1

BIG DATA AND GEOSPATIAL DATA MINING

3 DISTRIBUTED DATA PROCESSING

- *"You don't generate your own electricity. Why generate your own computing?"* ***--Jeff Bezos, CEO, Amazon, 2008***

* NIST word cloud

2

3.4.1 What is Cloud Computing

- Internet service offered by a supplier to multiple customers :
  a) Storage: It allows storing and accessing data stored in the cloud, usually in the form of files (although it can be a database hosted in the cloud).
  b) Computing: in the form of virtual machines with specified characteristics (RAM, CPU, disk, etc.) and configuration (OS).
- Leveraging the economy of large scale suppliers to offer cost savings to users.
- Payment for use, without initial investments (payment for CPU hours, for stored Gbyte and for downloaded Gbyte).

3

3.4.2 Example

- A start-up develops a mobile application with an innovative idea that requires computing and data storage capacity for production.
  - Option A: Acquire servers and perform the *housing* and *hosting* of the application in the company's own infrastructure (In-House).
  - Option B: Provision the necessary resources from a Cloud provider.

4

3.4.2 Example

Option A: In-House

- Rent and condition an office (refrigeration, wiring, UPS, etc.)
- Acquire hardware for computing and storage, update it periodically.
- Configure resources, update them.

5

3.4.2 Example

Option A: In-House

- If the application succeeds less than expected:
  - You are the victim of your own failure:
    - Investment in HW is not profitable and debts can cause the company to close.
- If the application succeeds more than expected:
  - You are a victim of your own success:
    - It is not possible to serve customer requests with the expected quality of service and customers leave.

6

3.4.2 Example

Option B (Cloud computing)

- Provision the initial computing resources necessary for commissioning.
- Configure the application to dynamically self-provision and release new resources (computation, storage) depending on the workload/users.

7

3.4.2 Example

Option B: Cloud computing

- If the application succeeds less than expected:
  - User reduction:
    - Unused resources are released and only paid for the consumption made.
- If the application succeeds more than expected:
  - Increase in users:
    - More resources from the Cloud provider are requested automatically to meet user requests.

8

3.4.3 Definition

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction

National Institute of Standards and Technology (NIST)
http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

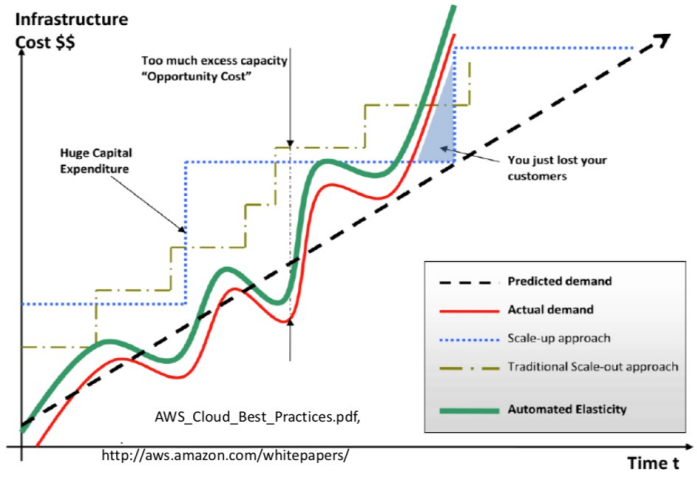
9

3.4.4 Need for a Cloud service

- Hardware investments become obsolete at high speed :
  - We must maximize the efficient use of resources
  - Unavailable for small centers/SMEs/Start-Ups
- The demand for resources (storage, computation) is very variable.
- The resource consumption must be adjusted to the needs of the applications dynamically and quickly.

10

3.4.4 Need for a Cloud service



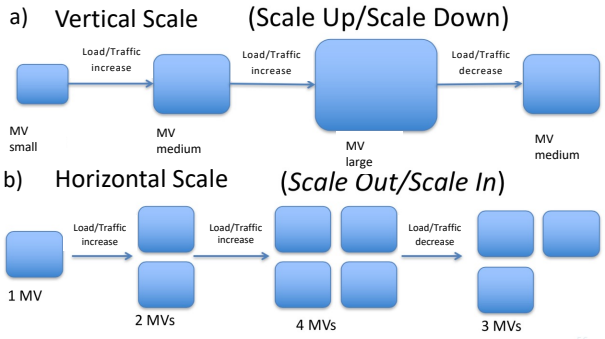
11

3.4.5 Cloud service characteristics

- Self-service on demand.
  - A consumer can unilaterally provide resources without interacting with service provider personnel.
- Access via Internet.
  - Capabilities are provided through the network with minimum requirements in the client.
- Elasticity.
  - The consumer can dynamically increase or decrease the number of resources in any moment, perceiving an illusion of infinite capacity.

12

3.4.5 Cloud service characteristics



13

3.4.5 Cloud service characteristics

- Service Through Payment for Use
  - The resources used are accounted for independently (storage, computing, bandwidth, etc.) and accurate to be able to implement the payment for use, taking the hour as a reference unit.

- Configurability
  - Rented resources must be highly configurable to adapt to the needs of different users. This will be more or less possible depending on the service model (IaaS, PaaS, SaaS).

14

3.4.5 Cloud service characteristics

- Separation
  - Cloud computing provides "for rent" resources under a pay-per-use model but does not expose the details of the infrastructure to customers or partners. Users use resources without knowing the details of the infrastructure of the suppliers.

- Isolation
  - Given the host nature of Cloud providers, consumers need mechanisms and guarantees that their applications are isolated from the rest of the clients hosted in the same infrastructure.

15

3.4.6 Basic Technologies: DataCenters

- Google, Amazon, Microsoft, etc. have large data centers worldwide (different locations to protect against failures).

