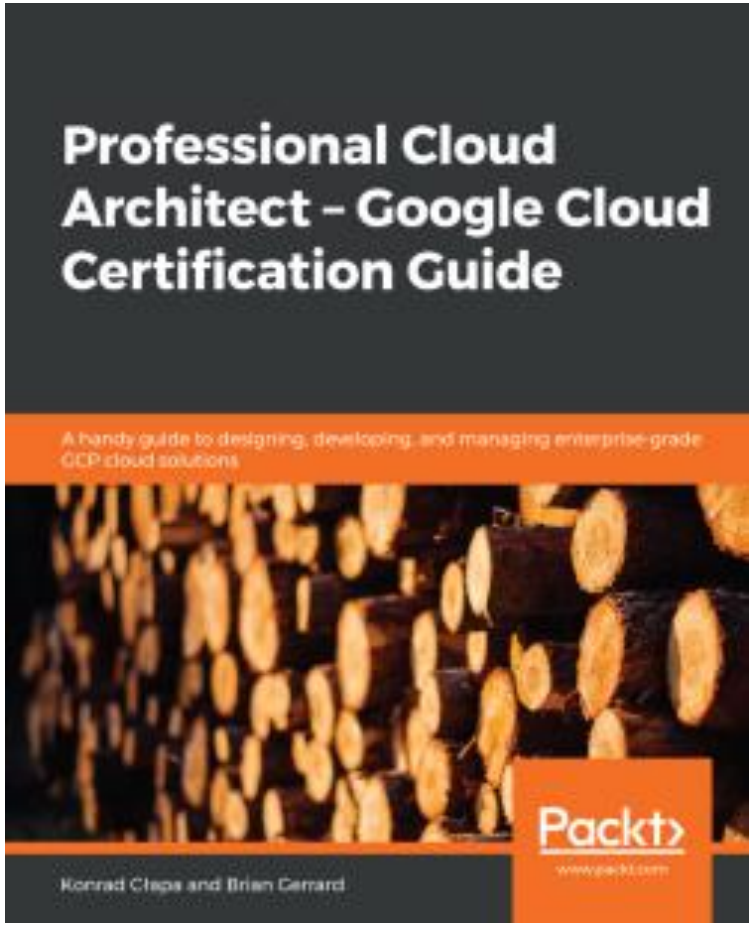
Ingeniería para el Procesado Masivo de Datos(MEXAVM)

Abel Coronado

Computo en la nube 3

Referencia



https://subscription.packtpub.com/book/cloud-and-networking/9781838555276/pref

Seis ventajas del cómputo en la nube

"La computación en la nube es un modelo para habilitar el acceso de red ubicuo, conveniente y bajo demanda a un grupo compartido de recursos informáticos configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios) que se pueden aprovisionar y liberar rápidamente con una gestión mínima. esfuerzo o interacción del proveedor de servicios. Este modelo de nube se compone de cinco características esenciales, tres modelos de servicio y cuatro modelos de implementación".

NIST https://csrc.nist.gov/publications/detail/sp/800-145/final

Características

- **Autoservicio a demanda**: los servicios se aprovisionan automáticamente sin la intervención manual del proveedor.
- **Amplio acceso a la red**: los recursos están disponibles a través de la red.
- **Agrupación de recursos**: los recursos se agrupan desde un grupo compartido, lo que le da al usuario una sensación de independencia de la ubicación. Para algunos de los recursos, la ubicación puede estar restringida.
- **Elasticidad rápida**: los servicios se pueden aprovisionar y desaprovisionar de forma elástica con la capacidad gestionada por el proveedor.
- **Servicio medido**: el uso de recursos se monitorea y se puede informar.

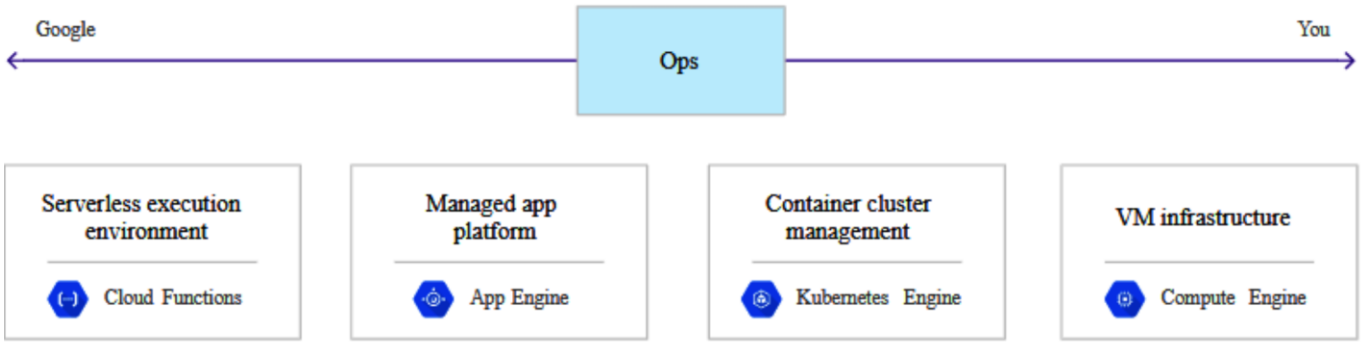
Modos de despligue

- **Nube privada**: utilizada por organizaciones específicas, pero puede ser administrada por terceros
- **Nube pública**: utilizada por el público en general
- **Nube comunitaria**: utilizada por comunidades específicas
- **Nube híbrida**: Compuesta por dos o más nubes diferentes

Modelos de servicio de Google Cloud Computing

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Container as a Service (CaaS)
- Function as a Service (FaaS)
- Managed services

Nivel de Responsabilidad



- Si necesitamos flexibilidad y control sobre nuestras máquinas virtuales (VM), simplemente usaríamos Compute Engine.
- Cloud Functions, todo lo que realmente nos importa es la codificación de una función en un lenguaje compatible con GCP. Una vez hecho y publicado, accedemos a él a través del protocolo HTTP.
- Finalmente, a medida que avanzamos hacia los servicios administrados, comenzamos simplemente a consumir servicios que nos brindan un valor comercial particular sin tener que preocuparnos por las partes subyacentes.

