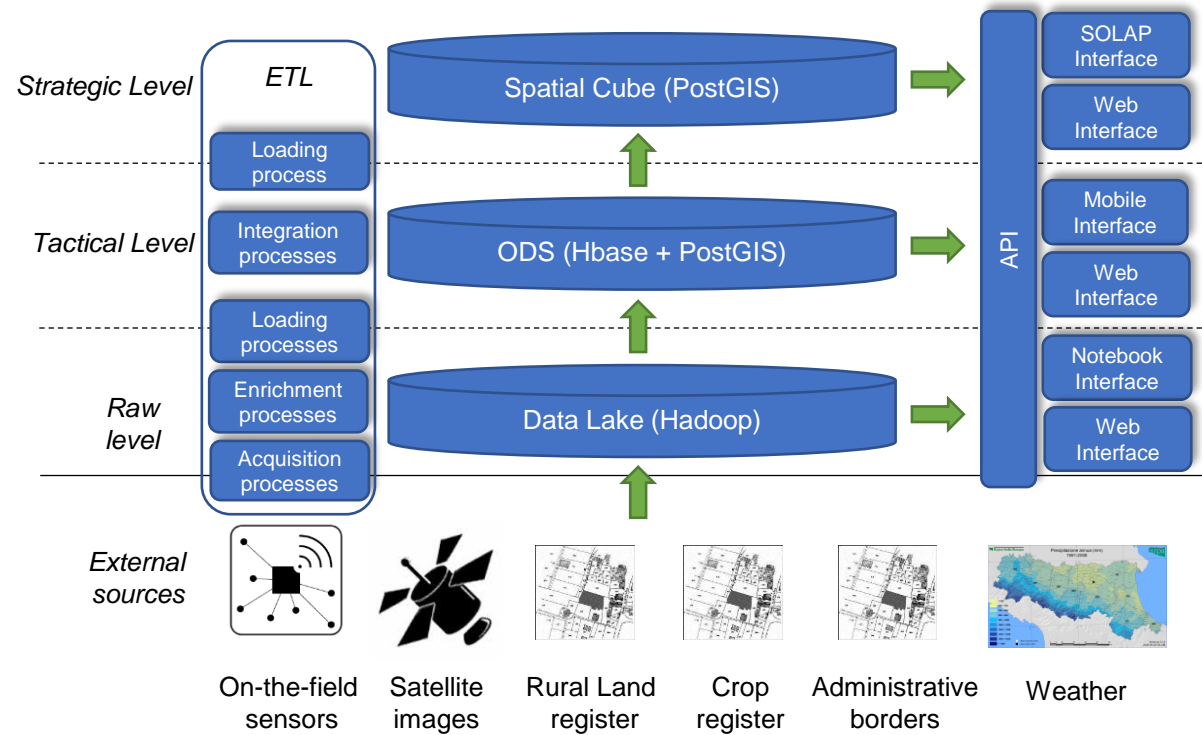
Cost	On-premises	On cloud
Hardware	?	?
Software	?	?