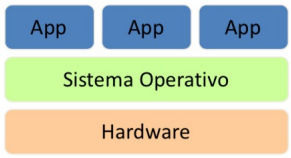
- Use of energy efficiency techniques to reduce the consumption derived from its operation: dynamically switching off/on nodes, proper air conditioning management etc.

Dublin microsoft data center with over 60.000 m^2 built

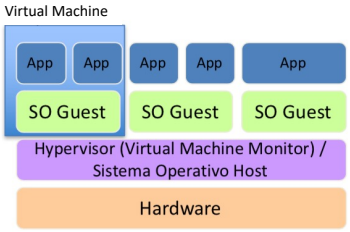
16

3.4.6 Basic Technologies: Virtualization

- Virtualization allows you to create a (or several) simulated environment (Virtual Machine, MV) that runs a guest OS. All this running on a host OS with the help of a hypervisor (or Virtual Machine Monitor).



Traditional Platform



Virtualized Platform

- An application can run on a specific version of OS, over modern hardware.
- Increase the hardware utilization rate by running more virtual machines on the same physical equipment.

17

3.4.7 Main cloud providers

- Amazon Web Services

Includes services for the dynamic provisioning of computing capacity as well as the efficient management and storage of data and the scalable design of applications in the cloud through a pay-per-use model.

18

3.4.7 Main cloud providers

- Windows Azure

Microsoft Cloud Platform. Development of hosted applications (.NET) that combine web, SQL databases, file storage, etc., on a Windows-based virtual infrastructure.

19

3.4.7 Main Cloud providers

- Google App Engine
  - Google solution to create and host Web applications (Java/Python) in the cloud.
  - Control of resources consumed by the application to fit the budget.

- Google Compute Engine
  - Google solution for provisioning computing infrastructure in the form of virtual machines (Linux).

20

3.4.8 Challenges of a Cloud infrastructure

- Challenge1: Service availability and data lock-in.
- Challenge2: Privacy of data and aspects of security.
- Challenge3: Non-deterministic benefits and bottlenecks.
- Challenge4: Distributed storage.
- Challenge5: Scalability, Interoperability and Standardization.
- Challenge 6: Software licenses
- Challenge 7: Reputation Sharing.

21

3.4.9 Trust and Reputation

- Outsource the computation and storage to a third party means trusting it.

- A provider can affect its survival and that of dependent companies.

- The supplier's reputation is a disruption in the service (outages).
  - Electric storm in July 2012 that affected AWS and customers (Netflix, etc.), bug in AWS caused a drop in service in October 2012 that affected Reddit, Foursquare, etc.
  - Windows Azure crash in February 2012 for several hours because of a security certificate.

22

3.4.10 Cloud risks

- Megaupload file storage service was closed in January 2012 by the FBI for copyright infringement.

- What happens to the data of legal users?

- Is the service offered really illegal?

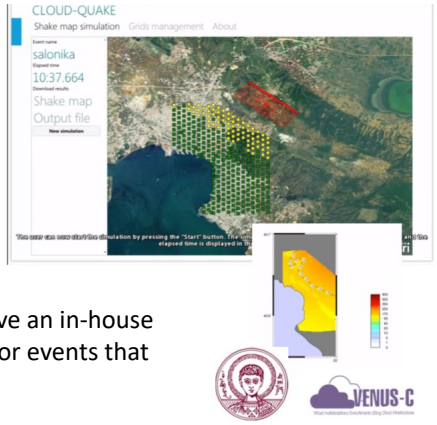
3.4.11 Critical opinions

- Control over data in the Cloud is lost
  - Data is transferred to resources with a level of unclear trust.
  - Access, modification or manipulation of stored data is potentially possible.
  - The operation and access to virtual machines is completely under the control of the provider.

"One reason you should not use web applications to do your computing is that you lose control. It's just as bad as using a proprietary program. Do your own computing on your own computer with your copy of a freedom-respecting program. If you use a proprietary program or somebody else's web server, you're defenseless. You're putty in the hands of whoever developed that software". - *Richard Stallman: Cloud computing is a trap, guardian.co.uk 29-09-2008*

3.4.12 Success Stories. Earthquake analysis

- Elastic use up to 100 Azure cores to simulate seismic propagation waves and their impact.

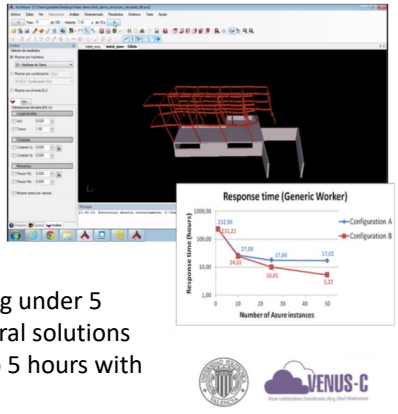


- Automatically capture data from seismic records offering cuasi real-time information about the affected areas.
- It would not make sense to have an in-house cluster of 100 dedicated nodes for events that happen with low frequency.

25

3.4.12 Success Stories. Structure calculation

- The simulation of the dynamic behavior of building structures allows to determine their response to earthquakes.
- This analysis must be carried out before different earthquakes and for different materials.



- A complex simulation of a 10-story building under 5 earthquake records and 10 different structural solutions goes from more tan 4 days to be resolved to 5 hours with 50 Azure cores (and less than € 50 cost).