Diferenciadores de Google Cloud Platform

- **La red de Google**: la red de Google es algo que diferencia a GCP de otros proveedores de nube. Google afirma que alrededor del 40% del tráfico de Internet del mundo lo realiza la red de Google, lo que la convierte en la red más grande del mundo. Esto permite que la red de Google proporcione respuestas con una latencia muy baja lo más cerca posible del usuario final.
- **Alcance global**: GCP se desarrolló teniendo en cuenta la disponibilidad global. Verá servicios como el balanceo de cargas disponibles globalmente en lugar de regionalmente, a diferencia de otros proveedores. Esto permite que el cliente se concentre en el desarrollo y adopte una alta disponibilidad y elasticidad listas para usar.
- **ML**: GCP ofrece una gran cantidad de servicios de ML tanto para científicos de datos como para desarrolladores regulares que tienen un conocimiento limitado del tema. ML permite utilizar modelos pre-entrenados, además de ofrecer servicios de AutoML. Este último le permite entrenar modelos ML sin saber cómo se crean realmente. La cartera de estos servicios está creciendo muy rápidamente. El objetivo clave de Google es permitir que las empresas tomen decisiones más rápidas e inteligentes con Machine Learning.

Diferenciadores de Google Cloud Platform

- **Precios**: las instancias de VM tienen un precio por segundo con un tiempo de ejecución mínimo de un minuto. Esto le permite hacer funcionar las máquinas para pruebas cortas y no tener que pagar por una hora completa de uso.
- **Acuerdo de nivel de servicio (SLA)**: los servicios de GCP proporcionan objetivos de nivel de servicio (SLO) de porcentaje de tiempo de actividad mensual. Si no se cumple el SLO, el cliente es elegible para créditos financieros. Tenga en cuenta que este porcentaje depende del servicio y que las características alfa y beta no están incluidas en ningún SLA.
- **Seguridad**: Google utiliza sus 15 años de experiencia en la ejecución de servicios como Gmail en GCP. Tus datos siempre están encriptados con una opción de Google o claves administradas por el cliente.
- **Carbono neutral**: esta podría no ser la característica más importante cuando se trata de funcionalidad, pero vale la pena conocerla. Los centros de datos de Google son neutros en carbono, lo que significa que el 100 % de la energía utilizada para alimentarlos proviene de energías renovables. Esto incluye los centros de datos de GCP.

Localizaciones GCP

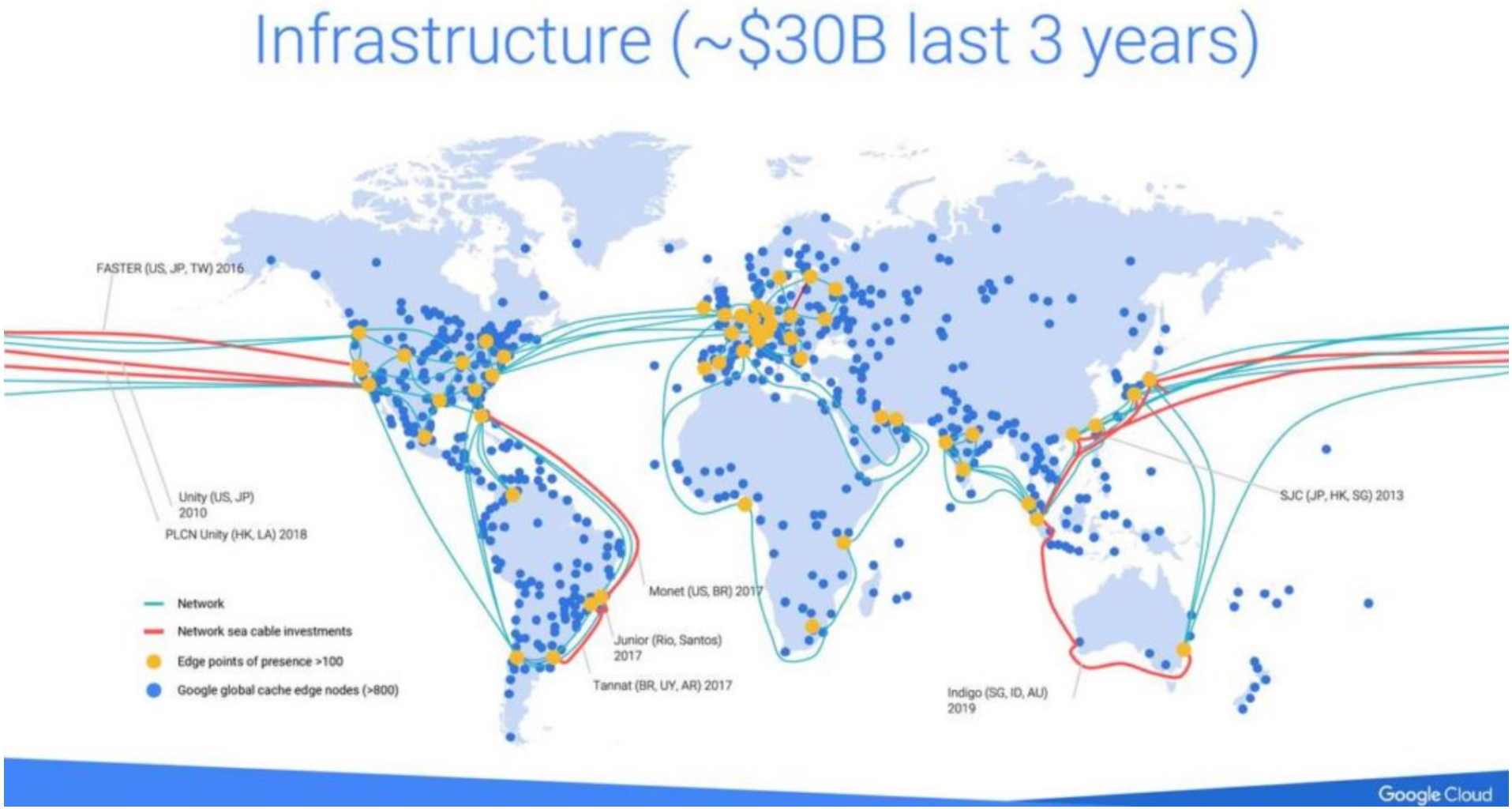
Google define una **región** como un área geográfica independiente que se divide en varias zonas. Las ubicaciones dentro de las regiones deben tener latencias de red de ida y vuelta inferiores a 1 ms en el 95 % de los casos.

