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

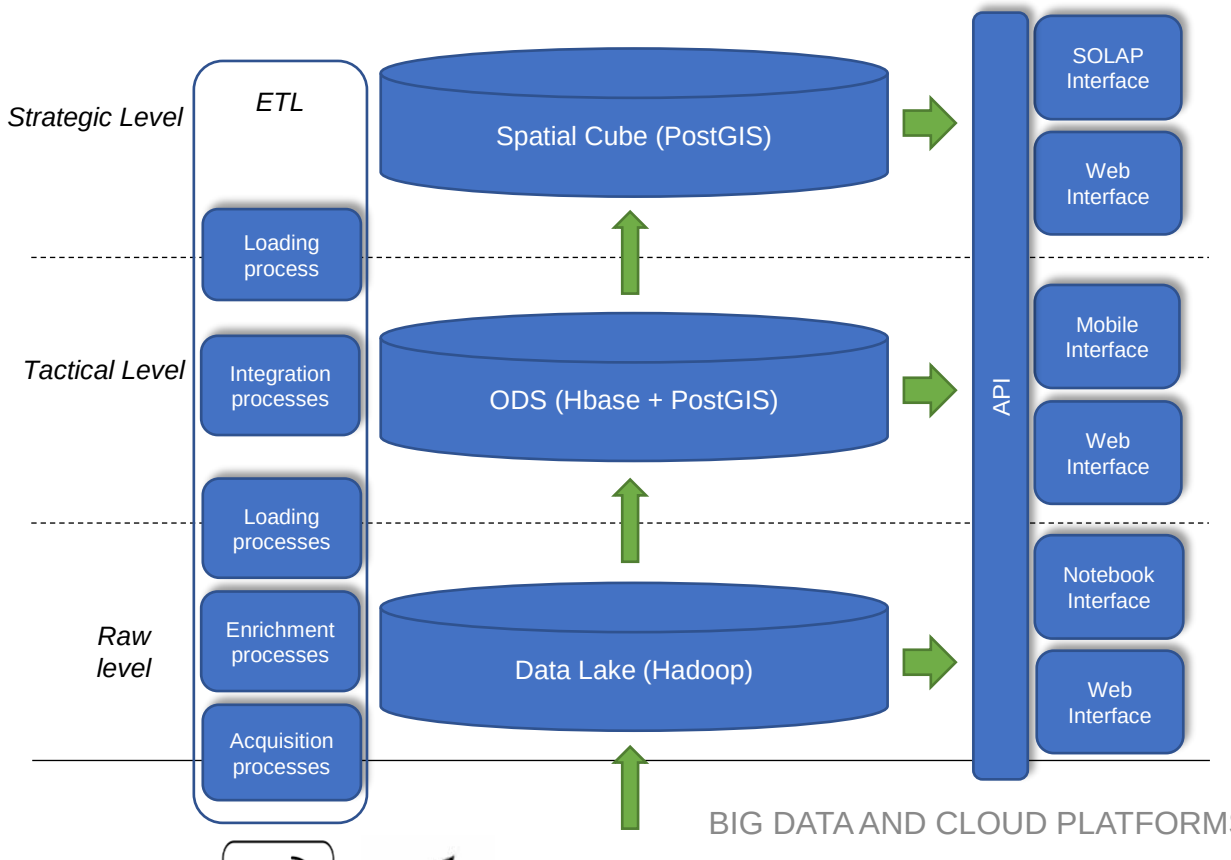
Cost	On-premises	On cloud
Hardware	?	?
Software	?	?

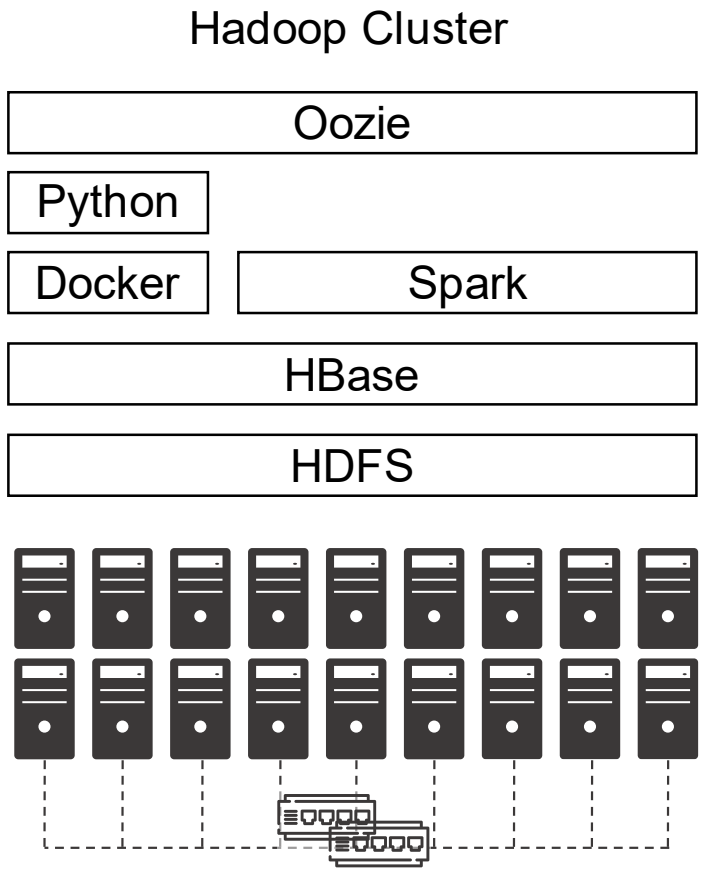
Case study

Business intelligence group

Migration

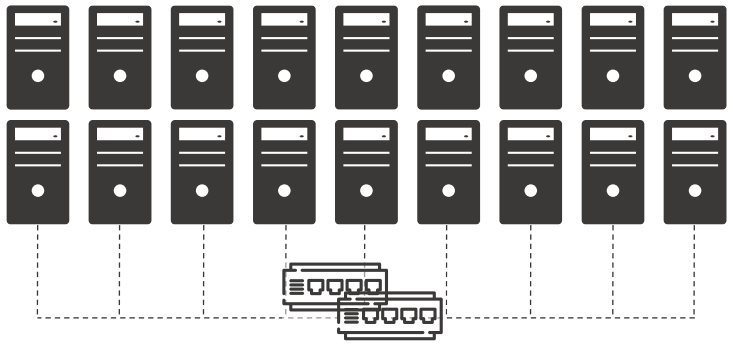
Reference architecture





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

SOL_{onprem}	On-premises	On cloud
Hardware	?	?
Software	?	?

