



1

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4 Cloud Computing

* NIST word cloud

2

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.1 Que es Cloud Computing

- Servicio, a través de Internet, que ofrece, por parte de un proveedor a múltiples clientes:
 - a) Almacenamiento: Permite almacenar y acceder a datos guardados en el cloud, generalmente en forma de ficheros (aunque puede ser una BD alojada en la nube).
 - b) Cómputo: en forma de máquinas virtuales con unas determinadas características (RAM, CPU, disco, etc.) y configuración (SO).
- Aprovechamiento de la economía de escala de grandes proveedores para ofrecer ahorro de costes a los usuarios.
- Pago por uso, sin inversiones iniciales (pago por horas de CPU, por Gbyte almacenado y por Gbyte descargado).

3.4 Cloud Computing

A. Martín

3

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.2 Ejemplo

- Una start-up desarrolla una aplicación para móviles con una idea innovadora que requiere capacidad de cómputo y almacenamiento de datos para su puesta en producción.
 - Opcion A: Adquirir servidores y realizar el *housing* y el *hosting* de la aplicación en la propia infraestructura de la empresa (In-House).
 - Opción B: Aprovisionar los recursos necesarios de un proveedor Cloud.

3.4 Cloud Computing

A. Martín

4

BIG DATA Y MINERÍA DE DATOS GEOESPAZIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.2 Ejemplo

Opción A: In-House

- Alquilar y acondicionar un local (refrigeración, cableado, SAIs, etc.).
- Dimensionar y adquirir hardware para cómputo y almacenamiento, actualizarlo periódicamente.
- Configurar los recursos, actualizarlos.

3.4 Cloud Computing

A. Martín

5

BIG DATA Y MINERÍA DE DATOS GEOESPAZIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.2 Ejemplo

Opción A: In-House

- Si la aplicación triunfa menos de lo esperado:
 - Eres víctima de tu propio fracaso:
 - La inversión en HW no se rentabiliza y las deudas pueden provocar el cierre de la empresa.
- Si la aplicación triunfa más de lo esperado:
 - Eres víctima de tu propio éxito:
 - No es posible servir las peticiones de los clientes con la calidad de servicio esperada y los clientes se van.

3.4 Cloud Computing

A. Martín

6

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.2 Ejemplo

Opción B (Cloud computing)

- Aprovisionar los recursos de cómputo iniciales necesarios para la puesta en producción.
- Configurar la aplicación para que auto-aprovisione y libere dinámicamente nuevos recursos (cómputo, almacenamiento) dependiendo de la carga de trabajo / usuarios de la misma.

3.4 Cloud Computing

A. Martín

7

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.2 Ejemplo

Opción B: Cloud computing

- Si la aplicación triunfa menos de lo esperado:
 - Reducción de usuarios:
 - Se liberan recursos no utilizados y únicamente se paga por el consumo realizado.
- Si la aplicación triunfa más de lo esperado:
 - Aumento de usuarios:
 - Se solicitan más recursos del proveedor Cloud de forma automática para satisfacer las peticiones de los usuarios.

3.4 Cloud Computing

A. Martín

8

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.3 Definición

La Computación en Nube es un modelo para permitir el **acceso ubicuo**, conveniente y **bajo demanda** mediante red a un conjunto compartido de **recursos de cómputo configurables** (i.e., redes, servidores, almacenamiento, aplicaciones y servicios) que pueden ser **rápidamente aprovisionados y liberados** con **mínimo esfuerzo** de gestión o interacción con el proveedor del servicio.

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 Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.4 Necesidad de un servicio Cloud

- Las inversiones en Hardware quedan obsoletas a gran velocidad:
 - Hay que maximizar el uso eficiente de los recursos
 - Inasequible para pequeños centros/Pymes/Start-Ups
- La demanda de recursos (almacenamiento, cómputo) es muy variable.
- Hay que ajustar el consumo de recursos a las necesidades de las aplicaciones de forma dinámica y rápida

3.4 Cloud Computing A. Martín

10

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.4 Necesidad de un servicio Cloud

The graph illustrates the cost of infrastructure over time for four different approaches:

- Predicted demand**: A dashed black line representing the estimated future demand.
- Actual demand**: A solid red line showing the actual fluctuating demand.
- Scale-up approach**: A dotted blue line representing a fixed capacity that is scaled up during peak demand.
- Traditional Scale-out approach**: A dashed green line representing a step-wise increase in capacity.
- Automated Elasticity**: A solid green line representing a smooth, dynamic response to actual demand.

Annotations on the graph highlight several issues with traditional approaches:

- Huge Capital Expenditure**: Indicated by a vertical arrow pointing to the initial high cost of the scale-up approach.
- Too much excess capacity "Opportunity Cost"**: Indicated by a vertical arrow pointing to the difference between predicted and actual demand at low times.
- You just lost your customers**: Indicated by a vertical arrow pointing to a sharp drop in demand followed by a loss of revenue.

Fuente: AWS_Cloud_Best_Practices.pdf, disponible en <http://aws.amazon.com/whitepapers/>

3.4 Cloud Computing A. Martín

11

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.5 Características de un servicio Cloud

- Auto-Servicio Bajo Demanda.
 - Un consumidor puede proveerse de forma unilateral de recursos sin interactuar con personal del proveedor del servicio.
- Acceso a Través de Internet
 - Las capacidades se proporcionan a través de la red con unos mínimos requerimientos en el cliente.
- Elasticidad.
 - El consumidor puede dinámicamente incrementar o decrementar el número de recursos en cualquier momento, percibiendo una ilusión de capacidad infinita.

