Google cloud  
Professional Data Engineer  
  
  
Módulo 1 - Clase 1

Tabla de contenido

[I. Introducción a Google Cloud Platform 2](#_Toc128838133)

[1. Fundamentos de Google Cloud Platform 2](#_Toc128838134)

[2. Tipos de servicios en la nube 2](#_Toc128838135)

[II. Certificaciones 8](#_Toc128838136)

[III. El Rol del Data Engineer 8](#_Toc128838137)

[IV. Big Data y Machine Learning 9](#_Toc128838138)

[V. Construir un Data Lake 9](#_Toc128838139)

[VI. Recursos Adicionales 9](#_Toc128838140)

1. **Introducción a Google Cloud Platform**
   1. **Fundamentos de Google Cloud Platform**

Google Cloud Platform (GCP) es un servicio de nube pública que ofrece algunas de las mismas tecnologías utilizadas por Google para entregar sus propios productos. Este módulo describe los componentes más importantes de GCP y analiza cómo difiere de la informática basada en el centro de datos local (Data Center On Premises).

* 1. **Tipos de servicios en la nube**

Los proveedores de la nube pública como Google, Amazon y Microsoft ofrecen una gama de servicios para implementar la informática, el almacenamiento, las redes y otras infraestructuras para ejecutar una amplia gama de servicios y aplicaciones empresariales.

Algunos usuarios de la nube son nuevas empresas que comienzan en la nube. Nunca han tenido su propio hardware y software. Otros clientes de la nube son empresas con múltiples centros de datos que usan nubes públicas para complementar sus centros de datos. Estos diferentes tipos de usuarios tienen diferentes requisitos.

Una empresa que comienza en la nube puede elegir los servicios que mejor se adapten a sus necesidades arquitectónicas y de aplicaciones sin tener que considerar la infraestructura existente. Por ejemplo, una startup podría usar los servicios de gestión de acceso e identidad en la nube de GCP para todas las necesidades de autenticación y autorización. Una empresa que ya ha invertido en una solución de Microsoft Active Directory para la gestión de identidades puede querer aprovechar ese sistema en lugar de trabajar únicamente con el sistema de gestión de identidad de la nube. Esto puede llevar a un trabajo adicional para integrar los dos sistemas y mantenerlos sincronizados.

Otra área de preocupación para las empresas con su propia infraestructura es establecer y mantener una red segura entre sus recursos locales y sus recursos de nube pública. Si habrá tráfico de red de gran volumen entre los sistemas locales y la nube pública, la empresa puede necesitar invertir en una red dedicada entre su centro de datos y una instalación del proveedor de la nube pública. Si el volumen de tráfico no justifica el costo de una conexión dedicada entre instalaciones, entonces la compañía puede usar una red privada virtual que se ejecuta a través de Internet pública. Esto requiere un diseño y gestión de red adicional que una empresa que se encuentre únicamente en la nube no debería abordar.

Los proveedores de nube pública ofrecen servicios que se dividen en cuatro amplias categorías.

1. Procesamiento
2. Almacenamiento
3. Redes
4. Servicios especializados como por ejemplo los servicios de Machine Learning (ML).

**Máquinas Virtuales**

Las máquinas virtuales son una unidad básica de recursos informáticos y un buen punto de partida para experimentar con la nube. Después de crear una cuenta con un proveedor de la nube y proporcionar información de facturación, puede usar un portal o herramientas de línea de comandos para crear máquinas virtuales. Google Cloud Platform ofrece una variedad de máquinas virtuales preconfiguradas con diferentes cantidades de vCPU y cantidades de memoria. También puede crear una configuración personalizada si las ofertas preconfiguradas no satisfacen sus necesidades.

Una vez que cree una VM, puede iniciar sesión y administrarla como desee. Tiene acceso completo a la VM, por lo que puede configurar sistemas de archivos, agregar almacenamiento persistente, parchear el sistema operativo o instalar paquetes adicionales. Usted decide qué ejecutar en la VM, quién más tendrá acceso a ella y cuándo cerrar la VM. Una máquina virtual que administra es como tener un servidor en su oficina para el que tiene todos los derechos de administrador.