26

"I think there is a world market for about five computers"
– Attributed to Thomas J. Watson, IBM

"… In a sense, says Yahoo Research Chief Prabhakar Raghavan, there are only five computers on earth. He lists Google, Yahoo, Microsoft, IBM, and Amazon. Few others, he says, can turn electricity into computing power with comparable efficiency …"
– Steven Baker, From Google and the wisdom of clouds

"… The World Wide Web is becoming ONE vast, programmable machine. As NYU's Clay Shirky likes to say, Watson was off by four …"
– Nicholas Carr, From Wired Magazine Q&A with Nicholas Carr

27

3.4.13 Amazon CLOUD. Amazon Web Service (AWS)

- Amazon Web Services (AWS) offers infrastructure services to run applications in the cloud.
- Pioneer in Cloud Computing since 2006 offering :
  - Low cost
  - Agility and instant elasticity
  - Open and Flexible
  - Safe

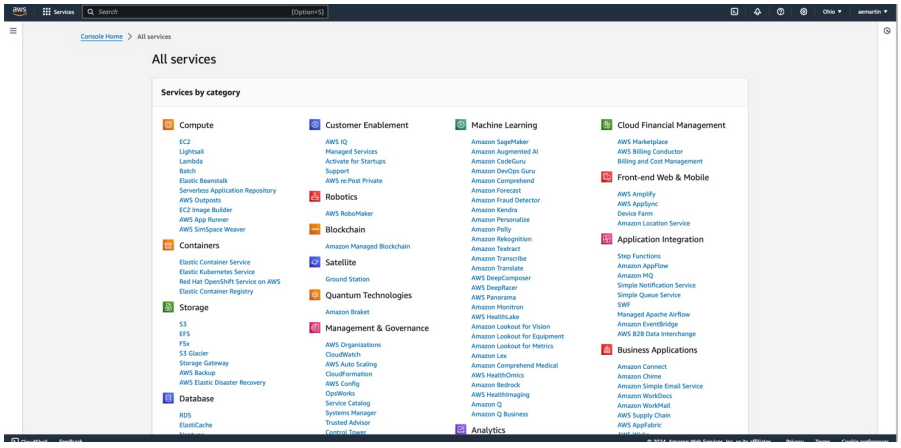
28

3.4.13 Amazon Web Service (AWS). ProductS

- Storage
  - Amazon Simple Storage Service (S3)
  - Amazon Glacier
  - Amazon Elastic Block Store (EBS)
- Data processing
  - Amazon Elastic Compute Cloud (EC2)
  - Amazon Elastic MapReduce
  - Auto Scaling
  - Elastic Load Balancing
- Data Bases
- Deployment and management
- Aplications Services
- Network
- Content Delivery
- Analytics
- Software
- •••••••••

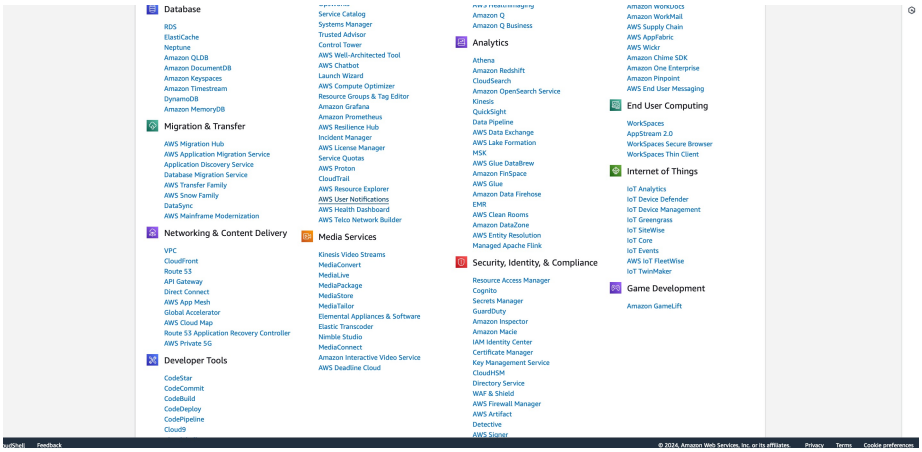
29

3.4.13 Amazon Web Service (AWS). Products



30

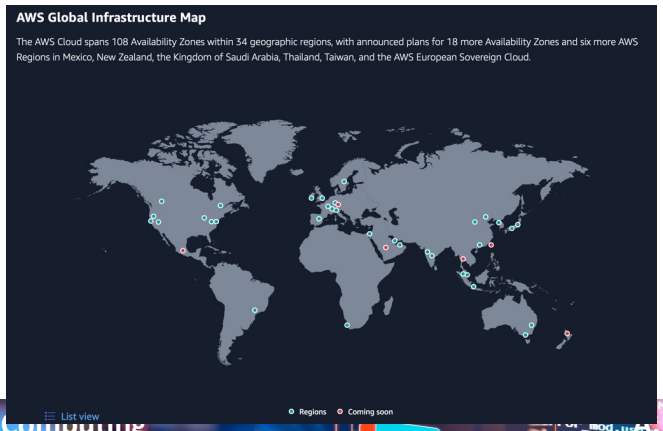
3.4.13 Amazon Web Service (AWS). Products



31

3.4.13 Amazon Web Service (AWS). Global Infrastructure

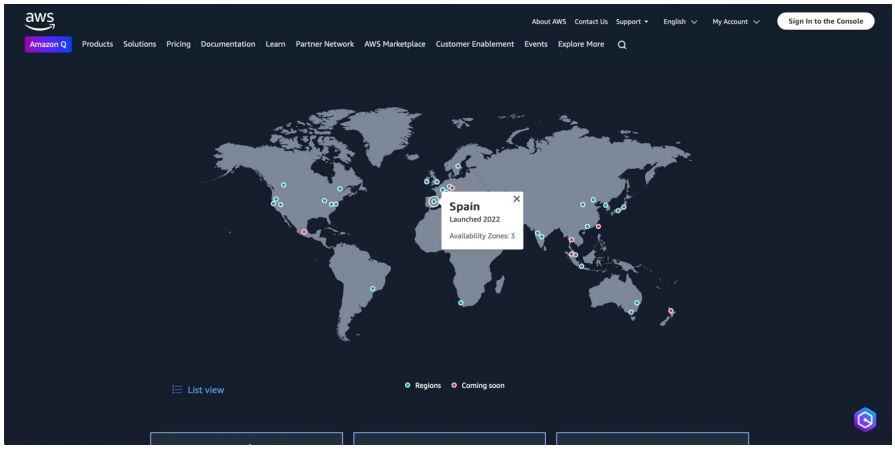
- AWS is available in more than 200 countries and its infrastructure is available throughout the world (34 regions) to facilitate proximity to customers. Latency reduction but variable costs by región