Hardware cost: ?

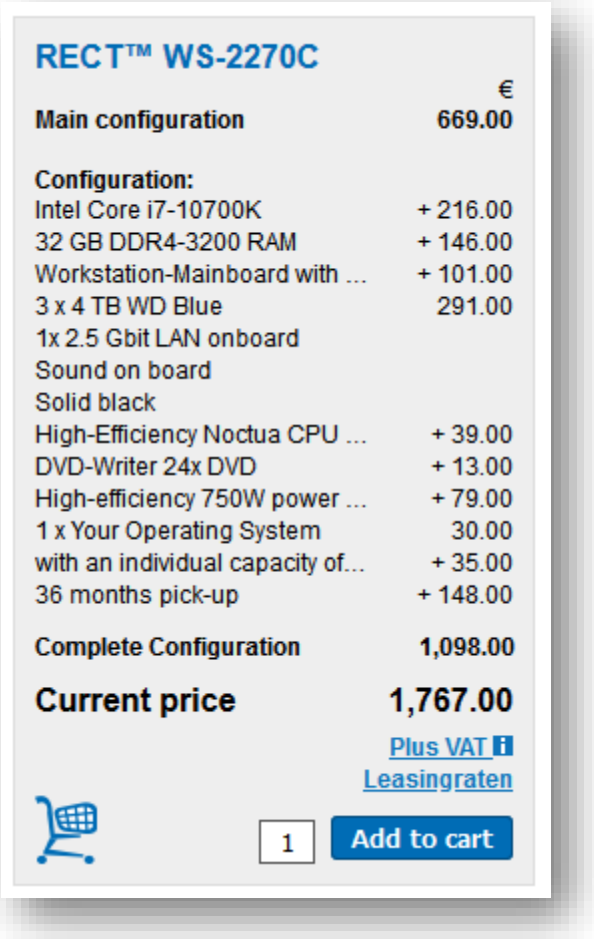
- Refer to https://www.rect.coreto-europe.com/en/search.html?clearsearch=1

On-premises

SOL_{onprem}	On-premises	On cloud
Hardware	10602€/year	?
Software	?	?

Hardware cost (up to Mar 05, 2021):
1767€ x 18 = 31806€

- Amortization over 3 years (i.e., 10602€/year)



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

On-premises

SOL_{onprem}	On-premises	On cloud
Hardware	10602€/year	?
Software	?	?

Software cost: ?

On-premises

SOL_{onprem}	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_{onprem}	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

SOL_{cloud1}	On-premises	On cloud
Hardware	10602€/year	?
Software	180000€/year	?

Migrating the cluster as-is: ?

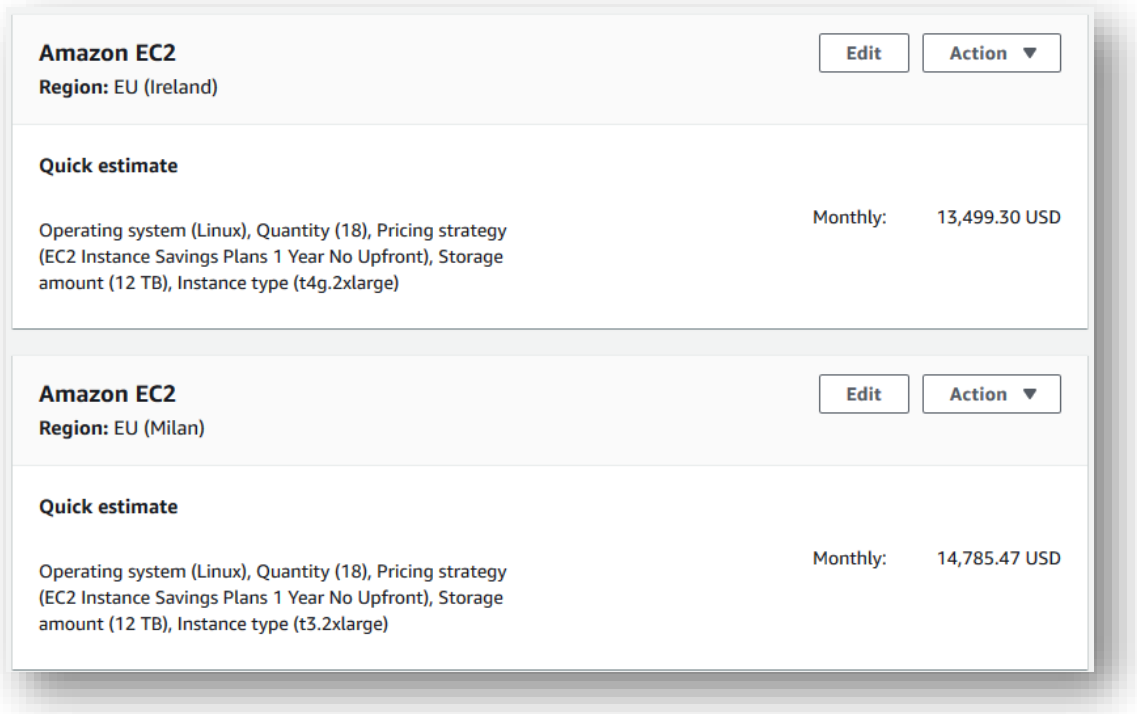
- Hint: add 18 EC2 instances satisfying the hardware requirements

On cloud v1

SOL_{cloud1}	On-premises	On cloud
Hardware	10602€/year	162000$/year
Software	180000€/year	?

