BIG DATA

Running a data platform

Migration

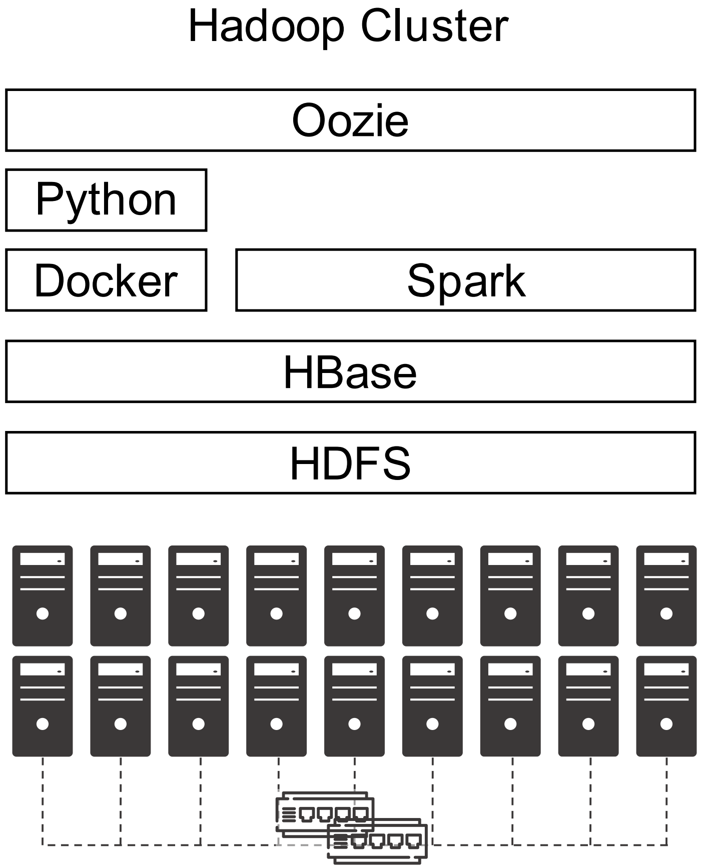
Goals

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

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

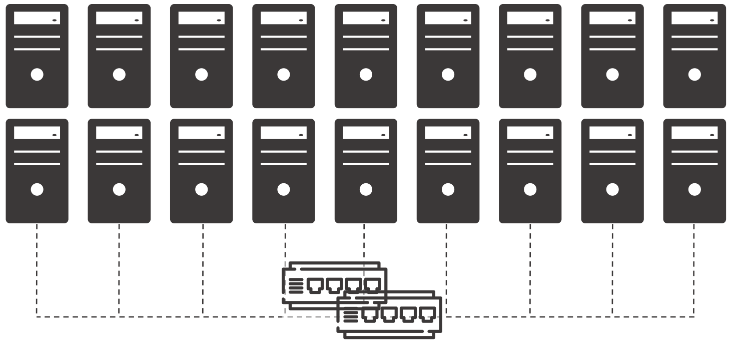
Migration

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: ?

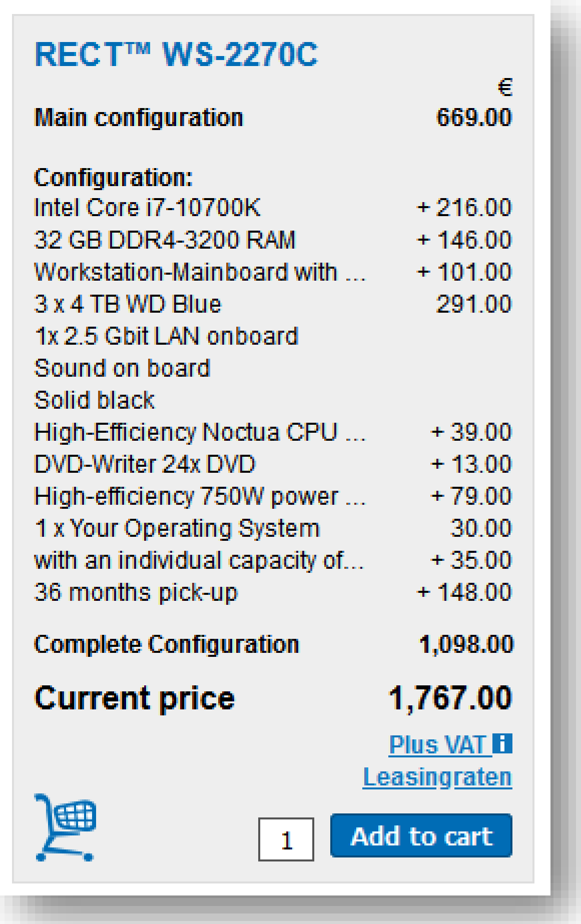
- Refer to https://www.rect.coreto-europe.com/en/search.html?clearsearch=1

Migration

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)



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

Migration

SOL_{onprem}	**On-premises**	**On cloud**
Hardware	10602€/year	?
Software	0€	?

