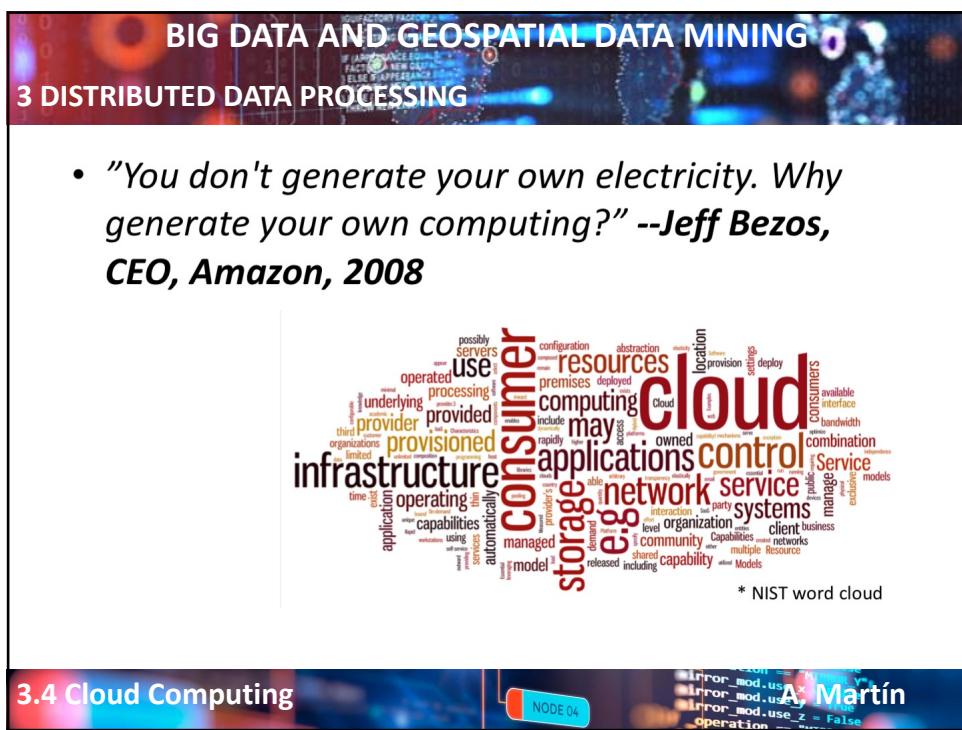




1



2

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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).

**3.4 Cloud Computing**

A. Martín

3

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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.

**3.4 Cloud Computing**

A. Martín

4

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

5

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

6

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

7

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

8

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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>

**3.4 Cloud Computing**

A. Martín

9

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

10

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 3.4.4 Need for a Cloud service

The graph illustrates the cost of infrastructure over time for different scaling approaches. The 'Traditional Scale-out approach' leads to high initial costs and significant waste ('Huge Capital Expenditure', 'Too much excess capacity "Opportunity Cost"'). The 'Scale-up approach' is less efficient than the 'Traditional Scale-out approach'. The 'Predicted demand' line represents the ideal scenario. The 'Automated Elasticity' approach (solid green line) is shown to be the most efficient, closely matching the actual demand curve.

AWS\_Cloud\_Best\_Practices.pdf,  
<http://aws.amazon.com/whitepapers/>

**3.4 Cloud Computing** A. Martín

11

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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.

**3.4 Cloud Computing** A. Martín

12

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 3.4.5 Cloud service characteristics

a) Vertical Scale      (Scale Up/Scale Down)

```

graph LR
    A[MV small] -- "Load/Traffic increase" --> B[MV medium]
    B -- "Load/Traffic increase" --> C[MV large]
    C -- "Load/Traffic decrease" --> D[MV medium]
    D -- "Load/Traffic decrease" --> C
  
```

b) Horizontal Scale      (Scale Out/Scale In)

```

graph LR
    A[1 MV] -- "Load/Traffic increase" --> B[2 MVs]
    B -- "Load/Traffic increase" --> C[4 MVs]
    C -- "Load/Traffic decrease" --> D[3 MVs]
    D -- "Load/Traffic decrease" --> C
  
```

**3.4 Cloud Computing**

A. Martín

13

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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).