Una **zona** es un área de implementación para los recursos de GCP. Tenga en cuenta que una zona no corresponde a un solo centro de datos; puede constar de varios edificios. Aunque una zona proporciona una cierta cantidad de protecciones contra fallas, una zona se considera un punto único de falla (SPOF). Por lo tanto, debe considerar colocar su aplicación en varias zonas para proporcionar tolerancia a fallas.

Las **ubicaciones de borde de red** son conexiones a los servicios de GCP ubicados en un área metropolitana particular.

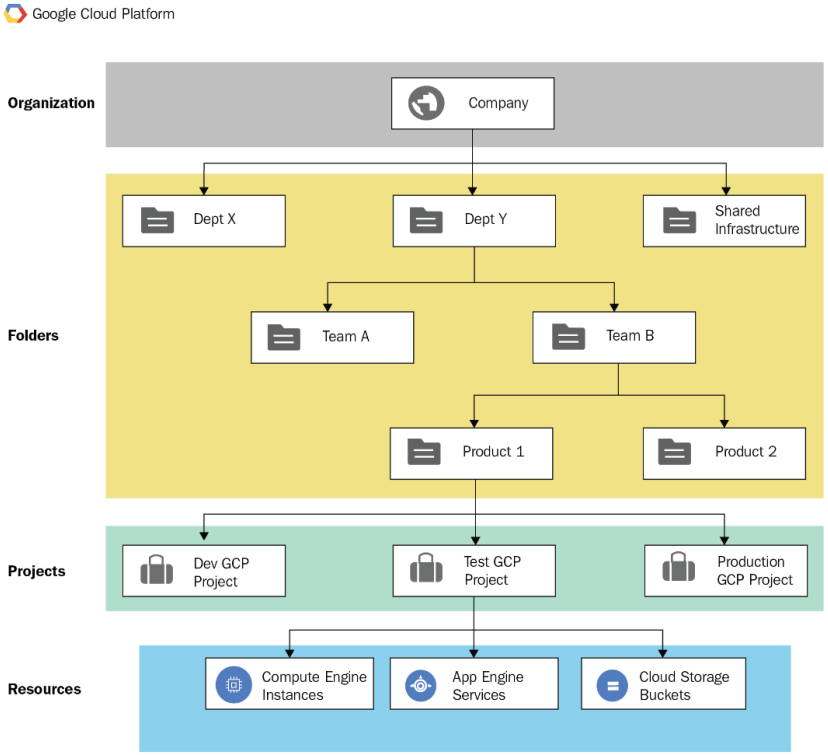
- 22 regions
- 67 zones
- 140 network edge locations

Localizaciones GCP



Gestor de Recursos

GCP consta de contenedores como organizaciones, carpetas y proyectos para agrupar jerárquicamente sus recursos. Esto le permite administrar su configuración y control de acceso. Los recursos se pueden administrar mediante programación mediante API. Google también proporciona herramientas como **Google Cloud Console** y utilidades de línea de comandos, que envuelven las llamadas a la API.



Recursos Globales

Direcciones: estas son direcciones IP externas reservadas y pueden ser utilizadas por balanceadores de carga globales.
Imágenes: Estas son predefinidas o personalizadas por el usuario. Se pueden usar para aprovisionar máquinas virtuales.
Instantáneas: Las instantáneas de un disco persistente permiten la creación de nuevos discos y máquinas virtuales. Tenga en cuenta que también puede exponer una instantánea a un proyecto diferente.
Plantillas de instancias: se pueden usar para la creación de grupos de instancias administrados.
Redes de nube privada virtual (VPC): estas son redes virtuales a las que puede conectar sus cargas de trabajo.
Cortafuegos: De hecho, están definidos por VPC, pero son accesibles globalmente.
Rutas: las rutas le permiten dirigir el tráfico de su red y se asignan a VPC, pero también se consideran globales.

