

Una Reseña Acerca de Servicios de Machine Learning en ambientes Cloud

Jorge Alejandro Kamlofsky^{1,2}

¹GIA: Grupo de Investigación en Inteligencia Artificial. Universidad Tecnológica Nacional, Facultad Regional Haedo

²CAETI: Centro de Altos Estudios en Tecnología Informática. Universidad Abierta Interamericana, Facultad de Tecnología Informática

Resumen:

Son varios los aspectos que ayudaron al desarrollo y a la expansión explosiva de la Inteligencia Artificial. Es muy destacable la creación del protocolo IPv6 que incorporó a miles de millones de dispositivos a Internet, generando nuevos datos. Esa nueva masividad de datos debía poder ser almacenada y procesada: la tecnología Cloud así lo permite. Y es aquí donde se proveen servicios de ML de todo tipo.

1. Introducción

Los modelos de nube pueden abordarse desde dos dimensiones: desde el tipo de despliegue o Implementación y desde el tipo de servicio ofrecido. Así, por un lado, se dispone modelos de nube públicas, privadas o híbridas, y por otro lado se presentan modelos de Software como servicio, Plataforma como servicio o Infraestructura como servicio: SaaS, PaaS y IaaS según sus siglas en inglés (Kamlofsky, 2022)

1.1. Computación en la Nube

Cloud computing, o computación en la nube, se refiere a la entrega de servicios de computación a través de internet. Incluye almacenamiento, procesamiento, bases de datos, redes, software y análisis.

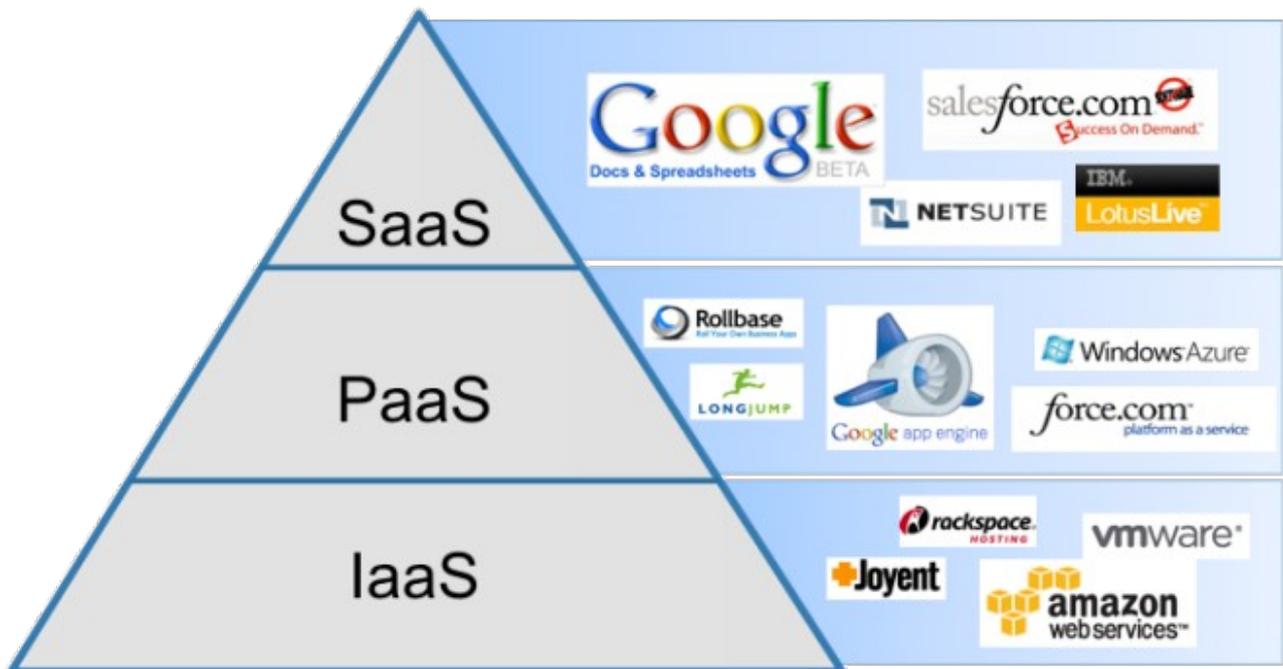


Imagen obtenida de: <https://nubecomputingblog.wordpress.com/servicios-o-usos-del-cloud-computing/>

En lugar de poseer y mantener infraestructura física, las organizaciones pueden acceder a estos recursos bajo demanda a través de proveedores de servicios en la nube como Amazon Web Services (AWS), Microsoft Azure y Google Cloud Platform (GCP).