Software cost: ?

Migration

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)

Migration

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

Migration

SOL$_{cloud1}$	**On-premises**	**On cloud**
Hardware	10602€/year	?
Software	180000€/year	?

Migrating the cluster as-is: ?

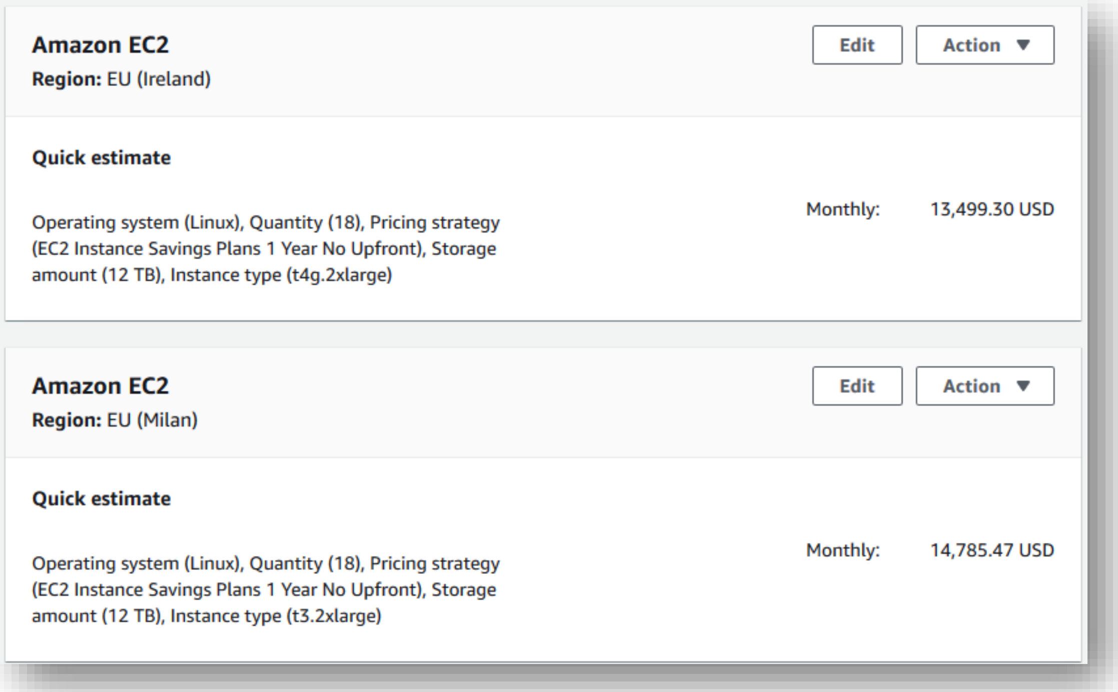
- Hint: add 18 EC2 instances satisfying the hardware requirements

Migration

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



Amazon EC2
Region: EU (Ireland)

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)

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)