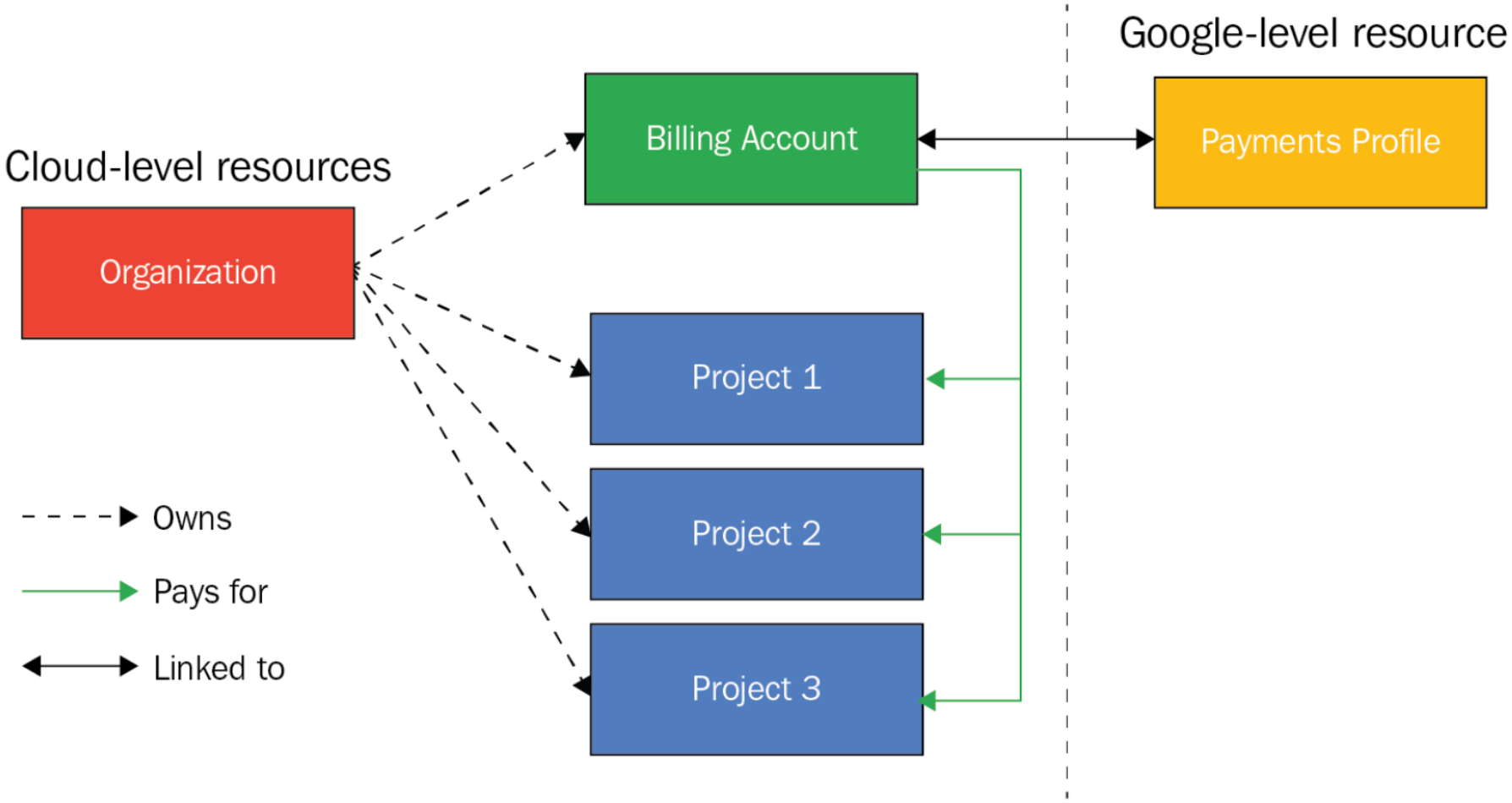
Recursos Regionales

Direcciones: las direcciones IP externas estáticas solo pueden ser utilizadas por instancias que se encuentran en la misma región.
Subredes: Están asociadas a redes VPC y permiten la asignación de direcciones IP a las VM.
Grupos de instancias administrados regionales: estos le permiten escalar grupos de instancias. El alcance se puede establecer en regiones o zonas.
Discos persistentes regionales: estos proporcionan almacenamiento persistente y replicado a las instancias de VM. También se pueden compartir entre proyectos para la creación de instantáneas e imágenes, pero no de archivos adjuntos de disco.

Recursos de Zona

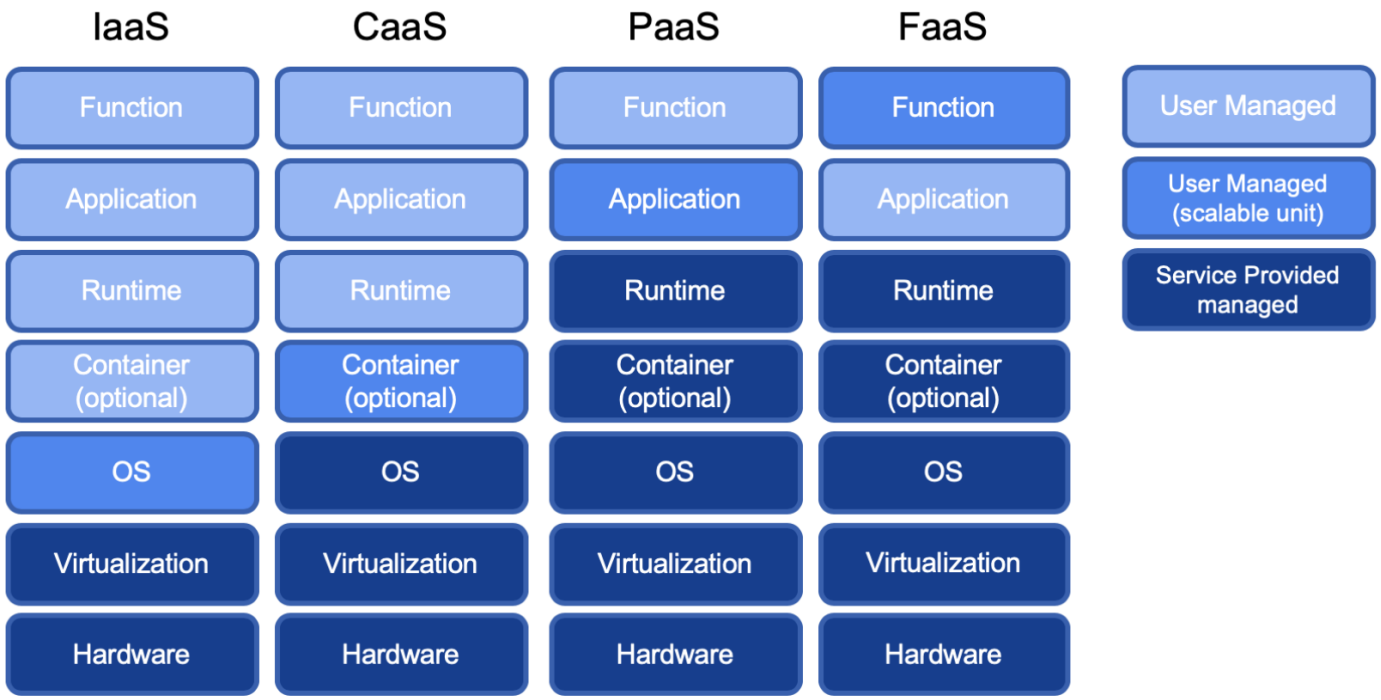
Instancias de VM: residen en una zona en particular.
Discos persistentes zonales: proporcionan almacenamiento persistente a las instancias de VM. También se pueden compartir como discos entre proyectos para la creación de instantáneas e imágenes, pero no como archivos adjuntos de disco.
Tipos de máquinas: estos definen la configuración de hardware para sus instancias de VM y se definen para cualquier zona en particular.
Grupos de instancias administrados zonales: estos le permiten escalar automáticamente grupos de instancias. El alcance se puede establecer en regiones o zonas.