SOL_{cloud1} 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

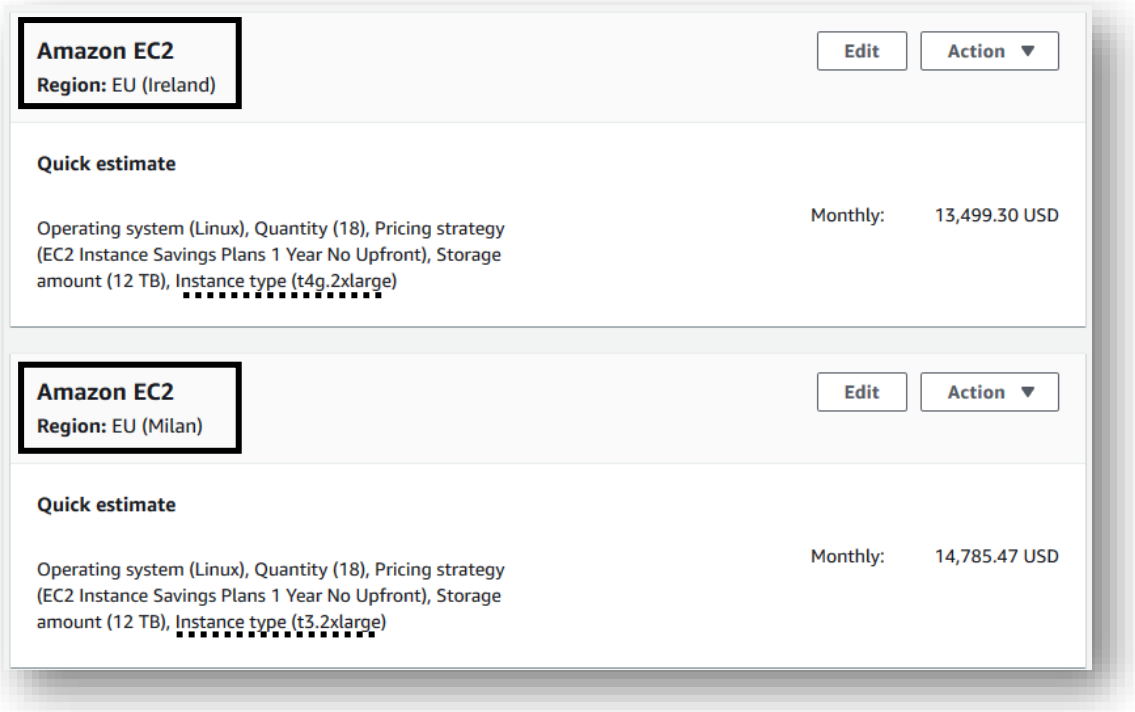


https://calculator.aws/#/estimate?id=7757afffccc3cafdcfdeb212b74623ef02ed5a36

Migration

Pay attention to the region

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



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

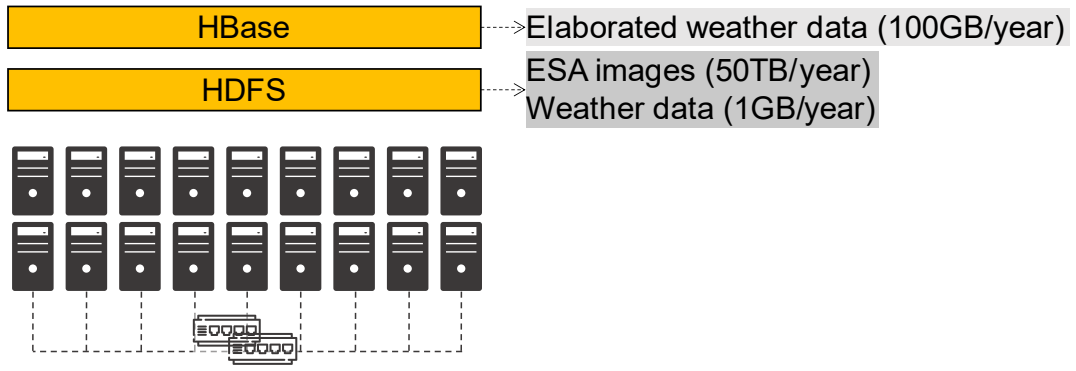
HDFS

- 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 << 50TB)

- For simplicity, we can omit it

Overall: 50TB storage/year



Migration

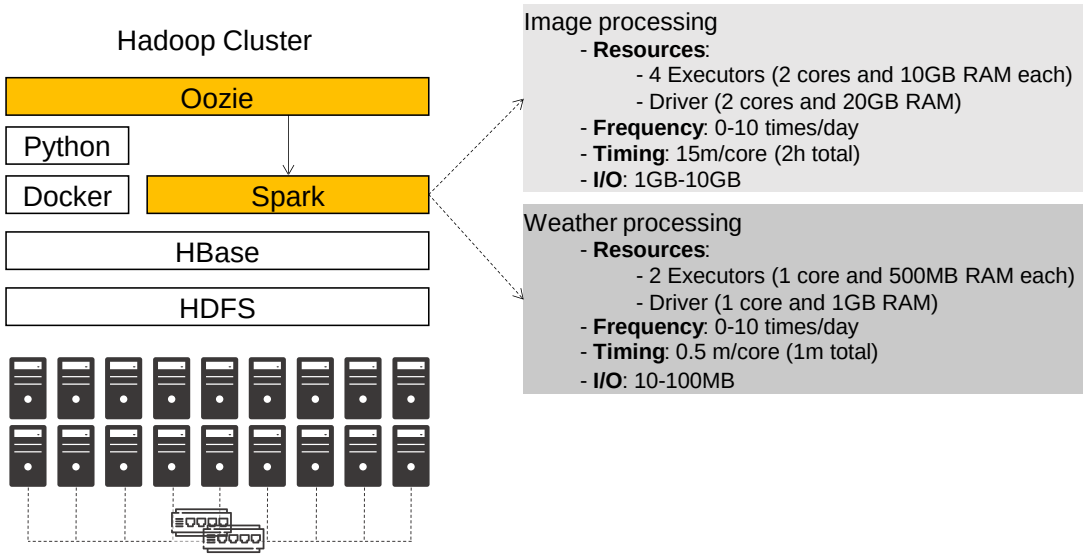
Processing takes place each time that ESA provides a satellite image