32

3.4.13 Amazon Web Service (AWS). Global Infrastructure

- https://aws.amazon.com/about-aws/global-infrastructure/?nc1=h_ls (2023)



33

3.4.13 Amazon Web Service (AWS). Global Infrastructure

- Each region includes different availability zones for fault tolerance

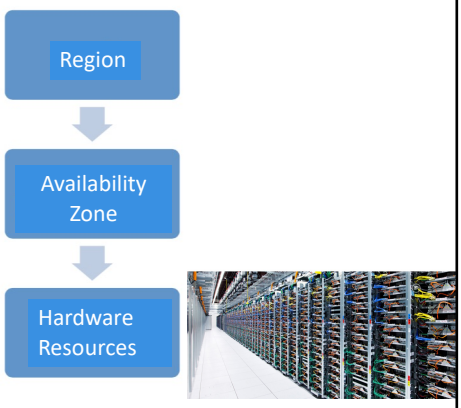
Code	Nombre
us-east-1	US East (N. Virginia)
us-east-2	EE.UU. Este (Ohio)
us-west-1	EE.UU. Oeste (Norte de California)
us-west-2	EE.UU. Oeste (Oregón)
ca-central-1	Canadá (Central)
eu-central-1	UE (Fráncfort)
eu-west-1	UE (Irlanda)
eu-west-2	UE (Londres)
eu-west-3	UE (París)
eu-north-1	UE Estocolmo
ap-east-1	Asia Pacífico (Hong Kong)
ap-northeast-1	Asia Pacífico (Tokio)
ap-northeast-2	Asia Pacífico (Seúl)
ap-northeast-3	Asia Pacífico (Osaka-local)
ap-southeast-1	Asia Pacífico (Singapur)
ap-southeast-2	Asia Pacífico (Sídney)
ap-south-1	Asia Pacífico (Mumbai)
sa-east-1	América del Sur (São Paulo)

USA Este
Norte Virginia (6)
Ohio (3)

USA Oeste
North California (3)
Oregon (4)

...

AWS GovCloud (USA)
West (3)
East (3)



34

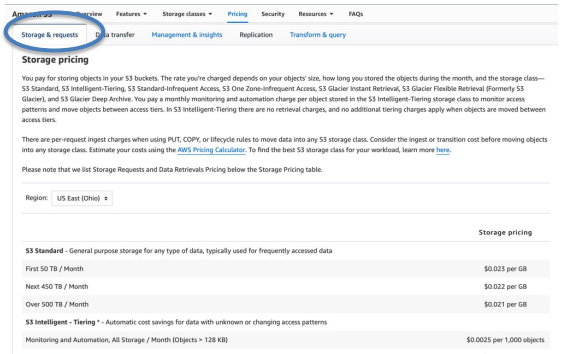
3.4.13 Amazon Web Service (AWS). Amazon S3

- "Internet storage. It is designed to provide developers with Web-scale computing"
- Consists of an unlimited number of objects storage system for the Cloud environment
- Accessible through standard protocols (HTTP)
- Reliability: 99.999999999% of storage durability and 99.99% of availability.
- Data stored as objects in buckets (deposits)
- A bucket is an object container.
  - It has access control set by the user (who can create, delete and list the bucket)
  - It can be audited (access log) and versioned
- An object
  - Is a file + description metadata(optional)
  - From 1 byte to 5 Terabytes

35

3.4.13 Amazon Web Service (AWS). Amazon S3

- Price scheme (Depending on the region):



36

3.4.13 Amazon Web Service (AWS). Amazon S3

- Price scheme (Depending on the region):

S3 Intelligent - Tiering * - Automatic cost savings for data with unknown or changing access patterns	
Monitoring and Automation, All Storage / Month (Objects > 128 KB)	$0.0025 per 1,000 objects
Frequent Access Tier, First 50 TB / Month	$0.023 per GB
Frequent Access Tier, Next 450 TB / Month	$0.022 per GB
Frequent Access Tier, Over 500 TB / Month	$0.021 per GB
Infrequent Access Tier, All Storage / Month	$0.0125 per GB
Archive Instant Access Tier, All Storage / Month	$0.004 per GB
S3 Intelligent - Tiering * - Optional asynchronous Archive Access tiers	
Archive Access Tier, All Storage / Month	$0.0036 per GB
Deep Archive Access Tier, All Storage / Month	$0.00099 per GB
S3 Standard - Infrequent Access ** - For long lived but infrequently accessed data that needs millisecond access	
All Storage / Month	$0.0125 per GB
S3 One Zone - Infrequent Access ** - For re-createable infrequently accessed data that needs millisecond access	
All Storage / Month	$0.01 per GB
S3 Glacier Instant Retrieval *** - For long-lived archive data accessed once a quarter with instant retrieval in milliseconds	
All Storage / Month	$0.004 per GB
S3 Glacier Flexible Retrieval (Formerly S3 Glacier) ***- For long-term backups and archives with retrieval option from 1 minute to 12 hours	
All Storage / Month	$0.0036 per GB
S3 Glacier Deep Archive *** - For long-term data archiving that is accessed once or twice in a year and can be restored within 12 hours	
All Storage / Month	$0.00099 per GB