3.4 Cloud Computing A. Martín

12

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.5 Características de un servicio Cloud

a) Escalado Vertical (Scale Up/Scale Down)

```

graph LR
    A[MV pequeña] -- "Incremento de Carga/Traffic" --> B[MV Mediana]
    B -- "Incremento de Carga/Traffic" --> C[MV Grande]
    C -- "Decremento de Carga/Traffic" --> D[MV Mediana]
  
```

b) Escalado Horizontal (Scale Out/Scale In)

```

graph LR
    A[1 MV] -- "Incremento de Carga/Traffic" --> B[2 MVs]
    B -- "Incremento de Carga/Traffic" --> C[4 MVs]
    C -- "Decremento de Carga/Traffic" --> D[3 MVs]
  
```

3.4 Cloud Computing **A. Martín**

13

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.5 Características de un servicio Cloud

- Servicio Mediante Pago por Uso
 - Los recursos utilizados se contabilizan de forma independiente (almacenamiento, computación, ancho de banda, etc.) y precisa para poder implementar el pago por uso, tomando como unidad de referencia típicamente la hora.
- Configurabilidad
 - Los recursos alquilados deben poder ser altamente configurables para adaptarse a las necesidades de los diferentes usuarios. Esto será más o menos posible dependiendo del modelo de servicio (IaaS, PaaS, SaaS).

3.4 Cloud Computing **A. Martín**

14

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.5 Características de un servicio Cloud

- Separación
 - Cloud computing proporciona recursos “en alquiler” bajo un modelo de pago por uso pero no expone los detalles de la infraestructura a los clientes o socios.
 - Los usuarios utilizan los recursos sin conocer los detalles de la infraestructura de los proveedores.
- Aislamiento
 - Dada la naturaleza de anfitrión de los proveedores de Cloud, los consumidores necesitan mecanismos y garantías de que sus aplicaciones se encuentran aisladas del resto de los clientes alojados en la misma infraestructura.

3.4 Cloud Computing

A. Martín

15

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.6 Tecnologías básicas: DataCenters

- Google, Amazon, Microsoft, etc. tienen grandes centros de datos por todo el mundo (diferentes localizaciones para protegerse frente a fallos)
- Uso de técnicas de eficiencia energética para reducir el consumo derivado de su operación: apagado/encendido de nodos de forma dinámica, gestión apropiada de la climatización etc.




Microsoft data center de Dublin, con más de 60.000 m² construidos.

3.4 Cloud Computing

A. Martín

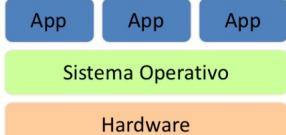
16

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

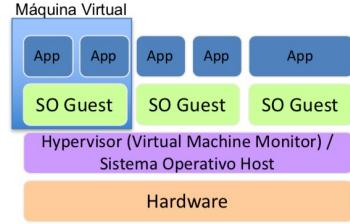
3.4.6 Tecnologías básicas: Virtualización

- La virtualización permite crear un (o varios) entorno simulado (Maquina Virtual, MV) que ejecuta un SO *guest* (invitado). Todo ello corriendo sobre un SO *host* (anfitrión) con ayuda de un hipervisor (o *Virtual Machine Monitor*).



Plataforma tradicional

Este diagrama muestra una jerarquía de tres niveles: en el nivel superior, tres cuadros azules rotulados como "App"; en el nivel medio, un cuadro verde rotulado como "Sistema Operativo"; en el nivel inferior, un cuadro naranja rotulado como "Hardware".



Máquina Virtual

Este diagrama muestra una jerarquía de cuatro niveles: en el nivel superior, cinco cuadros azules rotulados como "App"; en el segundo nivel, tres cuadros verdes rotulados como "SO Guest"; en el tercer nivel, un cuadro morado rotulado como "Hypervisor (Virtual Machine Monitor) / Sistema Operativo Host"; en el cuarto nivel, un cuadro naranja rotulado como "Hardware".

- Una aplicación puede ejecutarse sobre una versión específica de SO, por encima de un hardware moderno
- Permite incrementar la tasa de utilización del hardware al ejecutar más máquinas virtuales sobre el mismo equipo físico.

3.4 Cloud Computing

17

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.7 Principales proveedores Cloud

- Amazon Web Services
 - Incluye servicios para el aprovisionamiento dinámico de capacidad de cómputo así como la gestión y almacenamiento eficiente de datos y el diseño escalable de aplicaciones en la nube mediante un modelo de pago por uso.



amazon
webservices™

3.4 Cloud Computing

18

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.7 Principales proveedores Cloud

- Windows Azure
 - Plataforma Cloud de Microsoft. Desarrollo de aplicaciones (.NET) alojadas que combinan web, bases de datos SQL, almacenamiento de ficheros, etc., sobre una infraestructura virtual basada en Windows.



Windows Azure

3.4 Cloud Computing

A. Martín

19

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.7 Principales proveedores Cloud

- Google App Engine
 - Solución de Google para crear y alojar aplicaciones web (Java/Python) en la nube.
 - Control de recursos consumidos por la aplicación para ajustarse al presupuesto.
- Google Compute Engine
 - Solución de Google para el aprovisionamiento de infraestructura de cómputo en la forma de máquinas virtuales (Linux).



3.4 Cloud Computing

A. Martín

20

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.8 Desafíos de una infraestructura Cloud

- Desafío 1: Disponibilidad de servicio y *data lock-in*.
- Desafío 2: Privacidad de los datos y aspectos de seguridad.
- Desafío 3: Prestaciones no deterministas y cuellos de botella.
- Desafío 4: Almacenamiento distribuido.
- Desafío 5: Escalabilidad, Interoperabilidad y Estandarización.
- Desafío 6: Licencias de software
- Desafío 7: Compartición de Reputación.

3.4 Cloud Computing

A. Martín

21

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.9 Confianza y reputación

- Externalizar el cómputo y el almacenamiento a un tercero supone confiar en él.
- Un fallo en el proveedor puede afectar a su supervivencia y a la de las empresas dependientes.
- La reputación del proveedor está en juego ante disruptpciones en el servicio (*outages*).
 - Tormenta eléctrica en Julio de 2012 que afectó a AWS y clientes (Netflix, etc.), bug en AWS causó caída de servicio en Octubre de 2012 que afectó a Reddit, Foursquare, etc.
 - Caída de Windows Azure en Febrero de 2012 durante varias horas por culpa de un certificado de seguridad.