**3.4 Cloud Computing**

A. Martín

14

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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.

**3.4 Cloud Computing**

A. Martín

15

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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<sup>2</sup> built

**3.4 Cloud Computing**

A. Martín

16

## BIG DATA AND GEOSPATIAL DATA MINING

### 3 DISTRIBUTED DATA PROCESSING

#### 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).

The diagram illustrates the difference between a Traditional Platform and a Virtualized Platform. On the left, the Traditional Platform shows three separate application boxes ('App') stacked vertically within a 'Sistema Operativo' (Operating System) layer, which sits atop a 'Hardware' layer. On the right, the Virtualized Platform shows multiple application boxes ('App') within individual 'SO Guest' (Guest Operating System) layers, which are stacked within a 'Hypervisor (Virtual Machine Monitor) / Sistema Operativo Host' (Host Operating System) layer, all running on top of a 'Hardware' layer.

**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.

#### 3.4 Cloud Computing

A. Martín

17

## BIG DATA AND GEOSPATIAL DATA MINING

### 3 DISTRIBUTED DATA PROCESSING

#### 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.

The Amazon Web Services logo features a stylized cluster of orange cubes of varying sizes followed by the text "amazon web services™".

#### 3.4 Cloud Computing

A. Martín

18

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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.



Windows Azure

**3.4 Cloud Computing**

A. Martín

19

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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).



**3.4 Cloud Computing**

A. Martín

20

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

21

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

**3.4 Cloud Computing**

A. Martín

22

## BIG DATA AND GEOSPATIAL DATA MINING

### 3 DISTRIBUTED DATA PROCESSING

#### 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 Cloud Computing

A. Martín

23

## BIG DATA AND GEOSPATIAL DATA MINING

### 3 DISTRIBUTED DATA PROCESSING

#### 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 Cloud Computing

A. Martín

24

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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 quasi 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.

**3.4 Cloud Computing**

A. Martín

25

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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 than 4 days to be resolved to 5 hours with 50 Azure cores (and less than € 50 cost).

**3.4 Cloud Computing**

A. Martín

26

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

*"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](#)

**3.4 Cloud Computing**

A. Martín

27

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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

**aws**

**3.4 Cloud Computing**

A. Martín

28



### 3 DISTRIBUTED DATA PROCESSING

#### 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
- Applications Services
- Network
- Content Delivery
- Analytics
- Software
- .....