Por supuesto, puede crear varias máquinas virtuales que ejecuten diferentes sistemas operativos y aplicaciones. GCP también proporciona servicios, como equilibradores de carga, que proporcionan un único punto de acceso a un back-end distribuido. Esto es especialmente útil cuando necesita tener alta disponibilidad para su aplicación. Si una de las máquinas virtuales en un clúster falla, la carga de trabajo se puede dirigir a las otras máquinas virtuales en el clúster. Los escaladores automáticos pueden agregar o eliminar máquinas virtuales del clúster en función de la carga de trabajo. Esto se llama autoescalado . Esto ayuda a controlar el costo al no ejecutar más máquinas virtuales de las necesarias y garantizar que haya suficiente capacidad informática disponible cuando aumenten las cargas de trabajo.

**Clusters de Kubernetes Gestionados**

Google Cloud Platform le brinda todas las herramientas que necesita para crear y administrar clústeres de servidores. Muchos usuarios de la nube prefieren centrarse en sus aplicaciones y no en las tareas necesarias para mantener un grupo de servidores en funcionamiento. Para esos usuarios, los clústeres administrados son una buena opción.

Los clústeres gestionados hacen uso de contenedores. Un contenedor es como una máquina virtual liviana que aísla los procesos que se ejecutan en un contenedor de los procesos que se ejecutan en otro contenedor en el mismo servidor. En un clúster administrado, puede especificar la cantidad de servidores que desea ejecutar y los contenedores que deben ejecutarse en ellos. También puede especificar parámetros de autoescalado para optimizar el número de contenedores en ejecución.

En un clúster administrado, se supervisa el estado de los contenedores. Si un contenedor falla, el software de administración del clúster lo detectará e iniciará otro contenedor.

Los contenedores son buenas opciones cuando necesita ejecutar aplicaciones que dependen de múltiples micro servicios que se ejecutan en su entorno. Los servicios se implementan a través de contenedores, y el servicio de administración de clúster se encarga de la supervisión, la creación de redes y algunas tareas de administración de seguridad.

**Computación sin servidor (Serverless)**

Tanto las máquinas virtuales como los clústeres de kubernetes administrados requieren cierto nivel de esfuerzo para configurar y administrar recursos informáticos. La informática sin servidor es un enfoque que permite a los desarrolladores y administradores de aplicaciones ejecutar su código en un entorno informático que no requiere la configuración de máquinas virtuales o clústeres de kubernetes.

Google Cloud Platform tiene dos opciones de computación sin servidor (serverless): App Engine y Cloud Functions. App Engine se utiliza para aplicaciones y contenedores que se ejecutan durante largos períodos de tiempo, como un backend de un sitio web, un sistema de punto de venta o una aplicación comercial personalizada. Cloud Functions es una plataforma para ejecutar código en respuesta a un evento, como cargar un archivo o agregar un mensaje a una cola de mensajes. Esta opción sin servidor funciona bien cuando necesita responder a un evento ejecutando un proceso corto codificado en una función o llamando a una aplicación de ejecución más larga que podría estar ejecutándose en una VM, clúster administrado o App Engine.

**Almacenamiento**

Las nubes públicas ofrecen algunos tipos de servicios de almacenamiento que son útiles para una amplia gama de requisitos de aplicaciones. Estos tipos incluyen lo siguiente:

* Almacenamiento de objetos
* Almacenamiento de archivos
* Bloques de almacenamiento

**Cachés**

Los usuarios empresariales de servicios en la nube a menudo utilizarán una combinación de estos servicios.

**Almacenamiento de objetos**

El almacenamiento de objetos es un sistema que gestiona el uso del almacenamiento en términos de objetos o blobs. Por lo general, estos objetos son archivos, pero es importante tener en cuenta que los archivos no se almacenan en un sistema de archivos convencional. Los objetos se agrupan en cubos. Cada objeto es direccionable individualmente, generalmente por una URL.

El almacenamiento de objetos no está limitado por el tamaño de los discos o unidades de estado sólido (SSD) conectados a un servidor. Los objetos se pueden cargar sin preocuparse por la cantidad de espacio disponible en un disco. Se almacenan múltiples copias de objetos para mejorar la disponibilidad y durabilidad. En algunos casos, se pueden almacenar copias de objetos en diferentes regiones para garantizar la disponibilidad, incluso si una región se vuelve inaccesible.

Otra ventaja del almacenamiento de objetos es que no tiene servidor. No es necesario crear máquinas virtuales y adjuntarles almacenamiento. Se puede acceder al almacenamiento de objetos de Google Cloud Platform, llamado Cloud Storage, desde los servidores que se ejecutan en GCP, así como desde otros dispositivos con acceso a Internet.

