

Tarea 2 | Comparación proveedores de nube

Integrantes: Benjamín Daza, Daniel Maturana y Sebastián Von Kunowsky.

Proveedores para comparar: “Amazon Web Services” y “Google Cloud Platform”.

Servicios específicos

Categoría	AWS	Google Cloud
Plataforma de Machine Learning	Amazon SageMaker	Vertex AI Platform
Máquinas Virtuales (VMs)	Amazon EC2	Compute Engine
Contenedores (CaaS)	Amazon EKS	Google Kubernetes Engine
Procesamiento de datos Batch y Stream Serverless	AWS Lambda	Dataflow
Almacenamiento de Objetos	Amazon S3	Cloud Storage
Data Warehouse	Amazon Redshift	BigQuery
RDBMS (compatible con MySQL)	Amazon RDS	Cloud SQL

Características principales

Plataforma de Machine Learning	Característica 1	Característica 2
Amazon SageMaker	Estudio unificado con las herramientas de AWS para el desarrollo de modelos: IA generativa, procesamiento de datos y el análisis de SQL.	Amazon SageMaker Unified Studio incluye un conjunto de IDE administrados para desarrollo de ML, incluidos JupyterLab (y JupyterNotebooks).
Vertex AI Platform	Proporciona herramientas de MLOps específicas para que los científicos de datos y los ingenieros de AA automaticen, estandaricen y administren proyectos de AA.	Incluye notebooks de Vertex AI que se integran de forma nativa en BigQuery. Estos ofrecen un entorno integrado basado en Jupyter Notebooks llamado Vertex AI Workbench.
Máquinas Virtuales (VMs)	Característica 1	Característica 2

Amazon EC2	Tipos de instancia con configuraciones de: CPU, memoria, almacenamiento y redes.	Escalado automático según requisitos de la carga de trabajo.	
Compute Engine	Personalización de los tipos de máquinas.	Escalabilidad vertical y horizontal.	

Contenedores (CaaS)	Característica 1	Característica 2
Amazon EKS	Servicio gestionado de Kubernetes que simplifica la orquestación de contenedores.	Escalabilidad y alta disponibilidad para la gestión de contenedores.
Google Kubernetes Engine	Herramientas para la gestión y orquestación de contenedores que facilitan la integración continua	Escalabilidad automática y actualizaciones gestionada.

Procesamiento de datos Batch y Stream Serverless	Característica 1	Característica 2
AWS Lambda	Ejecución de código sin administración de servidores para el procesamiento de datos.	Posibilidad de integración con Apache Beam a través de conectores y/o el uso de otros servicios de AWS.
Dataflow	Servicio serverless unificado para el procesamiento de datos en lote y streaming.	Ejecución nativa de pipelines de Apache Beam.

Almacenamiento de Objetos	Característica 1	Característica 2
Amazon S3	Escalabilidad: Puede almacenar hasta exabytes de datos con S3, con un rendimiento inigualable. S3 es totalmente elástico y crece y se reduce de manera automática a medida que se agregan y eliminan datos.	Acceso y durabilidad: S3 está diseñado para ofrecer una durabilidad de datos del 99,999999999 % (11 nueves) y una disponibilidad del 99,99 % de forma predeterminada, con el respaldo de los SLA más sólidos de la nube.
Cloud Storage	Cloud Storage está diseñado para tener una durabilidad anual del 99.999999999%. Si el almacenamiento necesita aumentar, el	Mejor rendimiento con el uso de Anuwhere Cache, una caché de lectura inteligente, administrada por SSD para buckets. Multirregión para acercar

	proveedor de servicios en la nube iniciará más máquinas virtuales para controlar la carga.	los datos a tus TPU o GPU, logrando baja latencia y hasta 20 Tbps de velocidad para IA y análisis de datos.
--	--	---

Data Warehouse	Característica 1	Característica 2
Amazon Redshift	Hasta 3 veces más rendimiento de precio y 7 veces más rendimiento que otros almacenes de datos en la nube a medida que escala sus cargas de trabajo de análisis de datos. Sin servidor aprende de sus cargas de trabajo y escala automáticamente los recursos informáticos para gestionar sus necesidades analíticas cambiantes, de modo que pueda centrarse en descubrir información sin tener que administrar la infraestructura.	Petabytes de datos están disponibles para el análisis sin tener que crear y administrar canalizaciones complejas, lo que permite el acceso casi en tiempo real para los casos de uso de análisis. Integraciones sin ETL para mover sin problemas los datos transaccionales de bases de datos.
BigQuery	La arquitectura que separa el almacenamiento de los recursos de computación para analizar datos a escala de petabytes. Optimiza los costes gracias al almacenamiento comprimido y al auto escalado de la computación.	Gemini genera preguntas en lenguaje natural y las consultas en SQL para responderlas, según los metadatos de la tabla. Estas estadísticas te ayudan a descubrir patrones, evaluar la calidad de los datos y realizar análisis estadísticos.