3.4 Cloud Computing

A. Martín

22

BIG DATA Y MINERÍA DE DATOS GEOESPAZIALES

23

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

24

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.12 Casos de éxito. Análisis de terremotos

- Utiliza de forma elástica hasta 100 cores Azure para simular las ondas de propagación sísmicas y su impacto.
- Automáticamente captura los datos de los registros sísmicos ofreciendo en tiempo casi real información sobre las zonas afectadas.
- No tendría sentido tener un cluster de 100 nodos dedicado para eventos que suceden con baja frecuencia.



The screenshot shows the CLOUD-QUAKE software interface. It features a main map of a coastal city area with a green hexagonal grid overlay. Yellow dots are scattered across the map, indicating seismic activity. A small inset map is visible in the bottom right corner, and a color scale legend is at the bottom. The interface includes tabs for 'Shake map simulation', 'Grid management', and 'About'.

3.4 Cloud Computing

A. Martín

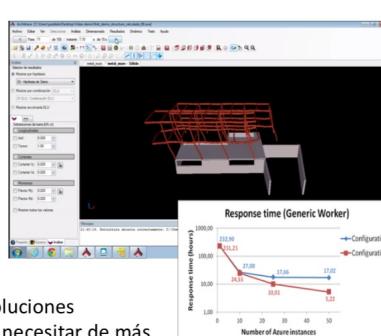
25

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.12 Casos de éxito. Cálculo de estructuras

- La simulación del comportamiento dinámico de estructuras de edificación permite determinar su respuesta ante terremotos.
- Este análisis se debe realizar ante diferentes terremotos y para diferentes materiales.
- Una simulación compleja de un edificio de 10 pisos ante 5 registros de terremotos y 10 soluciones estructurales diferentes puede necesitar de más de 4 días a resolverse en 5 horas con 50 cores Azure
 - Menos de 50€ de coste para este estudio.



The screenshot shows a software interface for structural engineering analysis. On the left, there's a 3D model of a building frame. To the right, a line graph plots 'Response time (Seconds)' on the y-axis (ranging from 0.00 to 1000.00) against the 'Number of Azure instances' on the x-axis (ranging from 0 to 50). Two data series are shown: Configuration A (blue line with circles) and Configuration B (red line with squares). Configuration A shows response times increasing from approximately 212.50 seconds at 5 instances to 374.02 seconds at 50 instances. Configuration B shows response times decreasing from approximately 212.50 seconds at 5 instances to 11.52 seconds at 50 instances.

3.4 Cloud Computing

A. Martín

26

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

"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 Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Cloud de Amazon. Amazon Web Service (AWS)

- Amazon Web Services (AWS) ofrece servicios de infraestructuras para ejecutar aplicaciones en la nube.
- Es pionero en Cloud Computing desde 2006 ofreciendo:
 - Bajo Coste
 - Agilidad y elasticidad instantánea
 - Abierto y flexible
 - Seguro



3.4 Cloud Computing  **A. Martín**

28

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

29

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Productos

Console Home > All services [Options]

All services

Services by category

- Compute**
 - EC2
 - Lambda
 - Batch
 - Elastic Beanstalk
 - Serverless Application Repository
 - AWS Outposts
 - EC2 Container Service
 - AWS App Runner
 - AWS Step Functions
- Containers**
 - Elastic Container Service
 - Elastic Kubernetes Service
 - Red Hat OpenShift Service on AWS
 - Elastic Container Registry
- Storage**
 - S3
 - EPS
 - FPS
 - S3 Glacier
 - Storage Gateway
 - AWS Backup
 - AWS Elastic Disaster Recovery
- Database**
 - RDS
 - ElastiCache
- Customer Enablement**
 - Aws IQ
 - Managed Services
 - Active for Startups
 - Support
 - AWS RePost Private
- Robotics**
 - AWS RoboMaker
- Blockchain**
 - Amazon Managed Blockchain
- Satellite**
 - Ground Station
- Quantum Technologies**
 - Amazon Braket
- Management & Governance**
 - AWS Organizations
 - CloudWatch
 - AWS Lambda Scaling
 - CloudFormation
 - AWS Config
 - DynamoDB
 - Service Catalog
 - Systems Manager
 - Trusted Advisor
 - CloudFront
- Machine Learning**
 - Amazon SageMaker
 - Amazon Augmented AI
 - Amazon CodeGuru
 - Amazon DevOps Guru
 - Amazon Forecast
 - Amazon Fraud Detector
 - Amazon Personalize
 - Amazon Polly
 - Amazon Rekognition
 - Amazon Transcribe
 - Amazon Translate
 - Amazon DeepLearning
 - AWS DeepCluster
 - AWS Panorama
 - Amazon Lookout
 - AWS HealthLake
 - Amazon Lookout for Vision
 - Amazon Lookout for Equipment
 - Amazon Lookout for Metrics
 - Amazon Lex
 - Amazon Personalized Medical
 - AWS HealthMetrics
 - Amazon Bedrock
 - AWS ImageIngesting
 - Amazon Q
 - Amazon Q Business
- Analytics**
- Cloud Financial Management**
 - AWS Marketplace
 - AWS Billing Conductor
 - Billing and Cost Management
- Front-end Web & Mobile**
 - AWS Amplify
 - AWS AppSync
 - Device Farm
 - Amazon Location Service
- Application Integration**
 - Step Functions
 - Amazon Connect
 - Amazon MQ
 - Simple Notification Service
 - Simple Queue Service
 - SWF
 - Managed Apache Airflow
 - Amazon EventBridge
 - AWS Lambda Message
- Business Applications**
 - Amazon Connect
 - Amazon Chime
 - Amazon Simple Email Service
 - Amazon WorkDocs
 - Amazon WorkMail
 - AWS Supply Chain
 - AWS AppFabric