# Case study

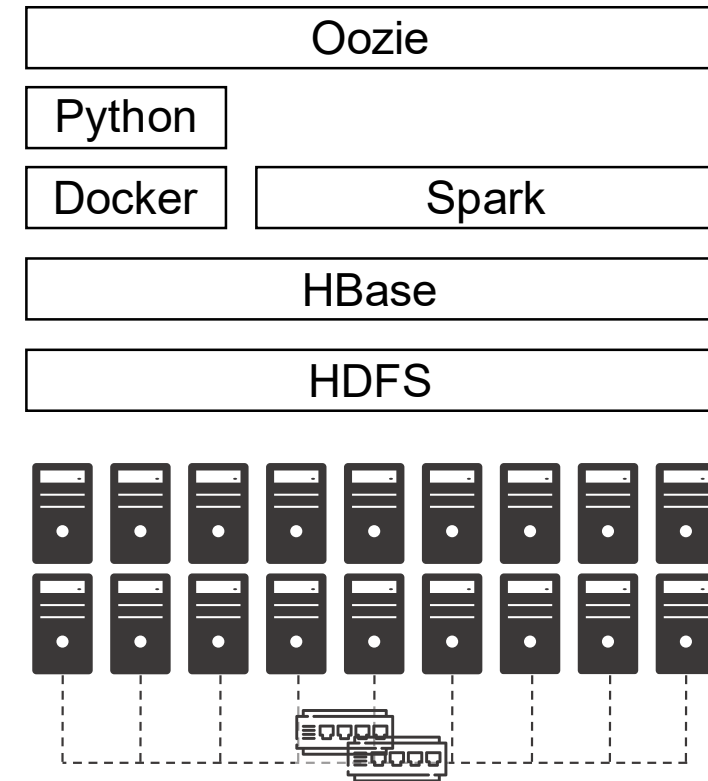
Business intelligence group

# Migration

## Reference architecture

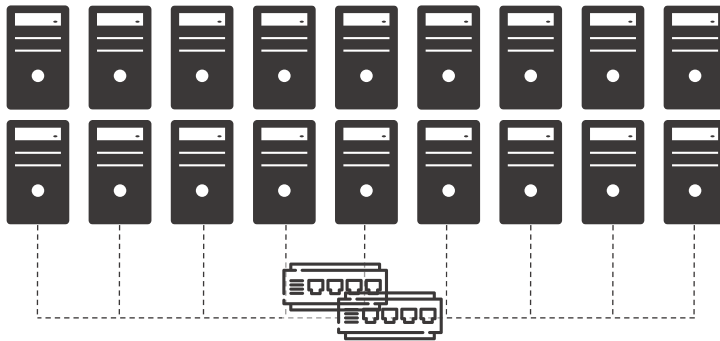


## Hadoop Cluster



# Migration

## Hardware



8 CPUs (144 total)  
- Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz  
32GB RAM (576GB total)  
- 2 x 16GB DIMM DDR4 2666 MHz  
12TB HDD Disk (216TB total)  
- 3 x 4TB ST4000DM004-2CV1

```
lshw -short -C cpu  
lshw -short -C memory  
lshw -short -C disk
```

## Software

- "Classic" Hadoop stack

# Migration

<b>SOL<sub>onprem</sub></b>	<b>On-premises</b>	<b>On cloud</b>
Hardware	?	?
Software	?	?

## Hardware cost: ?

- Refer to <https://www.rect.coreto-europe.com/en/search.html?clearsearch=1>

# On-premises

**Hardware cost (up to Mar 05, 2021):**

$1767\text{€} \times 18 = 31806\text{€}$

- Amortization over 3 years (i.e.,  $10602\text{€/year}$ )

SOL <sub>onprem</sub>	On-premises	On cloud
Hardware	$10602\text{€/year}$	?
Software	?	?

**RECT™ WS-2270C**  
Main configuration **669.00** €  
**Configuration:**  
Intel Core i7-10700K + 216.00  
32 GB DDR4-3200 RAM + 146.00  
Workstation-Mainboard with ... + 101.00  
3 x 4 TB WD Blue 291.00  
1x 2.5 Gbit LAN onboard  
Sound on board  
Solid black  
High-Efficiency Noctua CPU ... + 39.00  
DVD-Writer 24x DVD + 13.00  
High-efficiency 750W power ... + 79.00  
1 x Your Operating System 30.00  
with an individual capacity of... + 35.00  
36 months pick-up + 148.00  
**Complete Configuration 1,098.00**  
**Current price 1,767.00**  
[Plus VAT](#)   
[Leasingraten](#)  
  **Add to cart**

<https://www.rect.coreto-europe.com/en> (Accessed 2021-08-01)

# On-premises

<b>SOL<sub>onprem</sub></b>	<b>On-premises</b>	<b>On cloud</b>
Hardware	10602€/year	?
Software	?	?

**Software cost: ?**



# On-premises

SOL <sub>onprem</sub>	On-premises	On cloud
Hardware	10602€/year	?
Software	0€	?

## Software cost (up to 2020): 0€

- Free Cloudera Management System
- No software licensing (for research purpose)

# On-premises

SOL <sub>onprem</sub>	On-premises	On cloud
Hardware	10602€/year	?
Software	180000€/year	?