RDBMS (compatible con MySQL)	Característica 1	Característica 2
Amazon RDS	Permite a los clientes crear una nueva base de datos en cuestión de minutos y ofrece	Single-AZ: <ul style="list-style-type: none"> Amazon RDS para PostgreSQL

	flexibilidad para personalizar las bases de datos a fin de satisfacer sus necesidades en 8 motores y 2 opciones de despliegue.	<ul style="list-style-type: none"> • Amazon RDS for MySQL • Amazon RDS para MariaDB • Amazon RDS para SQL Server • Amazon RDS para Oracle • Amazon RDS para Db2
Cloud SQL	Escalabilidad automática para manejar las demandas de las aplicaciones, tanto en términos de almacenamiento como de procesamiento.	Facilidad de administración y mantenimiento, con características como parches automatizados, copias de seguridad y replicación.

Cobros, similitudes y diferencias

Plataforma de Machine Learning:

En Amazon SageMaker se cobra por solicitud, almacenamiento de metadatos y computación. Se cobra adicionalmente la posibilidad de obtener recomendaciones basadas en IA.

En Vertex AI se cobra por hora de entrenamiento, predicción y/o implementación, para cada tipo de datos utilizados en el modelo de machine learning (videos, imágenes, textos). Se cobra adicionalmente la implementación de IA, pipelines y el vector de búsqueda.

Similitud: ambos ofrecen herramientas específicas para automatizar procesos y para procesamiento de datos, en especial herramientas para IA generativa.

Diferencia: Amazon SageMaker incluye en su IDE JupyterLab con Jupyter Notebooks, en cambio Vertex AI ocupa Vertex AI Workbench que es un entorno integrado basado en Jupyter Notebooks.

Máquinas Virtuales (VMs):

Amazon EC2 cobra por uso en segundos o horas de cada instancia. Además, se cobra por capacidad computacional para las instancias.

Compute Engine cobra por cada instancia/servicio utilizado. Adicionalmente se cobra una tarifa por almacenamiento (disco persistente, hyperdisk, SSD local), por utilizar

red premium de google, por encriptación de datos y la cantidad de nodos físicos dedicados al proyecto.

Similitud: Tanto Amazon EC2 como Compute Engine soportan una variedad de sistemas operativos, incluyendo diversas distribuciones de Linux y Windows Server.

Diferencia: Amazon EC2 tiene escalado automático dependiendo de la carga de trabajo, en cambio Compute Engine ofrece escalabilidad vertical/horizontal.

Contenedores (CaaS):

En Amazon EKS tienen una tarifa por clúster por hora en función de la versión de Kubernetes del clúster. Teniendo tarifas extra por compatibilidad extendida y por el uso de nodos híbridos.

Google Kubernetes Engine cobra por la administración del ciclo de vida de los clústeres de manera automatizada, escalamiento automático y la visibilidad y la optimización automatizada de los costos de la infraestructura. Adicionalmente se cobra por los recursos de CPU, memoria y procesamiento que se aprovisionan y por cada instancia.

Similitud: Ambos se enfocan en la gestión y orquestación de contenedores

Diferencia: Google Kubernetes Engine gestiona completamente el plano de control de Kubernetes y ofrece reparación automática de nodos y escalado automático. EKS también proporciona un plano de control de Kubernetes pero con menos abstracción en la gestión de nodos de trabajo, requiriendo de configuración manual o uso de herramientas adicionales.

Procesamiento de datos Batch y Stream Serverless:

AWS Lambda aplica tarifas de tiempo de computación al mes y su uso, además una tarifa por volumen de datos.

Dataflow cobra por la utilización de recursos de computación como: CPU y memoria, unidades de computación de streaming y los datos tratados. Se cobra adicional si es necesario realizar tareas que requieren disco persistente, GPU y resúmenes.

Similitud: Ambos son serverless, lo que implica que no hay que aprovisionar, configurar o gestionar servidores. Ambos integran Apache Beam.

Diferencia: AWS Lambda está basado en eventos para ejecutar código en respuesta a estos, ideal para tareas de corta duración y arquitectura de microservicios. Dataflow

está diseñado para procesamiento de datos a gran escala, tanto por lotes como en streaming, ideal para ETL y análisis de datos en tiempo real.

Almacenamiento de Objetos:

Amazon S3 aplica una tarifa que se cobra dependiendo del tamaño de los objetos, de la cantidad de tiempo que los almacena durante el mes y de la clase de almacenamiento que utiliza, adicionalmente se cobra la transferencia de los archivos.

Cloud Storage cobra por almacenamiento de cada GiB al mes y la clase de almacenamiento, además cobra por transferencia de datos a otros servicios.

Similitud: Ambos almacenan datos como objetos dentro de contenedores (buckets). Están diseñados para una escalabilidad masiva, capaces de almacenar petabytes o incluso exabytes de datos. También ofrecen una durabilidad extremadamente alta.