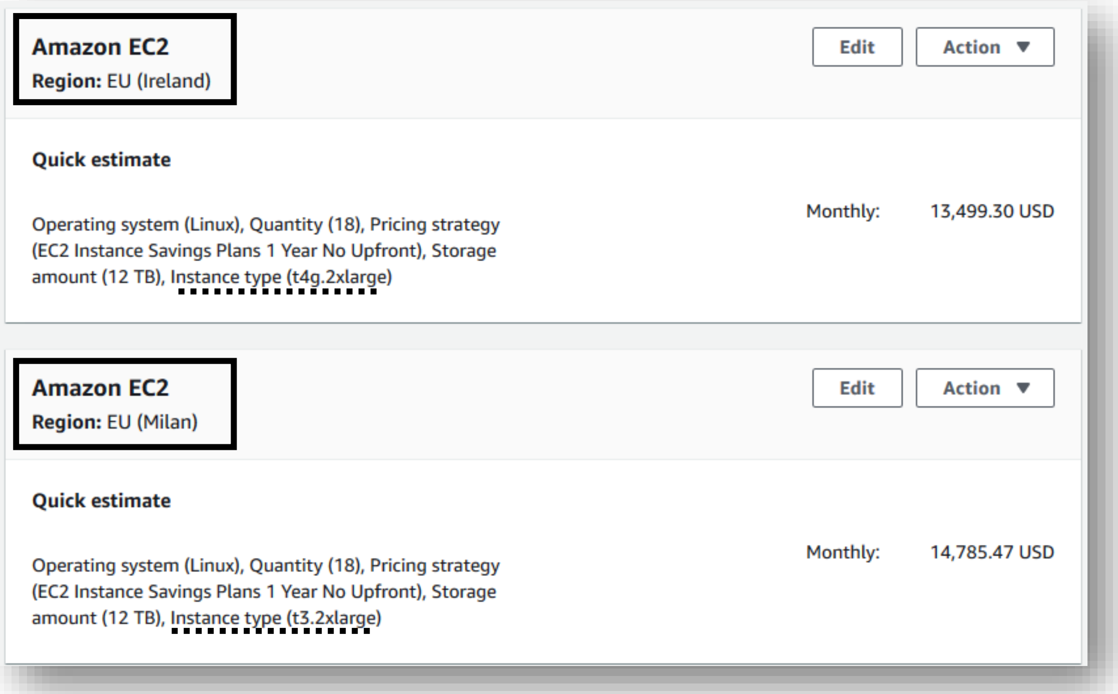
Monthly: 14,785.47 USD

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

Migration

Pay attention to the region

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



Amazon EC2
Region: EU (Ireland)

Quick estimate

Monthly: 13,499.30 USD

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

Amazon EC2
Region: EU (Milan)