**Software cost** (up to Mar 05, 2021): 10000€/year x 18 = 180000€/year

- Cloudera is no more free, 10K€ per node
- <https://www.cloudera.com/products/pricing.html#private-cloud-services>
- <https://www.cloudera.com/products/pricing/product-features.html>
- No license for research purpose

*“Houston we’ve had a problem!”*

- We cannot update/extend the cluster anymore
- What about migrating to the cloud? (we only consider AWS)

# Migration

Moving a Hadoop cluster to the cloud (we only consider AWS)

- AWS price calculator <https://calculator.aws/#/estimate>

How do we start?

- We have already defined the hardware and the software stack
- Start with coarse tuning, identify the dominating costs first
  - Is it computing, storage, or processing?
- Identify a suitable budget, implement, refine later
  - Wrong refinements can do a lot of damage

# On cloud v1

<b>SOL<sub>cloud1</sub></b>	<b>On-premises</b>	<b>On cloud</b>
Hardware	10602€/year	?
Software	180000€/year	?

## Migrating the cluster as-is: ?

- Hint: add 18 EC2 instances satisfying the hardware requirements

# On cloud v1

SOL <sub>cloud1</sub>	On-premises	On cloud
Hardware	10602€/year	162000\$/year
Software	180000€/year	?

SOL<sub>cloud1</sub> migrating the cluster as-is:  
13500\$/month = 162000\$/year

- 18 EC2 instances (t4g.2xlarge) with 12TB EBS storage each machine
- Still, we have no software configuration

**Amazon EC2**  
Region: EU (Ireland)

EditAction ▼

**Quick estimate**

Operating system (Linux), Quantity (18), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (12 TB), Instance type (t4g.2xlarge)

Monthly: 13,499.30 USD

**Amazon EC2**  
Region: EU (Milan)

EditAction ▼

**Quick estimate**

Operating system (Linux), Quantity (18), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (12 TB), Instance type (t3.2xlarge)

Monthly: 14,785.47 USD

<https://calculator.aws/#/estimate?id=7757affccc3cafdcfdeb212b74623ef02ed5a36>

# Migration

## Pay attention to the region

- Different regions, different prices
- Different regions, different services
- Remember the GDPR and data locality

<b>Amazon EC2</b> Region: EU (Ireland)	<a href="#">Edit</a>	<a href="#">Action ▼</a>
<b>Quick estimate</b>		
Operating system (Linux), Quantity (18), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (12 TB), Instance type (t4g.2xlarge)		Monthly: 13,499.30 USD
<b>Amazon EC2</b> Region: EU (Milan)	<a href="#">Edit</a>	<a href="#">Action ▼</a>
<b>Quick estimate</b>		
Operating system (Linux), Quantity (18), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (12 TB), Instance type (t3.2xlarge)		Monthly: 14,785.47 USD

# Migration

It makes no sense to move the cluster as-is

- More machines ensure better (on-prem) scalability but higher costs

How do we proceed with the migration?

- We need minimum software requirements
- Try to achieve the smallest migration impact
  - Find the most similar cloud-based solution to a Hadoop cluster
  - Rethink applications (later) when you got the know-how
- Identify a suitable budget, implement, refine later
  - Wrong refinements can do a lot of damage