30

BIG DATA Y MINERÍA DE DATOS GEOESPAZIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Productos

The image shows a screenshot of the AWS CloudFront dashboard. It lists several active and failed distributions. Active distributions include 'www.mercadolibre.com.ar' (status OK), 'www.mercadolibre.com' (status OK), 'www.mercadolibre.com.br' (status OK), 'www.mercadolibre.com.mx' (status OK), 'www.mercadolibre.com.pe' (status OK), 'www.mercadolibre.com.co' (status OK), 'www.mercadolibre.com.uy' (status OK), 'www.mercadolibre.com.ve' (status OK), 'www.mercadolibre.com.ec' (status OK), 'www.mercadolibre.com.ar' (status OK), and 'www.mercadolibre.com.uy' (status OK). Failed distributions include 'www.mercadolibre.com.ve' (status Failed) and 'www.mercadolibre.com.ve' (status Failed).

Nombre	Estado
www.mercadolibre.com.ar	OK
www.mercadolibre.com	OK
www.mercadolibre.com.br	OK
www.mercadolibre.com.mx	OK
www.mercadolibre.com.pe	OK
www.mercadolibre.com.co	OK
www.mercadolibre.com.uy	OK
www.mercadolibre.com.ve	OK
www.mercadolibre.com.ec	OK
www.mercadolibre.com.ar	OK
www.mercadolibre.com.uy	OK
www.mercadolibre.com.ve	Failed
www.mercadolibre.com.ve	Failed

Database

- RDS
- ElastiCache
- Nearstream
- Amazon QLDB
- Amazon DocumentDB
- Amazon Keyspaces
- Amazon TimeStream
- DynamoDB
- Amazon MemoryDB

Migration & Transfer

- AWS Migration Hub
- AWS Application Migration Service
- Application Discovery Service
- Database Migration Service
- AWS Transfer Family
- AWS Server Migration Service
- DataSync
- AWS Lambda
- AWS Mainframe Modernization

Networking & Content Delivery

- VPC
- CloudFront
- Route 53
- API Gateway
- Direct Connect
- AWS App Mesh
- Global Accelerator
- AWS Cloud Map
- Route 53 Analysis & Recovery Controller
- AWS Private 5G

Developer Tools

- CodeStar
- CodeCommit
- CodeBuild
- CodeDeploy
- CodePipeline
- Cloud9

Service Catalog

- Systems Manager
- Trusted Advisor
- CloudWatch
- AWS Well-Architected Tool
- AWS Chatbot
- Launch Wizard
- Resource Optimizer
- Resource Groups & Tag Editor
- Amazon Grafana
- Amazon Prometheus
- Amazon Metrics
- Amazon CloudWatch Metrics Insights
- Incident Manager
- AWS License Manager
- Service Quotas
- Amazon CloudWatch CloudTrail
- CloudWatch Metrics Insights
- AWS Resource Catalog
- AWS User Notifications
- Amazon CloudWatch Metrics Dashboard
- AWS Telos World Builder

Media Services

- Amazon CloudWatch Metrics Streams
- MediaConvert
- MediaLive
- MediaPackage
- MediaStore
- MediaTailor
- Elemental Appliances & Software
- Elastic Transcoder
- Nimble Studio
- Amazon CloudWatch Metrics Insights
- Amazon Interactive Video Service
- AWS Deadline Cloud

Analytics

- Athena
- Amazon Redshift
- CloudSearch
- Amazon OpenSearch Service
- Kinesis
- QuickSight
- Data Pipeline
- AWS Data Exchange
- AWS Lake Formation
- MSK
- AWS Glue DataBrew
- Amazon FinSpace
- AWS Glue
- Amazon Data Firehose
- EMR
- AWS Clean Rooms
- Amazon DataZone
- AWS Entity Resolution
- Managed Apache Flink

Security, Identity, & Compliance

- Resource Access Manager
- Cognito
- Secrets Manager
- GuardDuty
- Amazon Inspector
- Amazon Macie
- IAM Identity Center
- Identity Manager
- Key Management Service
- CloudHSM
- Directory Service
- WAF & Shield
- AWS Firewall Manager
- AWS Artifact
- Detective
- AWS Sismon

Amazon WorkDocs

- Amazon WorkMail
- AWS Shared Chain
- AWS WAF
- Amazon Cloud SMS
- Amazon One Enterprise
- Amazon Pinpoint
- AWS End User Messaging

End User Computing

- WorkSpaces
- AppStream 2.0
- WorkSpaces Secure Browser
- WorkSpaces Thin Client

Internet of Things

- IoT Analytics
- IoT Device Defender
- IoT Device Management
- IoT Greengrass
- IoT SiteWise
- IoT Core
- IoT Events
- AWS IoT FleetWise
- IoT TwinMaker

Game Development

- Amazon GameLift

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookies preferences

3.4 Cloud Computing

31

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

32

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Infraestructura Global

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

3.4 Cloud Computing A. Martín

33

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Infraestructura Global

- Cada región incluye diferentes zonas de disponibilidad por tolerancia a fallos

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 (Oregon)
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)
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)

Diagrama jerárquico:

```

graph TD
    Región[Región] --> ZDA[Zona de Disponibilidad]
    ZDA --> RH[Recursos Hardware]
    
```

Detalles de las regiones:

- USA Este**: Norte de Virginia (6), Ohio (3)
- USA Oeste**: Norte de California (3), Oregón (4)
- ...
- AWS GovCloud (EEUU)**: Oeste (3), Este (3)

Imagen de un corredor de servidores en un data center.