Pagos


Google-level resource
Billing Account
Payments Profile
Cloud-level resources
Organization
Project 1
Project 2
Project 3
Owns
Pays for
Linked to

Servicios principales de Google Cloud Platform

- Computing and hosting services
  - Infrastructure as a Service (IaaS): Google Compute Engine (GCE)
  - Container as a Service (CaaS): Google Kubernetes Engine (GKE)
  - Platform as a Service (PaaS): Google App Engine (GAE)
  - Function as a Service (FaaS): Cloud Functions



Servicios principales de Google Cloud Platform

Servicios de Almacenamiento: El almacenamiento es una parte esencial de la computación en la nube, ya que guarda los datos y el estado de sus aplicaciones. GCP ofrece una amplia variedad de almacenamiento, desde almacenamiento de objetos hasta bases de datos administradas. Los diferentes servicios de almacenamiento que veremos son los siguientes:

Servicios de Almacenamiento

Filestore: Cloud Filestore es un servicio de almacenamiento de archivos administrado. Permite a los usuarios aprovisionar un servicio de almacenamiento conectado a la red (NAS) que se puede integrar con GCE y GKE. Viene con dos niveles de rendimiento: estándar y premium, que ofrecen diferentes rendimientos y operaciones de entrada/salida por segundo (IOPS).

Cloud SQL: Cloud SQL es un servicio de base de datos relacional totalmente administrado que proporciona una base de datos MySQL o PostgreSQL. Ofrece replicación de datos, copias de seguridad, exportaciones de datos y monitoreo. Es ideal cuando necesita mover sus instancias actuales desde las instalaciones y desea delegar el mantenimiento de la base de datos a Google.

Cloud Datastore: Cloud Datastore es una base de datos no SQL completamente administrada. Es ideal para aplicaciones que dependen de datos estructurados de alta disponibilidad a escala. La escalabilidad y la alta disponibilidad se logran mediante una arquitectura distribuida y se abstraen del usuario. Solo hay una base de datos disponible por proyecto. Cloud Datastore ofrece un lenguaje similar a SQL para consultar sus datos.

Cloud Firestore: Cloud Firestore es la próxima generación de Cloud Datastore con varias funciones mejoradas. Puede ejecutarse en modo Nativo o Datastore. El primero es compatible con Cloud Datastore. Google ha anunciado que todos los clientes de Datastore se moverán automáticamente a Cloud Firestore sin tiempo de inactividad ni intervención del usuario. Todos los proyectos nuevos deben crearse en Cloud Firestore en lugar de Datastore.

Servicios de Almacenamiento

Cloud Spanner: Cloud Spanner es un servicio de base de datos totalmente administrado, distribuido globalmente y altamente consistente. Es una base de datos relacional fuerte y consistente con capacidades de escalado de bases de datos no relacionales. Los usuarios pueden definir un esquema y aprovechar el estándar de la industria ANSI 2011 SQL. Tiene un rendimiento muy alto, con un acuerdo de nivel de servicio (SLA) de disponibilidad del 99,999 %, lo que significa que casi no hay tiempo de inactividad aplicable. Los Cloud Spanners están destinados a casos de uso como el comercio financiero, los seguros, los centros de llamadas globales, las telecomunicaciones, los juegos y el comercio electrónico. La consistencia global lo hace ideal para aplicaciones accesibles globalmente.

Bigtable: Bigtable es una base de datos no SQL completamente administrada y de escala masiva con una latencia inferior a 10 ms. Google lo utiliza para ofrecer servicios como Gmail y Google Maps. Es ideal para casos de uso de almacenamiento de fintech, IoT y ML. Se integra fácilmente con familias de productos de big data como Dataproc y Dataflow. Se basa en Apache HBase de código abierto, lo que permite el uso de su API. El costo de Bigtable es mucho más alto que el de Datastore, por lo que la base de datos debe elegirse con mucho cuidado.

Bases de datos personalizadas: también puede optar por usar Compute Engine para instalar una base de datos de su elección, como MongoDB; sin embargo, sería un servicio no administrado.

Servicios principales de Google Cloud Platform

Servicios de Red: La red de GCP se basa en redes definidas por software (SDN), que permite a los usuarios ofrecer todos los servicios de red mediante programación. Todos los servicios están completamente administrados, dejando a los usuarios la tarea de configurarlos de acuerdo con sus requisitos. Los servicios de red que veremos son los siguientes:

Nube privada virtual (VPC): la VPC es la base de las redes de GCP. Cada proyecto de GCP tiene una red VPC predeterminada creada, pero el usuario también puede crear nuevas redes. Puede considerarlo como una versión en la nube de una red física. Una VPC puede contener una o más subredes regionales. Una VPC crea un límite lógico global que permite la comunicación entre máquinas virtuales dentro de la misma VPC. Para permitir la comunicación entre las VPC, el tráfico debe atravesar Internet o mediante la interconexión de VPC.