Diferencia: Amazon S3 se enfoca más en la escalabilidad, totalmente elástico creciendo y reduciéndose de manera automática. Cloud Storage ofrece consistencia fuerte de lectura después de escritura a nivel global, haciéndolo de manera más rápida.

Data Warehouse:

Amazon Redshift cobra dependiendo del nodo elegido. Es una tarifa mensual por GB fija de la región para cobrar datos almacenados en almacenamiento administrado. El uso del almacenamiento administrado se calcula por hora en función de los datos totales presentes en el almacenamiento administrado.

BigQuery cobra por el procesamiento, ya sea por hora de ranura o por TiB analizado, además cobra por el almacenamiento lógico y físico por cada GiB y la transferencia y/o extracción de datos.

Similitud: Ambos son altamente escalables y pueden manejar volúmenes masivos de datos, desde gigabytes hasta petabytes. Están diseñados para adaptarse a las crecientes necesidades de datos y pueden escalar sus recursos de cómputo y almacenamiento.

Diferencia: BigQuery utiliza una arquitectura sin servidor y desacoplada de almacenamiento y cómputo. Redshift Se basa en una arquitectura de clústeres con nodos de cómputo y almacenamiento acoplados

RDBMS (compatible con MySQL):

Amazon RDS cobra por instancia utilizada en horas, además del almacenamiento SSD y la transferencia de datos y por IPv4 públicos.

Cloud SQL cobra por la computación utilizada por hora de vCPU, además cobra por la cantidad y tipo de memoria usada, y por cada GB de almacenamiento (SSD o SDD local) utilizado.

Similitud: Ambas ofrecen bases de datos relacionales gestionadas en la nube, encargándose de las tareas tediosas (tiempo, hardware, etc.).

Diferencia: Amazon RDS tiene una variedad más amplia de motores de base de datos relacionales (Aurora, Oracle, SQL Server, etc.), en cambio Cloud SQL está enfocado en MySQL, PostgreSQL y SQL Server.

Conclusiones

Plataforma de Machine Learning:

En un inicio sería recomendable usar Amazon Sagemaker, ya que realiza cobros por solicitud, priorizando así una disminución en los costos y adaptándose a medida que la empresa lo amerite. Por otro lado, con respecto a la IA generativa, puede que en etapas tempranas de expansión no sea tan necesario.

Máquinas Virtuales (VMs):

Se recomienda usar Amazon EC2 tiene la capacidad de escalar de forma automática conforme a las necesidades computacionales, algo bastante beneficioso si se quiere expandir la empresa, ya que irá adaptándose a sus necesidades. En un futuro con las necesidades claras se podría elegir la contraparte ya que se tendría una base sólida del contexto para escalar de forma horizontal o vertical.

Contenedores (CaaS):

En este caso se recomienda Google Kubernetes Engine, ya que en un inicio es fácil de entender los costos, ya que estos irán aumentando a medida que aumenten los clústeres. Por otro lado, escala automáticamente respecto a la demanda, siendo compatible con cargas de trabajo variable en futuras expansiones.

Procesamiento de datos Batch y Stream Serverless:

Se recomienda Google Cloud Dataflow, adecuado para soportar crecimiento de grandes volúmenes de datos, tanto batch como streaming. A su vez, soporta ejecuciones de larga duración y, también soporta tareas complejas.

Almacenamiento de Objetos:

Se recomienda Google Cloud Storage, posee una consistente fuerte en todas las regiones, reduciendo problemas con inconsistencia eventual. A su vez, en toda etapa de una empresa la velocidad de acceso es fundamental, dándole un punto a favor a Storage. Y está diseñado para escalabilidad masiva.

Data Warehouse:

Se recomienda Google BigQuery, ya que es más simple su forma de cobro, ya que se centra en el uso computacional. También puede escalar los recursos de cómputo de forma dinámica.

RDBMS (compatible con MySQL):

Se recomienda Google Cloud SQL, ya que puede escalar de forma vertical sin la necesidad de reinicios. Por otro lado, su modelo de cobro está centrado en el uso de vCPU y memoria utilizada, siendo rentable en cargas de trabajo variable.

Finalmente, la inclinación indica que es hacia Google, principalmente por factores como:

- En la mayoría de sus servicios ofrece escalabilidad automática, permitiendo adaptarse a las necesidades crecientes de una empresa, centrándose explícitamente en el contexto actual y variable.
- Ofrece consistencia fuerte, garantizando una buena experiencia para los usuarios que utilicen la data, evitando inconsistencia eventual.
- La gran mayoría de sus modelos de cobro son centrados en el consumo, alineándose (al igual que la escalabilidad) al contexto expansible y/o variable de la empresa.
- Si se prefiere el uso de servicios de Google, en un futuro podrá utilizar herramientas de IA como en machine learning. Puede que en un inicio de expansión no sean necesarios, pero en un contexto futuro será de gran utilidad.

Cabe destacar que estas recomendaciones son respecto al breve contexto, ya que se debe tomar en cuenta factores económicos, experiencia del equipo y su fase de madurez, etc.