- 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

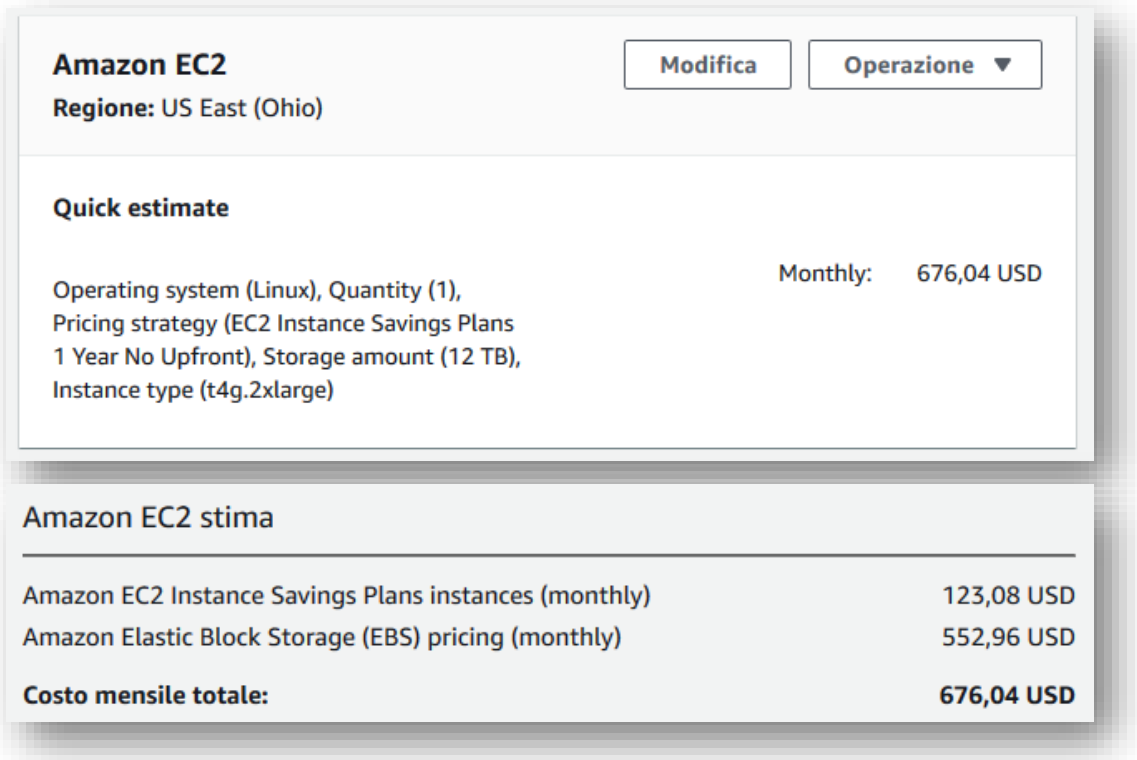
AWS

- 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?