Balanceador de carga: el balanceador de carga permite la distribución del tráfico entre sus cargas de trabajo. Está disponible para GCE, GAE y GKE. Para GCE, puede elegir balanceadores de carga con alcance global o regional. La elección también dependerá del tipo de red.

Red privada virtual (VPN): las VPN permiten una conexión entre su red local y la VPC de GCP a través de un túnel IPsec a través de Internet. Solo se admiten VPN de sitio a sitio. Para establecer una conexión VPN, debe haber dos puertas de enlace a cada lado del túnel. El tráfico en tránsito está encriptado. Se admite tanto el enrutamiento estático como el dinámico, y el primero requiere un enrutador en la nube. El uso de una VPN debe ser el primer método para conectar su entorno a GCP, ya que implica el costo más bajo. Si hay requisitos de baja latencia y alto ancho de banda, se debe considerar Cloud Interconnect.

Cloud Interconnect: si se necesita una latencia baja y una conexión de alta disponibilidad, se debe considerar la interconexión. En este caso, el tráfico no atraviesa Internet.

Servicios de Red

Cloud Router: Cloud Router es un servicio que permite el intercambio de enrutamiento dinámico entre Compute Engine, VPN y redes externas. Elimina la necesidad de crear rutas estáticas.

Cloud DNS: Cloud DNS es un servicio de DNS administrado con un SLA del 100 %. Traduce dominios a direcciones IP. Se pueden gestionar millones de zonas y registros. Cloud DNS también puede alojar zonas privadas a las que solo se puede acceder desde su red de GCP. Se puede integrar en las instalaciones, donde su DNS local está autorizado y Cloud DNS es responsable del almacenamiento en caché.

Red de entrega de contenido (CDN) en la nube: Cloud CDN es un servicio que permite el almacenamiento en caché de contenido equilibrado de carga HTTP(S), incluidos los objetos del depósito de Cloud Storage. El almacenamiento en caché reduce el tiempo y el costo de entrega de contenido. También puede protegerlo de un ataque de denegación de servicio distribuido (DDoS). Los datos se almacenan en caché en los puntos de borde distribuidos globalmente de Google. En la primera solicitud, cuando el contenido no se almacena en caché, los datos se recuperan de un servicio de back-end. Los datos de la próxima llamada se servirán directamente desde la memoria caché hasta que se alcance el tiempo de vencimiento.

Cloud NAT: Cloud NAT es un servicio regional que permite que las máquinas virtuales sin IP externas se comuniquen con Internet. Es un servicio completamente administrado con escalabilidad automática incorporada. Funciona con GCE y GKE. Es una mejor alternativa para las instancias NAT que deben ser administradas por los usuarios.

Firewall: GCP Firewall es un servicio que permite la microsegmentación. Las reglas de firewall se crean por VPC y se pueden basar en IP, rangos de IP, etiquetas y cuentas de servicio. Varias reglas de firewall se crean de forma predeterminada, pero se pueden modificar.

Identity Aware Proxy (IAP): IAP es un servicio que reemplaza la VPN cuando un usuario está trabajando desde una red que no es de confianza. Controla el acceso a su aplicación según la identidad del usuario, el estado del dispositivo y la dirección IP. Es parte del modelo de seguridad BeyondCorp de Google.

Cloud Armor: Cloud Armor es un servicio que permite la protección contra ataques DDoS a la infraestructura utilizando la infraestructura global y los sistemas de seguridad de Google. Se integra con balanceadores de carga HTTP(S) globales y bloquea el tráfico en función de rangos o direcciones IP. El modo de vista previa permite a los usuarios analizar el patrón de ataque sin interrumpir a los usuarios habituales.

Servicios principales de Google Cloud Platform

Los **servicios de big data** permiten al usuario procesar grandes cantidades de datos para brindar respuestas a problemas complejos. GCP ofrece muchos servicios que se integran estrechamente para crear una canalización de análisis de datos de extremo a extremo (E2E).

BigQuery: BigQuery es un almacén de datos en la nube altamente escalable y completamente administrado. Permite a los usuarios realizar operaciones de análisis con ML integrado. BigQuery no tiene servidor y puede alojar petabytes de datos. La infraestructura subyacente se escala sin problemas y permite el procesamiento de datos en paralelo. Los datos se pueden almacenar en BigQuery Storage, Cloud Storage, Bigtable, Sheets o Google Drive. El usuario define conjuntos de datos que contienen tablas. BigQuery usa SQL familiar compatible con ANSI para consultas y proporciona controladores ODBC y JDBC. Los usuarios pueden elegir entre dos tipos de modelos de pago: uno es flexible e implica pagar por almacenamiento y consultas, y el otro implica una tarifa plana con costos mensuales estables. Es ideal para casos de uso como análisis predictivo, IoT y análisis de registros, y se integra con la familia de productos de big data de GCP.

Pub/Sub: este es un servicio de mensajería asincrónica completamente administrado que le permite acoplar los componentes de su aplicación de manera flexible. No tiene servidor con disponibilidad global. Su aplicación puede publicar mensajes en un tema o suscribirse a él para recibir mensajes. Pub/Sub también puede enviar mensajes a Webhooks.

Servicios Big Data

Dataproc: Dataproc es un clúster de Apache Spark y Hadoop completamente administrado. Permite a los usuarios crear clústeres a pedido y usarlos solo cuando se necesita procesamiento de datos. Se factura por segundo. Permite a los usuarios mover clústeres locales ya existentes a la nube sin refactorizar el código. El uso de instancias prioritarias puede reducir aún más el costo.

Dataflow: Cloud Dataflow es un servicio totalmente administrado para procesar datos en secuencias y lotes. Se basa en Apache Beam de código abierto, no tiene ningún servidor y ofrece una capacidad casi ilimitada. Administrará los recursos y equilibrará el trabajo para el usuario. Se puede usar para casos de uso como análisis de fraude en línea, IoT, atención médica y logística.