3.4 Cloud Computing A. Martín

34

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon S3

- “Almacenamiento para Internet. Está diseñado para facilitar a los desarrolladores la informática a escala Web”
- Consiste en un sistema de almacenamiento de un número ilimitado de objetos para el entorno Cloud
- Accesible mediante protocolos estándar (HTTP)
- Confiabilidad (Reliability): 99.99999999% de durabilidad en almacenamiento y un 99.99% de disponibilidad.
- Datos almacenados como objetos en *buckets* (depósitos)
 - Un *bucket* es un contenedor de objetos.
 - Lleva control de acceso establecido por el usuario (quién puede crear, borrar y listar el *bucket*)
 - Puede ser auditado (registro de accesos) y versionado
- Un objeto
 - Es un fichero + metadatos de descripción (opcional)
 - Entre 1 byte y 5 Terabytes



3.4 Cloud Computing

A. Martín

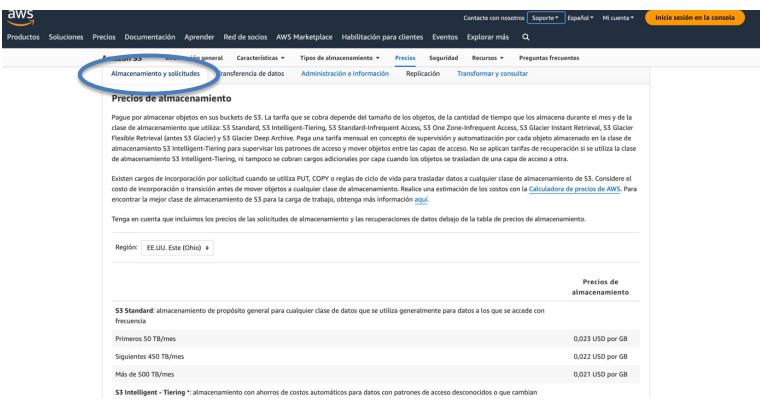
35

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon S3

- Esquema de precios (Dependen de la región):



3.4 Cloud Computing

A. Martín

36

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon S3

- Esquema de precios (Dependen de la región):

S3 Intelligent - Tiering ^{***} : almacenamiento con ahorros de costos automáticos para datos con patrones de acceso desconocidos o que cambian constantemente	
Monitoreo y automatización, todo el almacenamiento/mes (Objetos > 128 KB)	0,0003 USD por 1000 objetos
Capa de acceso frecuente, primeros 50 TB/mes	0,033 USD por GB
Capa de acceso frecuente, siguientes 450 TB/mes	0,022 USD por GB
Capa de acceso frecuente, más de 500 TB/mes	0,021 USD por GB
Nivel de acceso poco frecuente, todo el almacenamiento al mes	0,0125 USD por GB
Nivel de acceso instantáneo a archivos, todo el almacenamiento al mes	0,004 USD por GB
S3 Intelligent - Tiering ^{***} : niveles opcionales de acceso a archivos asincrónos	
Nivel de acceso a archivos, todo el almacenamiento al mes	0,0036 USD por GB
Nivel de acceso a archivo profundo, todo el almacenamiento al mes	0,00099 USD por GB
S3 Standard - Infrequent Access ^{***} : para almacenar datos de larga vida pero con poco acceso que requieren de acceso de milisegundos	
Todo el almacenamiento/mes	0,0125 USD por GB
S3 One Zone - Infrequent Access ^{***} : almacenamiento de datos recreables de acceso poco frecuente que requieren de acceso en milisegundos	
Todo el almacenamiento/mes	0,01 USD por GB
S3 Glacier Instant Retrieval ^{***} : almacenamiento de datos de archivo cada tres meses con recuperación instantánea en milisegundos	
Todo el almacenamiento/mes	0,004 USD por GB
S3 Glacier Flexible Retrieval (antes S3 Glacier) ^{***} : almacenamiento de copias de seguridad y archivos a largo plazo con opción de recuperación de 1 minuto a 12 horas	
Todo el almacenamiento/mes	0,0036 USD por GB
S3 Glacier Deep Archive ^{***} : almacenamiento de archivos de datos a largo plazo a los que se accede una o dos veces al año y que se pueden restaurar dentro de un plazo de 12 horas	
Todo el almacenamiento/mes	0,00099 USD por GB

3.4 Cloud Computing 

37

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon S3

- Esquema de precios (Dependen de la región):