Este modelo ofrece flexibilidad, escalabilidad y eficiencia de costos, permitiendo a las empresas adaptarse rápidamente a las necesidades cambiantes del mercado y centrarse en su núcleo de negocio sin preocuparse por la gestión de la infraestructura.

1.2. Introducción acerca de ML en cloud

Normalmente, el flujo de un proyecto de ML tiene cierto nivel de complejidad: desde la adquisición del conjunto de datos hasta el despliegue y puesta en producción de los modelos, hay muchos aspectos a considerar (Simon, 2020). La siguiente figura ilustra estas etapas.

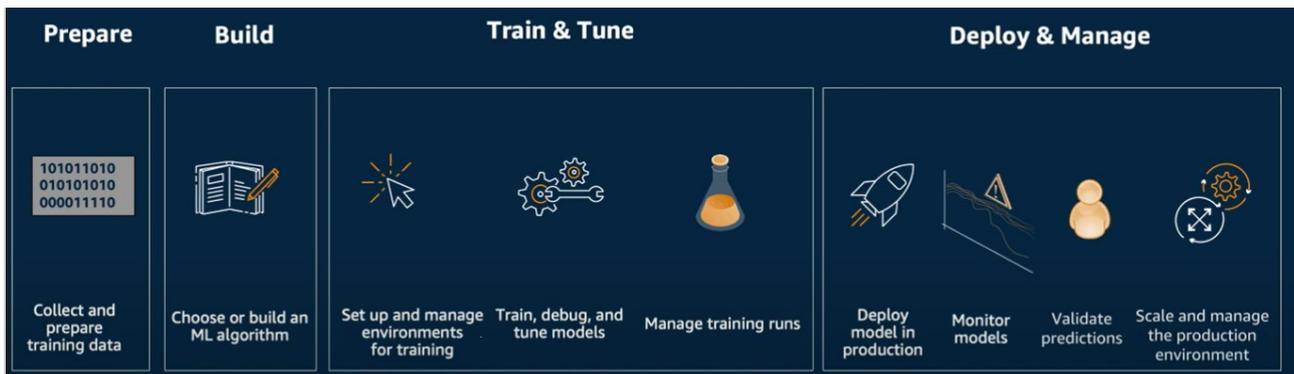


Figura: El flujo de tareas en un proyecto de ML

Para Machine Learning (ML), el ambiente cloud es de gran importancia debido a las siguientes razones:

1. Infraestructura para tratamiento de datos: El entrenamiento de modelos de ML a menudo requiere gran cantidad de recursos computacionales y almacenamiento de datos, que pueden ser costosos y difíciles de gestionar localmente. La nube proporciona acceso a potentes recursos de computación

2. Herramientas para implementar ML: Los servicios en la nube ofrecen herramientas y plataformas integradas, como Amazon SageMaker, Azure Machine Learning y Google AI Platform, que simplifican el desarrollo, entrenamiento, despliegue y monitoreo de modelos de ML. Esto no solo acelera el ciclo de vida del desarrollo de ML, sino que también facilita la colaboración y la implementación de prácticas de MLOps (Machine Learning Operations), asegurando que los modelos se mantengan actualizados y operativos en producción.

En este trabajo se presenta una breve descripción de los principales ambientes Cloud de los tres proveedores más destacados: Amazon, Microsoft y Google con las herramientas de más interés, y luego, una breve descripción de los servicios de ML y de IA que cada empresa ofrece.

1.3. Estructura del Trabajo

Este trabajo presenta un primer capítulo donde se introduce el problema, y en el segundo capítulo se lo aborda. Finalmente, se presentan las conclusiones.

2. Principales Ambientes en Cloud con ML

2.1. Amazon

2.1.1. Acerca de AWS

Amazon Web Services (AWS abreviado) es una colección de servicios de computación en la nube pública (también llamados servicios web) que en conjunto forman una plataforma de computación en la nube, ofrecidas a través de Internet por Amazon.com. Es considerado un pionero en este campo¹.