Quick estimate

Monthly: 14,785.47 USD

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

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

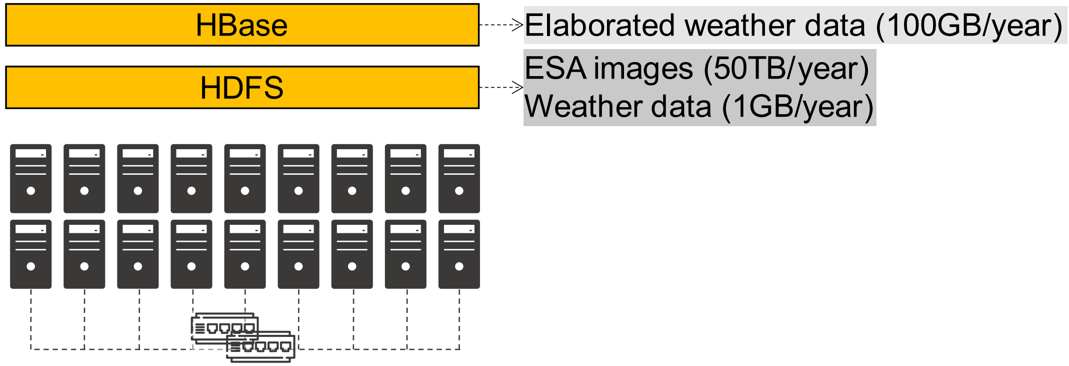
HDFS

- 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

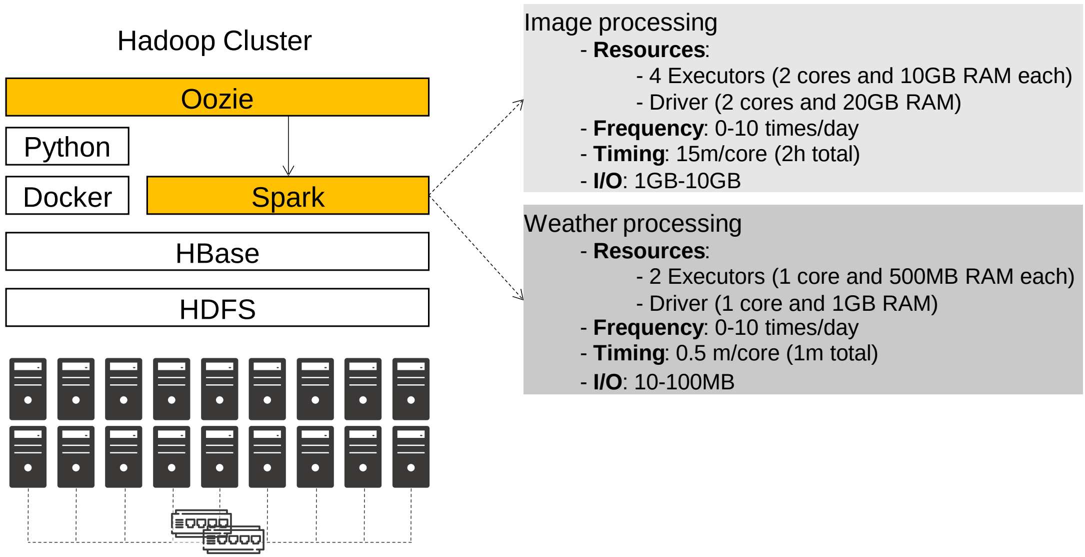


Image processing
- **Resources**:
  - 4 Executors (2 cores and 10GB RAM each)
  - Driver (2 cores and 20GB RAM)
- **Frequency**: 0-10 times/day
- **Timing**: 15m/core (2h total)
- **I/O**: 1GB-10GB

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

Migration

	On-premises	On cloud
Hardware	2356€/year	36000$/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

Migration

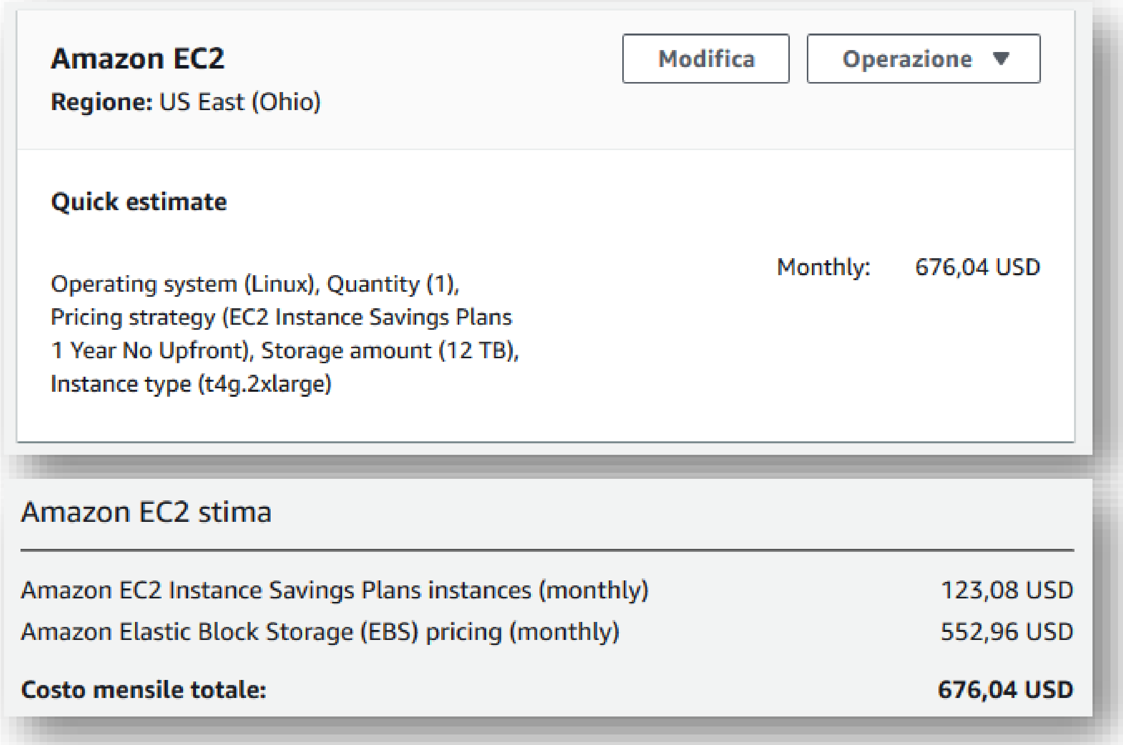
AWS

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

Migration

AWS

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



Amazon EC2

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)