Los controles de acceso se pueden aplicar a nivel de objeto. Esto permite a los usuarios de almacenamiento en la nube controlar qué usuarios pueden acceder y actualizar objetos.

**Almacenamiento de archivos**

Los servicios de almacenamiento de archivos proporcionan un sistema de almacenamiento jerárquico para archivos. El almacenamiento de sistemas de archivos proporciona sistemas de archivos compartidos en red. Google Cloud Platform tiene un servicio de almacenamiento de archivos llamado Cloud Filestore, que se basa en el sistema de almacenamiento del Sistema de archivos de red (NFS).

El almacenamiento de archivos es adecuado para aplicaciones que requieren un sistema operativo, como el acceso a archivos. El sistema de almacenamiento de archivos desacopla el sistema de archivos de máquinas virtuales específicas. El sistema de archivos, sus directorios y sus archivos existen independientemente de las máquinas virtuales o las aplicaciones que pueden acceder a esos archivos.

**Bloque de almacenamiento**

El almacenamiento en bloque utiliza una estructura de datos de tamaño fijo llamada bloque para organizar los datos. El almacenamiento en bloque se usa comúnmente en discos efímeros y persistentes conectados a máquinas virtuales. Con un sistema de almacenamiento en bloque, puede instalar sistemas de archivos en la parte superior del almacenamiento en bloque, o puede ejecutar aplicaciones que acceden a bloques directamente. Algunas bases de datos relacionales pueden diseñarse para acceder a bloques directamente en lugar de trabajar a través de sistemas de archivos.

En los sistemas de archivos de Linux, 4KB es un tamaño de bloque común. Las bases de datos relacionales a menudo escriben directamente en bloques, pero a menudo usan tamaños más grandes, como 8 KB o más.

El almacenamiento en bloque está disponible en discos conectados a máquinas virtuales en Google Cloud Platform. El almacenamiento en bloque puede ser persistente o efímero. Un disco persistente sigue existiendo y almacena datos incluso si se desconecta de un servidor virtual o el servidor virtual al que está conectado se apaga. Existen discos efímeros y almacenan datos solo mientras se esté ejecutando una VM. Los discos efímeros almacenan archivos del sistema operativo y otros archivos y datos que se eliminan cuando se apaga la máquina virtual. Los discos persistentes se usan cuando desea que existan datos en un dispositivo de almacenamiento en bloque independiente de una VM. Estos discos son buenas opciones cuando tiene datos que desea que estén disponibles independientemente del ciclo de vida de una VM, y admiten un acceso rápido a nivel de sistema operativo y de sistema de archivos.

El almacenamiento de objetos también mantiene los datos independientes del ciclo de vida de una VM, pero no admite el acceso a nivel del sistema operativo o del sistema de archivos; tienes que usar protocolos de nivel superior como HTTP para acceder a los objetos. Lleva más tiempo recuperar datos del almacenamiento de objetos que recuperarlos del almacenamiento en bloque. Es posible que necesite una combinación de almacenamiento de objetos y almacenamiento de bloques para satisfacer las necesidades de su aplicación. El almacenamiento de objetos puede almacenar grandes volúmenes de datos que se copian en un disco persistente cuando es necesario. Esta combinación ofrece la ventaja de grandes volúmenes de almacenamiento junto con el acceso basado en el sistema operativo y el sistema de archivos cuando sea necesario.

**Cachés**

Los cachés son almacenes de datos en memoria que mantienen un acceso rápido a los datos. El tiempo que lleva recuperar datos se llama latencia . La latencia de las tiendas en memoria está diseñada para ser inferior a milisegundos. Para darle una comparación, aquí hay algunas otras latencias:

Hacer una referencia de memoria principal toma 100 nanosegundos, o 0.1 microsegundos

Leer 4KB al azar desde un SSD toma 150 microsegundos

Leer 1MB secuencialmente de la memoria toma 250 microsegundos

Leer 1MB secuencialmente de un SSD toma 1,000 microsegundos, o 1 milisegundo

Leer 1MB secuencialmente desde el disco toma 20,000 microsegundos, o 20 milisegundos

Aquí hay algunas conversiones para referencia:

1,000 nanosegundos equivalen a 1 microsegundo.

1,000 microsegundos equivalen a 1 milisegundo.

1,000 milisegundos equivalen a 1 segundo.