**3.4 Cloud Computing**

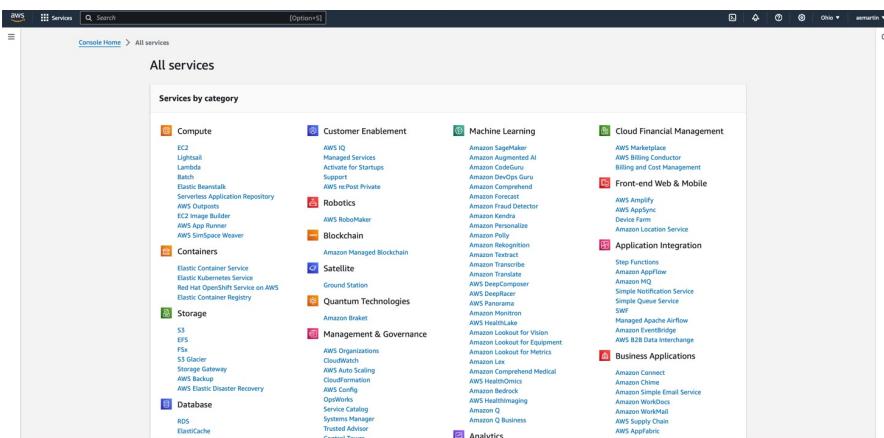
A. Martín

29



### 3 DISTRIBUTED DATA PROCESSING

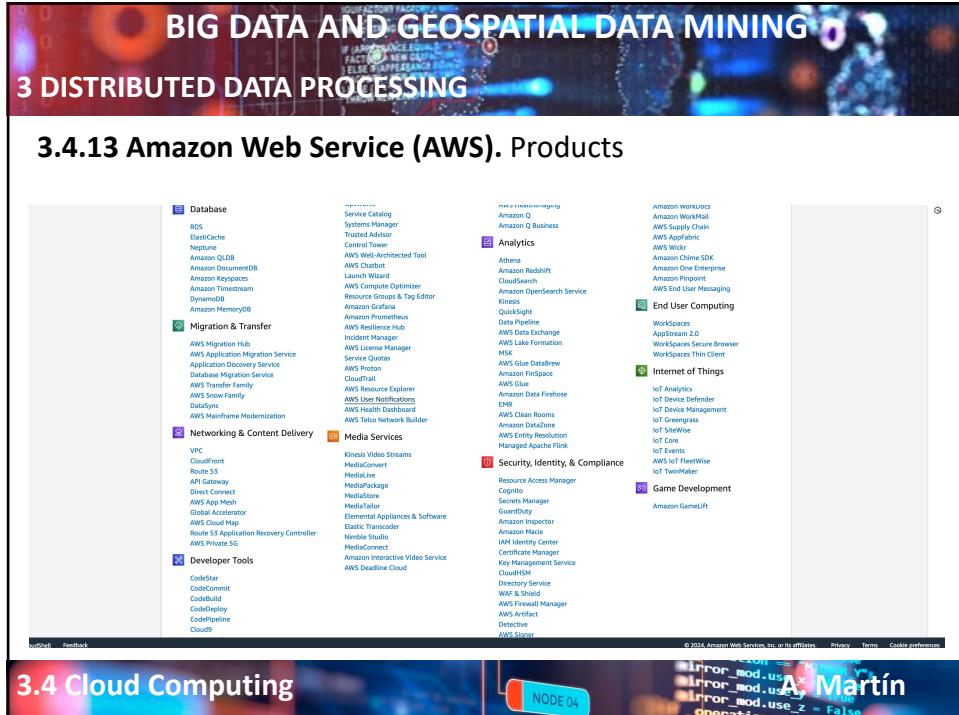
#### 3.4.13 Amazon Web Service (AWS). Products