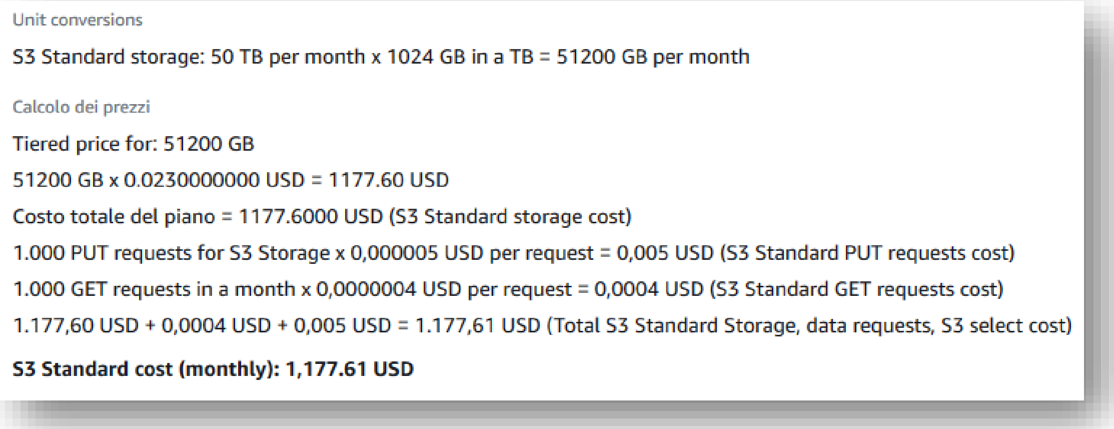
Monthly: 676,04 USD

Amazon EC2 stima

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

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)

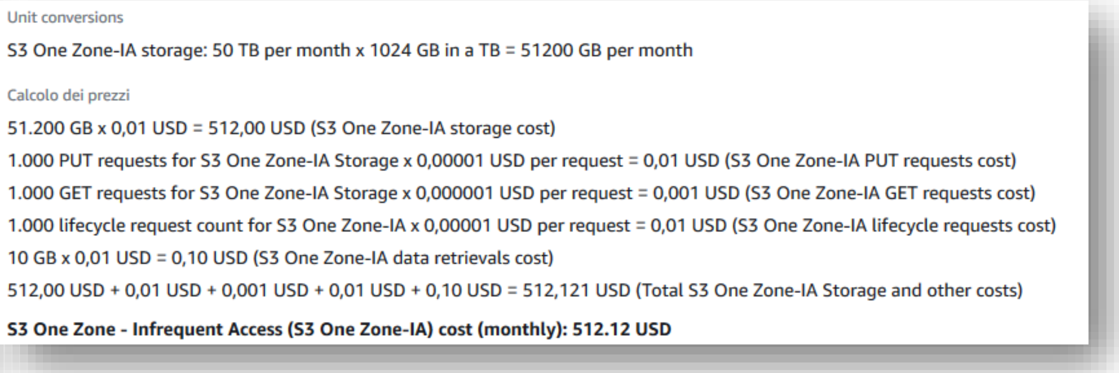
1.000 PUT requests for S3 Storage x 0,000005 USD per request = 0,005 USD (S3 Standard PUT requests cost)

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

Motivation

Amazon EMR (Elastic Map Reduce)

- Provides a managed Hadoop framework

Some features

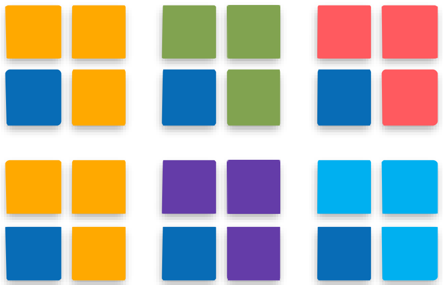
- Service integration
  - Automatically control EC2 instances
  - Transparently use S3 storage
- Pricing:
  - Low Hourly Pricing
  - Amazon EC2 Spot Integration