37

3.4.13 Amazon Web Service (AWS). Amazon S3

- Price scheme (Depending on the region):

Requests & data retrievals

You pay for requests made against your S3 buckets and objects. S3 request costs are based on the request type, and are charged on the quantity of requests as listed in the table below. When you use the Amazon S3 console to browse your storage, you incur charges for GET, LIST, and other requests that are made to facilitate browsing. Charges are accrued at the same rate as requests that are made using the API/SDK. Reference the S3 developer guide for technical details on the following request types: PUT, COPY, POST, LIST, GET, SELECT, Lifecycle Transition, and Data Retrievals. DELETE and CANCEL requests are free. LIST requests for any storage class are charged at the same rate as S3 Standard PUT, COPY, and POST requests. You pay for retrieving objects that are stored in S3 Standard – Infrequent Access, S3 One Zone – Infrequent Access, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive. Reference the S3 developer guide for technical details on Data Retrievals.

S3 Lifecycle Transition request pricing below represents requests to that storage class. For example, transitioning data from S3 Standard to S3 Standard-Infrequent Access will be charged $0.01 per 1,000 requests.

There are no retrieval charges in S3 Intelligent-Tiering. If an object in the infrequent access tier is accessed later, it is automatically moved back to the frequent access tier. No additional tiering charges apply when objects are moved between access tiers within the S3 Intelligent-Tiering storage class.

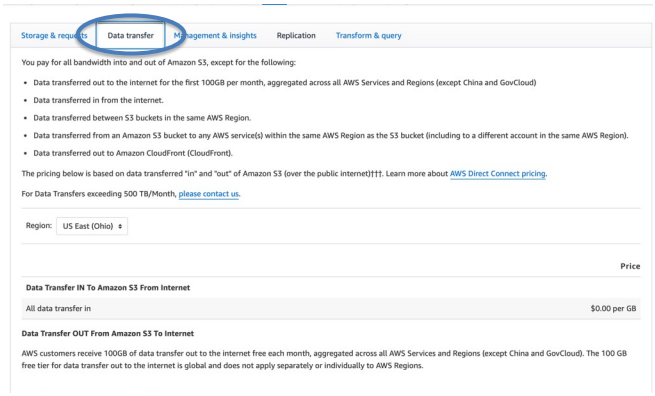
Region: US East (Ohio)

	PUT, COPY, POST, LIST requests (per 1,000 requests)	GET, SELECT, and all other requests (per 1,000 requests)	Lifecycle Transition requests into (per 1,000 requests)	Data Retrieval requests (per 1,000 requests)	Data retrievals (per GB)
S3 Standard	$0.005	$0.0004	n/a	n/a	n/a
S3 Intelligent-Tiering *	$0.005	$0.0004	$0.01	n/a	n/a
Frequent Access	n/a	n/a	n/a	n/a	n/a
Infrequent Access	n/a	n/a	n/a	n/a	n/a
Archive Instant	n/a	n/a	n/a	n/a	n/a
Archive Access, Standard	n/a	n/a	n/a	n/a	n/a
Archive Access, Bulk	n/a	n/a	n/a	n/a	n/a
Archive Access, Expedited	n/a	n/a	n/a	$10.00	$0.03

38

3.4.13 Amazon Web Service (AWS). Amazon S3

- Price scheme (Depending on the region):



Storage & requests | Data transfer | Management & insights | Replication | Transform & query

You pay for all bandwidth into and out of Amazon S3, except for the following:

- Data transferred out to the internet for the first 100GB per month, aggregated across all AWS Services and Regions (except China and GovCloud)
- Data transferred in from the internet.
- Data transferred between S3 buckets in the same AWS Region.
- Data transferred from an Amazon S3 bucket to any AWS service(s) within the same AWS Region as the S3 bucket (including to a different account in the same AWS Region).
- Data transferred out to Amazon CloudFront (CloudFront).

The pricing below is based on data transferred "in" and "out" of Amazon S3 (over the public internet)†††. Learn more about AWS Direct Connect pricing.

For Data Transfers exceeding 500 TB/Month, please contact us.

Region: US East (Ohio)

	Price
Data Transfer IN To Amazon S3 From Internet	
All data transfer in	$0.00 per GB

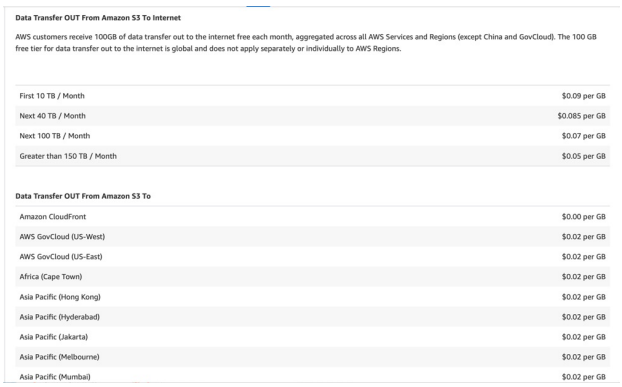
Data Transfer OUT From Amazon S3 To Internet

AWS customers receive 100GB of data transfer out to the internet free each month, aggregated across all AWS Services and Regions (except China and GovCloud). The 100 GB free tier for data transfer out to the internet is global and does not apply separately or individually to AWS Regions.

39

3.4.13 Amazon Web Service (AWS). Amazon S3

- Price scheme (Depending on the region):



Data Transfer OUT From Amazon S3 To Internet

AWS customers receive 100GB of data transfer out to the internet free each month, aggregated across all AWS Services and Regions (except China and GovCloud). The 100 GB free tier for data transfer out to the internet is global and does not apply separately or individually to AWS Regions.