Dataprep: esta es una herramienta que se puede usar para realizar la visualización y exploración de datos sin que se requiera ninguna habilidad de codificación. Los datos se pueden preparar de forma interactiva para su posterior análisis.

Datalab: Datalab es una herramienta integrada en Jupyter (anteriormente IPython) que permite a los usuarios explorar, analizar y transformar datos. También permite a los usuarios crear modelos de datos de ML y aprovecha Compute Engine.

Data Studio: esta es una herramienta que le permite consumir datos de fuentes y visualizarlos en forma de informes y paneles.

Cloud Composer: este es un servicio totalmente administrado basado en Apache Airflow de código abierto. Le permite crear y orquestar canalizaciones de big data.

Servicios principales de Google Cloud Platform

Servicios de Machine Learning: Uno de los puntos más fuertes de Google es su experiencia a largo plazo con ML. GCP ofrece varios servicios en torno a ML. Puede elegir entre un modelo pre-entrenado o entrenar el modelo usted mismo.

Cloud ML Engine: ML Engine es un servicio administrado que le permite entrenar y alojar sus modelos ML en GCP. Aprovecha la aplicación TensorFlow para el proceso de capacitación. Google administra la infraestructura subyacente, mientras que los usuarios pueden elegir entre diferentes opciones de hardware. Se puede acceder al modelo entrenado a través de las API para realizar predicciones.

API preentrenadas: las API de ML son servicios que le permiten aprovechar varios modelos preentrenados, lo que le permite analizar un video. Actualmente, las siguientes API están disponibles:

Google Cloud Video Intelligence
Google Cloud Speech
Google Cloud Vision
Google Cloud Natural Language
Google Cloud Translation

Servicios de Machine Learning

AutoML: AutoML es un servicio que los desarrolladores pueden usar para entrenar modelos sin tener un conocimiento extenso de la ciencia de datos. Por ejemplo, al proporcionar muestras etiquetadas a AutoML, se puede entrenar para que reconozca objetos que Vision API no reconoce. Las siguientes son las muestras etiquetadas de AutoML:

- Traducción de AutoML
- Visión y lenguaje natural de AutoML

Dialogflow: Este es un servicio que te permite construir aplicaciones de conversación que pueden interactuar con seres humanos. La interfaz puede interactuar con muchas plataformas compatibles, como Slack o Google Assistant. También puede integrarse con las funciones de Firebase para integrarse con plataformas de terceros utilizando API comunes.

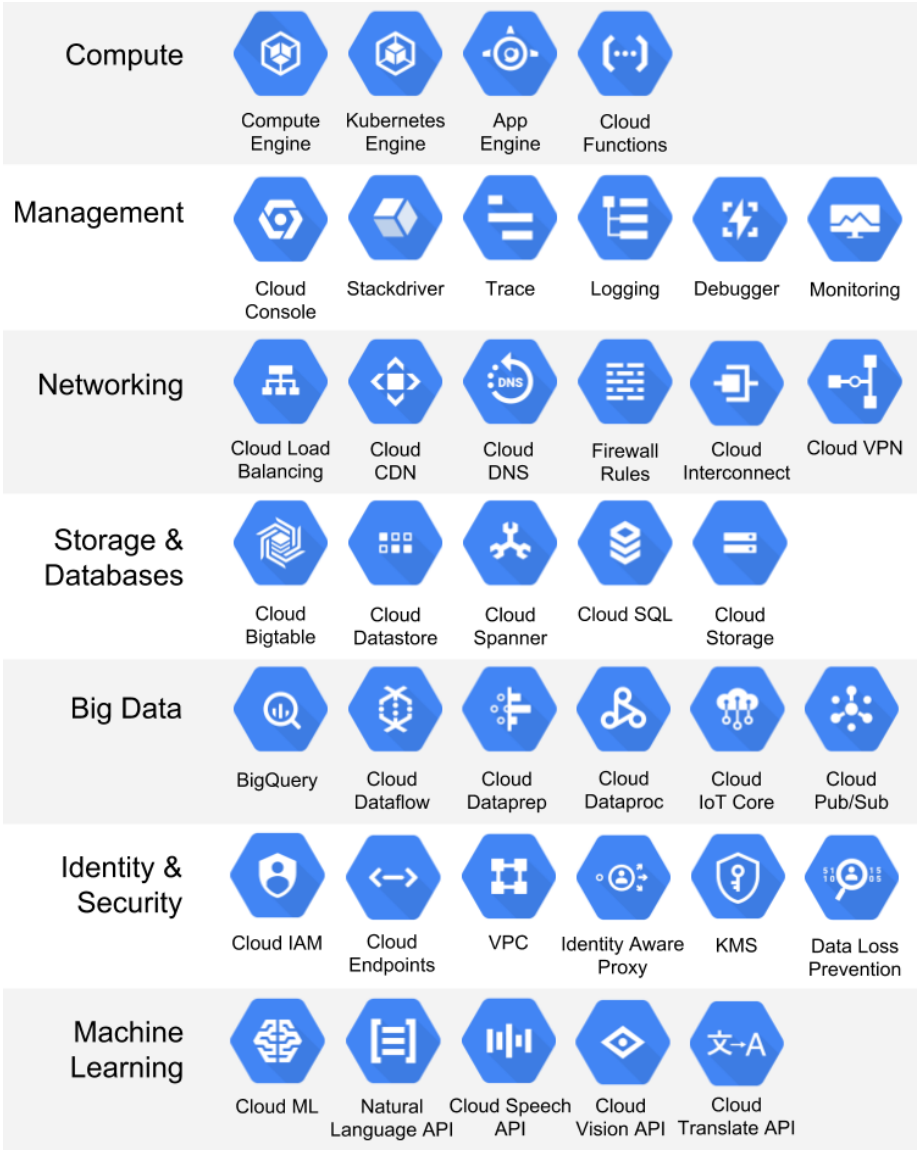
Servicios principales de Google Cloud Platform

La gestión de identidades y accesos (IAM) es uno de los aspectos más importantes de cualquier nube. Le permite controlar quién tiene acceso a la nube, pero también puede proporcionar servicios de identidad a sus aplicaciones. En resumen, esto se logra mediante una combinación de roles y permisos. Los roles se asignan a usuarios o grupos. Echemos un vistazo a las opciones que tenemos en GCP:

IAM: IAM permite que el administrador de GCP controle la autorización de los servicios de GCP. Los administradores pueden crear roles con permisos granulares. Luego, los roles se pueden asignar a los usuarios o, preferiblemente, a un grupo de usuarios.