https://aws.amazon.com/emr

EMR Cluster

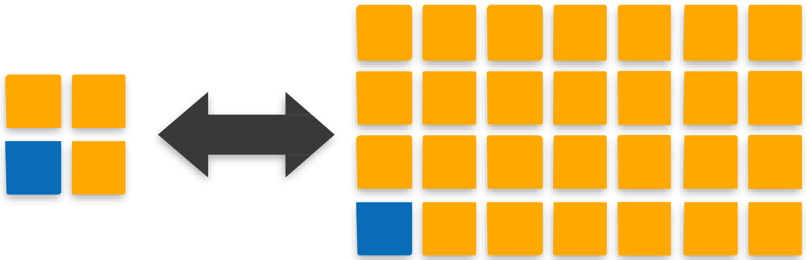
Provision as much capacity as you need

Deploy Multiple Clusters



Add or remove capacity at any time

Resize a Running Cluster



EMR Cluster

EMR cluster

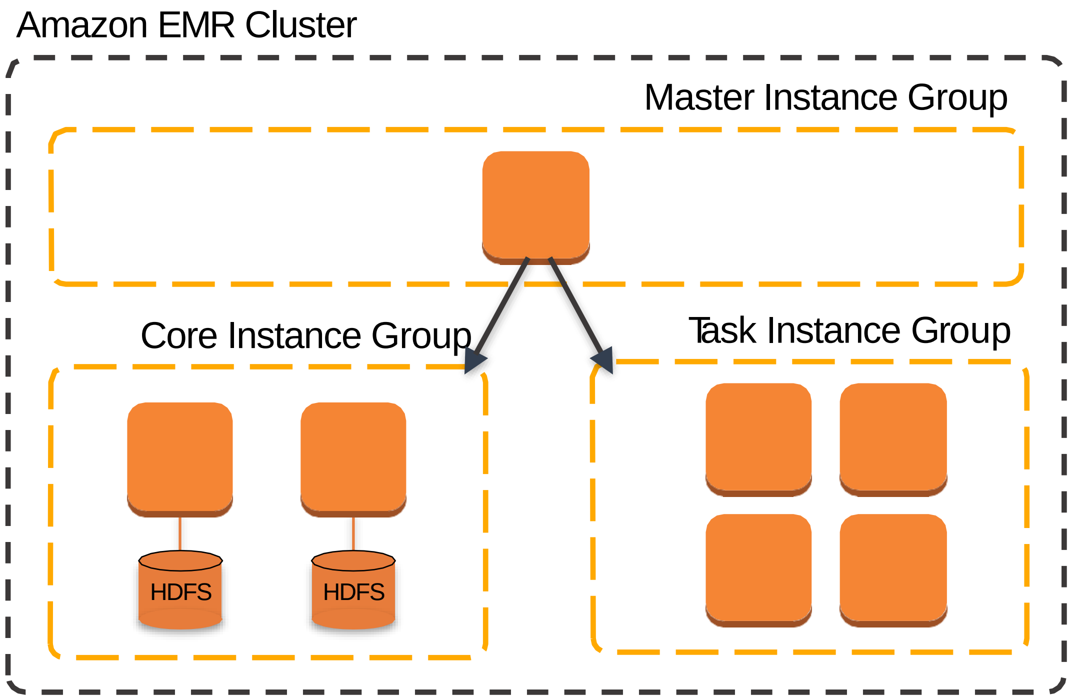
Master group controls the cluster

- Coordinate the work distribution
- Manage the cluster state

Core groups

- Core instances run Data Node daemons

(Optional) Task instances



EMR Cluster

The central component of Amazon EMR is the **cluster**

- A collection of **Amazon Elastic Compute Cloud (Amazon EC2)** instances
- Each instance is called a **node**

The **node type** identifies the role within the cluster

- **Master** node coordinates the distribution of data and tasks among other nodes
  - Every cluster has (at least) a master node
  - Always active
- **Core** node runs tasks and store data in the Hadoop Distributed File System (HDFS)
  - Multi-node clusters have at least one core node
  - Always active, contains the data node daemon
- **Task** node only runs tasks
  - Task nodes are optional
  - Decoupling processing and storage, we lose data locality

Migration

On-Demand Instance

- Pay for compute capacity by the hour (minimum of 60 seconds)
- No long-term commitments

Spot Instance

- Unused EC2 instance that is available for less than the on-demand price
- Hourly price is called *spot price*
  - Adjusted based on long-term supply and demand for spot instances
- Run the instance when capacity is available and price is below threshold
  - When data-center resources are low, spot instances are dropped
  - Mainly suitable for batch workloads

https://aws.amazon.com/ec2/pricing/

Migration

Spot Instance cost strategies

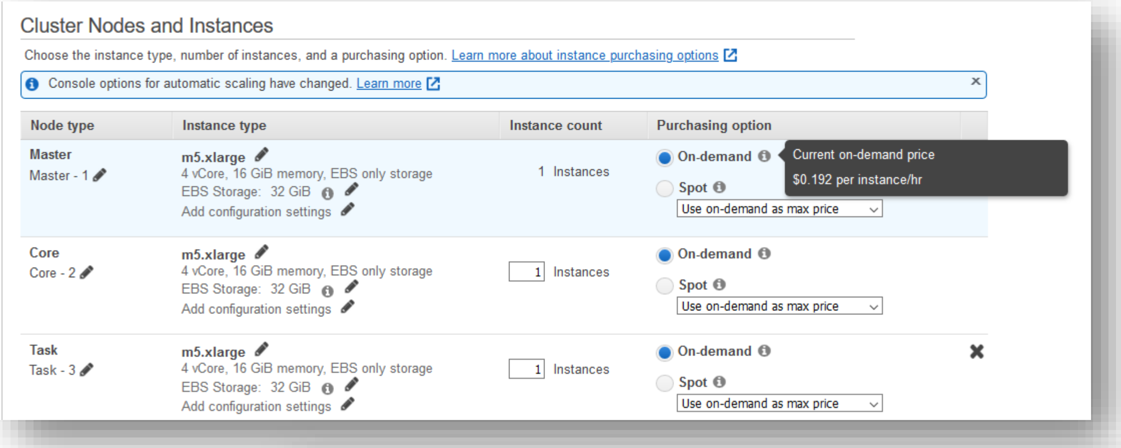
Capacity-optimized strategy

- Allocated instances into the most available pools
- Look at real-time capacity data, predict which are the most available
- Works well for workloads such as big data and analytics
- Works well when we have high cost of interruption