Solicitudes y recuperaciones datos					
Paga por las transacciones realizadas sin buckets y objetos de S3. Los costos de solicitud de S3 se basan en el tipo de solicitud y se cobran en función de la cantidad de solicitudes, como se describe en la siguiente tabla. Cuando se utiliza la consola de Amazon S3 para buscar su almacenamiento, se aplicarán cargos por las solicitudes GET, LIST u otras que se realizan para facilitar la tarea. Los cargos se acumulan a la misma velocidad que las solicitudes que crean la API o el SDK. Consulte la guía para desarrolladores de S3 para obtener descripciones detalladas de las solicitudes PUT, COPY, POST, LIST, GET, SELECT, HEAD, DELETE, HEAD, COPY, POST, LIST, GET, SELECT, HEAD, DELETE y HEAD.					
Las solicitudes DELETE y CANCEL son gratuitas. A las solicitudes LIST de cualquier clase de almacenamiento se les aplica la misma tarifa que a las solicitudes PUT, COPY y POST en S3 Standard. Paga por la recuperación de objetos que se almacenan en S3 Standard - Infrequent Access, S3 One Zone - Infrequent Access, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval o S3 Glacier Deep Archive. Consulte la guía para desarrolladores de S3 para obtener detalles técnicos sobre las Recuperaciones de datos.					
Los siguientes precios de las solicitudes de transición de ciclo de vida de S3 corresponden a las solicitudes de esa clase de almacenamiento. Por ejemplo, la transición de datos de S3 Standard a S3 Standard - acceso poco frecuente se cobrará a 0,01 USD por cada 1000 solicitudes.					
No hay cargos de recuperación en S3 Intelligent-Tiering. Si luego se accede a un objeto de la capa de acceso poco frecuente, automáticamente regresa a la capa de acceso frecuente. No se aplican cargos adicionales a las capacidades cuando los objetos se desplazan entre las capas de acceso dentro del tipo de almacenamiento S3 Intelligent-Tiering.					
Región	EE.UU. Este (Ohio) *				
	Solicitudes PUT, COPY, POST y LIST (por 1000 solicitudes)	GET, SELECT y el resto de las solicitudes (por 1000 solicitudes)	Solicitudes de transición de ciclo de vida (por 1000 solicitudes)	Solicitudes de recuperación de datos (por 1000 solicitudes)	Recuperaciones de datos (por GB)
S3 Standard	0,005 USD	0,0004 USD	n/a	n/a	n/a
S3 Intelligent-Tiering [*]	0,005 USD	0,0004 USD	0,01 USD	n/a	n/a
Acceso frecuente	n/a	n/a	n/a	n/a	n/a
Acceso poco frecuente	n/a	n/a	n/a	n/a	n/a
Archivo instantáneo	n/a	n/a	n/a	n/a	n/a
Acceso a archivos, Estándar	n/a	n/a	n/a	n/a	n/a
Acceso a archivos, Masivo	n/a	n/a	n/a	n/a	n/a
Acceso a archivos, Acelerado	n/a	n/a	n/a	10,00 USD	0,03 USD
Acceso a Deep Archive, Standard	n/a	n/a	n/a	n/a	n/a

3.4 Cloud Computing 

38

BIG DATA Y MINERÍA DE DATOS GEOESPAZIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon S3

- Esquema de precios (Dependen de la región):

The screenshot shows the AWS Pricing page for Amazon S3. It highlights the 'Transferencia Entrante' (Incoming Transfer) and 'Transferencia Saliente' (Outgoing Transfer) sections. The 'Transferencia Entrante' section shows a flat rate of 0,00 USD por GB for all traffic. The 'Transferencia Saliente' section shows rates starting at 0,09 USD por GB for the first 10 TB/mes, decreasing to 0,02 USD por GB for traffic to specific regions like Amazon CloudFront, AWS GovCloud, and Asia-Pacific (Hong Kong).

39

BIG DATA Y MINERÍA DE DATOS GEOESPAZIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon S3

- Esquema de precios (Dependen de la región):

The screenshot shows the AWS Pricing page for Amazon S3. It highlights the 'Transferencia Entrante' (Incoming Transfer) and 'Transferencia Saliente' (Outgoing Transfer) sections. The 'Transferencia Entrante' section shows a flat rate of 0,00 USD por GB for all traffic. The 'Transferencia Saliente' section shows rates starting at 0,09 USD por GB for the first 10 TB/mes, decreasing to 0,02 USD por GB for traffic to specific regions like Amazon CloudFront, AWS GovCloud, and Asia-Pacific (Hong Kong).

40

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

- Amazon EC2 es un servicio que proporciona recursos de cómputo en la nube. Permite el despliegue de máquinas virtuales ó instancias de tamaño predefinido para disponer de cómputo bajo demanda mediante un modelo de pago por uso.

The diagram illustrates the Amazon EC2 architecture. On the left, a user icon interacts with a green rounded rectangle labeled "Amazon EC2 API". Inside this box, there are two horizontal arrows: one pointing right labeled "Desplegar VMs" and one pointing left labeled "Terminar VMs". To the right of the API box is a large white area labeled "Amazon EC2". Inside this area, there are several blue server racks. Orange circles containing the letters "VM" represent virtual machines running on these servers. The entire diagram is set against a background of blurred digital data and code snippets.

3.4 Cloud Computing

A. Martín

41

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

- AWS utiliza el término *instancia* para referirse a una máquina virtual (MV), una instancia proporciona una cantidad predecible de capacidad de cómputo dedicada.
- El despliegue de una instancia requiere indicar:
 - Tipo de instancia, hay tipos predefinidos de instancia cuyo precio por hora varía según las prestaciones.
 - AMI (Amazon Machine Image), imagen de la Máquina Virtual, determina el SO y las aplicaciones disponibles de la instancia cuando arranca.
 - Grupo de Seguridad (SG), es la configuración de cortafuegos de la instancia (qué tráfico puede recibir la instancia)
 - Par de claves (*keypair*) para permitir la conexión a la instancia mediante SSH sin contraseña.
 - Región y zona de disponibilidad (opcional). La región por defecto es us-east-1 (Virginia, USA)

3.4 Cloud Computing

A. Martín

42

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

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

The screenshot shows the AWS EC2 instance types page. At the top, there's a navigation bar with links like 'Productos', 'Soluciones', 'Precios', etc. Below it, a breadcrumb trail shows 'Productos / Informática / Amazon EC2 / ...'. The main title is 'Tipos de instancias de Amazon EC2'. A large orange cube icon is on the right. The page content describes the M5 family as the latest generation of general-purpose instances with Intel Xeon processors. It lists several instance types: A1, T3, T3a, T2, M6g, M5, M5a, M5n, and M4. The M5 series is highlighted with a yellow background. A note at the bottom says 'Las instancias M5 son la última generación de instancias de uso general con la tecnología de los procesadores Intel Xeon®'.

3.4 Cloud Computing

43

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

INSTANCIAS ESTÁNDAR				
NOMBRE	vCPU	MEMORIA	ALMACENAMIENTO	PRECIO/HORA
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

The screenshot shows the AWS EC2 instance types page. At the top, there's a navigation bar with links like 'Productos', 'Soluciones', 'Precios', etc. Below it, a breadcrumb trail shows 'Productos / Informática / Amazon EC2 / ...'. The main title is 'Tipos de instancias de Amazon EC2'. A large orange cube icon is on the right. The page content describes the M5 family as the latest generation of general-purpose instances with Intel Xeon processors. It lists several instance types: A1, T3, T3a, T2, M6g, M5, M5a, M5n, and M4. The M5 series is highlighted with a yellow background. A note at the bottom says 'Las instancias M5 son la última generación de instancias de uso general con la tecnología de los procesadores Intel Xeon®'.