# Migration

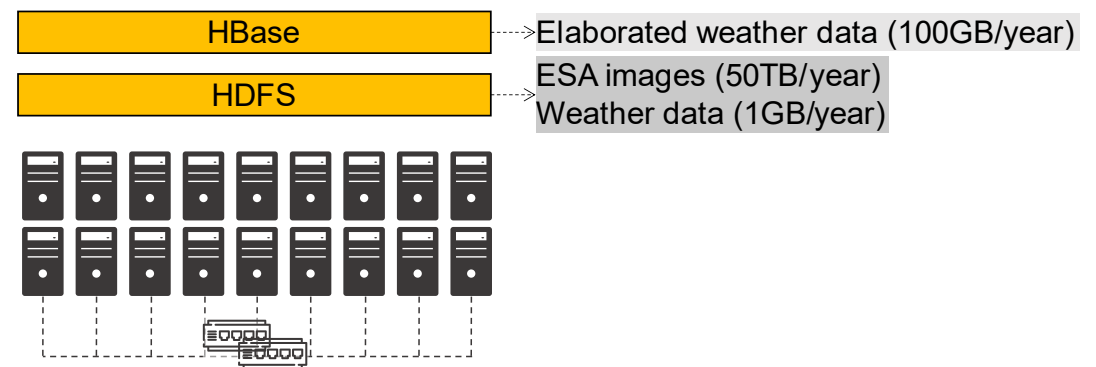
## HDFS

- How much durability do we need?
  - $HP_0$ : three replicas (we stick to this)
  - $HP_1$ : decrease replicas for cold data
  - $HP_2$ : move cold data to glacier or delete id
  - ...

**HBase** has marginal effects on the pricing (100GB  $\ll$  50TB)

- For simplicity, we can omit it

**Overall:** 50TB storage/year





# Migration

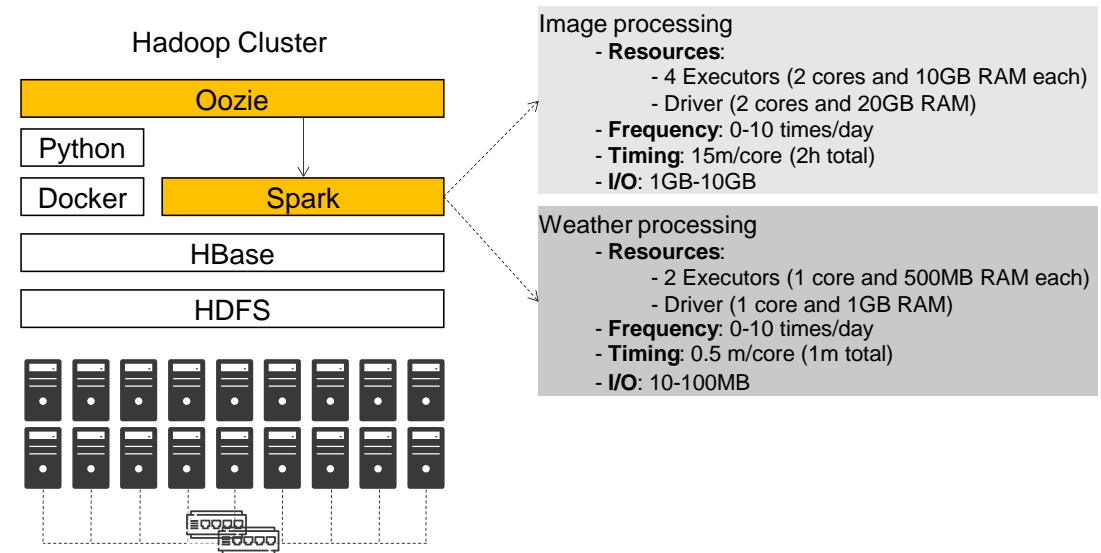
Processing takes place each time that ESA provides a satellite image

- Some days no images are available
- Some days up to 10 images are available
- Spark jobs are always executed with the same parameters

## Image processing

- 4 machines, 2 cores, 10GB RAM at least

## Weather processing is negligible



# On cloud v2

	On-premises	On cloud
Hardware	2356€/year	38000\$/year
Software	100000€/year	?

## Assuming 1 Executor = 1 Machine

- Compare 4 machines on-premises vs on cloud

## On-premises

- 4 machines: 10602€/year / 18 machines x 4 machines = 2356€/year
- Cloudera requires at least 10 nodes: 100000€/year

## AWS

- 4 EC2 instances: 162000\$/year / 18 machines x 4 machines = 36000\$/year
  - Plus the resources for master services = 2000\$/year
- Problems
  - Still no software stack
  - A lot of storage cost
  - Machines are up-and-running even when no computation is necessary (just to persist data)

# On cloud v2

## AWS

- Still, we have no software stack configuration
- Which is the major cost?

# On cloud v2

## AWS

- Still, we have no software stack configuration
- Which is the major cost?