2.1.2. Algunos Componentes importantes en el ecosistema AWS

Amazon SageMaker Studio

Amazon SageMaker es un servicio de machine learning completamente administrado. Con Amazon SageMaker, los científicos de datos y los desarrolladores pueden crear y entrenar rápidamente modelos de machine learning e implementarlos directamente en un entorno alojado listo para producción².

AWS Lambda

Con AWS Lambda se puede ejecutar código sin aprovisionar ni administrar servidores. Ideal para armado de servicios server-less. El servicio de Lambda ejecuta la función solo cuando es necesario. Normalmente puede iniciarse luego de una carga en el almacenamiento o luego de un pedido vía API³.

Amazon S3

Amazon Simple Storage Service (Amazon S3) es almacenamiento para Internet. Puede utilizar Amazon S3 para almacenar y recuperar cualquier cantidad de datos en cualquier momento y desde cualquier lugar de la web⁴.

AWS Glue

AWS Glue es un servicio de integración de datos sin servidor, escalable que facilita la detección, preparación y combinación de datos de análisis, machine learning, y desarrollo de aplicaciones⁵.

1 Acerca de AWS (Wikipedia): https://es.wikipedia.org/wiki/Amazon_Web_Services

2 Fuente: https://docs.aws.amazon.com/es_es/sagemaker/?icmpid=docs_homepage_ml

3 Fuente: https://docs.aws.amazon.com/es_es/lambda/

4 Fuente: <https://docs.aws.amazon.com/s3/>

5 Fuente: https://docs.aws.amazon.com/es_es/glue/

2.1.3. AWS Machine Learning Services

Machine learning services at AWS



Amazon Rekognition



Amazon Textract



Amazon Comprehend



Amazon Transcribe



Amazon Translate



Amazon Lex



Amazon SageMaker

Textract

Textract es un servicio que detecta y extrae automáticamente texto y datos de documentos escaneados. Va más allá del simple reconocimiento óptico de caracteres (OCR) para identificar también el contenido de los campos de los formularios y la información almacenada en las tablas⁶.

Rekognition

Es un servicio de análisis de imágenes y videos basado en deep learning. Funciones clave: Detección de objetos, rostros, análisis de contenido, reconocimiento facial, y etiquetado de imágenes. Muy útil para desarrolladores que requieren reconocimiento visual sin entrenamiento de modelos personalizados.

Comprehend

Es un servicio de análisis de texto y procesamiento de lenguaje natural (NLP). Funciones clave: Extracción de sentimientos, entidades, análisis de tono y organización de texto en categorías. Ideal para crear aplicaciones de análisis de texto o clasificaciones de documentos.

Lex

Permite la Creación de chatbots y asistentes virtuales conversacionales. Funciones clave: Reconocimiento de voz y texto, manejo de intenciones y contextos conversacionales. Se integra fácilmente con Amazon Connect para soluciones de centros de llamadas.

Polly (Transcribe)

Realiza la Conversión de texto a voz (Text-to-Speech). Funciones clave: Convierte texto en voz natural en varios idiomas y estilos de voz, ideal para aplicaciones que necesitan interacción de voz.

Translate

Es un servicio de traducción automática de textos en diferentes idiomas. Su principal función es permitir la traducción en tiempo real entre múltiples idiomas, útil para globalizar aplicaciones de contenido.

6 Fuente: <https://aws.amazon.com/es/textract/>

Forecast

AWS Forecast es un Servicio de predicción de series temporales. Principalmente permite la Generación de pronósticos automáticos en ventas, demanda, planificación de inventario, entre otros. Utiliza algoritmos avanzados de Amazon y ajusta modelos automáticamente.

2.2. Microsoft Azure

2.2.1. Acerca de Azure

Azure es la plataforma en la nube de Microsoft. Azure cuenta con más de 200 productos y servicios en la nube como ser: máquinas y escritorios virtuales, bases de datos, almacenamiento servicios de machine learning, copilot, entre otros.

Según el sitio de la empresa *“De las empresas listadas en Fortune 500, el 95% confía en Azure para obtener servicios de nube confiables. Empresas de todos los tamaños y niveles de madurez utilizan Azure en su transformación digital”*⁷.



2.2.2. Dominios en la Oferta de Azure

Azure ofrece una amplia gama de servicios cloud. Estas ofertas abarcan múltiples dominios, incluidos el cómputo, el almacenamiento, las redes, las bases de datos, la inteligencia artificial y otros (Borra, 2022).