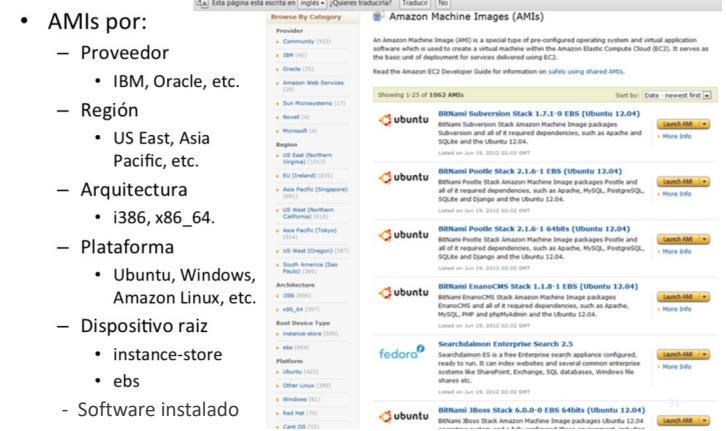
3.4 Cloud Computing

44

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2



The screenshot shows the AWS Lambda console with a list of functions. The functions listed are:

- BRINAMI Subversion Stack 1.7.1-0 EBS (Ubuntu 12.04)
- BRINAMI Poodle Stack 2.1.6-1 4-GBtbs (Ubuntu 12.04)
- BRINAMI Poodle Stack 2.1.6-1 6-GBtbs (Ubuntu 12.04)
- BRINAMI EloquentCMS Stack 1.1.8-1 EBS (Ubuntu 12.04)
- Searchdemons Enterprise Search 2.5
- BRINAMI JBoss Stack 6.0.0-0 EBS 6-GBtbs (Ubuntu 12.04)

Each function entry includes a "Launch AMI" button and a "More Info" link.

3.4 Cloud Computing

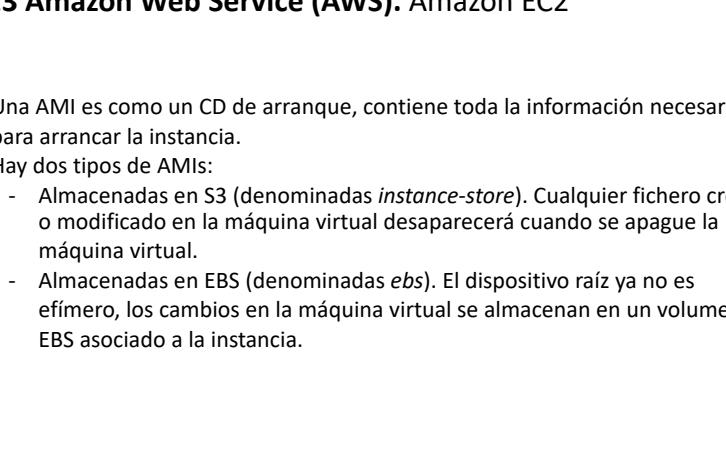
A. Martín

45

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2



The screenshot shows the AWS Lambda console with a list of functions. The functions listed are:

- Una AMI es como un CD de arranque, contiene toda la información necesaria para arrancar la instancia.
- Hay dos tipos de AMIs:
 - Almacenadas en S3 (denominadas *instance-store*). Cualquier fichero creado o modificado en la máquina virtual desaparecerá cuando se apague la máquina virtual.
 - Almacenadas en EBS (denominadas *ebs*). El dispositivo raíz ya no es efímero, los cambios en la máquina virtual se almacenan en un volumen EBS asociado a la instancia.

3.4 Cloud Computing

A. Martín

46

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

- Amazon Elastic Block Store (EBS) proporciona volúmenes orientados a bloques (como si fuera un disco duro externo o un USB) para ser conectados a instancias de EC2.
 - Se crean para una zona de disponibilidad concreta y sólo pueden conectarse a instancias desplegadas en la misma zona de disponibilidad.
 - Solo puede estar conectado a una única instancia en un momento determinado.
 - S3 es un servicio de objetos (ficheros) y EBS de volúmenes (discos)
 - La puesta en marcha de una AMI basada en EBS provoca la creación automática de un volumen EBS para almacenar los datos de la misma
 - El volumen queda conectado (y los datos preservados) aunque se detenga la instancia (no es el caso si se termina la instancia)
 - Al detener la instancia ya no se produce gasto por consumo de horas de instancia, pero sí por el almacenamiento del volumen de datos EBS asociado

3.4 Cloud Computing

A. Martín

47

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

- Los precios dependen de la región.
- Las instancias basadas en Windows cuestan más que las basadas en Linux.
- Instancias reservadas: a 1 año y 3 años, permite obtener reducción significativa (más del 50%) del precio por hora de instancias.
- Instancias dedicadas o compartidas (las dedicadas son más caras)
- <https://calculator.aws/#/estimate>
 - Región de Ohio:
 - 3 instancias m5x.large activas durante todo el día (gasto de la instancia más el coste del almacenamiento de las tres AMIs EBS asociadas de 30 Gigas mínimo): 420.48 dólares las instancias EC2 + 7.2 dólares las AMIs EBS al mes (730 horas).
 - Almacenamiento de 20 Gigas en S3 donde se supone una descarga de 5 GB/mes a cualquier usuario: 0.46 dólares el almacenamiento + 0.45 dólares la transferencia.
 - Coste total (un mes son 730 horas): 428.59 dólares/mes

3.4 Cloud Computing

A. Martín

48

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

The diagram illustrates a distributed data processing architecture using Amazon EC2. It shows an Application Load Balancer (ELB) on the left receiving traffic from multiple clients. The ELB routes traffic to four instances, which are grouped into two availability zones. Availability Zone 1 contains Instance 1 (blue) and Instance 2 (orange). Availability Zone 2 contains Instance 3 (red) and Instance 4 (green). This setup ensures redundancy and fault tolerance.