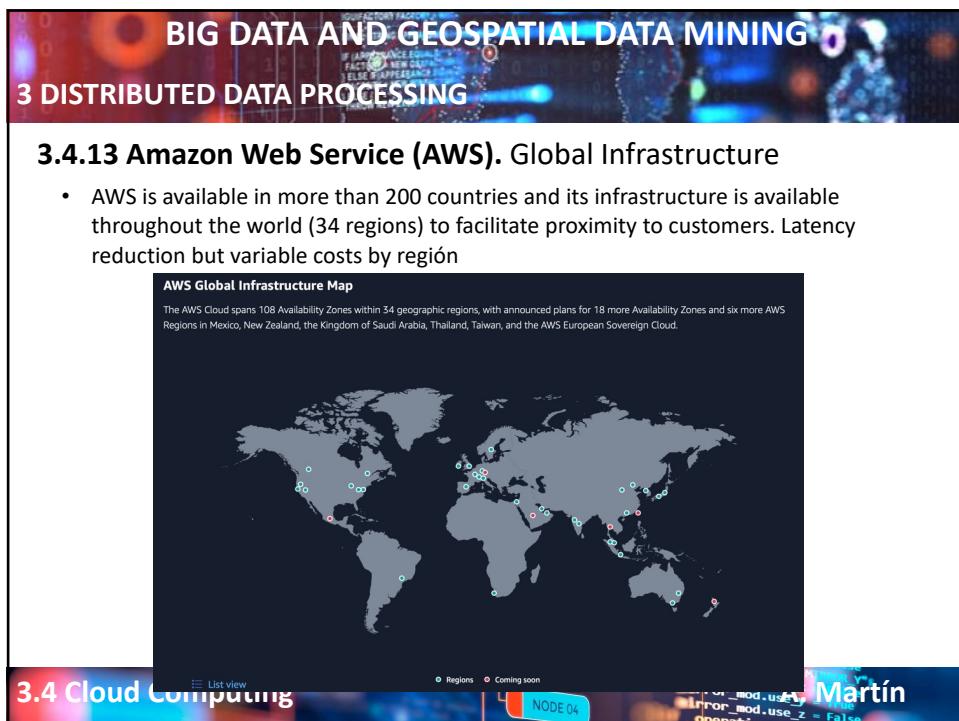
**3.4 Cloud Computing**

A. Martín

30



31



32

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

#### 3.4.13 Amazon Web Service (AWS). Global Infrastructure

- [https://aws.amazon.com/about-aws/global-infrastructure/?nc1=h\\_ls](https://aws.amazon.com/about-aws/global-infrastructure/?nc1=h_ls) (2023)

**3.4 Cloud Computing**

A. Martín

33

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

#### 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)

**3.4 Cloud Computing**

A. Martín

34

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

**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: 99.99999999% 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



**3.4 Cloud Computing**

A. Martín

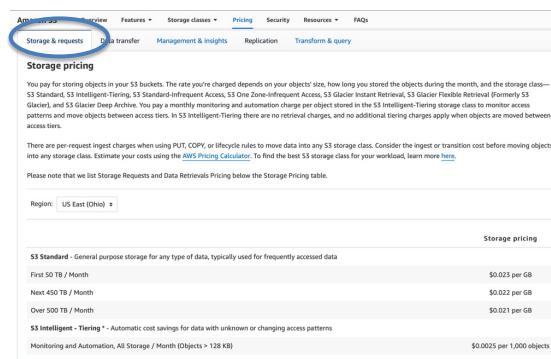
35

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

**3.4.13 Amazon Web Service (AWS). Amazon S3**

- Price scheme (Depending on the region):



Storage class	Price per GB
S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data	\$0.025 per GB
Next 50 TB / Month	\$0.022 per GB
Over 500 TB / Month	\$0.021 per GB
S3 Intelligent - Tiering - Automatic cost savings for data with unknown or changing access patterns	\$0.0025 per 1,000 objects

**3.4 Cloud Computing**

A. Martín

36

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 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
Inrequent Access Tier, All Storage / Month	\$0.0125 per GB
Archive Instant Access Tier, All Storage / Month	\$0.004 per GB
<b>S3 Intelligent - Tiering * - Optional asynchronous Archive tiers</b>	
Archive Access Tier, All Storage / Month	\$0.0036 per GB
Deep Archive Tier, All Storage / Month	\$0.00099 per GB
<b>S3 Standard - Infrequent Access ** - For long lived and infrequently accessed data that needs millisecond access</b>	
All Storage / Month	\$0.0125 per GB
<b>S3 One Zone - Infrequent Access *** - For re-creatable infrequently accessed data that needs millisecond access</b>	
All Storage / Month	\$0.01 per GB
<b>S3 Glacier Instant Retrieval *** - For long-lived archive data accessed once a quarter with instant retrieval in milliseconds</b>	
All Storage / Month	\$0.004 per GB
<b>S3 Glacier Flexible Retrieval (Formerly S3 Glacier) **** - For long-term backups and archives with retrieval option from 1 minute to 12 hours</b>	
All Storage / Month	\$0.0036 per GB
<b>S3 Glacier Deep Archive *** - For long-term data archiving that is accessed once or twice a year and can be restored within 12 hours</b>	
All Storage / Month	\$0.00099 per GB

**3.4 Cloud Computing**

A. Martín

37

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

### 3.4.13 Amazon Web Service (AWS). Amazon S3

- Price scheme (Depending on the region):

Requests & data retrievals

You pay on 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. For reference, S3 developer guide for technical details on the following request types: PUT, COPY, POST, LIST, GET, HEAD, and DELETE. Requests for S3 Batch Operations, S3 Batch GET, and CANCEL requests are not charged. Requests for S3 Batch Operations 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. Refer to 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
Inrequent 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	\$0.05