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

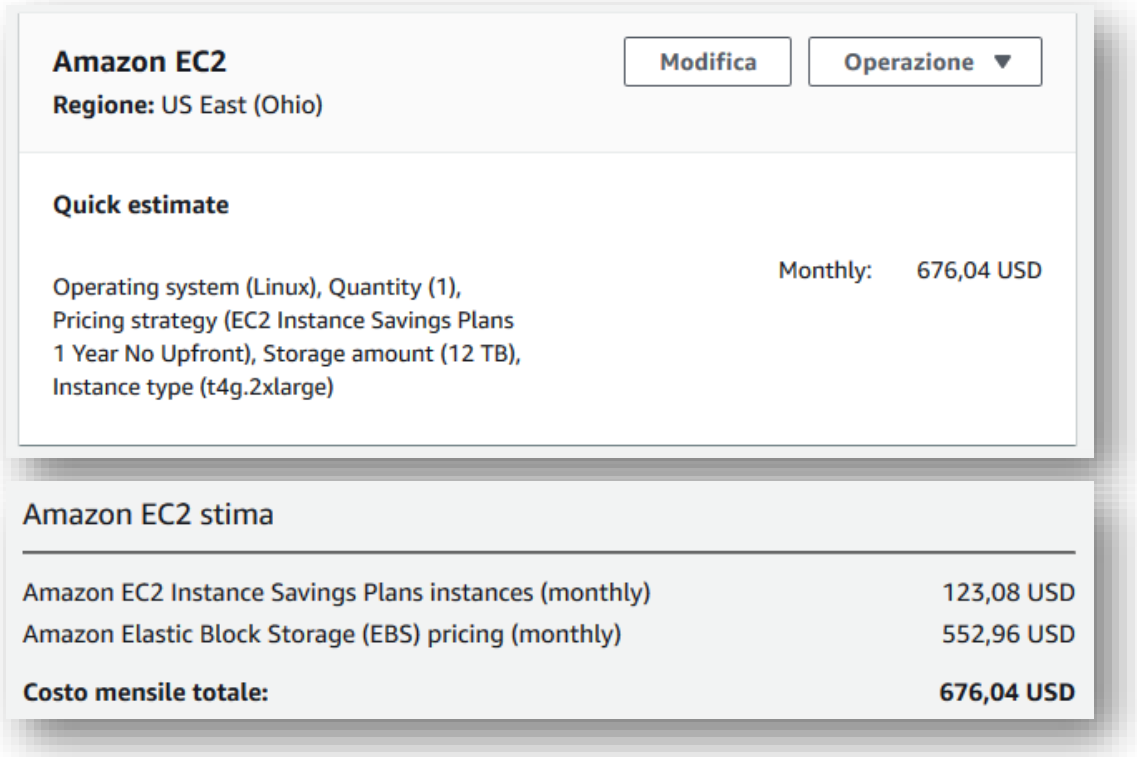
AWS Storage

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



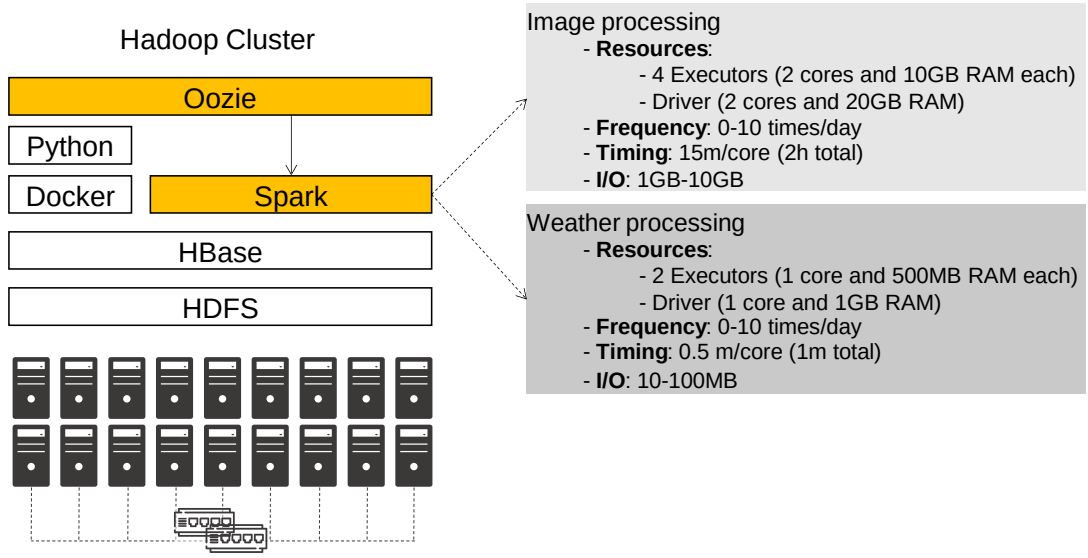
Migration

	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)



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)

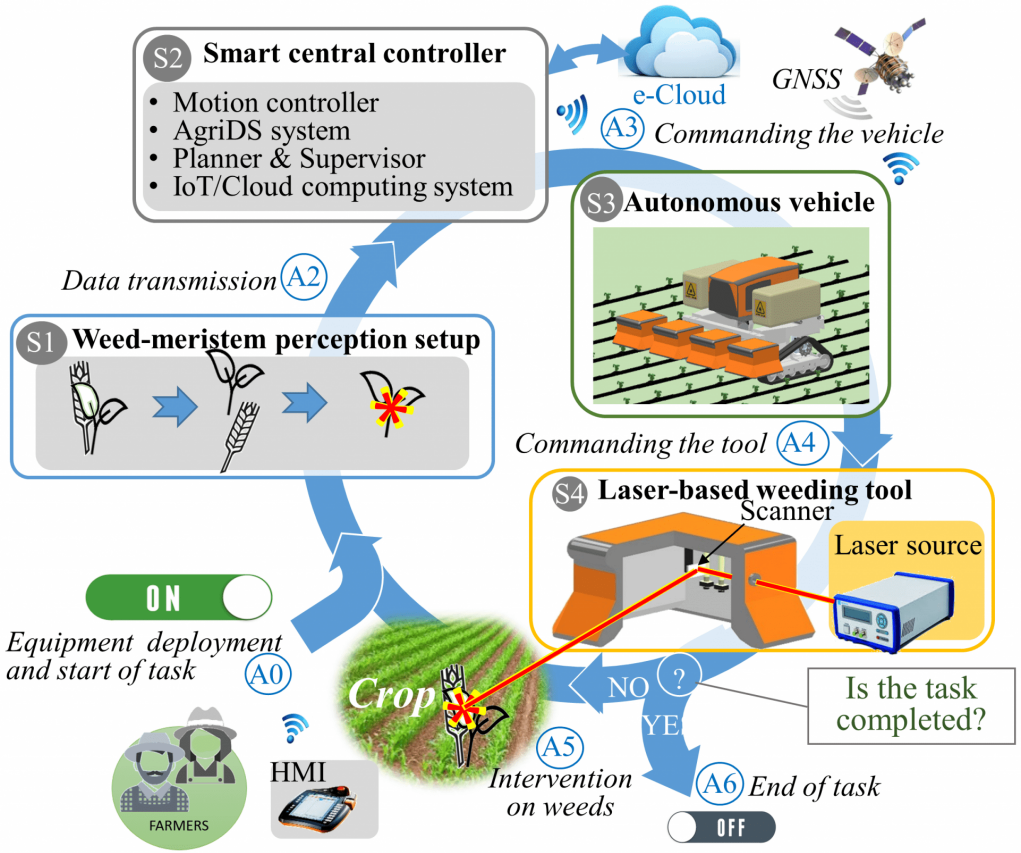
Case study

WeLASER

The WeLASER project

Project description

The increased use of pesticides and fertilisers damages the environment, destroys non-target plants and beneficial insects for the soil and harms human and animal health. Most seeds develop herbicide-resistant properties, rendering pesticides ineffective. Mechanical automatic systems that are studied as alternatives to pesticides deteriorate soil features, damage beneficial soil organisms and offer limited results for in-row weeding. The EU-funded WeLASER project will develop a non-chemical solution for weed management based on pioneering technology consisting of the application of lethal doses of energy on the weed meristems through a high-power laser source. An AI-vision system separates crops from weeds, identifying the weed meristems and pointing the laser at them. A smart controller based on IoT and cloud computing techniques coordinates the system, which is transfered all over the field by an autonomous vehicle.