**Amazon EC2**

ModificaOperazione ▼

**Regione:** US East (Ohio)

**Quick estimate**

Operating system (Linux), Quantity (1), Pricing strategy (EC2 Instance Savings Plans 1 Year No Upfront), Storage amount (12 TB), Instance type (t4g.2xlarge)

Monthly: 676,04 USD

**Amazon EC2 stima**

Amazon EC2 Instance Savings Plans instances (monthly)	123,08 USD
Amazon Elastic Block Storage (EBS) pricing (monthly)	552,96 USD
<b>Costo mensile totale:</b>	<b>676,04 USD</b>

# Migration

## S3 standard

### Unit conversions

S3 Standard storage: 50 TB per month x 1024 GB in a TB = 51200 GB per month

### Calcolo dei prezzi

Tiered price for: 51200 GB

51200 GB x 0.0230000000 USD = 1177.60 USD

Costo totale del piano = 1177.6000 USD (S3 Standard storage cost)

1.000 PUT requests for S3 Storage x 0,000005 USD per request = 0,005 USD (S3 Standard PUT requests cost)

1.000 GET requests in a month x 0,0000004 USD per request = 0,0004 USD (S3 Standard GET requests cost)

1.177,60 USD + 0,0004 USD + 0,005 USD = 1.177,61 USD (Total S3 Standard Storage, data requests, S3 select cost)

**S3 Standard cost (monthly): 1,177.61 USD**

## S3 Infrequent Access

### Unit conversions

S3 One Zone-IA storage: 50 TB per month x 1024 GB in a TB = 51200 GB per month

### Calcolo dei prezzi

51.200 GB x 0,01 USD = 512,00 USD (S3 One Zone-IA storage cost)

1.000 PUT requests for S3 One Zone-IA Storage x 0,00001 USD per request = 0,01 USD (S3 One Zone-IA PUT requests cost)

1.000 GET requests for S3 One Zone-IA Storage x 0,000001 USD per request = 0,001 USD (S3 One Zone-IA GET requests cost)

1.000 lifecycle request count for S3 One Zone-IA x 0,00001 USD per request = 0,01 USD (S3 One Zone-IA lifecycle requests cost)

10 GB x 0,01 USD = 0,10 USD (S3 One Zone-IA data retrievals cost)

512,00 USD + 0,01 USD + 0,001 USD + 0,01 USD + 0,10 USD = 512,121 USD (Total S3 One Zone-IA Storage and other costs)

**S3 One Zone - Infrequent Access (S3 One Zone-IA) cost (monthly): 512.12 USD**

# Migration

## AWS Storage

### HDFS on EC2

- Heavy price
- Machine must be always on to guarantee data persistency
- Data locality

### S3

- Much cheaper
- Does not require machines for data storage
- Data locality is lost

**Amazon EC2** Modifica Operazione ▼  
Regione: US East (Ohio)

**Quick estimate**  
  
Operating system (Linux), Quantity (1),  
Pricing strategy (EC2 Instance Savings Plans  
1 Year No Upfront), Storage amount (12 TB),  
Instance type (t4g.2xlarge)  
  
Monthly: 676,04 USD

**Amazon EC2 stima**

Amazon EC2 Instance Savings Plans instances (monthly)	123,08 USD
Amazon Elastic Block Storage (EBS) pricing (monthly)	552,96 USD
<b>Costo mensile totale:</b>	<b>676,04 USD</b>

# Migration

	On-premises	On cloud
Hardware	2356€/year	
Software	100000€/year	?

Migrating cluster to EMR: ?

Given the software requirements, we need

- 1 x Master Node (to manage the cluster)
- 1 x Core node (with HDFS/EBS)
- 4 x Task Nodes (to compute)

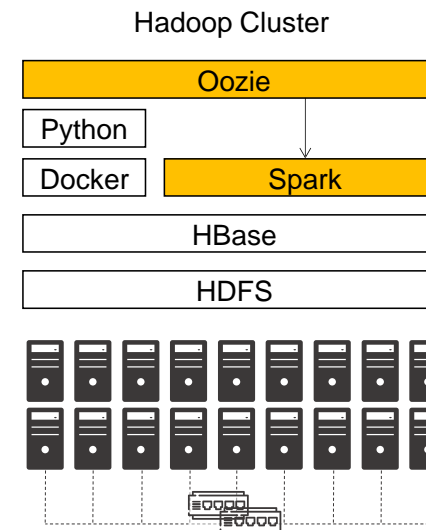


Image processing

- **Resources:**

- 4 Executors (2 cores and 10GB RAM each)
- Driver (2 cores and 20GB RAM)