Cómputo:

En Azure, los usuarios pueden acceder a varias opciones para implementar aplicaciones y cargas de trabajo en la nube, como máquinas virtuales, contenedores, computación sin servidor y servicios informáticos especializados como Azure Batch y Azure Kubernetes Service (AKS).

Almacenamiento:

Azure Storage ofrece soluciones de almacenamiento escalables y seguras para datos, incluido el almacenamiento de blobs para datos no estructurados, el almacenamiento de archivos para recursos compartidos de archivos en la nube y el almacenamiento de tablas y colas para datos estructurados y colas de mensajes.

Networking:

En el ámbito de Azure Networking, los usuarios pueden crear redes seguras y de alto rendimiento mediante servicios como Virtual Network para entornos de red aislados, Azure Load Balancer para la distribución del tráfico y Azure VPN Gateway para conexiones seguras entre redes locales y en la nube.

Bases de datos

Las bases de datos de Azure ofrecen servicios de bases de datos administradas que cubren varios tipos de bases de datos relacionales y no relacionales, incluidas Azure Database for MySQL y PostgreSQL, Azure SQL Database y Azure Cosmos DB.

7 Sitio de Microsoft Azure: <https://azure.microsoft.com/es-mx/resources/cloud-computing-dictionary/what-is-azure>

Machine Learning e IA

Para iniciativas de inteligencia artificial y aprendizaje automático, Azure proporciona herramientas y modelos prediseñados para desarrollar aplicaciones inteligentes, lo que permite a los desarrolladores entrenar e implementar modelos de aprendizaje automático personalizados de manera eficiente.

Además de esto, Azure ofrece servicios para IoT, DevOps, análisis, seguridad, identidad y más, posicionándose como una plataforma de nube integral para el desarrollo, implementación y administración de aplicaciones.

2.2.3. Acerca de Machine Learning e IA en Azure

Azure Artificial Intelligence (IA) presenta una amplia gama de servicios y herramientas diseñados para facilitar el desarrollo, la implementación y la gestión de soluciones de IA en el entorno de la nube. A través de ofertas como Azure Machine Learning, las organizaciones pueden crear y perfeccionar modelos de aprendizaje automático utilizando diversos conjuntos de datos. Azure Cognitive Services ofrece capacidades de IA listas para usar, que incluyen visión artificial, manejo del lenguaje natural y reconocimiento de voz, lo que permite a los desarrolladores integrar sin problemas funcionalidades avanzadas de IA en las aplicaciones (Borra, 2022).

Azure Machine Learning ofrece una amplia gama de herramientas y servicios que incluyen:

Azure Machine Learning Designer:

Una interfaz visual para crear, probar y desplegar modelos de Machine Learning sin necesidad de escribir código.

Azure Machine Learning Notebooks:

Soporte para Jupyter Notebooks para el desarrollo de modelos de Machine Learning utilizando Python.

Automated Machine Learning (AutoML):

Herramientas para automatizar el proceso de selección de modelos y ajuste de hiperparámetros.

MLOps:

Herramientas para la gestión del ciclo de vida de los modelos, incluyendo el despliegue, monitoreo y reentrenamiento de modelos.

Integración con otros servicios de Azure:

Como Azure Databricks, Azure Synapse Analytics, y más.

2.3. Google Cloud Platform



2.3.1. Acerca de Google Cloud Platform

Los proveedores de servicios de computación en la nube ofrecen a sus clientes escalabilidad, rentabilidad y flexibilidad. Así, lograron revolucionar la forma en que las empresas administran y usan sus recursos informáticos. En particular, Google Cloud Platform (GCP) se consolidó como un

proveedor líder global en este rubro, caracterizado por la disponibilidad de una enorme infraestructura y servicios de cómputo (Smith and Brown, 2020).

2.3.2. Sus Orígenes

El primer antecedente de Google en el mercado de computación en la nube puede rastrearse en 2008 con el lanzamiento de Google App Engine, lo que marcó su primer paso para ofrecer soluciones escalables de infraestructura en la nube. Desde su creación, creció significativamente, aprovechándose en su extensa infraestructura global y la destreza tecnológica para ofrecer una amplia gama de servicios adaptados a las necesidades comerciales modernas (Borra, 2024a).