49

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EC2

The diagram illustrates scaling strategies in Amazon EC2. It shows two main types of scaling:

- Escalado automático.** (Automatic scaling)
- Escalado Vertical (Scale Up / Scale Down).** This diagram shows a single instance transitioning between different sizes (Small, Medium, Large, Medium) in response to traffic increases and decreases.
- Escalado Horizontal (Scale Out / Scale In).** This diagram shows the addition and removal of instances to handle varying traffic loads. It starts with 1 instance, scales up to 2 instances, reaches a peak of 4 instances, and then scales down to 3 instances.

50

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EMR



- Permite el despliegue dinámico de clusters de procesamiento distribuido para procesar datos que pueden estar almacenados en diferentes servicios AWS.
- Ofrece configuración automática de diferentes frameworks y herramientas.
- El cluster puede ser redimensionado (elasticidad) en tiempo de ejecución.
- Acceso *root* a las instancias de EC2.
- Gestión del ciclo de vida del cluster.
- Permite integrarse con otros servicios de AWS para la lectura y almacenamiento de datos (Base de datos –DynamoDB, RDS- o servicio de almacenamiento de ficheros -S3, HDFS-).

3.4 Cloud Computing

A. Martín

51

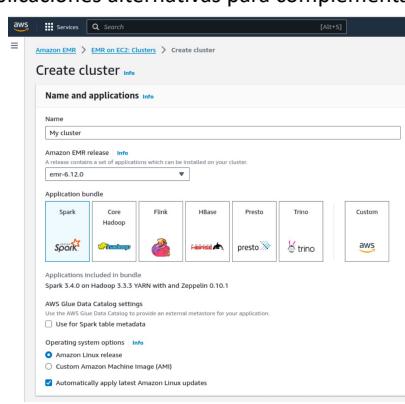
BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EMR



- Es posible elegir aplicaciones alternativas para complementar la instalación.



3.4 Cloud Computing

A. Martín

52

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EMR



• Tipos de nodos:

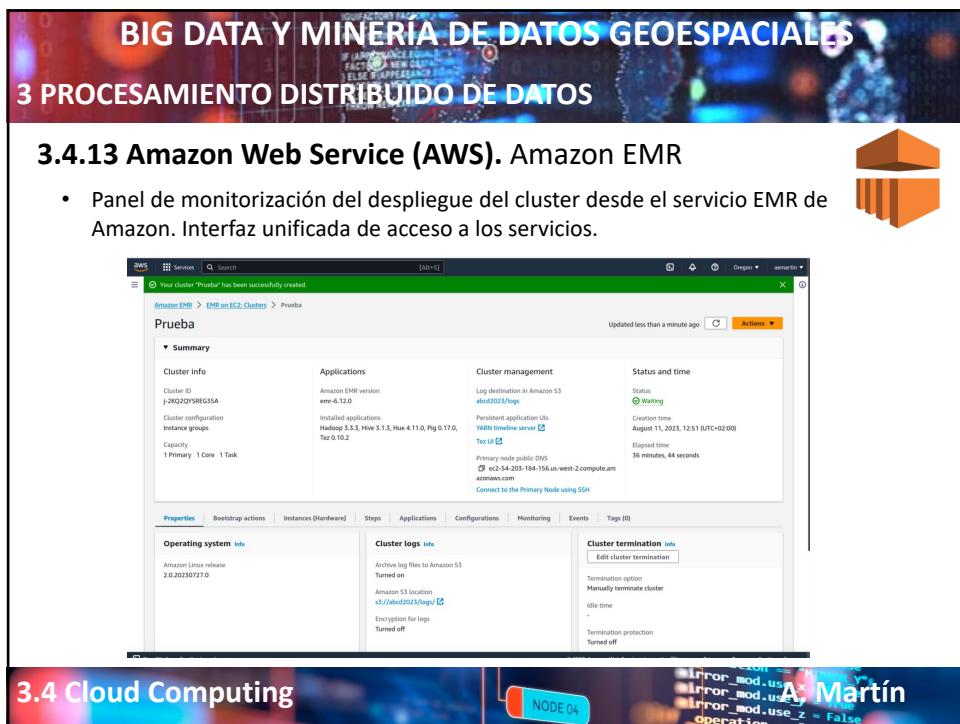
- Master: Único. Control
- Core: Ejecución de tareas + HDFS.
- Task: Ejecución de tareas

53

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EMR



• Panel de monitorización del despliegue del cluster desde el servicio EMR de Amazon. Interfaz unificada de acceso a los servicios.

54

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Amazon EMR

The screenshot shows the AWS EC2 Management Console. A modal window titled "Connect to the primary node using SSH" is open, providing instructions to connect via SSH and giving the command to run on the terminal. Below this, a terminal window titled "Windows" shows the command being run. The main interface displays the "Prueba" cluster details, including its status as "Running" and the number of instances (1 Primary, 1 Task). The "Network and security" section shows the public IP address: 172.31.7.153.

3.4 Cloud Computing

A. Martín

55

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

56

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Acceso desde Python, librería boto3

- Esta librería proporciona una sencilla API de acceso directo a los siguientes recursos de AWS:
 - S3
 - EC2
 - CloudWatch
 - SQS (Simple Queue Service)

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

57

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

3 PROCESAMIENTO DISTRIBUIDO DE DATOS

3.4.13 Amazon Web Service (AWS). Acceso desde 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
```

3.4 Cloud Computing **A. Martín**

58