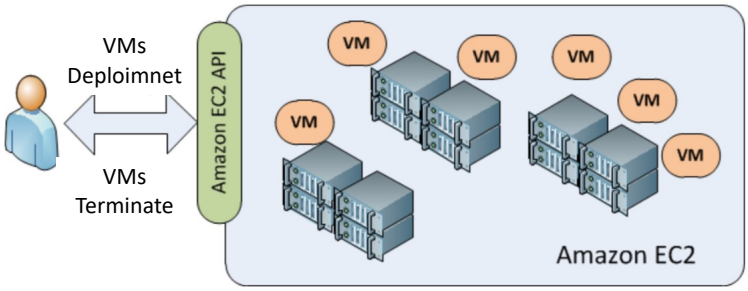
First 10 TB / Month	$0.09 per GB
Next 40 TB / Month	$0.085 per GB
Next 100 TB / Month	$0.07 per GB
Greater than 150 TB / Month	$0.05 per GB

Data Transfer OUT From Amazon S3 To	
Amazon CloudFront	$0.00 per GB
AWS GovCloud (US-West)	$0.02 per GB
AWS GovCloud (US-East)	$0.02 per GB
Africa (Cape Town)	$0.02 per GB
Asia Pacific (Hong Kong)	$0.02 per GB
Asia Pacific (Hyderabad)	$0.02 per GB
Asia Pacific (Jakarta)	$0.02 per GB
Asia Pacific (Melbourne)	$0.02 per GB
Asia Pacific (Mumbai)	$0.02 per GB

40

3.4.13 Amazon Web Service (AWS). Amazon EC2

- Amazon EC2 is a service that provides cloud computing resources. It allows the deployment of virtual machines or predefined size instances to have computation on demand through a pay-per-use model.



41

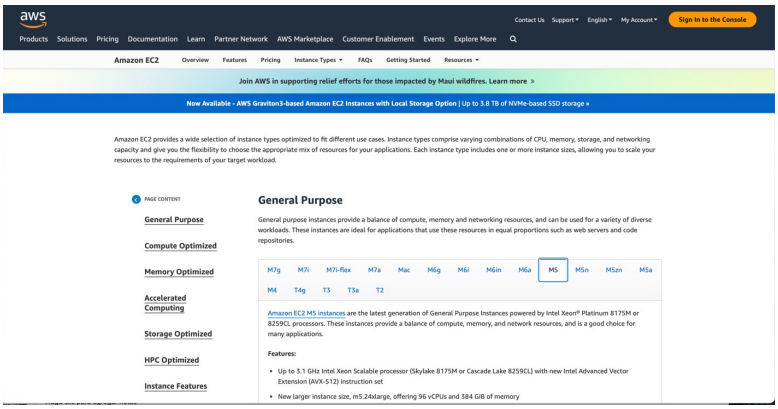
3.4.13 Amazon Web Service (AWS). Amazon EC2

- AWS uses the term instance to refer to a virtual machine (VM), an instance provides a predictable amount of dedicated computing capacity.
- The deployment of an instance requires indicating :
  - Type of instance, there are predefined types of instance whose price per hour varies according to the benefits.
  - AMI (Amazon Machine Image), image of the Virtual Machine, determines the OS and the available applications of the instance when it starts.
  - Security Group (SG), is the firewall configuration of the instance (what traffic the instance can receive)
  - Keypair to allow connection to the instance via SSH without password.
  - Region and availability zone (optional). The default region is us-east-1 (Virgina, USA).

42

3.4.13 Amazon Web Service (AWS). Amazon EC2

https://aws.amazon.com/ec2/instance-types/?nc1=h_ls



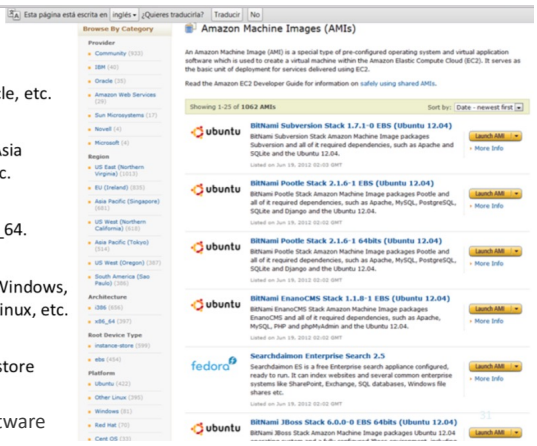
3.4.13 Amazon Web Service (AWS). Amazon EC2

STANDARD INSTANCES				
NAME	vCPU	MEMORY	STORAGE	PRICE/HOUR
m5.large	2	8 GB	75 GB	$ 0.096
m5.xlarge	4	16 GB	150 GB	$ 0.192
m5.2xlarge	8	32 GB	300 GB	$ 0.384
m5.4xlarge	16	64 GB	600 GB	$ 0.768
m5.8xlarge	32	128 GB	1.2 TB	$ 1.536
m5.12xlarge	48	192 GB	1.8 TB	$ 2.304
m5.16xlarge	64	256 GB	2.4 TB	$ 3.072
m5.24xlarge	96	384 GB	3.6 TB	$ 4.608

Amazon EC2 instances support multithreading, which enables multiple threads to run concurrently on a single CPU core. Each thread is represented as a virtual CPU (vCPU) on the instance. An instance has a default number of CPU cores, which varies according to instance type. For example, an m5.xlarge instance type has two CPU cores and two threads per core by default—four vCPUs in total. EEUU W (Oregon región) prices.