Cloud Identity: Cloud Identity es una oferta de identidad como servicio (IDaaS). Se encuentra fuera de GCP, pero se puede integrar fácilmente con GCP. Le permite crear organizaciones, grupos y usuarios, y administrarlos de forma centralizada. Si ya tiene un catálogo de usuarios existente, puede sincronizarlo con Cloud Identity.

Servicios de Google Cloud Platform



TEMA 1. INTRODUCCIÓN A LAS TECNOLOGÍAS BIG DATA	1.1. Las Tres V's del Big Data	**Volumen, Velocidad y Variedad** son las características principales del Big Data. Entenderlas es clave para identificar un proyecto de Big Data.
		Volumen: Se refiere a la gran cantidad de datos que se generan y necesitan ser procesados.
		Velocidad: La rapidez con la que se generan y procesan los datos.
		Variedad: La diversidad de los tipos de datos, estructurados y no estructurados.
	1.2. Definición de Big Data y sus Aplicaciones	Big Data son conjuntos de tecnologías diseñadas para manejar grandes volúmenes de datos heterogéneos.
		Apache Kafka: Es un sistema de mensajería que permite manejar flujos de datos en tiempo real.

TEMA 2. HDFS Y MAPREDUCE	2.1. HDFS (Hadoop Distributed File System)	HDFS es un sistema de archivos distribuido que permite el almacenamiento y procesamiento de grandes volúmenes de datos a través de múltiples máquinas.	
		Comandos HDFS relevantes:	**mkdir**: Crea un nuevo directorio en HDFS.
			rm: Elimina archivos en HDFS, y puede hacerlo de manera recursiva para carpetas.
			chmod: Cambia los permisos de un archivo en HDFS.
	2.2. MapReduce y Apache Spark	**MapReduce**: Es un paradigma de programación para procesar datos distribuidos en clústeres. Se basa en las funciones Map y Reduce.	
		Apache Spark: Mejora las limitaciones de MapReduce al permitir operaciones en memoria y un procesamiento más rápido.	
		Funciones Map y Reduce: Spark permite realizar operaciones de mapeo y reducción en paralelo, distribuyendo tareas entre nodos.	

TEMA 3. SPARK I	3.1. Aprendizaje Automático en Big Data	Los algoritmos de Machine Learning se benefician de Spark al permitir múltiples iteraciones sobre los datos.
		RDD (Resilient Distributed Dataset): Es una estructura de datos clave en Spark que permite almacenar y procesar datos en paralelo.
	3.2. Métodos Principales en Spark ML	El método **fit** es el principal en los estimadores de Spark ML para ajustar los modelos a los datos.
		Transformers y **Pipelines**: Después del entrenamiento, los modelos utilizan estas interfaces para procesar nuevos datos.
TEMA 4. SPARK II	4.1. DataFrames y SQL en Spark	**DataFrames**: Estructuras de datos distribuidas que permiten manipular datos de forma eficiente en Spark.
		Spark SQL: Interfaz que permite ejecutar consultas SQL sobre DataFrames.
TEMA 5. SPARK III	5.1. Spark MLlib y Structured Streaming	**MLlib**: Biblioteca de aprendizaje automático de Spark que incluye algoritmos como regresión, clasificación y clustering.
		Spark Structured Streaming: Módulo para procesar flujos de datos en tiempo real.

TEMA 6. APACHE KAFKA	6.1. Introducción a Apache Kafka	Kafka es una plataforma de mensajería distribuida que permite la publicación, suscripción y procesamiento de flujos de datos en tiempo real.
		Producers y Consumers: Los productores envían mensajes a Kafka, mientras que los consumidores los leen para su procesamiento.
TEMA 7. HIVE E IMPALA	7.1. Apache Hive y Apache Impala	**Hive**: Herramienta para manejar datos en HDFS usando SQL.
		Impala: Motor de consultas SQL que permite ejecutar consultas en HDFS con alta eficiencia.
TEMA 8. CLOUD COMPUTING I	8.1. Introducción a Cloud Computing	**Cloud Computing**: Modelo de computación que permite el acceso a recursos de computación y almacenamiento a través de la nube.
		Ventajas del Cloud Computing: Escalabilidad, reducción de costos y flexibilidad en el acceso a los recursos.
	8.2. Servicios en la Nube	**Tipos de Nube**: Pública, privada, híbrida.
		Servicios en la Nube: IaaS, PaaS, SaaS.
		Microsoft Azure: Plataforma en la nube que ofrece una amplia gama de servicios.

TEMA 9. CLOUD COMPUTING II	9.1. Amazon Web Services (AWS)	**AWS**: Plataforma de servicios en la nube que ofrece soluciones de computación, almacenamiento, bases de datos y más.
		Servicios de AWS: Computación, almacenamiento, redes, bases de datos, seguridad y más.
TEMA 10. CLOUD COMPUTING III	10.1. Google Cloud Platform (GCP)	**GCP**: Plataforma de servicios en la nube de Google que ofrece soluciones de computación, almacenamiento, bases de datos, y análisis de Big Data.
		Servicios de GCP: Computación, almacenamiento, redes, bases de datos, y machine learning.

¡GRACIAS!

unir
LA UNIVERSIDAD
EN INTERNET