**3.4 Cloud Computing**

A. Martín

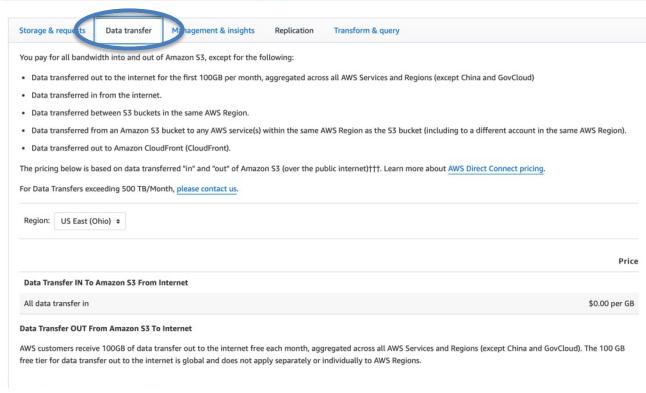
38



### 3 DISTRIBUTED DATA PROCESSING

#### 3.4.13 Amazon Web Service (AWS). Amazon S3

• Price scheme (Depending on the region):



The screenshot shows the AWS S3 Pricing page for the US East (Ohio) region. It highlights the 'Data transfer' tab. The page explains that you pay for all bandwidth into and out of Amazon S3, except for the first 100GB per month. It lists several types of data transfers and their descriptions. Below this, it shows the price for Data Transfer IN To Amazon S3 From Internet, which is \$0.00 per GB. It also notes that 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.

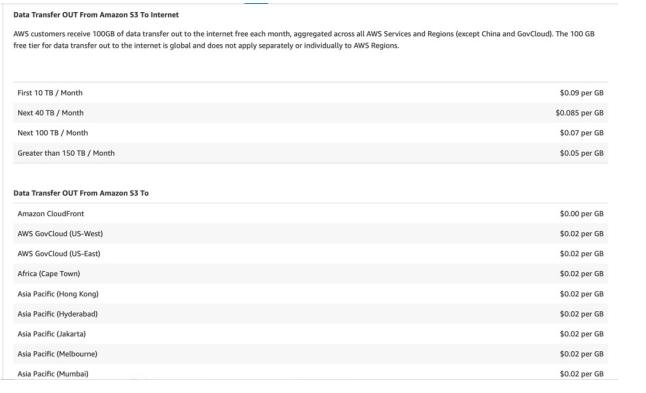
39



### 3 DISTRIBUTED DATA PROCESSING

#### 3.4.13 Amazon Web Service (AWS). Amazon S3

• Price scheme (Depending on the region):



The screenshot shows the AWS S3 Pricing page for the US East (Ohio) region, focusing on the 'Data Transfer OUT From Amazon S3 To' section. It details the pricing for data transfer out to the internet, noting that 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. The table below shows the price per GB for data transfer out to various destinations.

Destination	Price
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

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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.

The diagram illustrates the Amazon EC2 architecture. On the left, a user icon interacts with a green vertical bar labeled "Amazon EC2 API". Two arrows point from the user to the API: one labeled "VMs Deployment" pointing right, and another labeled "VMs Terminate" pointing left. To the right of the API is a light blue rounded rectangle containing several server racks. Each rack is labeled "Amazon EC2" and contains multiple orange circles labeled "VM".

**3.4 Cloud Computing**

A. Martín

41

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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 (Virginia, USA).

**3.4 Cloud Computing**

A. Martín

42

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

**3.4.13 Amazon Web Service (AWS). Amazon EC2**

[https://aws.amazon.com/ec2/instance-types/?nc1=h\\_ls](https://aws.amazon.com/ec2/instance-types/?nc1=h_ls)