Estos y otros datos de tiempo útiles están disponibles en los "Números de latencia que todo programador debe saber" de Jonas Bonér en: https://gist.github.com/jboner/2841832 .

Analicemos un ejemplo de lectura de 1 MB de datos. Si tiene los datos almacenados en una memoria caché en memoria, puede recuperar los datos en 250 microsegundos, o 0,25 milisegundos. Si esos mismos datos se almacenan en un SSD, tardará cuatro veces más en recuperarlos en 1 milisegundo. Si recupera los mismos datos de una unidad de disco duro, puede esperar 20 milisegundos, u 80 veces más que leer desde un caché en memoria.

Los cachés son bastante útiles cuando necesita mantener la latencia de lectura al mínimo en su aplicación. Por supuesto, ¿a quién no le gustan los tiempos de recuperación rápidos? ¿Por qué no siempre almacenamos nuestros datos en cachés? Hay tres razones

La memoria es más cara que el almacenamiento SSD o unidad de disco duro (HDD). En muchos casos, no es práctico tener tanto almacenamiento en memoria como almacenamiento de bloque persistente en SSD o HDD.

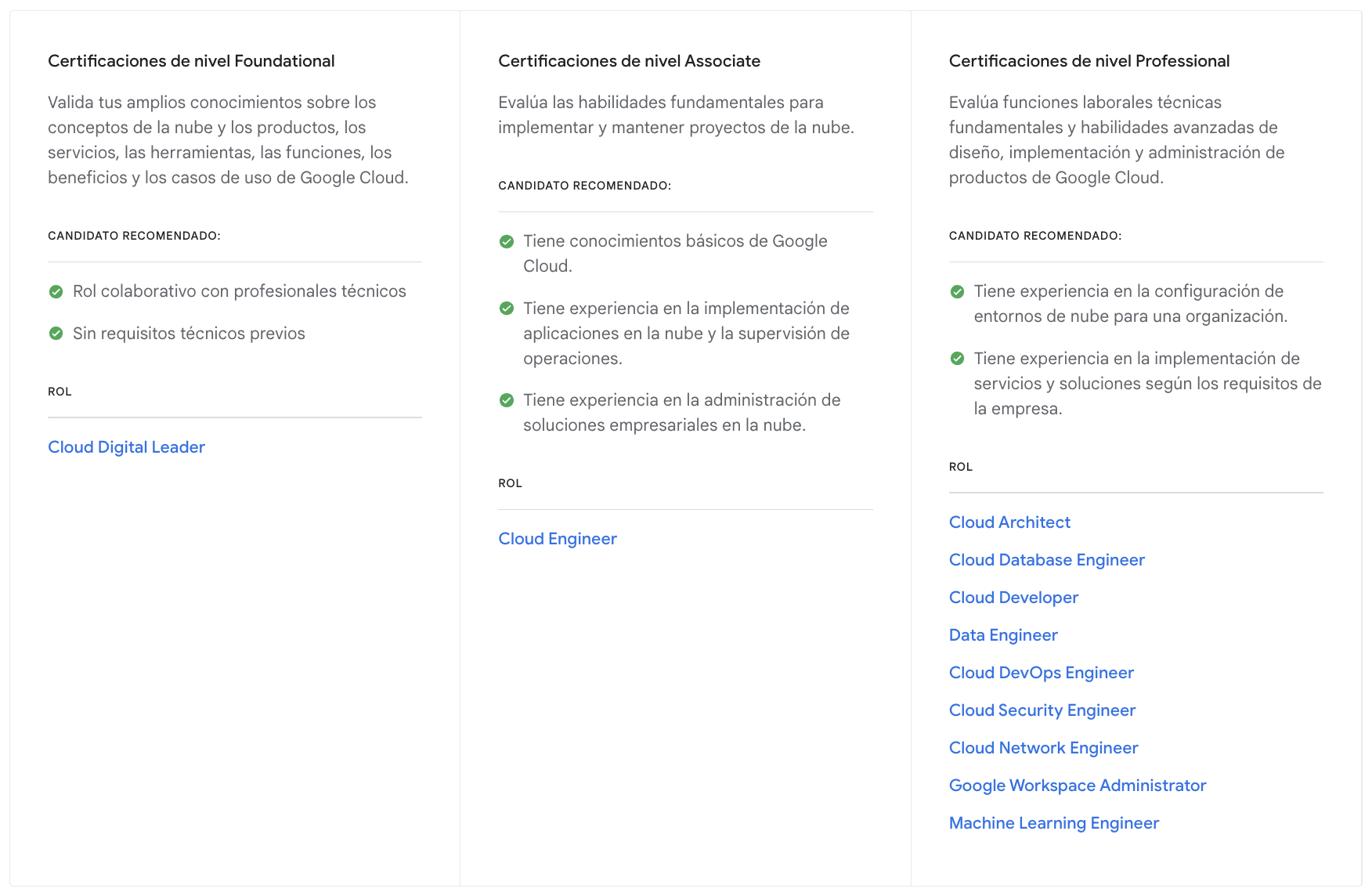
Los cachés son volátiles; pierde los datos almacenados en la memoria caché cuando se pierde la energía o se reinicia el sistema operativo. Puede almacenar datos en un caché para un acceso rápido, pero nunca debe usarse como el único almacén de datos que mantiene los datos. Se debe utilizar alguna forma de almacenamiento persistente para mantener un "sistema de verdad" o un almacén de datos que siempre tenga la versión más reciente y precisa de los datos.