- **Frequency:** 0-10 times/day

- **Timing:** 15m/core (2h total)

- **I/O:** 1GB-10GB

Weather processing

- **Resources:**

- 2 Executors (1 core and 500MB RAM each)
- Driver (1 core and 1GB RAM)

- **Frequency:** 0-10 times/day

- **Timing:** 0.5 m/core (1m total)

- **I/O:** 10-100MB

# On cloud v3

	On-premises	On cloud
Hardware	2356€/year	14710€/year
Software	100000€/year	

## Migrating cluster to EMR: 14710€/year

- S3 Infrequent Access storage (50 TB per month): 640€
- 1 x Master EMR nodes, EC2 (m4.xlarge), Utilization (75 h/month): 4.5€
  - 75 h/month = 15min/task x 10task/day x 30day/month / 60min/hour
- 1 x Core EMR nodes, EC2 (m4.xlarge), Utilization (75 h/month): 4.5€
- 4 x Task EMR nodes, EC2 (m4.4xlarge), Utilization (75 h/month): 72€
- 4 x EC2 on demand (task node): 174.83€
  - Storage amount (30 GB)
  - Workload (Daily, Duration of peak: 0 Hr 15 Min)
  - Instance type (m4.xlarge)
- 2 x EC2 on demand (master and core nodes): 330€
  - Storage amount (30 GB)
  - Instance type (m4.xlarge)



# On cloud v3

	On-premises	On cloud
Hardware	2356€/year	13445€/year
Software	100000€/year	

## Migrating cluster to EMR: 13445€/year

- S3 Infrequent Access storage (50 TB per month): 640€
- 1 x Master EMR nodes, EC2 (m4.xlarge), Utilization (75 h/month): 4.5€
  - 75 h/month = 15min/task x 10task/day x 30day/month / 60min/hour
- 1 x Core EMR nodes, EC2 (m4.xlarge), Utilization (75 h/month): 4.5€
- 4 x Task EMR nodes, EC2 (m4.4xlarge), Utilization (75 h/month): 72€
- 4 x EC2 **spot (task node): 69.55€**
  - Storage amount (30 GB)
  - Workload (Daily, Duration of peak: 0 Hr 15 Min)
  - Instance type (m4.xlarge)
- 2 x EC2 on demand (master and core nodes): 330€
  - Storage amount (30 GB)
  - Instance type (m4.xlarge)

<https://calculator.aws/#/estimate?id=c3780b12bb43b593d05def5a1d5218d9764b8a65>

# Migration

Summing up (cloud options)

Machine uptime	Storage	Software	Feasible?	Cost per year
Constant	EC2	Manual	YES: but high storage cost	~36K€
Constant	EC2	EMR	YES: but high storage cost	~37K€
Constant	S3	Manual	YES: but still manual provisioning	~17K€
Constant	S3	EMR	YES	~18K€
Pay-per-use	EC2	Manual	NO: pay-per-use + EC2 = Data unpersisted	-
Pay-per-use	EC2	EMR	NO: pay-per-use + EC2 = Data unpersisted	-
Pay-per-use	S3	Manual	ISH: repetitive manual provisioning	-
Pay-per-use	S3	EMR	YES	~14K€

# Migration

## Summing up

- We estimated the cluster costs
  - On-premises solution with 18 machines: no go
  - Cloud solution with 18 EC2 instances: no go
- We reduced the solution based on software requirements
  - On-premises solution with 4 machines: no go
  - Cloud solution with 4 EC2 instances: no go, we miss the software configuration
- We moved the cluster to AWS EMR + spot instances + S3 storage

## Can we do better?

- Pick ad-hoc cloud services (AWS Lambda e AWS Batch)
- ... to re-think the applications (food for thoughts)