The screenshot shows the AWS EC2 instance types page. The top navigation bar includes links for Products, Solutions, Pricing, Documentation, Learn, Partner Network, AWS Marketplace, Customer Enablement, Events, Explore More, Contact Us, Support, English, My Account, and Sign In to the Console. The main content area is titled "Amazon EC2" and "General Purpose". It lists various instance types: M2g, M7i, M7i-Plus, M7a, Mac, M6g, M6i, M6in, M6a, M4, T4g, T3, T3a, T2, and M5. The M5 instance is highlighted with a blue border. A sidebar on the left provides links for PAGE CONTENT, General Purpose, Compute Optimized, Memory Optimized, Accelerated Computing, Storage Optimized, HPC Optimized, and Instance Features. A note at the bottom states: "Amazon EC2 MS instances are the latest generation of General Purpose Instances powered by Intel Xeon® Platinum 8175M or 8259CL processors. These instances provide a balance of compute, memory, and network resources, and is a good choice for many applications." It also lists features such as up to 3.1 GHz Intel Xeon Scalable processor (Skylake 8175M or Cascade Lake 8259CL) with new Intel Advanced Vector Extension (AVX-512) instruction set, and larger instance sizes like m5.24xlarge offering 96 vCPUs and 384 GB of memory.

**3.4 Cloud Computing**

A. Martín

43

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

**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 Cloud Computing**

A. Martín

44

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

#### 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

45

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

#### 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.

46

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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 volume.

**3.4 Cloud Computing**

A. Martín

47

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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

**3.4 Cloud Computing**

A. Martín

48

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

**3.4.13 Amazon Web Service (AWS). Amazon EC2**

The diagram illustrates the architecture of Amazon EC2. It shows an ELB (Elastic Load Balancer) at the bottom left, represented by a blue rounded rectangle with three arrows pointing towards it from the left. These arrows are colored blue, red, and green, representing different traffic paths. The ELB is connected to four instances, each labeled 'Instancia' followed by a number (1, 2, 3, or 4). These instances are arranged in two groups: 'Disponibilidad Zone 1' (Availability Zone 1) contains 'Instancia 1' (blue) and 'Instancia 2' (orange); 'Disponibilidad Zone 2' (Availability Zone 2) contains 'Instancia 3' (red) and 'Instancia 4' (green). Each instance is represented by a teal-colored rounded rectangle.

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

**3.4 Cloud Computing**

A. Martín

49

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

**3.4.13 Amazon Web Service (AWS). Amazon EC2**

The diagram illustrates scaling strategies for Amazon EC2 instances. It is divided into two main sections: 'Vertical Scale (Scale Up/Scale Down)' and 'Horizontal Scale (Scale Out/Scale In)'.

- Vertical Scale (Scale Up/Scale Down):** This section shows a sequence of four blue rounded rectangles representing instances. The first instance is labeled 'MV small'. An arrow labeled 'Load/Traffic increase' points to the second instance, labeled 'MV medium'. Another arrow labeled 'Load/Traffic increase' points to the third instance, labeled 'MV large'. A final arrow labeled 'Load/Traffic decrease' points to the fourth instance, labeled 'MV medium'.
- Horizontal Scale (Scale Out/Scale In):** This section shows a sequence of instances starting with '1 MV'. An arrow labeled 'Load/Traffic increase' leads to '2 MVs'. Another arrow labeled 'Load/Traffic increase' leads to '4 MVs'. A final arrow labeled 'Load/Traffic decrease' leads to '3 MVs'.

- Automatic scaling.

**3.4 Cloud Computing**

A. Martín

50

**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**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-).

**3.4 Cloud Computing**

A. Martín

51

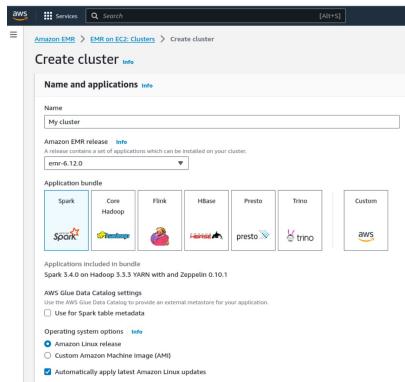
**BIG DATA AND GEOSPATIAL DATA MINING**

**3 DISTRIBUTED DATA PROCESSING**

**3.4.13 Amazon Web Service (AWS). Amazon EMR**



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



**3.4 Cloud Computing**

A. Martín

52

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

#### 3.4.13 Amazon Web Service (AWS). Amazon EMR

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

Instance groups

**Primary**

Choose EC2 instance type  
m5.large  
4 vCore - 16 GB memory  
On-Demand price: \$0.192 per instance/hour  
Lower Spot price: \$0.089 (us-west-2d)