https://cordis.europa.eu/project/id/101000256 (accessed 2020-08-01)

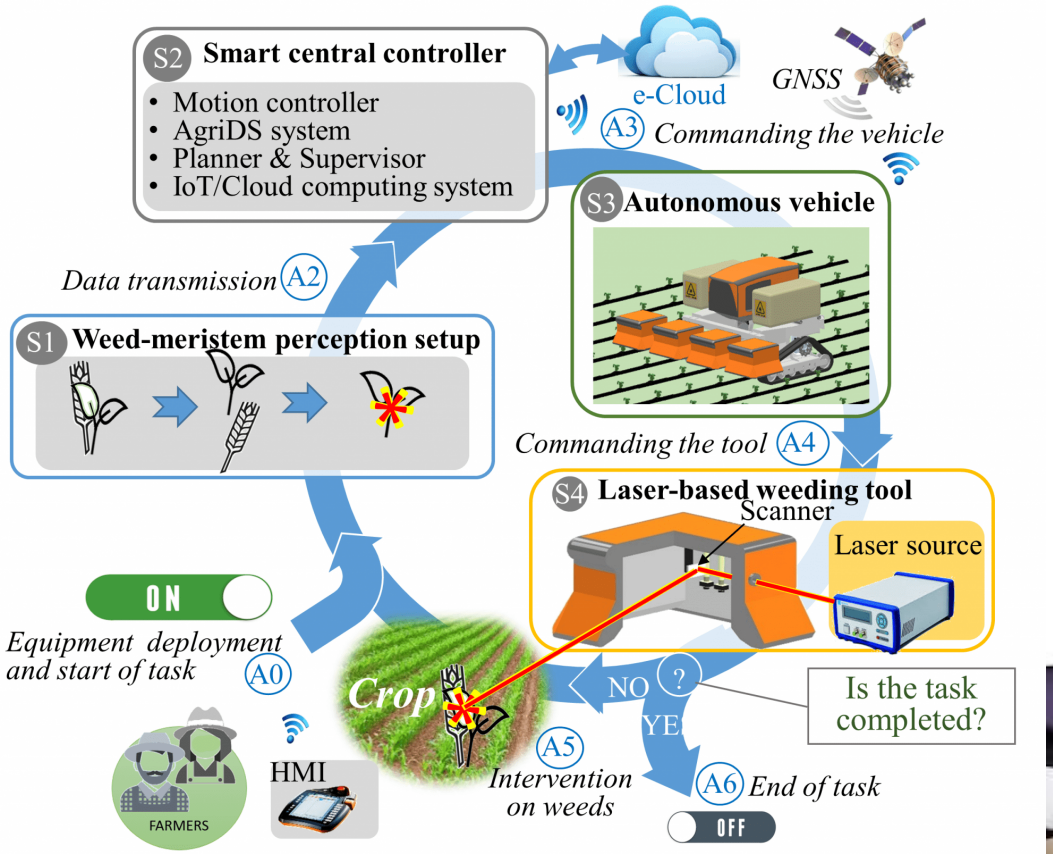
The WeLASER project

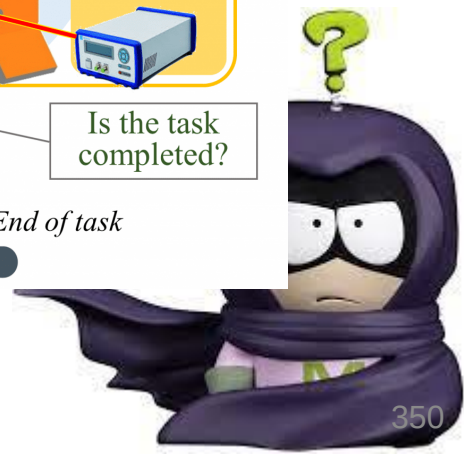
Which requirements do you foresee?

Can we define a tentative (service) architecture for the WeLASER project?

Assumptions

- Do not consider the collection of weed/crop images & training/deploying of the CV algorithm





https://cordis.europa.eu/project/id/101000256 (accessed 2020-08-01)

Data sources

- 8 cameras on the field, average image size 3MB; two services:
  - Alerting: a camera sends an image when smt enters the field (at most 1 img/5 min); assuming 20 alerts/day
    - $20\frac{alert}{day} \cdot 1\frac{image}{alert} \cdot 8\frac{camera}{field} = 160\frac{image}{day \cdot field}$
    - $160\frac{image}{day \cdot field} \cdot 3\frac{MB}{image} = 480\frac{MB}{day \cdot field} \sim 500\frac{MB}{day \cdot field}$
  - Monitoring the crop/weed; assuming 2 images/day
    - $2\frac{image}{day} \cdot 8\frac{camera}{field} = 16\frac{image}{day \cdot field}$
    - $16\frac{image}{day \cdot field} \cdot 3\frac{MB}{image} = 48\frac{MB}{day \cdot field} \sim 50\frac{MB}{day \cdot field}$
- Weather station: 24 measurement/day for humidity/solar radiation/temperature/wind; 1KB/measurement
  - $24\frac{sample}{day} \cdot 4\frac{measurement}{sample} \cdot 1\frac{weather\ station}{field} = 96\frac{measurement}{day \cdot field}$
  - $96\frac{measurement}{day \cdot field} \cdot 1\frac{KB}{measurement} = 96\frac{KB}{day \cdot field} \sim 0.1\frac{MB}{day \cdot field}$
- Robot mission: lasts 4 hours, 100 measurement/s from sensor systems; 1KB/measurement
  - $3600\frac{second}{hour} \cdot 4\frac{hour}{mission} \cdot 100\frac{measurement}{second} \cdot 1\frac{mission}{day \cdot field} = 1\,440\,000\frac{measurement}{day \cdot field}$
  - $1\,440\,000\frac{measurement}{day \cdot field} \cdot 1\frac{KB}{measurement} = 1.44\ \frac{GB}{day \cdot field} \sim 2\frac{GB}{day \cdot field}$
- Historic data; worst case of 1KB/document: 10^3 Json documents describing the farm $\sim 1\frac{MB}{farm}$

https://docs.google.com/spreadsheets/d/17zEr62CzyqeIy0vU-DcjEUoxf6bMd3ziLSSeIXvk4Lg/edit?usp=sharing