Los cachés pueden salir de la sincronización con el sistema de verdad. Esto puede suceder si el sistema de verdad se actualiza pero los nuevos datos no se escriben en el caché. Cuando esto sucede, puede ser difícil para una aplicación que depende del caché detectar el hecho de que los datos en el caché no son válidos. Si decide usar un caché, asegúrese de diseñar una estrategia de actualización de caché que cumpla con sus requisitos de consistencia entre el caché y el sistema de verdad. Este es un problema de diseño tan desafiante que se ha memorizado en la conocida broma de Phil Karlton: “Solo hay dos cosas difíciles en informática: invalidación de caché y nombrar cosas”.

<https://cloud.google.com/>

1. **Certificaciones**

Google Cloud Platform separa sus certificaciones en 3 niveles, que son Foundational, Associate, y Professional certifications.

****

<https://cloud.google.com/certification>

1. **El Rol del Data Engineer**

Toda organización puede beneficiarse de los datos, utilizados de manera adecuada en coordinación con los objetivos de la organización. El proceso de hoy de acumular enormes cantidades de información de diferentes fuentes (*big data* en lenguaje común) es como un poderoso telescopio que le permite ver más adentro del universo y aprender sobre tendencias que antes no podía rastrear.

El término *ingeniero de datos* es relativamente nuevo, y el papel aparece solo esporádicamente en la literatura técnica. Son los científicos de datos los que son exaltados como los héroes de hoy: los pilotos de combate de alto nivel de big data que crean modelos de aprendizaje automático para predicciones y otros análisis. Pero ningún avión podría volar sin un pequeño ejército de mecánicos y otro personal capacitado para preparar los aviones. En esta analogía, el ingeniero de datos es como el mecánico de aviones, preparando los datos que permiten al científico de datos llevar a cabo los objetivos de la organización. Cuando suma todas las tareas que se ajustan a este rol (obtener los datos, limpiarlos y crear versiones mejoradas), los observadores a menudo afirman que la ingeniería de datos comprende del 80 al 90% del trabajo de las organizaciones con los datos.

¿Por qué la ingeniería de datos se volvió tan importante? Durante la década de 1990, el mundo experimentó una transformación trascendental que alcanzó cada vez más profundamente en nuestras vidas y trajo impactos en cascada en casi todas las carreras. La transformación fue impulsada por una mezcolanza de tendencias:

1. **Big Data y Machine Learning**

<https://cloud.google.com/solutions/ai>

1. **Construir un Data Lake**

<https://cloud.google.com/solutions/build-a-data-lake-on-gcp?hl=es-419>

1. **Recursos Adicionales**

A continuación se listan algunos recursos adicionales para guiarlo en su camino a ser un Professional Data Engineer.

* The Professional Data Engineer Certification Exam Guide

<https://cloud.google.com/certification/guides/data-engineer>

* Preguntas Frecuentes (FAQs)

<https://cloud.google.com/certification/faqs/>

* Examen de Práctica Professional Data Engineer

<https://cloud.google.com/certification/practice-exam/data-engineer>

* Documentación Oficial de Google Cloud Platform

<https://cloud.google.com/docs/>

* Cursos on-demand de Coursera.org

<https://www.coursera.org/professional-certificates/gcp-data-engineering>