**Core**

Choose EC2 instance type  
m5.large  
4 vCore - 16 GB memory  
On-Demand price: \$0.192 per instance/hour  
Lower Spot price: \$0.089 (us-west-2d)

**Task 1 of 1**

Name: Task-1

Choose EC2 instance type  
m5.large  
4 vCore - 16 GB memory  
On-Demand price: \$0.192 per instance/hour  
Lower Spot price: \$0.089 (us-west-2d)

A. Martín

53

**BIG DATA AND GEOSPATIAL DATA MINING**

### 3 DISTRIBUTED DATA PROCESSING

#### 3.4.13 Amazon Web Service (AWS). Amazon EMR

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

Your cluster "Prueba" has been successfully created.

Prueba

Summary

Cluster info

Cluster ID: j-2KQ2Q58EG35A

Cluster configuration

Instance groups

Capacity: 1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version: emr-6.12.0

Installed applications: Hadoop 3.3.5, Hive 3.1.0, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2

Cluster management

Status and time

Status: Waiting

Creation time: August 11, 2023, 12:51 (UTC+02:00)

Elapsed time: 36 minutes, 44 seconds

Log destination in Amazon S3: abd2023/logs

Persistent application UI: YARN timeline server, Tez UI

Primary node public DNS: ec2-54-203-184-156.us-west-2.compute.amazonaws.com

Connect to the Primary Node using SSH

Properties Bootstrap actions Instances (Hardware) Steps Applications Configurations Monitoring Events Tag (0)

Operating system info

Amazon Linux release: 2.0.2020727.0

Cluster logs info

Archive log files to Amazon S3: Turned off

Amazon S3 location: s3://awslogs/logs/

Encryption for logs: Turned off

Cluster termination info

Termination option: Manually terminate cluster

Idle time: Turned off

Termination protection: Turned off

A. Martín

54

# BIG DATA AND GEOSPATIAL DATA MINING

## 3 DISTRIBUTED DATA PROCESSING

### 3.4.13 Amazon Web Service (AWS). Amazon EMR

The screenshot shows the AWS EC2 Management Console with the following details:

- Properties > Prueba > Instances > EC2 Manager**
- Instances**: 1 instance
- Region**: us-west-2
- Cluster ID**: j-1H2Q2Y53RGC5A
- Name**: Prueba
- Summary**: Cluster info, Cluster logs, Cluster metrics, Cluster status, Cluster properties, Operating system, Network and security.
- Connect to the primary node using SSH**: A modal window is open with the following content:
  - Instructions: "You can connect to the Amazon EMR primary node using SSH to perform actions like running interactive queries, examining log files, submit Linux commands, and view web-interface hosted on Amazon EMR clusters. Learn more"
  - Terminal type: Windows (selected) or Mac/Linux
  - SSH command: "1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found under Applications > Accessories > Terminal."  
"2. To establish a connection to the primary node, enter the following command. Replace ~/awsroot/.aws/keys/keypair.pem with the location and filename of the private key file you used to launch the cluster."  
"ssh -i /home/ec2-user/.aws/keys/keypair.pem hadoop@ec2-54-203-184-156.us-west-2.compute.amazonaws.com"
  - Warning: "Enter yes to dismiss the security warning."
- View web interfaces hosted on Amazon EMR clusters**: A link to <https://us-west-2.console.aws.amazon.com/AmazonEMR/clusters/j-1H2Q2Y53RGC5A>

## 3.4 Cloud Computing

A. Martín

55

# BIG DATA AND GEOSPATIAL DATA MINING

## 3 DISTRIBUTED DATA PROCESSING

### 3.4.13 Amazon Web Service (AWS). Amazon EMR

The logo for Amazon Web Services (AWS) features a stylized orange 3D block icon.

- 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

## 3.4 Cloud Computing

The logo for Cloud Computing features a blue button-like shape with the text "NODE 04" and a small orange icon.