Workload

Nothing special

- Every night compute aggregated indexes on the collected data (2h/day)

On-premises (HDFS cluster)

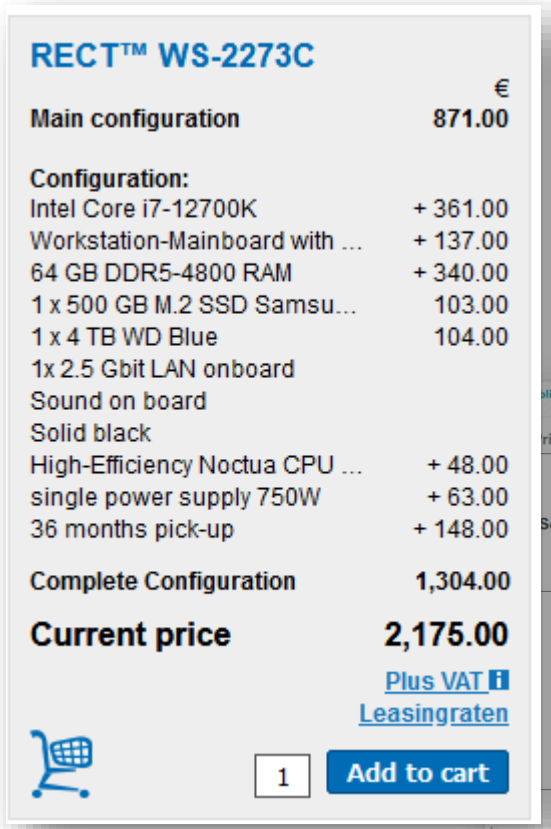
- How many machines do we need?
- With which resources?

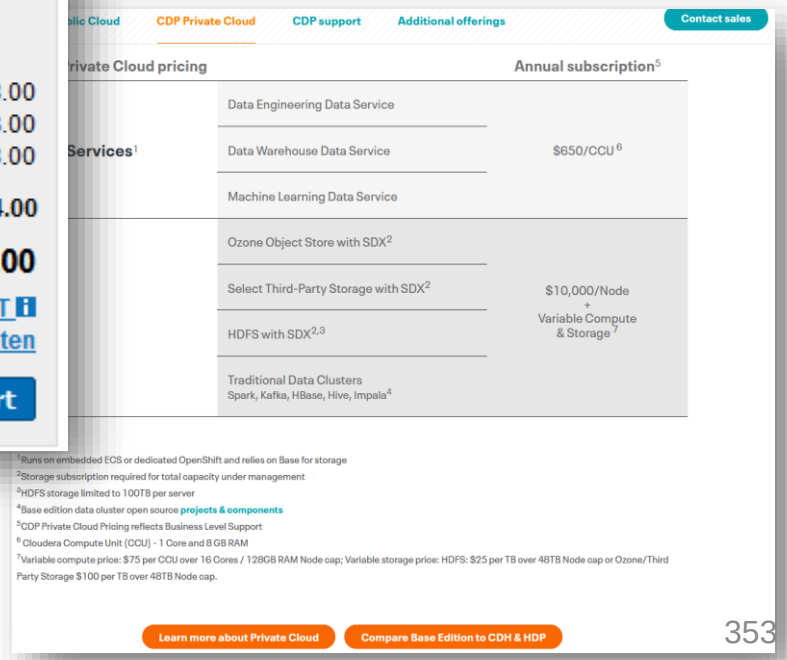
On-premises

	On-premises	On cloud
Hardware	2900€/year	?
Software	40000€/year	

On-premises

- How many machines do we need?
  - 4: 1 master node + 3 HDFS data nodes
- With which resources?
  - Assuming a HDFS replication factor of 3, we need at least 1TB of disk overall (not that much)
  - Think bigger: at least 8 cores, 64GB RAM, 500GB SSD + 4TB HDD, no GPU
- 8700€ / 3 years = 2900€





https://www.rect.coreto-europe.com/en (accessed 2022-09-01)
https://www.cloudera.com/products/pricing.html (accessed 2022-09-01)