3.4.13 Amazon Web Service (AWS). Amazon EC2

- **AMIs by**:
  - Provider
    - IBM, Oracle, etc.
  - Region
    - US East, Asia Pacific, etc.
  - Architecture
    - i386, x86_64.
  - Platform
    - Ubuntu, Windows, Amazon Linux, etc.
  - Root device
    - instance-store
    - ebs
  - Installed Software



3.4.13 Amazon Web Service (AWS). Amazon EC2

- An AMI is like a bootable CD, it contains all the information necessary to boot the instance.
- There are two types of AMIs :
  - Stored in S3 (called instance-store). Any file created or modified in the virtual machine will disappear when the virtual machine is turned off.
  - Stored in EBS (called EBS). The root device is no longer ephemeral, changes in the virtual machine are stored in an EBS volume associated with the instance.

3.4.13 Amazon Web Service (AWS). Amazon EC2

- Amazon Elastic Block Store (EBS) provides block-oriented volumes (as if it were an external hard drive or USB) to be connected to instances of EC2.
  - They are created for a specific availability zone and can only be connected to instances deployed in the same availability zone.
  - It can only be connected to a single instance at a given time.
  - S3 is a service of objects (files) and EBS of volumes (disks).
  - The start-up of an EBS-based AMI causes the automatic creation of an EBS volume to store data from it.
  - The volume is connected (and the data preserved) even if the instance is stopped (this is not the case if the instance is terminated)
  - Stopping the instance is no longer an expense for consumption of instance hours, but for the storage of the associated EBS data volumen.

47

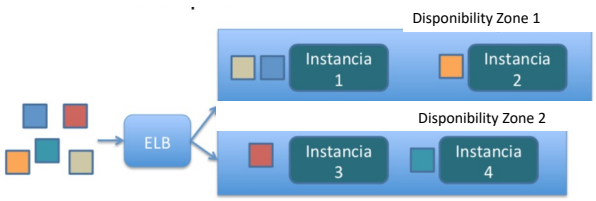
3.4.13 Amazon Web Service (AWS). Amazon EC2

- Prices depend on the region.
- Windows-based instances cost more than those based on Linux.
- Reserved instances: at 1 year and 3 years, it allows to obtain a significant reduction (more than 50%) of the price per hour of instances.
- Dedicated or shared instances (dedicated are more expensive)
- https://calculator.aws/#/estimate:
  - Ohio Region:
    - 3 m5x.large instances active throughout the day (cost of the instance plus the storage cost of the three associated 30 Giga –minimum- EBS AMIs): $ 420.48 for EC2 instances + $ 7.2 for EBS AMIs a month (730 hours).
    - Storage of 20 GB in S3 where a download of 5 GB/month is supposed to any user: $ 0.46 storage + $ 0.45 transfer.
    - Total cost (one month is 730 hours): $ 428.59/month

48

3.4.13 Amazon Web Service (AWS). Amazon EC2

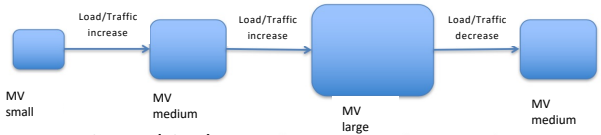
- Services associated with EC2:
  - Elastic Load Balancer Service (ELB).



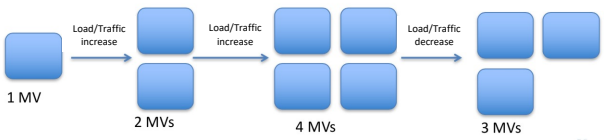
3.4.13 Amazon Web Service (AWS). Amazon EC2

- Automatic scaling.

- Vertical Scale (Scale Up/Scale Down)



- Horizontal Scale (*Scale Out/Scale In*)



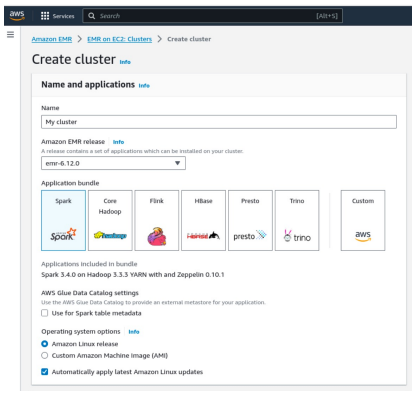
3.4.13 Amazon Web Service (AWS). Amazon EMR

- Allows dynamic deployment of distributed processing clusters to process data that may be stored in different AWS services.
- Offers automatic configuration of different frameworks and tools.
- The cluster can be resized (elasticity) at runtime.
- Root access to EC2 instances.
- Cluster life cycle management.
- It allows to integrate with other AWS services for reading and storing data (Database –DynamoDB, RDS- or file storage service -S3, HDFS-).

51

3.4.13 Amazon Web Service (AWS). Amazon EMR

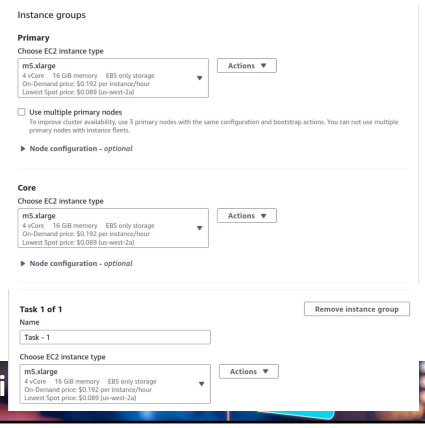
- It is possible to choose alternative applications to complement the installation.