2.3.3. Dominios de GCP

GCP tiene organizados sus servicios en los siguientes dominios críticos que incluyen: computación, almacenamiento, bases de datos, networking, big data, IA y Machine Learning y herramientas para desarrollo. Estas ofertas están diseñadas para brindar soluciones escalables, seguras y eficientes, lo que permite a las empresas aprovechar la computación en la nube de manera eficaz para la innovación y la mejora operativa.

Cómputo:

El corazón de la propuesta de cómputo de GCP se encuentra Google Compute Engine (GCE), que proporciona máquinas virtuales escalables para ejecutar aplicaciones en cloud. GCE se complementa con Google Kubernetes Engine (GKE), que facilita la gestión y el mantenimiento de aplicaciones en contenedores para una mayor escalabilidad y eficiencia. Además, App Engine permite a los desarrolladores crear y desplegar aplicaciones web escalables sin tener que gestionar la infraestructura subyacente (Borra, 2024a).

En este dominio puede también encontrarse a la computación cuántica: Pocas empresas en el mundo poseen esta capacidad. Google posee una de las computadoras cuánticas más rápidas y poderosas del mundo, que promete la mayor capacidad de cómputo masivo y paralelo. Sin embargo, se probó que estos equipos, que prometen capacidades disruptivas, aún son desafiadas por computadores convencionales (Garay, 2024).

Almacenamiento:

GCP tiene dentro de su portafolio soluciones de almacenamiento robustas creadas para dar escalabilidad y durabilidad. Google Cloud Storage proporciona almacenamiento de objetos para varios tipos de datos, con opciones de redundancia regional o multirregional para garantizar la resiliencia de los datos. Persistent Disks propone a sus clientes ofertas de almacenamiento en bloque fiables y adaptadas a los requisitos específicos de las aplicaciones, compatibles con instancias de máquinas virtuales de alto rendimiento.

Networking:

Las opciones de red de GCP incluyen Virtual Private Cloud (VPC) para crear entornos de red aislados con control sobre rangos de IP y de subredes. Cloud Load Balancing garantiza gran disponibilidad y alto rendimiento debido a su capacidad de distribuir el tráfico entre instancias o servicios (Borra, 2024a).

Servicios de Bases de datos

Los servicios de bases de datos administrados desde esta plataforma cumplen con una amplia gama de requisitos de gestión de datos. Google Cloud SQL proporciona bases de datos relacionales totalmente gestionadas, como MySQL, PostgreSQL y SQL Server, garantizando alta disponibilidad y copias de seguridad automáticas. Google Cloud Spanner, por su lado, propone una base de datos distribuida globalmente y escalable horizontalmente para aplicaciones que requieren una fuerte

consistencia entre regiones diferentes. Firestore por su lado, entrega una base de datos de documentos NoSQL flexible para el desarrollo de aplicaciones escalables con actualizaciones en tiempo real y soporte offline (Borra, 2024a).

Big Data

GCP es una herramienta sólida para tratamiento y uso de datos de gran porte. BigQuery permite tratar sencillamente fragmentos SQL de datos de varios terabytes para su almacenamiento y análisis. El flujo de datos está disponible para los análisis de datos por lotes y flujos de trabajo y análisis de tiempo completos y flujos de trabajo ETL. En Dataproc se incluye Apache Spark y Hadoop-Cluster para la integración de cargas de trabajo de Big Data (Borra, 2024a).

2.3.4. Machine Learning e IA en Google Cloud Platform

Google Cloud Platform (GCP) se ha convertido en un líder en aprendizaje automático (ML) e inteligencia artificial (IA), conocido por sus tecnologías de vanguardia y accesibilidad inclusiva. GCP no solo impulsa la innovación, sino que también democratiza el acceso a potentes herramientas de ML e IA, lo que permite a las organizaciones de todos los tamaños aprovechar los conocimientos basados en datos para mejorar la innovación, la eficiencia y el crecimiento escalable. El impacto de GCP trasciende los avances tecnológicos y representa un cambio significativo en la transformación digital en diversas industrias (Borra, 2024b).

Google Cloud Platform (GCP) ofrece una amplia gama de herramientas de aprendizaje automático (ML) e inteligencia artificial (IA) diseñadas para satisfacer diversos requisitos comerciales en varios sectores (Borra, 2024b): Google AI Platform, AutoML, BigQueryML, AI Building Blocks.