On cloud v1

	On-premises	On cloud
Hardware	2900€/year	~40000$/year
Software	40000€/year	

Moving the Hadoop cluster as IAAS

EC2

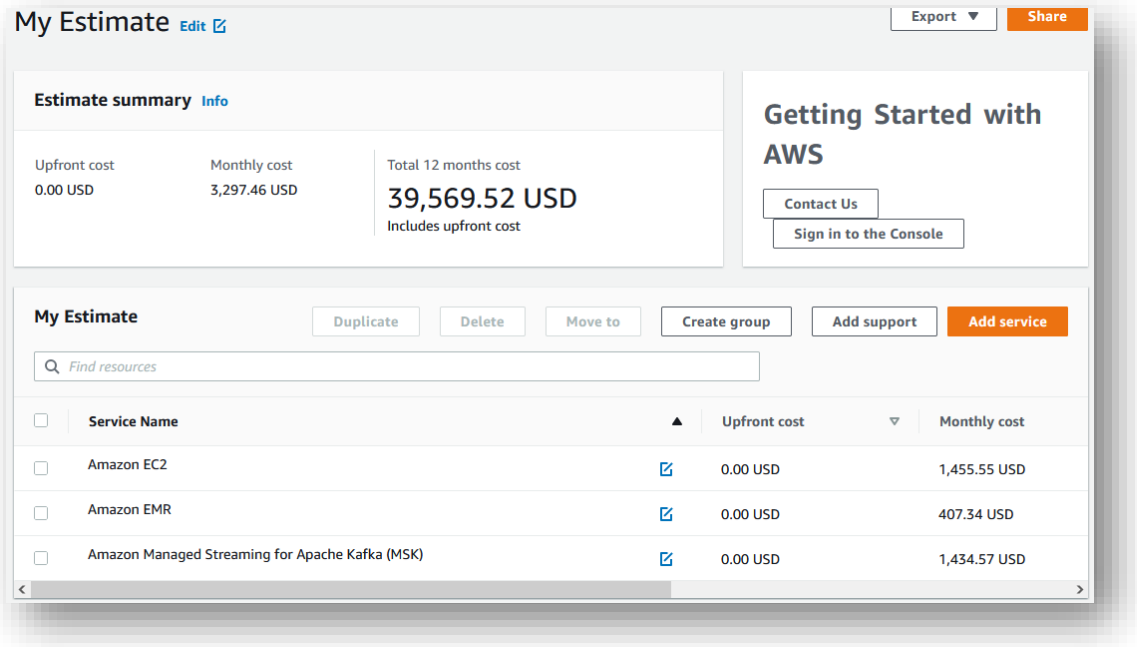
- Quantity (4), Pricing strategy (EC2 Instance Savings Plans 3 Year No Upfront), Storage amount (4 TB), Instance type (r6g.2xlarge)

EMR

- Number of master EMR nodes (1), EC2 instance (r5.2xlarge), Utilization (100 %Utilized/Month) Number of core EMR nodes (3), EC2 instance (r5d.2xlarge), Utilization (100 %Utilized/Month)

MKS (KAFKA)

- Storage per Broker (10 GB), Number of Kafka broker nodes (3), Compute Family (m5.2xlarge)



https://calculator.aws/#/estimate?id=05965ca7de23fd9e7d2ab2cd0175fe8c01822c9c (accessed 2022-09-01)

On cloud v2

	On-premises	On cloud
Hardware	2900€/year	~4000$/year
Software	40000€/year	

Moving the Hadoop cluster as PAAS

EC2

- Quantity (4), Pricing strategy (On-Demand Instances), Storage amount (30 GB), Instance type (r6g.2xlarge)

EMR

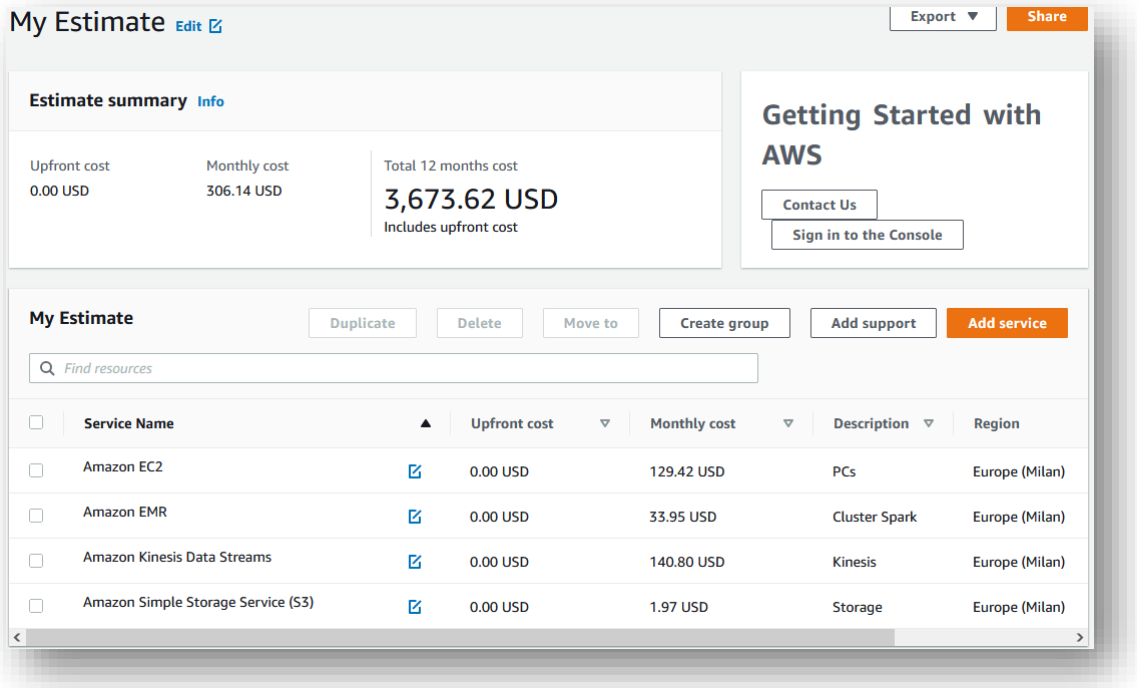
- Number of master EMR nodes (1), EC2 instance (r5.2xlarge), Utilization (2 Hours/Day) Number of core EMR nodes (3), EC2 instance (r5d.2xlarge), Utilization (2 Hours/Day)

S3

- Standard storage (60 GB per month)

Kinesis

- Days for data retention (1 days), Records (100 per second), Consumer Applications (3)



https://calculator.aws/#/estimate?id=53f60ff0412a18877dc8e1274f7d9875aa3bf665 (accessed 2022-09-01)

Cost vs price

How would you evaluate the cost and the price?

Cost vs price

Price is the amount a customer is willing to pay for a product or service

Cost is the expense incurred for creating a product or service

- Hardware
- Development
- Maintenance

Profit is the difference between price paid and costs incurred is profit

- If a customer pays $10 for a product that costs $6 to make and sell, the company earns $4