52

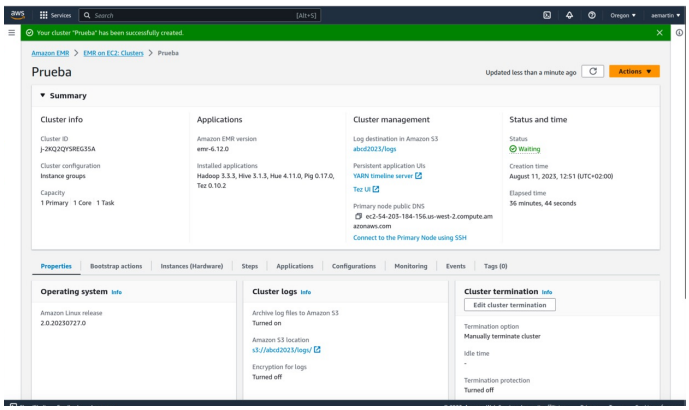
3.4.13 Amazon Web Service (AWS). Amazon EMR

- Node type:
  - Master: Only one. Control
  - Core: Tasks execution+ HDFS.
  - Task: Tasks execution

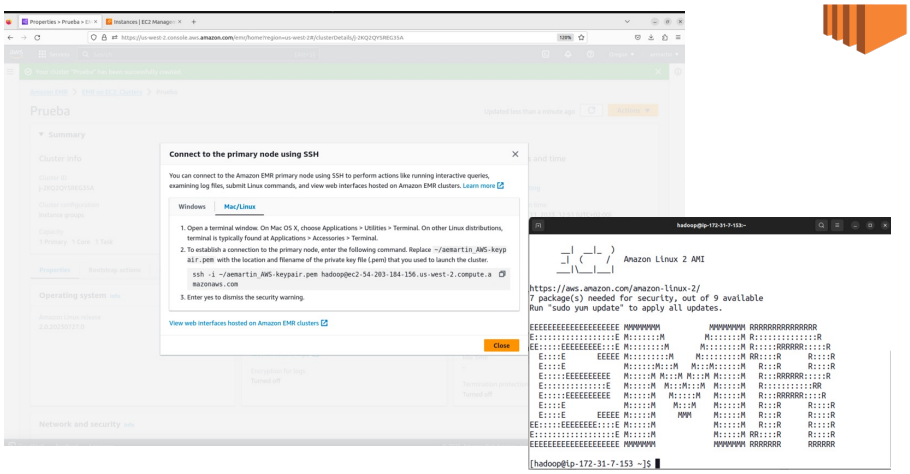


3.4.13 Amazon Web Service (AWS). Amazon EMR

- Cluster deployment monitoring panel from the Amazon EMR service. Unified service access interface.



3.4.13 Amazon Web Service (AWS). Amazon EMR



55

3.4.13 Amazon Web Service (AWS). Amazon EMR

- The smallest type of instance currently eligible is m5.xlarge (16 GB of RAM).
- It is possible to scale a *Hadoop cluster* (increase or decrease the number of Core or Task nodes) manually or automatically (*autoscaling* group based on different metrics).
- At the end of the work you have to finish the cluster to stop attending expenses.
- Amazon EMR carries an associated fee (in addition to the consumption of EC2 instances) depending on the type of instance:

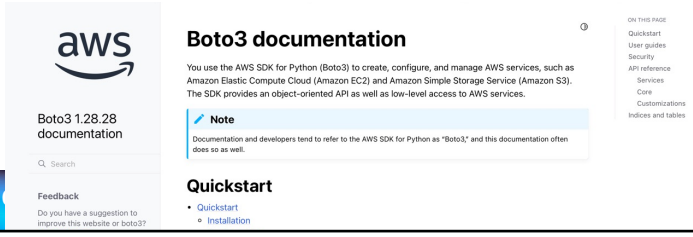
STANDARD INSTANCES		
NAME	**PRICE/HOUR**	**EMR PRICE/HOUR**
m5.large	$ 0.096	Does Not Exist
m5.xlarge	$ 0.192	$ 0.048
m5.2xlarge	$ 0.384	$ 0.096
m5.4xlarge	$ 0.768	$ 0.192
m5.8xlarge	$ 1.536	$ 0.270
m5.12xlarge	$ 2.304	$ 0.270
m5.16xlarge	$ 3.072	$ 0.270
m5.24xlarge	$ 4.608	$ 0.270

56

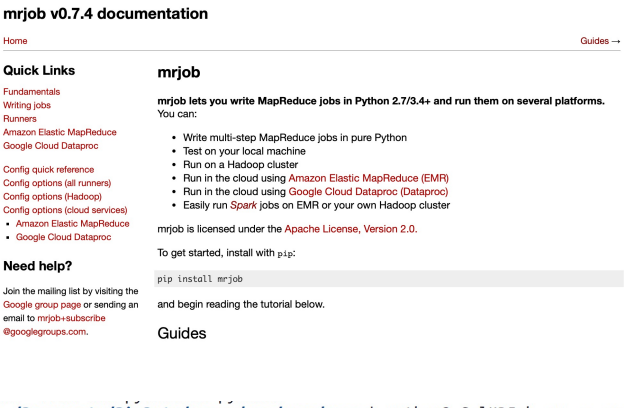
3.4.13 Amazon Web Service (AWS). Access from Python, boto3 Library

- This library provides a simple direct API access to the following AWS resources :
  - S3
  - EC2
  - CloudWach
  - SQS (Simple Queue Service)

https://boto3.amazonaws.com/v1/documentation/api/latest/index.html



3.4.13 Amazon Web Service (AWS). Access from Python, MRJob




```
angel@angel-VM:~/Documents/BigData/mapreduce/purchases$ python3 SolMRJob.py -r emr --num-core-instan
ces=1 --core-instance-type=m5.xlarge --master-instance-type=m5.xlarge --conf-path=mrjob.conf --outpu
t-dir=s3://abcd2023/output/ s3://abcd2023/input/purchases.txt
```