Lowest-price strategy

- Allocates instances in pools with lowest price at time of fulfillment

Creating the cluster



Cluster Nodes and Instances

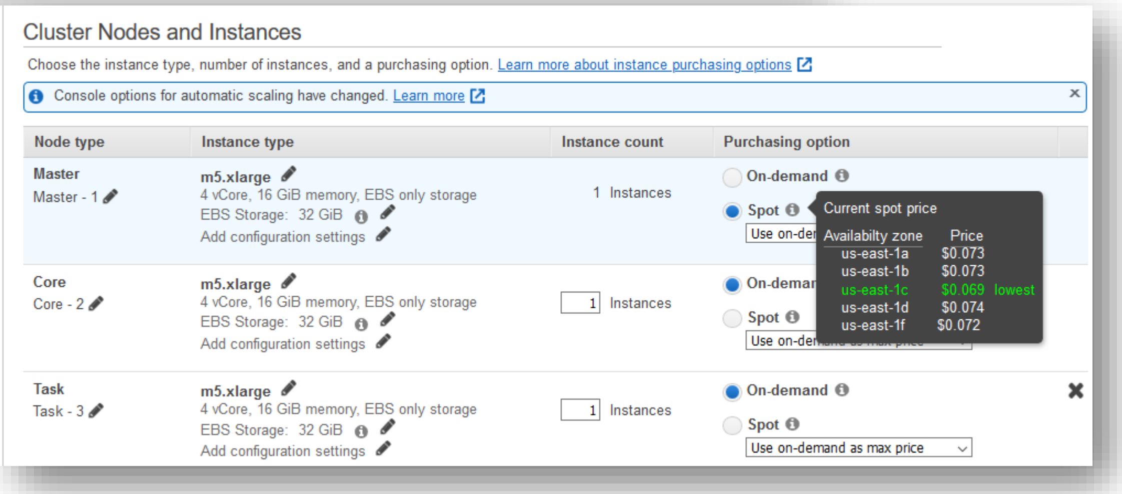
Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more ×

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price ✖

Current on-demand price

$0.192 per instance/hr



Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more ×

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	○ On-demand ● Spot Use on-der
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demar ○ Spot Use on-de
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price ✖

Current spot price

Availabilty zone	Price
us-east-1a	$0.073
us-east-1b	$0.073
us-east-1c	$0.069 lowest
us-east-1d	$0.074
us-east-1f	$0.072