Google AI Platform

Google AI Platform simplifica todo el ciclo de vida del aprendizaje automático, desde la preparación de datos y el entrenamiento de modelos hasta la implementación y el escalamiento. Propone un ambiente colaborativo para que los científicos de datos y los desarrolladores experimenten con diferentes algoritmos y marcos de aprendizaje automático, lo que garantiza un desarrollo y una implementación eficientes de los modelos. La robusta infraestructura de la plataforma admite cálculos de gran complejidad, lo que lo hace adecuado para administrar trabajos de análisis y proceso de datos de alto volumen.

AutoML

AutoML permite que usuarios con pocos conocimientos en ML puedan crear modelos de aprendizaje automático personalizados de forma automática. Al automatizar la selección, el entrenamiento y la optimización de modelos, AutoML acelera el desarrollo de aplicaciones de IA, como el reconocimiento de imágenes y el procesamiento del lenguaje natural.

BigQueryML

BigQuery ML, está integrado al almacén de datos BigQuery de Google. BigQuery ML mejora el análisis de datos tradicional al incorporar el aprendizaje automático directamente en las consultas SQL. Esta integración permite a los analistas de datos y desarrolladores de SQL crear y ejecutar modelos de aprendizaje automático fácilmente en sus flujos de trabajo de análisis de datos. Al aprovechar la infraestructura escalable de Google, BigQuery ML permite la generación de análisis predictivos y de información en tiempo real sin necesidad de mover datos entre sistemas y recursos.

Building Blocks ML

Building Blocks de IA de GCP ofrecen APIs pre-entrenadas para tareas de visión, lenguaje y traducción. Estas APIs, que incluyen Vision API, Translation API y Natural Language API, brindan a los desarrolladores herramientas fáciles de integrar para incorporar funcionalidades avanzadas de IA

en las aplicaciones. Abstraen las complejidades del desarrollo de IA en llamadas API simples, lo que permite a las empresas mejorar las interacciones con los clientes, automatizar procesos y obtener información útil a partir de datos no estructurados de manera eficiente.

3. Conclusiones

En este trabajo se presentaron las infraestructuras de las principales infraestructuras Cloud. Todas poseen: Almacenamiento, Cómputo, bases de datos, y networking. Quizás a todos ellos les resultó natural evolucionar a partir de aquí (gracias a los privilegios de poseer datos de los usuarios), en plataformas de IA y Machine Learning, que hoy están en la vanguardia tecnológica.

Entre las plataformas presentadas, en común se destaca la presencia de modelos de aprendizaje profundo que ofrecen a sus clientes para lectura de imágenes, textos y demás. En general, se destaca que estos servicios se entregan sin necesidad de desarrollar código alguno.

Finalmente, todas las plataformas (entre las aquí presentadas) presentan herramientas para el desarrollo de modelos de Machine Learning y uso de modelos LLM y de IA. Este apartado queda abierto, ya que a futuro se irán incorporando más modelos que enriquecerán las IA que se disponen.

Referencias

Kamlofsky Jorge, (2022). “Computación en la Nube: Fundamentos, Críticas y Desafíos”, Revista Abierta de Informática Aplicada/vol 6 N°2: 3-30, ISSN 2591-5320.

Simon Julien (2020). “Building, training and deploying machine learning models with Amazon SageMaker”, Youtube. En línea: <https://www.youtube.com/watch?v=sOUhLi85sU>

Borra, Praveen. (2022). “Exploring Microsoft Azure's Cloud Computing: A Comprehensive Assessment. International Journal of Advanced Research”. In Science, Communication and Technology (IJARSCT) Volume, 2.

Borra, Praveen. (2024). “A Survey of Google Cloud Platform (GCP): Features, Services, and Applications”. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume, 4*.

Borra, Praveen. (2024). The Evolution and Impact of Google Cloud Platform in Machine Learning and AI. *Available at SSRN 4914163*.

Smith, J., & Brown, A. (2020). “The Impact of Cloud Computing on Business Operations”. *Journal of Cloud Technology*, 8(2), 45-56.

Garay, Jorge. (2024). “La computadora cuántica de Google acaba de ser humillada por una red de ordenadores convencionales chinos”. *Revista Wired*. En línea: <https://es.wired.com/articulos/computadora-cuantica-de-google-humillada-por-ordenadores-convencionales>. Consultado: 14/11/2024