A. Martin

56

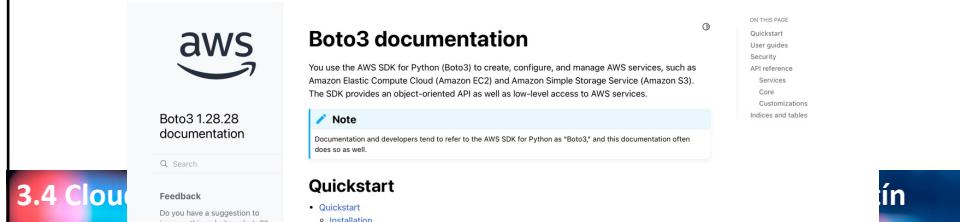


### 3 DISTRIBUTED DATA PROCESSING

#### 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
  - CloudWatch
  - SQS (Simple Queue Service)

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



57



### 3 DISTRIBUTED DATA PROCESSING

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

##### mrjob v0.7.4 documentation

Home      Guides →

**Quick Links**

- Fundamentals
- Writing jobs
- Runners
- Amazon Elastic MapReduce
- Google Cloud Dataproc
- Config quick reference
- Config options (all runners)
- Config options (Hadoop)
- Config options (cloud services)
- Amazon Elastic MapReduce
- Google Cloud Dataproc

**mrjob**

mrjob lets you write MapReduce jobs in Python 2.7/3.4+ and run them on several platforms.

You can:

- Write multi-step MapReduce jobs in pure Python
- Test on your local machine
- Run on a Hadoop cluster
- Run in the cloud using [Amazon Elastic MapReduce \(EMR\)](#)
- Run in the cloud using [Google Cloud Dataproc \(Dataproc\)](#)
- Easily run [Spark](#) jobs on EMR or your own Hadoop cluster

mrjob is licensed under the [Apache License, Version 2.0](#).

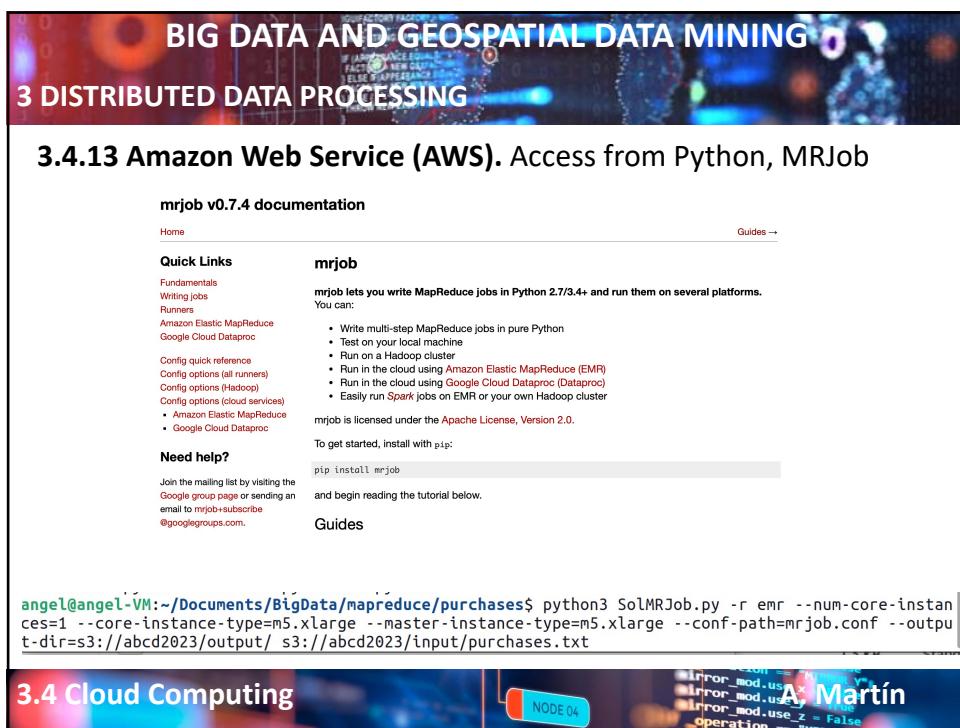
To get started, install with pip:

```
pip install mrjob
```

and begin reading the tutorial below.

**Guides**

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



58