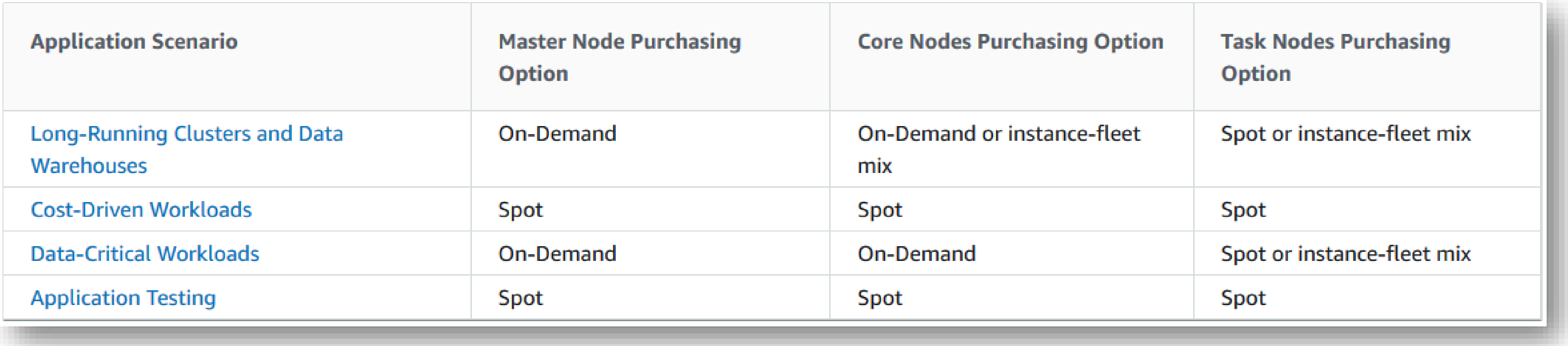
EMR

Choose to launch **master**, **core**, or **task** on Spot Instances

- The **master** node controls the cluster
  - When terminated, the cluster ends
  - Use *spot instances* if you are running a cluster where sudden termination is acceptable
- **Core** nodes process data and store information using HDFS
  - When terminated, data is lost
  - Use *spot instances* when partial HDFS data loss is tolerable
- **Task** nodes process data but do not hold persistent data in HDFS
  - When terminated, computational capacity is lost
  - The effect of spot instances on the cluster is "minimal"

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

EMR



Application Scenario	Master Node Purchasing Option	Core Nodes Purchasing Option	Task Nodes Purchasing Option
Long-Running Clusters and Data Warehouses	On-Demand	On-Demand or instance-fleet mix	Spot or instance-fleet mix
Cost-Driven Workloads	Spot	Spot	Spot
Data-Critical Workloads	On-Demand	On-Demand	Spot or instance-fleet mix
Application Testing	Spot	Spot	Spot

Add some storage

Amazon EMR provides two main file systems

- **HDFS** and **EMRFS**, specify which file system to use by the prefix
- `hdfs://path (or just `path`)`
  - HDFS is used by the master and core nodes
  - AWS EBS volume storage is used for HDFS data
  - Is fast, best used for caching the results produced by intermediate job-flow steps, why?
  - It's ephemeral storage which is reclaimed when the cluster ends
- `s3://DOC-EXAMPLE-BUCKET1/path` (EMRFS)
  - An implementation of the Hadoop file system atop Amazon S3
  - We can avoid EBS storage

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-storage.html

Migration

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

Migrating cluster to EMR: ?

Given the software requirements, we need

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

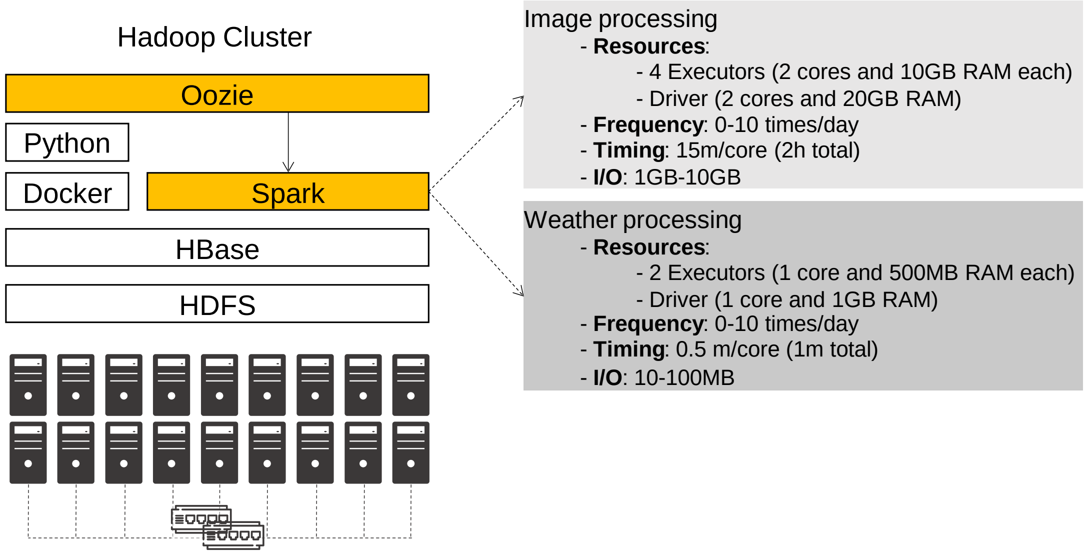


Image processing
- **Resources**:
  - 4 Executors (2 cores and 10GB RAM each)
  - Driver (2 cores and 20GB RAM)
- **Frequency**: 0-10 times/day
- **Timing**: 15m/core (2h total)
- **I/O**: 1GB-10GB

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

Migration

	On-premises	**On cloud**
Hardware	2356€/year	8000€/year
Software	100000€/year	

Migrating cluster to EMR: ~8000€/year

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

Migration

	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)

Migration

	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)

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 miss the software configuration
- We reduced the cluster 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)